Studien zum Physik- und Chemielernen

H. Niedderer, H. Fischler, E. Sumfleth [Hrsg.]

287

Markus Sebastian Feser

Physiklehrkräfte korrigieren Schülertexte

Eine Explorationsstudie zur fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung im Physikunterricht

λογος

Studien zum Physik- und Chemielernen

Herausgegeben von Hans Niedderer, Helmut Fischler und Elke Sumfleth

Diese Reihe im Logos-Verlag bietet ein Forum zur Veröffentlichung von wissenschaftlichen Studien zum Physik- und Chemielernen. In ihr werden Ergebnisse empirischer Untersuchungen zum Physik- und Chemielernen dargestellt, z. B. über Schülervorstellungen, Lehr-/Lernprozesse in Schule und Hochschule oder Evaluationsstudien. Von Bedeutung sind auch Arbeiten über Motivation und Einstellungen sowie Interessensgebiete im Physik- und Chemieunterricht. Die Reihe fühlt sich damit der Tradition der empirisch orientierten Forschung in den Fachdidaktiken verpflichtet. Die Herausgeber hoffen, durch die Herausgabe von Studien hoher Qualität einen Beitrag zur weiteren Stabilisierung der physik- und chemiedidaktischen Forschung und zur Förderung eines an den Ergebnissen fachdidaktischer Forschung orientierten Unterrichts in den beiden Fächern zu leisten.

Hans Niedderer Helmut Fischler Elke Sumfleth

Studien zum Physik- und Chemielernen

Band 287

Markus Sebastian Feser

Physiklehrkräfte korrigieren Schülertexte

Eine Explorationsstudie zur fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung im Physikunterricht

Logos Verlag Berlin

λογος

Studien zum Physik- und Chemielernen
Hans Niedderer, Helmut Fischler, Elke Sumfleth [Hrsg.]

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

ISBN 978-3-8325-5020-2
ISSN 1614-8967

Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47,
D-10243 Berlin

Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
https://www.logos-verlag.de

Physiklehrkräfte korrigieren Schülertexte

Eine Explorationsstudie zur fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung im Physikunterricht

Dissertation zur Erlangung des akademischen Doktors der Philosophie am Fachbereich Didaktik der gesellschaftlichen und mathematisch-naturwissenschaftlichen Fächer, Fakultät für Erziehungswissenschaft, Universität Hamburg

vorgelegt von:
Markus Sebastian Feser
Geboren am 14. Februar 1989 in Schweinfurt

Gutachter_innen:
Prof. Dr. Dietmar Höttecke
Universität Hamburg, Fakultät für Erziehungswissenschaft

Prof. Dr. Timo Ehmke
Leuphana Universität Lüneburg, Fakultät Bildung

Prof. Dr. Dagmar Killus
Universität Hamburg, Fakultät für Erziehungswissenschaft

Hamburg, November 2019

Danksagung

Eine Vielzahl von Menschen und Institutionen hat mich während der Arbeit an meinem Dissertationsprojekt inspiriert, herausgefordert und gefördert. Ohne diese Menschen und Institutionen wäre die vorliegende Arbeit nicht zustande gekommen. Ihnen allen möchte ich an dieser Stelle herzlich danken. Mein Dank gebührt dabei insbesondere den im Folgenden namentlich genannten. Ich danke...

... meinem Betreuer Prof. Dr. Dietmar Höttecke für die geduldige, konstruktive und wertschätzende Begleitung meiner Arbeit. Ein besseres Betreuungsverhältnis hätte ich mir nicht wünschen können.

... Prof. Dr. Timo Ehmke und Prof. Dr. Dagmar Killus für ihre Bereitschaft meine Arbeit zu begutachten, sowie ihre wertvolle Unterstützung in ganz unterschiedlichen Stadien meines Dissertationsprojekts.

... allen (ehemaligen oder assoziierten) Mitgliedern der Arbeitsgruppe Physikdidaktik der Universität Hamburg, die ich über die Jahre hinweg kennenlernen durfte: Prof. Dr. Dietmar Höttecke, Dr. Andreas Henke, Dr. Olaf Uhden, Jan Ruhrig, Prof. Dr. Ricardo Karam, Dr. Janne Langhof, Dr. Hannes Sander, Carina Wöhlke, Nadezda Strunk, Annemarie Klemp, Johanna Ratzek, Timo Hackemann und Nele Kroll. Bei euch möchte ich mich für die vielen fachdidaktischen Diskussionen, für den Austausch über mein Dissertationsprojekt, sowie für das gewissenhafte und fleißige Korrekturlesen meiner Arbeit herzliche bedanken.

... Heidrun Kraußer für ihr stets offenes Ohr und ihre umfängliche Unterstützung in Verwaltungsangelegenheiten.

... allen Mitgliedern und Mitarbeiter_innen der *Arbeitsgruppe Fach und Sprache*. Der interdisziplinäre und universitätsübergreifende Austausch im Rahmen der regelmäßigen Treffen dieser Arbeitsgruppe hat zur Qualität meiner Arbeit wesentlich beigetragen. Vor allem möchte ich mich an dieser Stelle bei Prof. Dr. Lena Heine und Prof. Dr. Knut Schwippert für die methodischen und methodologischen Anregungen und Hinweise zu meiner empirischen Untersuchung bedanken.

... den zahlreichen Mitgliedern der *Arbeitsgemeinschaft qualitative Inhaltsanalyse* der Graduiertenschule der Fakultät für Erziehungswissenschaft der Universität Hamburg. Die regelmäßigen, sowie gleichzeitig sehr produktiven, kritisch-konstruktiven und bestärkenden Diskussionen im Rahmen dieser Arbeitsgemeinschaft haben meine Arbeit umfänglich bereichert.

... Dr. Carola Großmann, Prof. Dr. Jenna Koenen, Britta Lübke, Dr. Thomas Plotz und Prof. Dr. Tanja Tajmel für die vielzähligen Anmerkungen und Denkanstöße zu meinem Dissertationsprojekt, sowie den inspirierenden fachdidaktischen Austausch.

... Prof. Dr. Andreas Borowski dafür, dass er mir ein Lernvideo zur Methode des lauten Denkens, das in seiner Arbeitsgruppe entwickelt wurde, für meine empirische Untersuchung zur Verfügung stellte.

... Prof. Dr. Drorit Lengyel dafür, dass sie mir Materialien und Vorarbeiten der FörMig-Initiative zur Verfügung stellte.

... allen Personen, die mich bei der Transkription und Codierung meiner erhobenen Daten in der Entwicklungs- und Hauptstudie tatkräftig unterstützt haben und die im Rahmen der vorliegenden Arbeit anonym bleiben sollen.

... allen Schüler_innen und allen angehenden, sowie im Schuldienst aktiven Physiklehrer_innen, die sich breit erklärt haben, an meiner empirischen Untersuchung teilzunehmen und ohne die die vorliegende Arbeit nicht möglich gewesen wäre.

... der *Fakultät für Erziehungswissenschaft* der Universität Hamburg, die mir durch ihre Reisemittelunterstützung ermöglichte an der *GDCP*-Jahrestagung 2016 in Zürich teilzunehmen.

... dem *Deutschem Akademischen Austauschdienst*, der mir durch seine Förderung die Teilnahme an der *ESERA*-Konferenz 2017 in Dublin möglich machte.

... allen Mitarbeiter_innen im Projekt *Hein & Fiete*. Die ehrenamtliche Zusammenarbeit mit euch hat mich regelmäßig zurück ins „richtige Leben“ geholt und gleichzeitig den nötigen Ausgleich gegeben, um die vorliegende Arbeit abschließen zu können.

Da ich das große Glück hatte im Rahmen meines Dissertationsprojekts mit so vielen Menschen und Institutionen zusammenzuarbeiten, dass ich sie hier nicht alle namentlich nennen kann, ist auch die eben aufgeführte Liste weit entfernt von jeglicher Vollständigkeit. Allen Personen und Institution, die ich nicht namentlich genannt habe, möchte ich daher abschließend noch einmal meinen herzlichen Dank aussprechen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

> „Ich setze mich an den Tisch, entkorke eine rote Tinte, mach mir dabei die Finger tintig und ärgere mich darüber. [...] Sechsundzwanzig blaue Hefte liegen neben mir, sechsundzwanzig Buben, so um das vierzehnte Jahr herum, hatten gestern [...] einen Aufsatz zu schreiben[.] [...] Draußen scheint noch die Sonne, fein muß es sein im Park! Doch Beruf ist Pflicht, ich korrigiere die Hefte und schreibe in mein Büchlein hinein, wer etwas taugt oder nicht.“ (von Horváth, 2017, S. 7-8)

Im aufgeführten Zitat aus dem Roman *Jugend ohne Gott* wird eine Szene beschrieben, die jedem_jeder Lehrer_in wohl bekannt ist. Aufgrund der Alltäglichkeit dieser oder ähnlicher Handlungsepisoden für Lehrer_innen verwundert es nicht, dass sich in der erziehungswissenschaftlichen Literatur bereits intensiv und aus verschiedensten Perspektiven mit schulischer Leistungsfeststellung und -beurteilung auseinandergesetzt wurde (vgl. Fast & Klein, 1998, S. 39; Beutel & Vollstädt, 2000, S. 13; Ingenkamp & Lissmann, 2008, S. 130 u. f.; U. Maier, 2015, S. 7). Insbesondere wird der Anspruch gelten gemacht, dass (Fach-)Lehrkräfte mehr als andere bezüglich schulischer Leistungsfeststellung und -beurteilung grundgebildet sein sollen (vgl. Schafer, 1993, 124 u. f.; Kultusministerkonferenz , 2018, S. 4). Physiklehrkräfte sollten dabei nicht nur bezüglich der Feststellung und Beurteilung von fachlich-konzeptuellen Schülerleistungen grundgebildet sein, sondern auch bezüglich der von sprachlichen Schülerleistungen. Grund hierfür ist, dass es für Schüler_innen gilt, neben fachlich-konzeptuellen, auch nicht zu vernachlässigende sprachliche Anforderungen zu meistern, um im Physikunterricht erfolgreich zu sein (vgl. Wellington & Osborne, 2001, S. 2; Höttecke, 2017, S. 107; Tajmel, 2017b, S. 199 u. f.). Umso erstaunlicher ist daher, dass sich den folgenden beiden Fragen in der (physikdidaktischen) Forschung bislang nur unzureichend gewidmet wurde:

(F1) Welche Ressourcen[1] werden von Physiklehrkräften bei schriftlichen, aus einer Klassenarbeit stammenden Schülerleistungen zur Genese fachlich-konzeptueller und sprachlicher Leistungsurteile eingesetzt?

(F2) Inwieweit findet im Rahmen einer Klassenarbeit bei der Feststellung und Beurteilung von schriftlichen Schülerleistungen durch Physiklehrkräfte eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile statt?

Die vorliegende Arbeit hat daher die Exploration dieser beiden Fragen zum Anliegen. Hierzu ist sie in zwei Teile gegliedert:

[1] Im Rahmen der vorliegenden Arbeit dient der Ressourcenbegriff – in Anlehnung an die Terminologie von Vogelsang & Reinhold (2013) – als möglichst neutraler Globalbegriff für Lehrerwissen und -können im Kontext schulischer Leistungsfeststellung und -beurteilung (vgl. Unterkapitel 4.1).

Teil I beginnt mit grundlegenden Überlegungen zu schulischer Leistungsfeststellung und -beurteilung (vgl. Kapitel 1). Dem folgt eine umfassende Aufarbeitung des bisherigen Forschungsstands über Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung, sowie die detaillierte Erörterung des Konzepts einer Assessment Literacy, das in der vorliegenden Arbeit als heuristischer Referenzrahmen dient (vgl. Kapitel 2). Der Schwerpunkt der Darstellung liegt hierbei auf dem bisherigen Erkenntnisstand der naturwisssenschaftsdidaktischen, speziell der physikdidaktischen Forschung. In Kapitel 3 erfolgt schließlich eine ausführliche Erörterung der Bedeutung von Sprache im Rahmen von Leistungsfeststellung und -beurteilung im Physikunterricht. Dies erfolgt sowohl auf theoretischer Ebene, wie auch auf Ebene bisheriger empirischer Forschungsarbeiten, wobei der Fokus hierbei auf der Perspektive von Physiklehrer_innen liegt.

In Teil II wird zu Beginn das Erkenntnisinteresse und die methodische Grundlegung der empirischen Untersuchung der vorliegenden Arbeit dargelegt. Hierzu wird in Kapitel 4 – basierend auf der Darstellung in Teil I – begründet, dass es sich bei den benannten Forschungsfragen (F1) und (F2) um Desiderate physikdidaktischer Forschung handelt. Daraufhin befasst sich Kapitel 5 mit der Entwicklungsstudie, die im Rahmen der vorliegenden Arbeit Zwecks der Eruierung eines gegenstandsangemessenen methodischen Vorgehens durchgeführt wurde. Kapitel 6 widmet sich schließlich der ausführlichen Darstellung der Hauptstudie der vorliegenden Arbeit, die der Beantwortung von Forschungsfrage (F1) und (F2) diente und auf Basis der Erkenntnisse der Entwicklungsstudie erfolgte. Im Rahmen dieser Hauptstudie wurde eine Gelegenheitsstichprobe von N = 21 Hamburger Physiklehrkräfte in einer Laborsituation gebeten, laut denkend kontrastierende Schülerlösungstexte zu einer Klassenarbeitsaufgabe zu korrigieren. Des Weiteren wurden die Teilnehmer_innen im Anschluss an die Korrektur retrospektiv befragt. Die dabei gewonnenen Daten wurden durch verschiedene Techniken der (qualitativen) Inhaltsanalyse ausgewertet und, sofern dies möglich bzw. zulässig ist, mit Hilfe nicht-parametrischer statistischer Methoden analysiert. Kapitel 6 endet damit, dass die durch verschiedene Analysen gewonnenen Teilbefunde zu einem kaleidoskopartigen Gesamtbild der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile der Teilnehmer_innen der Hauptstudie zusammengesetzt werden. Abschließend erfolgt ein Resümee über die empirischen Erkenntnisse, die in den vorherigen Teilen der Arbeit gewonnen wurden (vgl. Kapitel 7). Die Darstellung gliedert sich dabei in eine kritische Betrachtung des Vorgehens im Rahmen der vorliegenden Arbeit, sich aus dieser Arbeit ableitende Implikationen für die physikdidaktische Forschung, sowie jene Implikationen, die für die Physiklehrkräfteaus- und -weiterbildung resultieren.

Teil I:

Theoretischer Hintergrund und Stand der Forschung

„Ja, natürlich stellt sich mir immer wieder die Frage, gibt's da eigentlich Regeln. [...] Natürlich gibt es richtig und falsch oder richtiger und falscher, bei in der Art und Weise, wie man an die Korrektur herangeht. Aber äh ich vermute fast, das hören Sie häufig. Ähm ich kann mich nicht erinnern, zum Beispiel im Referendariat, dazu ausführlich ähm einen, also (..) Informationen oder Übungen gemacht zu haben. [...] Und ich muss auch sagen, im äh Fachkollegenkreis ist das äh durchaus auch nicht unbedingt Thema. Sondern nur dann, wenn wir uns (.) äh mit Vergleichsarbeiten beschäftigen. [...] Dann kommen wir sehr intensiv ins Gespräch."

Herr Uckermark (Transkiptsegment 742)

1. Grundlegende Überlegungen zu Leistungsfeststellung und -beurteilung in der Schule

Leistungsfeststellungen und -beurteilungen gehören zum Schulalltag. Lehrkräfte schätzen die mündliche Mitarbeit ihrer Schüler_innen ein, sie beurteilen fachspezifisches praktisches Tun im laufenden Unterrichtsgeschehen, führen Klassenarbeiten durch usw. „[Sie gelten dabei] nicht selten als undankbare Aufgabe [...] [und sind] [z]uallererst [...] oft zeitintensiv“ (vgl. Friege, 2017, S. 2-3). Verschiedenen Studien zufolge beschäftigen sich Lehrkräfte in ca. 20 bis 30 % ihrer Arbeitszeit mit Aktivitäten, die unmittelbar mit der Leistungsfeststellung und -beurteilung von Schüler_innen zusammenhängen (z. B. Comber & Keeves, 1973, S. 80 u. f.; Stiggins, 1988; Böhm-Kasper, 2004, S. 181). Sie ist damit kein Randthema, sondern gehört für Lehrkräfte zum Kerngeschäft.

Es verwundert daher nicht, dass schulische Leistungsfeststellungen und -beurteilungen auch regelmäßig in praxisorientierter pädagogischer Literatur diskutiert werden. Tabelle 1.1 liefert hierzu ein illustratives Beispiel. In dieser sind die Ausgaben der Zeitschrift *Unterricht Physik*[2] aufgelistet, die das Thema Leistungsfeststellung und -beurteilung in den Vordergrund rücken. Betrachtet man hier die Beitragstitel, so fällt zunächst auf, dass sich ein überwiegender Teil dieser mit der Frage auseinandersetzen, welche Formen der Leistungsfeststellung und -beurteilung geeigneten und praktikablen sind, bzw. wie sich diese optimieren lassen. Für ein an der Praxis orientiertes Magazin wie der *Unterricht Physik* erscheint dies zunächst nicht ungewöhnlich, zumal Form und Funktion von schulischen Leistungsfeststellungen und -beurteilungen nicht vollkommen unabhängig voneinander, sondern „bis zu einem bestimmten Grad interdependent sind“ (Füller, 1975, S. 16). Allerdings vermag eine ausschließliche Debatten über die Optimierung von und über unterschiedlicher Möglichkeiten der Leistungsfeststellung und -beurteilungen nicht die Fragen zu beantworten, warum in der Schule Leistungsfeststellungen und -beurteilungen überhaupt vorgenommen werden, warum derart viele Ressourcen in sie investiert werden, und insbesondere was die Funktionen dieser Praxis sind. Betrachtet man hierzu den Diskurs in der deutschsprachigen naturwissenschaftsdidaktischen Forschung, so fällt auf, dass dieser seit der Jahrtausendwende, insbesondere seit dem sog. „PISA-Schock“ im Jahr

[2] Als Grundlage für Tabelle 1.1 diente die entsprechende Auflistung von Gunnar Friege (vgl. Friege, 2017, S. 2). In dieser sind auch die Vorläuferzeitschriften *Naturwissenschaften im Unterricht – Physik* und *Naturwissenschaften im Unterricht – Physik/Chemie* mit berücksichtigt. Zu beachten ist dabei, dass die Themenhefte vor dem Jahr 1990 einer eigenen Nummerierung folgen.

Jahr	Nr.	Hefttitel	Beitragstitel
2017	158	Leistung transparent bewerten	Leistungsbewertung – eine ungeliebte Aufgabe; Klassenarbeiten vorbereiten, durchführen und bewerten; Wie beurteilt man eine Schülererklärung?; Bewertung experimenteller Leistungen; Mündliche Physikprüfungen vorbereiten und durchführen; „Mündliche Mitarbeit“: von der 2Q- zur 3k-Bewertung; Gemeinsames Lernen – individuelle Leistung?; Eine Norm für alle?
2002	70/71	Experimente als Lernerfolgskontrolle	Lernerfolgskontrolle mit Experimenten; Schülerexperimente als Testsituation; Einstieg in eine Lernerfolgskontrolle mit Experimenten; Experimente im Unterricht bewerten; Experimente als Teil komplexer Aufgaben; Black-Box-Aufgaben mit elektrischen Widerständen; Andere Länder – andere Tests; Experimentelle Praktika im Physikunterricht; Experimentelles Praktikum; Naturwissenschaftliches experimentelles Praktikum; Experimentieraufgaben mit mehreren Lösungswegen
1997	38	Unterricht bewerten	Unterricht vielfältig bewerten; Bewertungsmethoden; Concept Mapping; Die Portfoliomethode; Individualisierte Leistungsbewertung; Ein Bild sagt mehr als tausend Worte...; Lern- und Unterrichtsklima im Physikunterricht
1988	38	Schülerbewertung	Schülerbeurteilung im Physikunterricht; Zum Problem der Gewinnung von Zeugnisnoten für den Physikunterricht; Anleitung zur Entwicklung „Informeller Tests“; Planung und Auswertung einer Physikarbeit; Die Bewertung mündlicher Schüleräußerungen; Förderung der Leistungsbereitschaft durch Schülerbeteiligung und Beurteilung praktischen Tuns; Beurteilung selbstständiger Schülerleistungen im Physikunterricht

Tabelle 1.1.: Ausgaben der Praxiszeitschrift *Unterricht Physik* und deren Vorgängerzeitschriften zum Thema Leistungsfeststellung und -beurteilung. Die Titel der Beiträge sind durch ein Semikolon voneinander getrennt und in der selben Reihenfolge wie im entsprechenden Themenheft gelistet.

2001, stark vom Kompetenzbegriff dominiert ist und damit auch der Frage nach outputorientierten Standards und deren Erfassung verstärkt nachgegangen wurde (vgl. Tajmel, 2017b, S. 99 u. f.). Allerdings sind auch hier, worauf Tillmann und Vollstädt bereits im Jahr 2000 hinweisen,...

> „[... offenbar die in den 1980er-Jahren] geführten Diskussionen über Sinn und Unsinn des Leistungsprinzips [...] allzu sehr in Vergessenheit geraten [...] [, ebenso wie] die empirisch begründeten Argumente von *Ingenkamp* zur Fragwürdigkeit der Zensurengebung[.]“ (Tillmann & Vollstädt, 2000, S. 27, Hervorhebungen im Original)

In diesem Kapitel soll daher eine Bestandsaufnahme über die Funktionen und die Güte schulischer Leistungsfeststellungen und -beurteilungen erfolgen. Zunächst gilt es allerdings die grundlegende Terminologie des Diskurses um schulische Leistungsfeststellungen und -beurteilungen aufzuarbeiten. Warum hier ein Klärungsbedarf besteht, deutet sich ebenfalls bereits in Tabelle 1.1 an: Mal wird hier von Leistungsbewertung, mal von Leistungsbeurteilung gesprochen. Auch aus diesem Grund soll in der nachfolgenden Darstellung zunächst ausschließlich an dem Begriffspaar „Leistungsfeststellungen und -beurteilungen“ festgehalten werden.

1.1. Terminologie des Diskurses um schulische Leistungsfeststellungen und -beurteilungen

Über Leistung wurde und wird im erziehungs- und gesellschaftswissenschaftlichen Diskurs viel gesprochen (vgl. Ingenkamp & Lissmann, 2008, S. 130 u. f.). Der Begriff selbst gilt dabei als umstrittener Terminus, der sowohl vielseitig, als auch unterschiedlich eng bzw. weit gefasst werden kann (vgl. Schröder, 1985, S. 176 u. f.). Ferner ist Leistung das Bestimmungswort einer Reihe von Komposita und Ausdrücken, wie beispielsweise Leistungsfeststellung oder Leistungsbewertung, von denen einige im Verlauf dieses Unterkapitels definiert werden sollen. Es ist daher sinnvoll zunächst den Leistungsbegriff selbst näher zu bestimmen.

1.1.1. Der Begriff der Leistung im schulischen Kontext

Wenn im Kontext von Schule über Leistung gesprochen wird, ist damit in den meisten Fällen das Lernen der Schüler_innen relativ zu einem Gütemaß gemeint (vgl. Bos, Voss, & Goy, 2009, S. 563). Erst der entsprechende Gütemaßstab ermöglicht es eine Bewertung[3] vorzunehmen, z. B. ob das Lernen eines bestimmten deklarativen Wissensbestands erfolgreich oder weniger erfolgreich stattgefunden hat. Dies betonen auch Heller & Hany (2001). Sie gehen allerdings noch weiter und merken an, dass die Leistung von Schüler_innen nicht nur von ihnen alleine, „sondern ebenso vom didaktischen und pädagogischen Geschick des Lehrers und verschiedenen Merkmalen des Lernumfeldes abhäng[t]“ (ebd., S. 88). Dementsprechend ist der Begriff Leistung zwar fokussiert auf das Lernen von Schüler_innen, kann aber gleichzeitig nur als von „Schule initiiertet“ gedacht werden (vgl. Ingenkamp & Lissmann, 2008, S. 131).

Aufgrund des eben Gesagten, hängt das Verständnis von Leistungsbegriff davon ab, wie weit der Begriff des Lernens gefasst wird. Viele empirische Untersuchungen beschränken ihren Leistungsbegriff aus jeweils unterschiedlichen Gründen auf kognitive Aspekte des Lernens (z. B. Jung, Reul, & Schwedes, 1977; Kauertz, 2008). Dementsprechend ist auch Merzyn (2006) zu verstehen: Er verortet Lernerfolge allgemein im kognitiven Bereich und unterscheidet diese von affektiven Einstellung, wie z. B. dem Fachinteresse oder dem Selbstvertrauen der Schüler_innen (vgl. ebd.). Gleichzeitig betont er aber, dass zwischen Lernerfolg und Schülereinstellungen starke wechselseitige Zusammenhänge bestehen (vgl. ebd.). Dieses Unterscheidung findet sich auch bei anderen Autor_innen, vor allem im Bereich der pädagogischen Psychologie[4]. Insbesondere werden Schülereinstellungen meist als Determinanten von Leistung aufgefasst und hierdurch von dieser abgegrenzt. Zu einer derartigen Unterscheidung merken Duit & Häußler (1997) allerdings kritisch an:

> „Die Forschung hat klar gezeigt, dass der Erwerb längerfristig „haltbaren Wissens“ eng mit dem Interesse an einer Sache, mit der Einstellung zum Fach und zum Fachunterricht zu tun

[3] Auf den Begriff der Leistungsbewertung wird in Abschnitt 1.1.2 eingegangen.
[4] Eine Zusammenschau hierzu ist z. B. bei Helmke & Weinert (1997, S. 111 u. f.) zu finden.

> hat. Aber nicht nur in dieser vermittelnden Funktion ist der affektive Bereich wichtig, er hat durchaus seinen Eigenwert.“ (ebd., S. 4)

Duit & Häußler (1997) verweisen hier darauf, dass Lernen im Allgemeinen neben kognitiven auch emotionale, psychomotorische und soziale Aspekte, sowie Werte und Haltungen umfasst (vgl. Klafki, 1995, S. 983 u. f.) und dass es auch bei einem breiten Lernbegriff möglich ist, Lernen einen Wert bezogen auf ein entsprechendes Gütemaß zuzuweisen. Aus diesem Grund empfehlen Bos et al. (2009, S. 563) einem verallgemeinerten Leistungsbegriff einen Lernbegriff zugrunde zu legen, der neben Erkenntnissen und Wissen, auch Einstellungen, Fähigkeiten, Fertigkeiten umfasst.

Die Enge bzw. Weite des Leistungsbegriffs wird aber nicht nur vom Verständnis des Lernbegriffs bestimmt. Der Begriff selbst kann in zweierlei Hinsicht aufgefasst werden. Zum einen adressiert er das Ergebnis individuellen Aneignens, Anstrengens oder Übens (Ravitch, 2007, S. 9 u. f.). Zum anderen kann er aber auch prozesshaft verstanden werden (vgl. Heller & Hany, 2001), also mehr die Genese selbst anstatt das Endprodukt eines Verhaltens in den Vordergrund rücken. Schröder (1985, S. 177) unterscheidet dementsprechend einen statischen, sowie einen dynamischen Leistungsbegriff. Er betont dabei, dass Leistungen von Schüler_innen oftmals verengt auf den statischen Leistungsbegriff betrachtet werden, es seiner Ansicht nach in schulischen Kontexten aber gilt, den dynamischen Leistungsbegriff in besonderem Maße zu berücksichtigen (vgl. ebd.).

Im letzten Jahrzehnt sind schulische Leistungen allerdings vor allem im Rahmen der Kompetenzdebatte diskutiert worden. Nach Hartig & Klieme (2006) kann Kompetenz verstanden als „kontextspezifische kognitive Leistungsdisposition, die sich funktional auf bestimmte Klassen von Situationen von Aufgaben bezieht“ (ebd., S. 128). Kompetenz ist nach dieser Definition also das bereichsspezifische Vermögen zu einer bestimmten Art der Performanz, wohingegen Leistungen Performanzen in konkreten jeweils unterschiedlichen Situationen sind. In der probabilistischen Testtheorie bezeichnet man Kompetenzen daher auch als latente Variablen die Performanzen (manifeste Variablen) plausibel erklären können (vgl. Rost, 1996, S. 17 u. f.). Hier wird insgesamt also eine wohldefinierte Unterscheidung zwischen Kompetenzen und Leistung vorgenommen. Die Begriffe stehen hierbei nebeneinander, da Kompetenzen als Erklärung von Leistungen heranzogen werden und umgekehrt Leistungen von Schüler_innen zurückgeführt werden auf entsprechend modellierte Kompetenzen. Allerdings kann der Kompetenzbegriff ähnlich wie der Leistungsbegriff unterschiedlich ausgelegt werden[5]. Je breiter der Kompetenzbegriff gefasst wird, desto fließender wird hierdurch auch der Übergang zum Begriff der Leistung. Ferner werden Kompetenzen insbesondere als etwas Erlern- bzw. Erwerbbares verstanden (vgl. Weinert, 2001b, S. 26 u. f.). Kompetenzen, sowie der Aneignungsprozess von Kompetenzen sind damit selbst eine statische bzw. dynamische Form von Leistung. In einem weiten Begriffsverständnis kann der Kompetenzbegriff daher dem der Leistung unter- anstatt nebengeordnet werden. Helmke & Weinert (1997) kommen daher zu folgendem Schluss:

[5]Eine Übersicht über unterschiedlicher Kompetenzbegriffe findet sich z. B. bei Weinert (2001a).

> „Selbst wenn man nur kognitive Aspekte berücksichtigt, kann es sich bei der Schulleistung um den Erwerb, die kurz- wie langfristige Verfügbarkeit und/oder die Nutzung von fachspezifischem deklarativen [...] Wissen, prozedurale Fertigkeiten [...] oder metakognitive Kompetenzen [...] handeln[.]“ (ebd., S. 75)

In der Quintessenz ist Leistung im Kontext von Schule also ein Begriff, der im Allgemeinen sehr weit gefasst werden muss. Die vorliegende Ausarbeitung macht deutlich, dass sich die Breite des Begriffsverständnisses jeweils daraus ergibt, dass Leistung sich auf den facettenreichen Begriff des Lernens bezieht, der Begriff selbst prozesshaft und/oder ergebnisorientiert aufgefasst werden kann und dass aus Perspektive der Kompetenzdebatte zwischen latentem Vermögen und manifester Performanz unterschieden werden kann. Gleichzeitig wird damit deutlich, dass es sich bei Leistung um einen vielseitigen Begriff handelt. Es wundert daher nicht, dass verschiedene Autorin_innen oftmals unterschiedliche Dinge meinen, wenn sie über Leistung im schulischen Kontext sprechen. Systematisiert lassen sich die Erkenntnisse aus diesem Abschnitt für den weiteren Verlauf der vorliegenden Arbeit folgendermaßen zusammenfassen:

Zusammenfassung

Das von Schule initiierte Lernen von Schüler_innen in Relation zu einem Gütemaß bezeichnet man als *Leistung*. Dem Begriff lassen sich drei wesentliche Bestimmungsmerkmale zuordnen, durch die er wiederum jeweils enger bzw. weiter gefasst werden kann:

1. Eng gefasst adressiert Leistung ausschließlich kognitive Aspekte schulischen Lernens. Zwischen affektiven und kognitiven Aspekten bestehen allerdings wechselseitige Zusammenhänge, weswegen im engen Begriffsverständnis erstere als Determinanten schulischer Leistung verstanden werden. Die Bedeutung des Begriffs weitet sich, wenn neben kognitiven auch andere Formen schulischen Lernens zugrunde gelegt werden. Im sehr weiten Sinn bezieht sich Leistung daher sowohl auf Erkenntnisse und Wissen, als auch auf Einstellungen, Fähigkeiten und Fertigkeiten.

2. Im Allgemeinen versteht man unter Leistung sowohl den Prozess individuellen Aneignens, Anstrengens, oder Übens (dynamischer Leistungsbegriff), als auch das entsprechende Endergebnis (statischer Leistungsbegriff). Ein enger Leistungsbegriff beschränkt sich auf letzteres.

3. Leistung kann als Manifestation eines latenten Leistungsvermögens (Kompetenz) aufgefasst werden. Im engen Begriffsverständnis sind Leistungen daher von Kompetenzen zu unterscheiden. Im weit gefassten Sinn werden Kompetenzen dem Leistungsbegriff allerdings untergeordnet, da ihr Erwerb selbst als statische und/oder dynamische Form von Leistung gedeutet werden kann.

Auf Grundlage dieser Zusammenfassung folgt nun im nächsten Abschnitt die Aufarbeitung der Terminologie, mit deren Hilfe in der Literatur die Facetten des Begriffspaars Leistungsfeststellung und -beurteilung näher bestimmen werden.

1.1.2. Facetten des Begriffspaars Leistungsfeststellung und -beurteilung

Frei formuliert ist das Begriffspaar Leistungsfeststellung und -beurteilung ein Oberbegriff für schulische Verfahren, die dazu dienen Leistungen von Schüler_innen einzuschätzen. Durch sie wird also das Lernen der Schüler_innen in Relation zu einem Gütemaß gestellt. In der erziehungswissenschaftlichen Literatur finden sich zu diesem Begriffspaar eine ganze Reihe von synonym verwandten Begriffspaaren oder auch Einzelbegriffen, z. B. Leistungsmessung und -bewertung, Leistungsdiagnose, Assessment, Evaluation, usw. Daneben gibt es zu jedem dieser Oberbegriffe eine ganze Reihe von Unterbegriffen. Deren gemeinsame Systematik wird im Folgenden dargestellt.

1.1.2.1. Facette der Durchführungsform

Eine naheliegende Facette des Begriffspaars ergibt sich daraus Leistungsfeststellungen und -beurteilungen anhand ihrer Durchführung zu unterscheiden, also ob sie in Form von einer Klassenarbeit, eines mündlichen Unterrichtsgesprächs, eines Portfolios, usw. erfolgen. Füller (1975) unterscheidet hier drei prinzipiell verschiedene Formen der Leistungsfeststellung und -beurteilung: schriftlich, mündlich und praktisch (vgl. ebd., S. 14). Nach Kühberger (2014) ist jedoch die hinter einer Leistungsfeststellung und -beurteilung liegende Intention[6] und nicht die Form ihrer Durchführung ein entscheidendes strukturgebendes Merkmal, da unterschiedliche Absichten nicht zwingend verschiedene Formen der Leistungsfeststellung und -beurteilung erforderlich machen (vgl. ebd., S. 10). Beispielsweise kann sowohl ein schriftlicher Test als auch ein Portfolio eine Form der Eingangs- und/oder Abschlussprüfung darstellen, wobei in beiden Fällen mögliche Intentionen die Steuerung des Zugangs zu bestimmten Lehr-Lern-Angeboten oder die Begutachtung vergangenen Lernens sein können. Es gilt also „verschiedene Arten des Umgangs mit Schülerleistungen, die über Tools der Leistungsfeststellung zu erheben sind, zu unterscheiden“ (ebd., S. 9).

1.1.2.2. Facette der zeitlichen Stellung im Lehr-Lern-Prozess

Für eine solche Taxonomierung von Leistungsfeststellungen und -beurteilungen, wie Kühberger (2014) sie vorschlägt, finden sich in der Literatur unterschiedlichste Begrifflichkeiten. Dies sorgt zunächst für den Eindruck, verschiedene Autor_innen würden sehr unterschiedliche Intentionen mit schulischer Leistungsfeststellungen und -beurteilungen verbinden und/oder hervorheben. Dies trifft aber nur zu einem bestimmten Grad zu, da die den unterschiedlichen Taxonomien zugrundeliegende gemeinsame Struktur erkennbar wird, wenn man die verschiedenen Begriffe danach sortiert, welchen zeitlichen Abschnit-

[6] Diese Intentionen hängen mit den Funktionen schulischer Leistungsfeststellungen und -beurteilungen insofern zusammen, als dass sie sich aus diesen ableiten lassen. Auf diese Funktionen wird daher in Unterkapitel 1.2 genauer eingegangen.

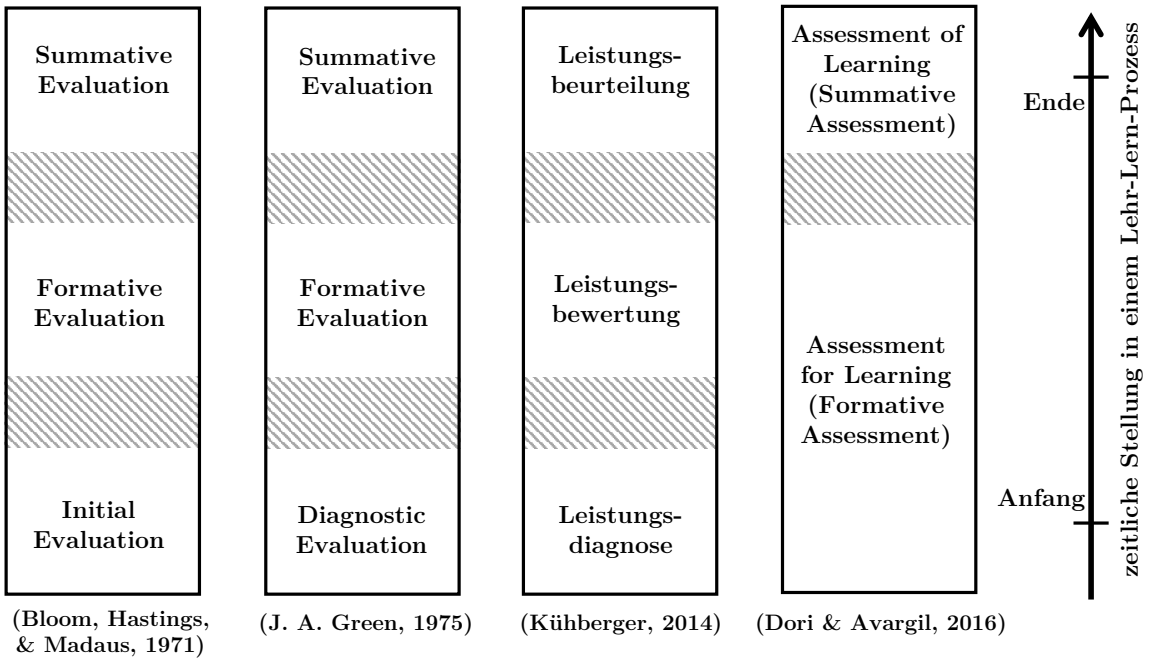


Abbildung 1.1.: Facette des Begriffs der Leistungsfeststellung und -beurteilung bezüglich der zeitlichen Stellung in einem Lehr-Lern-Prozess.

ten eines Lehr-Lern-Prozess sie von den entsprechenden Autor_innen zugeordnet worden sind. In Abbildung 1.1 ist dies beispielhaft geschehen. In dieser sind die Begrifflichkeiten aus vier erziehungswissenschaftlichen Beiträgen entlang des zeitlichen Verlaufs eines beliebigen Lehr-Lern-Prozesses (z. B. einer mehrwöchigen Unterrichtseinheit) sortiert. Die dargestellten Beiträge adressieren zum Teil sehr unterschiedliche Zielgruppen[7], sind zu verschiedenen Zeitpunkten entstanden und von jeweils anderen Autor_innen verfasst worden. Diese Auswahl wurde bewusst vorgenommen, um den diesbezüglich heterogene Charakter der Begriffswahl im erziehungswissenschaftlichen Diskurs sichtbar zu machen und in besonderer Weise hervor zu heben. Sie darf daher nicht als vollständiges Abbild des Diskurses verstanden werden, sondern vielmehr als repräsentative Auswahl kontrastierender Beispiele. Daneben verdeutlichen die schraffierten Bereiche in Abbildung 1.1, dass die Übergänge zwischen den Begrifflichkeiten fließend sind und daher auch Überschneidungen möglich sind.

Betrachtet man Abbildung 1.1 fällt zunächst auf, dass in der Literatur entweder begriffliche Zwei- (z. B. Dori & Avargil, 2016, S. 1033) oder Dreiteilungen (z. B. Bloom, Hastings, & Madaus, 1971, S. 14; J. A. Green, 1975, S. 49; Kühberger, 2014, S. 11) vorgenommen werden. Dies ergibt sich wiederum daraus, ob ein Lehr-Lern-Prozess als Ganzes seinem Abschluss gegenüber gestellt wird, oder ob ein Lehr-Lern-Prozess in Anfang, Ende und Verlauf unterteilt wird.

[7] Z. B. adressiert der Beitrag von Kühberger (2014) die Praxis des Geschichtsunterrichts, wohingegen Dori & Avargil (2016) aus der Perspektive der naturwissenschaftsdidaktischen Forschung schreiben.

Zunächst zu den begrifflichen Zweiteilungen: Die primäre Absicht hinter Leistungsfeststellungen und -beurteilungen am Anfang oder in einem laufenden Lehr-Lern-Prozess ist die, Einschätzungen im Dienste des geplanten oder gerade stattfindenden Lernens von Schüler_innen vorzunehmen. In der englischsprachigen Literatur findet sich daher auch der Begriff eines „Assessment for Learning", der oftmals gleichgesetzt wird mit formativem Assessment (vgl. Dori & Avargil, 2016, S. 1033). Im Gegensatz dazu steht ein „Assessment of Learning" bzw. ein summatives Assessment am Ende eines Lehr-Lernprozesses, wobei hier die Einschätzung über das vergangene Lernen von Schüler_innen die primäre Absicht ist (ebd.). Füller (1975) schreibt hierzu folgendes, wobei er sich bei seiner Wortwahl an Terminologie von (Bloom et al., 1971, S. 14) orientiert:

> „[Bei der summativen Evaluation] wird all das geprüft, was nach Ansicht des Lehrenden repräsentativ für den Lernprozeß und die impliziten oder expliziten Lernziele war. [...] [Sie] muß von relativ wenigen Aufgaben auf relativ detallierte Fähigkeiten schließen. Es geht um das Prüfen allgemeinerer Lernziele, als bei der formativen Evaluation." (ebd., S. 14)

Die Intention ist also zu überprüfen, inwieweit das beabsichtige Lernen bei den Schüler_innen stattgefunden hat und inwiefern die vorgenommenen Instruktionen bzw. hergestellten Lernsettings die erhoffte Wirkung gezeigt haben. In der Literatur wird summatives Assessment oftmals mit der Vergabe von Noten verbunden oder wie z. B. bei Kühberger (2014) sogar gleichgesetzt. Insbesondere liefern Leistungsfeststellungen und -beurteilungen am Ende eines Lehr-Lern-Prozesses aber auch Informationen darüber, was von Seiten der Schüler_innen noch gelernt werden muss (Bloom et al., 1971, S. 14). Im Sinne der Terminologie von Kühberger (2014, S. 11) können damit Leistungsbeurteilungen am Ende eines Lehr-Lern-Prozesses gleichzeitig auch als Leistungsdiagnosen[8] für einen darauf folgenden Lehr-Lern-Prozesses verstanden werden. Die Darstellung in Abbildung 1.1 ist daher als ein Ausschnitt aus einer Folge mehrerer Lehr-Lehr-Sequenzen zu verstehen, deren Enden sich jeweils überlappen. Mit diesem Gedankengang befindet man sich allerdings schon nicht mehr bei einer zeitlichen Zwei-, sondern einer Dreiteilung eines Lehr-Lern-Prozesses. Die Intentionen von Leistungsfeststellungen und -beurteilungen am Anfang und während eines laufenden Lehr-Lern-Prozesses werden daher nun noch einmal getrennt voneinander betrachtet:

Leistungsfeststellungen und -beurteilungen zu Beginn eines Lehr-Lern-Prozessen dienen dazu, die Lernvoraussetzungen bzw. Lernstände von Schüler_innen zu erheben (vgl. Füller, 1975, S. 13). Sie geben also Auskunft darüber, an welcher Stelle Schüler_innen relativ zu einem angestrebten Ziel stehen (vgl. Kühberger, 2014, S. 11). Aufnahme- und Eingangsprüfungen, Aufzeichnungen über bereits erbrachte Leistungen der Schüler_innen oder Vorjahresnoten sind Beispiele für derartige Leistungsfeststellungen und -beurteilungen (Bloom et al., 1971, S. 14). Die hierbei gewonnenen Informationen werden dazu genutzt, den geplanten Lehr-Lern-Prozess entsprechend dieser Eingangsvoraussetzungen (z. B. durch Differenzierung) zu optimieren bzw. abzustimmen (ebd.), den Schüler_innen

[8] Einem verallgemeinerten Diagnosebegriff ist schulische Leistungsfeststellung und -beurteilung jedoch unterzuordnen, da hierunter die Feststellung und Bestimmung von Personenmerkmalen im Allgemeinen, die nicht notwendigerweise mit Lernen in Verbindung stehen, zu verstehen ist (vgl. von Aufschnaiter et al., 2015, S. 740 u. f.).

Orientierung über ihre Ausgangslage zu verschaffen (Kühberger, 2014, S. 11) und/oder Schüler_innen, deren Lernstand nicht einer festgelegten Mindestnorm entsprechen von Lehr-Lern-Angeboten auszuschließen (vgl. Füller, 1975, S. 13).

Dagegen handelt es sich bei Leistungsfeststellungen und -beurteilungen in einem laufende Lehr-Lern-Prozess um...

> „[...] Gutachten über erbrachte punktuelle oder prozessual angelegte Leistungen der Schüler/innen [...][.] [...] Oftmals werden dafür eigene Bewertungsmodelle herangezogen, um sich von einer summativen Notengebung abzusetzen und um dennoch Einzelaspekte einer Leistung zu bewerten (z. B. „+“, „o“, „-“ oder 1 bis 10 Punkte).“ (Kühberger, 2014, S. 11)

Lehrkräfte führen derartige Feststellungen und Beurteilungen durch, um einzelne Lernsettings sinnvoll planen und um Schüler_innen Rückmeldung über (nicht) erfolgreich genommene Lernschritte geben zu können. Diese Intentionen überschneiden sich also mit denen von Leistungsfeststellungen und -beurteilungen am Anfang eines Lehr-Lern-Prozesses. Des Weiteren ermöglichen sie Lehrkräften die Optimierung der anfängliche Planung des Lehr-Lern-Prozesses, da sie Informationen darüber bereitstellen, inwieweit ihre ursprüngliche Planung um alternative Lernzugänge oder zusätzliche Unterstützungsmaßnahmen ergänzt bzw. angepasst werden sollte (Bloom et al., 1971, S. 14). Ferner sind auch Zwischenprüfungen Beispiel für Leistungsfeststellungen und -beurteilungen in einem laufenden Lehr-Lern-Prozess. Hierdurch wird deutlich, dass auch an dieser Stelle Selektion und Allokation mögliche Absichten sind (vgl. Füller, 1975, S. 13) und damit dass es erneut Intentionsüberschneidungen gibt, dieses mal allerdings sowohl mit den Intentionen von Leistungsfeststellungen und -beurteilungen am Anfang eines Lehr-Lern-Prozesses, als auch mit denen, die am Ende eines Lehr-Lern-Prozesses stehen.

1.1.2.3. Facette des Feststellungs- und Beurteilungsprozesses selbst

Neben der Facette des Begriffspaars der Leistungsfeststellung und -beurteilung bezüglich der zeitlichen Stellung in einem Lehr-Lern-Prozess, ergibt sich eine weitere Facette, wenn dieser Oberbegriff bezogen auf den Feststellungs- und Beurteilungsprozesses selbst in Teilaspekte untergliedern wird. Hierzu gibt es in der erziehungswissenschaftlichen Literatur ebenfalls unterschiedliche Begrifflichkeiten, wie in Abbildung 1.2 dargestellt ist. So findet sich beispielsweise bei Neuweg (2009) folgende Definition:

> „*Leistungsfeststellung* ist das Ermitteln der Schülerleistung durch die Messung von Lernergebnissen unter Anwendung eines Messinstruments (Feststellung der Mitarbeit, besondere mündliche, schriftliche, praktische oder graphische Formen der Leistungsfeststellung). [...] *Leistungsbeurteilung* ist die im Anschluss an Leistungsfeststellung vorgenommene Bewertung des Messergebnisses durch den Vergleich mit einem Beurteilungsmaßstab. Das Ergebnis der Leistungsbeurteilung wird durch die vom Gesetzgeber definierten Beurteilungsstufen (Noten) ausgedrückt.“ (ebd., S. 9-10, Hervorhebungen im Original)

Was bei Neuweg (2009, S. 3) die Leistungsbeurteilung ist, bezeichnen z. B. Fischer & Malle (1989, S. 305) oder Wodzinski (2007, S. 70) als Leistungsbewertung. Nach Jürgens (1997, S. 39 u. f.) werden beide Begriffe von manchen Autor_innen auch synonym ver-

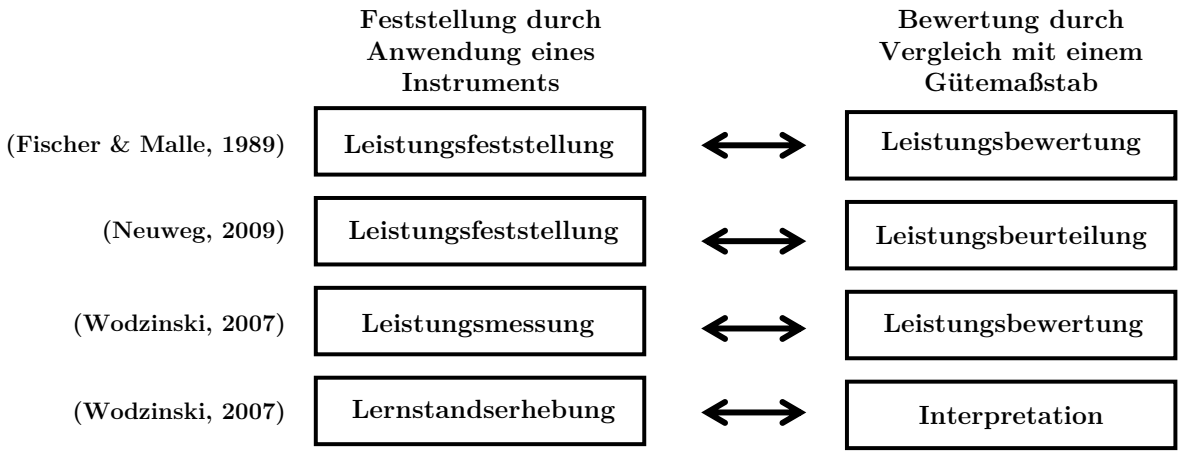


Abbildung 1.2.: Facette des Begriffspaars der Leistungsfeststellung und -beurteilung bezüglich des Feststellungs- und Beurteilungsprozesses selbst.Wodzinski (2007, S. 70) gibt zwei mögliche Begriffspaare an: Lernstandserhebung und Interpretation im Rahmen von Lerndiagnosen und Leistungsmessung und Leistungsbewertung bei Leistungsbeurteilungen. Worin der Unterschied zwischen Lerndiagnose und Leistungsbeurteilung besteht, lässt der Autor allerdings offen.

wendet. Allgemein liegen den in Abbildung 1.2 dargestellten Begrifflichkeiten das folgende Unterscheidungsprinzip zu Grunde: Einer der beiden Begriffe adressiert die Erhebung in einer konkreten Lehr-Lern-Situation, z. B. einer Klassenarbeit, sowie die Feststellung inwieweit die „gegebenen Antworten“ der Schüler_innen korrekt und vollständig sind. Der jeweils andere Begriff meint dagegen die Bewertung oder Beurteilung, die sich daraus ergibt, dass die (zuvor) festgestellten Leistungen in Relation zu einem Gütemaß gesetzt und gegebenenfalls in Zensuren übersetzt werden. Nach Jürgens (1997, S. 39 u. f.) sind solche begrifflichen Zweiteilungen allerdings eher die Ausnahme. Im Allgemeinen wird kaum zwischen z. B. Leistungsfeststellung und -beurteilung unterschieden (ebd.). Ferner weißt Benischek (2006) darauf hin, dass derartige begriffliche Unterscheidungen zwar prinzipiell möglich, allerdings auch problematisch sind:

> „Man kann die Messung der Schülerleistung zwar grundsätzlich von der Beurteilung dieser Leistung unterscheiden, in der Praxis sind aber Leistungsfeststellungen und Leistungsbeurteilung nicht immer voneinander zu trennen, weil es schwierig ist, eine Leistung zu messen, ohne sie in irgendeiner Form zugleich zu bewerten. [...] Man könnte jedoch auch versucht sein zu meinen, dass Leistungsfeststellung der objektive Teil des Beurteilungsprozesses sei, während in der anschließenden Beurteilung subjektive Momente zum Tragen kommen. Dies stimmt so nicht, denn schon die die Leistungsfeststellung enthält subjektive Komponenten. Besonders bei Schülerleistungen, die über eine bloße Reproduktion hinausgehen, können Lehrerurteile über den Grad der Richtigkeit der Schülerantworten voneinander abweichen (Neuweg, 2000, S. 1-4).“ (ebd., S. 96-97)

Insofern kann festgehalten werden, dass eine Trennung des Begriffspaars Leistungsfeststellung und -beurteilung nicht missverstanden werden darf, als eine Gliederung in zwei streng voneinander getrennte Teilschritte. Vielmehr werden durch beide Begriffe unterschiedliche

Aspekte fokussiert, die im Allgemeinen große Überschneidungen aufweisen können und in manchen Feststellungs- und Beurteilungssituation kaum voneinander zu unterscheiden sind. Aus diesem Grund sind die Begriffspaare in Abbildung 1.2 auch mit Doppelpfeilen verbunden.

1.1.2.4. Zusammenfassung und Begriffsfestlegung

Es lässt sich festhalten, dass die Terminologie in der erziehungswissenschaftlichen Literatur zum Themenkomplex der Leistungsfeststellung und -beurteilung als überaus heterogen charakterisiert werden muss. Welcher Oberbegriff verwendet wird und was die zugehörigen Unterbegriffe sind unterscheidet sich von Autor_in zu Autorin_in. Die verschiedenen Begrifflichkeiten lassen sich allerdings in einem groben Raster bezüglich der zeitlichen Stellung im Lehr-Lern-Prozess und bezogen auf den Feststellungs- und Beurteilungsprozesses selbst ordnen. Hierdurch werden gemeinsame Strukturen sichtbar, die den unterschiedlichen Terminologien zugrunde liegen, aber auch, dass die Übergänge innerhalb dieser Strukturen fließend sind und zum Teil verschwimmen. Ferner sind die beiden strukturgebenden Rasterdimensionen („zeitlichen Stellung im Lehr-Lern-Prozess“ und „Prozess Leistungsfeststellung- und -beurteilung selbst“) konkreten Formen der Leistungsfeststellung und -beurteilung übergeordnet. Unterschiedliche Durchführungsformen sind damit lediglich ein oberflächliches und daher kein wesentliches Charakterisierungsmerkmal für die Facetten des Begriffspaars Leistungsfeststellung und -beurteilung.

Des Weiteren ist deutlich geworden, dass verschiedene Autor_innen zum Teil unterschiedliches meinen, obwohl sie dieselben Begriffe verwenden. Beispielsweise ist mit „Leistungsbewertung“ bei Fischer & Malle (1989, S. 305) die Bewertung einer Schülerleistung durch Vergleich mit einem Gütemaßstab gemeint, wohingegen Kühberger (2014, S. 11) diesen Begriff definiert als Leistungsfeststellung und -beurteilung im laufenden Lehr-Lern-Prozess. Um Missverständnissen vorzubeugen werden daher für den weiteren Verlauf folgende Begriffsfestlegungen vorgenommen:

Begriffsfestlegung

1. Aus pragmatischen Gründen wird am Begriffspaar *Leistungsfeststellung und -beurteilung* als Oberbegriff für schulischen Verfahren, die dazu dienen, Leistungen von Schüler_innen einzuschätzen, festgehalten. Der in der englischsprachigen Literatur weit verbreitete Begriff *Assessment* wird als hierzu synonym festgelegt.

2. Bezüglich der zeitlichen Stellung von Leistungsfeststellung und -beurteilung in einem Lehr-Lern-Prozess wird sich im Folgenden an den Begriffen von Bloom, Hastings, & Madaus (1971, S. 14) und Dori & Avargil (2016, S. 1033) orientiert und diese sprachlich vereinheitlicht. Dementsprechend wird bei einer gedanklichen Zweiteilung eines Lehr-Lern-Prozesses von *Assessment for Learning* und *Assessment of Learning* gesprochen, bei einer Dreiteilung hingegen von *Initial*, *Formative* und *Summative Assessment*.
3. Bezogen auf den Feststellungs- und Beurteilungsprozesses selbst wird in Anlehnung an die Definition von Neuweg (2009, S. 3) zwischen *Leistungsfeststellung* und *Leistungsbeurteilung* unterschieden und die Begriffe *Leistungsmessung* bzw. *Leistungsbewertung* als hierzu entsprechende Synonyme definiert.

Für den weiteren Verlauf der vorliegenden Arbeit ist damit die grundlegende Terminologie des Diskurses um schulische Leistungsfeststellung und -beurteilung hinreichen aufgearbeitet. Das nun anschließende Unterkapitel widmet sich wie angekündigt einer Bestandsaufnahme über die unterschiedlichen Funktionen schulischer Leistungsfeststellungen und -beurteilungen.

1.2. Funktionen schulischer Leistungsfeststellungen und -beurteilungen

In der Literatur finden sich Anmerkungen zu Funktionen schulischer Leistungsfeststellungen und -beurteilungen meist nur beiläufig (z. B. Friege, 2017; Wodzinski, 2007; Duit & Häußler, 1997; Simon, 1979). Die Autor_innen weben ihre Funktionsbestimmungen oftmals in einen anderen, übergeordneten Aspekt ein, erwähnen daher nur eine Auswahl von Funktionen und bedienen sich dementsprechend einer eigenen Terminologie (vgl. Füller, 1975, S. 17). Im Folgenden soll eine möglichst umfassende Auflistung und nähere Bestimmung unterschiedlicher Funktionen schulischer Leistungsfeststellungen und -beurteilungen erfolgen. Der Begriff Funktionen umfasst hierbei sowohl Zielsetzungen, als auch Aufgabenstellungen und Wirkungserwartungen (vgl. Tillmann & Vollstädt, 2000, S. 29). Des Weiteren adressiert er Zusammenhänge von Leistungsfeststellungen und -beurteilungen mit den Zwecken der Institution Schule[9] und mit Zielen konkreter Lehr-Lern-Sequenzen.

[9] Nach Jung (1983) lassen sich die Zwecke der Institution Schule grob unterteilen in Sozialisation, Enkulturation und Individuation. „Sozialisation meint die Formung zum Gesellschaftswesen, Enkulturation die Einführung in die kulturellen Traditionen und gegenwärtigen Strömungen, Individuation meint die mit allem einhergehende und verbundene Entfaltung eines individuellen Charakters. [...] [Zwecke sind damit] Prozesse, die die Institution ingangsetz[t] und [erhält] [...] [und] unterscheiden sich von Zielen [...] dadurch, daß sie nicht einfach einen Zustand oder eine Qualität oder eine Qualifikation beschreiben, sondern all dies in einem umfassenden Zusammenhang, in einer Ganzheit.“ (ebd., S. 28)

Die Darstellung orientiert sich dabei an der Unterteilung von Füller (1975). Im weiteren Verlauf wird daher eine Gruppierung in vier Funktionsbereiche von schulischen Leistungsfeststellungen und -beurteilungen vorgenommen:

1. Die pädagogischen Funktionen schulischer Leistungsfeststellungen und -beurteilungen
2. Die psychologischen Funktionen schulischer Leistungsfeststellungen und -beurteilungen
3. Die Repräsentationsfunktionen schulischer Leistungsfeststellungen und -beurteilungen
4. Die sozialen Funktionen schulischer Leistungsfeststellungen und -beurteilungen

Anzumerken ist dabei, dass diese Bereiche nicht trennscharf sind. Sie weisen zum Teil inhaltliche Ähnlichkeiten auf bzw. können einander überschneiden.

1.2.1. Pädagogische Funktionen schulischer Leistungsfeststellungen und -beurteilungen

Die pädagogischen Funktionen von Leistungsfeststellungen und -beurteilungen widmen sich der Aufgabe diagnostische oder prognostische Rückmeldungen für alle Personen zur Verfügung zu stellen, die direkt oder indirekt am Unterricht beteiligt sind. Dies sind Schüler_innen, Lehrer_innen und Erziehungsberechtigte. Für jede dieser Personengruppen ergeben sich wiederum individuelle pädagogische Funktionen von schulischen Leistungsfeststellungen und -beurteilungen. Was all diesen Funktionen gemein ist, ist dass bei ihnen die Optimierung von Lehr-Lern-Prozessen im Vordergrund steht:

- Für Lehrer_innen liefern Leistungsfeststellungen und -beurteilungen Informationen, um das Erreichen curricularer Vorgaben zu kontrollieren und geben Aufschluss über Erfolg bzw. Misserfolg der eigenen Unterrichtskonzeption. Befunde über die Leistungen von Schüler_innen dienen also in erste Linie dazu, die im eigenen Unterricht eingesetzten Inhalte, Methoden, Medien usw. kritisch zu hinterfragen, um ggf. Anpassungen vorzunehmen, sowie Förder- oder Differenzierungsangebote zu optimieren (vgl. Tillmann & Vollstädt, 2000, S. 30). Ferner liefern sie die diagnostische und prognostische Grundlage für Lernberatungsgespräche mit Schüler_innen und/oder deren Erziehungsberechtigten, sowie für Absprachen im Kollegenkreis (vgl. Häußler, Bünder, Duit, Gräber, & Mayer, 1998, S. 67 u. f.; Friege, 2017, S. 3).
- Für Schüler_innen sind Leistungsfeststellungen und -beurteilungen im pädagogischen Sinn als Rückmeldungen zu verstehen, die über das eigene Lernen informieren. Im Idealfall liefern sie ihm_ihr Aufschluss über seine_ihre Leistungen im Vergleich zu gesetzten Lernzielen, Normen und/oder anderen Schüler_innen (vgl. Füller, 1975, S. 18). Für Schüler_innen fungieren sie daher dazu, bisheriges Lernen zu kontrollieren und als Ausgangsbasis um zukünftiges Lernen effektiv gestalten zu können (vgl. Tillmann & Vollstädt, 2000, S. 30). Daneben stellen Leistungsfeststellungen und -beurteilungen für Schüler_innen eine Konfrontation mit dem Prinzip der Leis-

tungsgerechtigkeit dar und sind daher Lerngelegenheiten, die ihnen eine Auseinandersetzung mit diesem Prinzip ermöglichen (vgl. Salzmann, 1971, S. 357 u. f.).

- Für Erziehungsberechtige haben Leistungsfeststellungen und -beurteilungen die Funktion der Berichtserstattung. Sie sollen „Aufschluss über die Lernsituation ihrer Kinder geben, damit sie notwendige Hilfestellungen geben und die richtigen Bildungsentscheidungen treffen können“ (Füller, 1975, S. 18).

1.2.2. Psychologische Funktionen schulischer Leistungsfeststellungen und -beurteilungen

Anders als bei den pädagogischen Funktionen steht bei den psychologischen Funktionen nicht die Optimierung von Lehr-Lern-Prozessen unmittelbar im Vordergrund. Stattdessen werden Bereitschaften, Empfindungen und Verhaltensweisen von Schüler_innen fokussiert, denen ein positiver Wert beigemessen wird und die jeweils in bestimmter Art und Weise mit schulischen Leistungsfeststellungen und -beurteilungen im Zusammenhang stehen. Der Begriff „positiver Wert“ bezieht sich hier auf den Zusammenhang mit Zwecken der Institution Schule und/oder Zielen konkreter Lehr-Lern-Situationen. Man könnte hier auch von einem Lernen im erweiterten Sinn sprechen (vgl. Abschnitt 1.1.1). Unbestritten gibt es auch Konstrukte und Konzepte, die Problematiken im Kontext schulischer Leistungsfeststellungen und -beurteilungen aufzeigen, z. B. die in pädagogisch-psychologischen Kontexten diskutierte Prüfungsangst (vgl. Fehm & Fydrich, 2011). Es soll nicht verheimlicht werden, dass diese schulische Leistungsfeststellungen und -beurteilungen berechtigterweise kritik- bzw. fragwürdig erscheinen lassen. Derartige Überlegungen werden an dieser Stelle allerdings weitgehend ausgespart, da sich aus ihnen im Sinne der hier gewählten Terminologie eher Fehlfunktionen als Funktionen schulischer Leistungsfeststellungen und -beurteilungen ableiten lassen.

Beschrieben und erklärt werden Bereitschaften, Empfindungen und Verhaltensweisen von Schüler_innen durch entsprechende psychologische Konstrukte, z. B. das der Leistungsmotivation. Der positiv gedachte Zusammenhang zwischen schulischen Leistungsfeststellungen und -beurteilungen und dem jeweiligen Konstrukt bestimmt dann eine psychologische Funktion schulischer Leistungsfeststellungen und -beurteilungen im Sinne einer Wirkungserwartung. Ein solcher positiver Zusammenhang kann für eine Vielzahl von Konstrukten vor allem aus der pädagogischen Psychologie konstatiert werden. Diese können an dieser Stelle allerdings weder alle genannt, noch detailliert dargestellt werden. Stattdessen werden anhand einer Auswahl prominenter Beispiele psychologische Funktionen schulischer Leistungsfeststellungen und -beurteilungen grob skizziert bzw. veranschaulicht:

- Leistungsfeststellungen und -beurteilungen können Schüler_innen zu weiterer Auseinandersetzung mit einem Lerngegenstand disziplinieren und/oder eine vertiefte Auseinandersetzung anregen (vgl. Tillmann & Vollstädt, 2000, S. 30). Sie haben also die Funktion positiv auf die Lernmotivation und -volition zu wirken.

- Man geht heute davon aus, dass das schulische Selbstkonzept von Schüler_innen nicht nur ihre schulischen Leistungen beeinflusst, sondern dass auch umgekehrt schulische Leistungen auf diesen Teil ihres Selbstkonzepts rückwirken (vgl. Helmke & van Anken, 1995). Schulische Leistungsfeststellungen und -beurteilungen nehmen daher die Funktion wahr, zur Entwicklung des schulischen Selbstkonzepts der Schüler_innen beizutragen.
- „Die subjektive Könnenserfahrung in der Prüfung verleiht [...] [schulischen Leistungsfeststellungen und -beurteilungen] eine bedeutsame psychische Entlastungsfunktion. Die von außen kommende Anerkennung und Sanktionierung des Lernfortschritts führt zu subjektiver Entlastung und macht den Lernenden frei für weiterführende, neue Selbstbeanspruchung“ (Füller, 1975, S. 21).
- Schulische Leistungsfeststellungen und -beurteilungen konfrontieren Schüler_innen mit der Erfahrung des Könnens und der des Nichtkönnens und regen diese damit zu subjektiven Ursachenzuschreibungen an (vgl. Weiner, 1985). Für Schüler_innen bieten Leistungsfeststellungen und -beurteilungen daher eine Gelegenheit sowohl angemessenes, als auch selbstwertdienliches Zuschreiben von Ursachen zu üben. Kurz gesagt haben sie damit die Funktion zum Erwerb von Attributionsstilen beizutragen und diese zu festigen.

1.2.3. Repräsentationsfunktionen schulischer Leistungsfeststellungen und -beurteilungen

Die Abiturprüfungen in unterschiedlichen Fächern oder eine Klassenarbeit zum Themengebiet der Wärmelehre in Jahrgangsstufe 6 sind Beispiele für Instrumente zur schulischen Leistungsfeststellungen und -beurteilungen. Diese dienen aber nicht nur dazu, ein summatives Abbild des Lernstandes eines_einer Schüler_in zu einem bestimmten Zeitpunkt oder über einen gewissen Zeitraum zu generieren, sondern stehen auch stellvertretend für die Anforderungen eines bestimmten Schulfachs oder der Schule als gesellschaftliche Institution. Daher weist Füller (1975) schulischen Leistungsfeststellungen und -beurteilungen in Form von Prüfungen die Funktion der Repräsentation zu und begründet dies folgendermaßen:

> „[D]ie Prüfung steht in ihrer konkreten Form und ihren speziellen Ansprüchen stellvertretend für die Inhalte eines wissenschaftlichen Gebiets [...] [oder kann] auch Institutionen vertreten. So repräsentiert beispielsweise das Abitur die Institution des Gymnasiums. Das Abitur umfaßt die Sachgebiete aller Gymnasialfächer und ist Ausdruck eines bestimmten Anspruchsniveaus, nämlich der Hochschulreife.“ (ebd., S. 21-22)

1.2.4. Soziale Funktionen schulischer Leistungsfeststellungen und -beurteilungen

Leistungsfeststellungen und -beurteilungen im schulischen Kontext dienen dazu, soziale Beziehungen und Systeme, die die Gesellschaft konstituieren, zu reproduzieren, zu stabilisieren und Heranwachsende in diese zu integrieren (vgl. Capelle, 1969, S. 258; Fend, 1980, S. 15 u. f.; Jung, 1983, S. 26 u. f.). Ergebnisse von Leistungsfeststellungen und -beurteilungen (z. B. die Durchschnittsnote des Abiturjahrgangs einer Schule) werden deshalb auch zum Zweck des Bildungsmonitoring oder der Schulevaluation herangezogen. Sie geben in diesem Zusammenhang eine Rückmeldung darüber, wie gut im Vergleich gelernt wird (vgl. Friege, 2017, S.3) und können daher auch verstanden werden als „Mittel zur Verhaltens- und Erwartungsstabilisierung, deren sich eine Gruppe (Gesellschaft) bedient, um Übernahme und Internalisierung von Verhaltensweisen, Kulturinhalten und Normen durch ihre Mitgl[ieder] durchzusetzen“ (Salzmann, 1971, S. 357).

Dies führt gleichzeitig auf die zweite soziale Funktion schulischer Leistungsfeststellungen und -beurteilungen. Sie haben die Funktion der sozialen Kontrolle. Sie dienen als Berechtigungsnachweis und werden zur Selektion herangezogen. Insbesondere Zeugnisnoten als kondensiertes Abbild vergangener Leistungsfeststellungen und -beurteilungen werden dazu verwendet, den Zugang zu zukünftige Ausbildungswegen zu regulieren und rechtlich abzusichern (vgl. Häußler et al., 1998, S. 68). Eine solche Auslese findet aber nicht nur auf der Grundlage der bisherigen Leistungen der Schüler_innen statt, sondern wird auch von begrenzten gesellschaftlichen Kapazitäten bedingt z. B. in der Lehre oder am Arbeitsmarkt (vgl. Füller, 1975, S. 20). Im Zusammenhang mit der Selektionfunktion von Schule führt Fend bereits 1980 an, dass das Schulsystem „kein „Rüttelsieb“ [ist], das eine vollkommene Neuverteilung der Lebenschancen zwischen den Generationen vornimmt[,] [...] [sondern] einige schulimmanente und schulexterne Barrieren [existieren], die einer idealen Realisierung [von Bildungsgerechtigkeit] im Wege stehen“ (ebd., S. 37-39). Solche Barrieren bestehen bis heute, worauf im naturwissenschaftsdidaktischen Diskurs beispielsweise Tajmel (2017b, S. 125 u. f.) ausführlich hinweist. Auf der anderen Seite gilt es nach Füller (1975) in diesem Kontext folgendes mitzubedenken:

> „Man könnte zugestehen, daß Bewertung und Vergleich der Leistung von Lernenden, sowie die damit verbundenen Prognosen und Selektion weitgehend entbehrlich sind. [...] Wenn man Prüfungen als Institution abschaffen wollte, so würden deren soziale Funktionen sofort von anderen Instanzen der Gesellschaft übernommen werden. [...] [Insbesondere] würde dies dazu führen, daß andere der öffentlichen Kontrolle entzogene Rekrutierungskriterien an Bedeutung gewinnen würden.“ (ebd., S. 17-20)

Insgesamt war und ist damit eine kritische Auseinandersetzung mit Selektionsmechanismen im schulischen Kontext berechtigt, bezogen auf die Selektionsfunktion von schulischen Leistungsfeststellungen und -beurteilungen muss allerdings mit bedacht werden, dass die Gesellschaft an die Schule die Erwartung stellt „die «leistungsgerechte» Zuweisung für die nachschulischen Positionen vorzubereiten“ (Tillmann & Vollstädt, 2000, S. 30). Ge-

mäß dieser Erwartung müssen der Schule daher auch zu einem gewissen Grad Verfahren zugestanden werden, um der Erfüllung dieser Anforderung gerecht zu werden.

1.2.5. Zwischenfazit

In diesem Unterkapitel ist Folgendes deutlich geworden: An die Frage, warum in der Schule Leistungsfeststellungen und -beurteilungen überhaupt vorgenommen werden, lässt sich aus unterschiedlichen Perspektiven herangehen. Jede dieser Sichtweisen liefert allerdings immer nur eine spezifischen Teilantwort. Folglich lässt sich eine Vielzahl von Funktionen für Leistungsfeststellungen und -beurteilungen im schulischen Kontext konstatieren, die sich gegenseitig ergänzen oder miteinander konkurrieren (z. B. die pädagogischen Funktionen mit der Selektionsfunktion schulischer Leistungsfeststellungen und -beurteilungen), weswegen erst eine Vielfalt der Perspektiven dieser komplexen Frage gerecht wird. Es verwundert daher auch nicht, dass es in der Literatur als strittig gilt, ob schulische Leistungsfeststellungen und -beurteilungen all diesen Funktionen nachkommen können (Tillmann & Vollstädt, 2000, S. 30). Skepsis äußert z. B. Ingenkamp (1995), indem er Zensurengebung als „fragwürdig“ betitelt und an andere Stelle Funktionen schulischer Leistungsfeststellungen und -beurteilungen wie folgt kommentiert:

> „Manche dieser Funktionen sind kaum miteinander vereinbar, und es ist schwer verständlich, wie man glauben konnte, die Zensur könne so unterschiedliche Aufgaben gleichzeitig erfüllen.“ (Ingenkamp, 1985, S. 177)

Bezogen auf das Handeln von Lehrkräften lässt sich an dieser Stelle festhalten, dass sie es sind, die Leistungsmessungen und -bewertungen innerhalb der Institution Schule vornehmen. Der Großteil der höchst unterschiedlichen Zielsetzungen, Aufgabenstellungen und Wirkungserwartungen an schulische Leistungsfeststellungen und -beurteilungen rahmen und bedingen folglich ihr Berufshandeln in diesem Kontext. Auf diesen Gedanken wird in Kapitel 2 noch vertieft eingegangen. Zunächst gilt es allerdings noch, wie zu Beginn dieses Kapitels angekündigt, die Frage nach der Güte schulischer Leistungsfeststellungen und -beurteilungen zu klären.

1.3. Güte schulischer Leistungsfeststellungen und -beurteilungen

1.3.1. Die Gütekriterien der Testtheorie im Kontext schulischer Leistungsfeststellungen und -beurteilungen

Hinter dem Begriff der Güte steckt die Idee Leistungsfeststellungen und -beurteilungen als eine Form der „Messung“ zu konzipieren. Dass dies zunächst plausibel ist, macht bereits Wagenschein (1970b) deutlich, indem er eine gedankliche Verbindung zwischen schulischer Zensurengebung und der Messverfahrens-Idee der Physik herstellt:

> „Niemand wird anfechten, daß eine Zahlenskala angebracht ist, um die Temperatur des Raumes zu ordnen, oder um die Menschen einer Stadt nach ihrer Körperlänge zu sortieren. Schwieriger wird es schon, wollte man sie ordnen nach ihrer (auch nur körperlichen) „Leistungsfähigkeit“, weil die ja wieder von mehreren Faktoren abhängt. – Ein physikalisches Beispiel scheint zu zeigen, daß auch in solchen Fällen (einer „Funktion von mehreren Veränderlichen“) eine „Gesamtnote“ möglich ist. Ein Gummiballon, gefüllt mit dicht gepreßtem, aber kaltem Gas, wird denselben Druck äußern können wie ein gleicher Ballon, in dem das Gas dünn, aber heißt ist.“ (ebd., S. 264)

Folgt man diesem Gedanken, so ist es legitim an schulischen Leistungsfeststellungen und -beurteilungen dieselben Qualitätsanforderungen zu stellen, wie sie auch bei anderen Messverfahren gelten. Eine Vielzahl von Autor_innen (z. B. Sacher, 1996, S. 24 u. f.; Jürgens, 1997, S. 61 u. f.; Kühberger, 2014, S. 10) teilt daher die Ansicht, dass bei schulischen Leistungsfeststellungen und -beurteilungen, welcher Form auch immer, die Gütekriterien der Testtheorie zu beachten sind. Meist wird dabei auf die von Lienert (1967, S. 12 u. f.) vorgeschlagenen Hauptgütekriterien *Objektivität*, *Reliabilität* und *Validität* zurückgegriffen[10]. Diese Gütekriterien beschreiben „das Ausmaß, in dem durch die Anwendung eines bestimmten Meßverfahrens gewonnenen Meßwerte mit Fehlern behaftet sind“ (Hartmann, 1991, S. 20). Die Genauigkeit einer konkreten Messung (hier einer Leistungsfeststellung und -beurteilung eines_einer einzelnen Schülers_Schülerin) drückt sich hingegen im „Messfehler“ aus, der beschreibt, inwieweit der gemessene Wert von einem erwarteten/postulierten Wert abweicht (vgl. ebd.).

Zum Zweck einer weitere Auseinandersetzung sollen die drei Hauptgütekriterien kurz vorgestellt werden. Die Darstellung orientiert sich dabei an jener, die sich in diversen Werken zur Testtheorie finden lässt, insbesondere an den Lehrbüchern von Lienert (1967) und Döring & Bortz (2016). Die Kurzcharakteristika der Gütekriterien werden dabei auf auf den Kontext schulischer Leistungsfeststellungen und -beurteilungen entsprechend angepasst:

1.3.1.1. Objektivität

Unter Objektivität versteht man den Grad, bis zu dem das Ergebnis schulischer Leistungsfeststellungen und -beurteilungen unabhängig von der begutachtenden Lehrkräften ist (vgl. Lienert, 1967, S. 13). Dementsprechend würde vollkommene Objektivität genau dann vorliegen, wenn verschiedene Lehrkräfte bei demselben_derselben Schüler_in zu gleichen Leistungsfeststellungen und -beurteilungen kommen würden (vgl. ebd.). Objektivität ist damit ein formales Kriterium, das nichts über die Inhalte, die ein Verfahren zur Leistungsfeststellung und -beurteilung abprüft, aussagt (vgl. ebd.). Ferner lassen sich zwei Subformen unterscheiden:

[10] In der Testtheorie werden neben diesen drei Hauptgütekriterien auch sog. Nebengütekriterien genannt. Ein ausführlichen Überblick über diese findet sich bspw. bei Döring & Bortz (2016, S. 449 u. f.). Die beiden Autor_innen unterscheiden insgesamt sieben Nebengütekriterien, namentlich die *Skalierbarkeit*, die *Normierung*, die *Ökonomie*, die *Nützlichkeit*, die *Zumutbarkeit*, die *Nicht-Verfälschbarkeit* und die *Fairness* eines Verfahrens (vgl. ebd.).

- **Durchführungsobjektivität**:
  Bei der Durchführungsobjektivität steht die Erhebung in einer Leistungsfeststellungs- und -beurteilungssituation im Vordergrund. Sie gibt an, inwieweit Leistungsfeststellungen und -beurteilungen unabhängig davon sind, welche Lehrkraft die Erhebungen im Rahmen der Leistungsfeststellungen vorgenommen hat (vgl. Döring & Bortz, 2016, S. 443). Hergestellt werden kann dies beispielsweise durch Vereinheitlichung von Verfahrensregeln, Instruktionen, erlaubten Hilfsmitteln, usw. (vgl. ebd.).
- **Auswertungs- und Interpretationsobjektivität:**
  Perfekte Auswertungs- und/oder Interpretationsobjektivität liegt genau dann vor, wenn Leistungsfeststellungen und -beurteilungen unabhängig davon sind, welche Lehrkraft die Auswertung bzw. Interpretation der Schülerleistungen vornimmt (vgl. ebd.). Wie in Abschnitt 1.1.2 bereits erwähnt, ist bei schulischen Leistungsfeststellungen und -beurteilungen, im Gegensatz zu standardisierten Testverfahren, eine Trennung in Auswertung und Interpretation nicht immer oder nur schwer möglich. Daher sind hier, im Gegensatz zur der Darstellung wie sie zumeist in der testtheoretischen Literatur zu findest ist, Auswertungs- und Interpretationsobjektivität zusammengefasst.

1.3.1.2. Reliabilität

Ebenso wie die Objektivität ist auch die Reliabilität ein formales Gütekriterium, das nichts über die das entsprechende Leistungsfeststellungs- und -beurteilungsverfahren umfassenden Inhalte aussagt (vgl. Lienert, 1967, S. 14 u. f.). Reliabilität, setzt allerdings Objektivität voraus (vgl. ebd.). Unter ihr versteht man den Grad an Exaktheit mit dem bestimmte Leistungen von Schüler_innen erfasst werden (vgl. ebd.). Dabei ist zunächst gleichgültig, ob das gewählte Instrument zur Leistungsfeststellung und -beurteilung für sich beanspruchen kann, die entsprechende Leistung auch tatsächlich zu messen (vgl. ebd.). Dies ist eine Frage der Validität, auf die im weiteren Verlauf noch eingegangen wird. Ein einfaches Beispiel soll dies verdeutlichen: Man vermisst die Durchmesser von Äpfeln mit Hilfe eines Lineals. Unabhängig davon, wie oft man die Äpfel verfahrenstechnisch korrekt vermisst, ist dieses Verfahren nur bis zu einem bestimmten Grad exakt, da ein Lineal üblicherweise höchsten über eine Millimeterskala verfügt. Erhöhen lässt sich die Exaktheit des Verfahrens allerdings dadurch, dass man anstatt eines Lineals beispielsweise einen Messschieber mit Nonius verwendet. Hierdurch wird die Messgenauigkeit vom Millimeter- auf den Zehntelmillimeterbereich erhöht bzw. die Reliabilität des Messverfahrens nimmt zu. Durch Verwendung weiterer, noch exaktere Längenmessinstrumente könnte dies noch gesteigert werden, dennoch wäre keines dieses Messverfahren valide, wenn man beispielsweise behauptet mit diesen die Färbung der Äpfel bestimmen zu können.

Auch bei der Reliabilität werden in der Literatur Subformen unterschieden, die gleichzeitig Verfahrensweisen beschreiben, mit denen die Reliabilität bestimmt werden kann. Meist

wird dabei zwischen der Paralleltest-, der Re-Test- und der Split-Half-Reliabiltät, sowie der inneren Konsistenz unterschieden (vgl. Döring & Bortz, 2016, S. 444). Nach Sacher (1996, S. 25) ist deren Übertragbarkeit auf den Kontext von schulischen Leistungsfeststellungen und -beurteilungen allerdings überaus beschränkt. Hierauf wird im weiteren Verlauf dieses Unterkapitels noch eingegangen, weswegen auf eine weitere Ausführung der eben genannten Subformen an dieser Stelle verzichtet wird.

1.3.1.3. Validität

Frei formuliert gibt die Validität an, ob ein Verfahren zur Leistungsfeststellung und -beurteilung gültig ist. Nach Lienert (1967) ist dies für ein Verfahren genau dann vollkommen erfüllt, „wenn seine Ergebnisse einen unmittelbaren und fehlerfreien Rückschluß auf den Ausprägungsgrad des zu erfassenden [...] [Merkmals] zulassen“ (ebd., S. 16). Eine hierfür notwendige, aber nicht hinreichende Voraussetzung ist, dass das zu validierende Verfahren reliabel und damit auch objektiv ist (vgl. ebd.). Für schulische Leistungsfeststellungen und -beurteilungen gibt Sacher (1996) vier Aspekte bzw. Subformen von Validität, die aus seiner Sicht in diesem Kontext zu nennen sind:

- **inhaltliche Validität**: Ein Verfahren zur schulischen Leistungsfeststellung und -beurteilung ist inhaltsvalide, wenn es Leistungen erfasst, zu denen die Schüler_innen aufgrund des bisherigen Unterrichts auch in der Lage sind (vgl. ebd., S. 27). Hierzu zählt auch die sog. curriculare Validität, also inwieweit die vom Verfahren geprüften Inhalte sich mit denen des entsprechenden Lehrplans decken (vgl. ebd., S. 28).
- **prognostische Validität**: „*Prognosevalidität* ist dann gegeben, wenn man aus den Meßergebnissen zutreffende Schlüsse auf Ergebnisse zukünftiger Messungen vornehmen kann“ (ebd., S. 29, Hervorhebung im Original). Da sich das Begriffspaar Leistungsfeststellung und -beurteilung in schulischen Kontexten auf das Lernen von Schüler_innen bezieht (vgl. Abschnitt 1.1.1) sind mit zutreffenden Schlüssen also Prognosen auf zukünftiges Lernen gemeint. Eine unmittelbare Konsequenz hieraus ist, dass der prognostischen Validität eine besonders große Bedeutung beizumessen ist, wenn ein Leistungsfeststellungs- und -beurteilungsverfahren zum Zweck eines Assessment for Learning eingesetzt wird (vgl. Abschnitt 1.1.2).
- **Konstruktvalidität**: Allgemein gibt Konstruktvalidität an, inwieweit ein Verfahren mit den Erwartungen gemäß eines theoretischen Modells (Konstrukts) übereinstimmen (vgl. Döring & Bortz, 2016, S. 446). Sacher (1996) schreibt diesem Aspekt von Validität eine wenig bedeutende Rolle für schulische Leistungsfeststellung und -beurteilung zu und begründet dies damit, dass es „noch kaum elaborierte Modelle für Komponenten von Schulleistungen gibt“ (ebd., S. 30). Allerdings wurde seine Arbeit im Jahr 1996 veröffentlich und damit zeitlich vor der sog. „empirischen Wende“ im Rahmen der Kompetenzdebatte ab dem Jahr 2001 (vgl. Tajmel, 2017b, S. 99 u. f.). Aus heutiger Sicht ist daher Sachers Aussage zumindest dahingehend zu relativie-

ren, als dass die Bedeutung von Konstruktvalidität für den Kontext schulischer Leistungsfeststellungen und -beurteilungen seit dem Beginn der Kompetenzdebatte zugenommen hat.

- **Übereinstimmungsgültigkeit**: Diese liegt vor, wenn verschiedene Verfahren zur Leistungsfeststellung und -beurteilung zu übereinstimmenden Resultaten kommen (vgl. Sacher, 1996, S. 30). In der Testtheorie spricht man auch von konvergenter Validität und meint damit, dass die Ergebnisse, die das zu validierende Verfahren liefert, im Einklang stehen sollten mit denen verwandten Verfahren, deren Validität bereits als gesichert gilt (vgl. Döring & Bortz, 2016, S. 446).

1.3.2. Erkenntnisstand zur Güte schulischer Leistungsfeststellungen und -beurteilungen und sich hieraus ergebende Konsequenzen

Die eben genannten drei Hauptgütekriterien sind vor allem seit den 1970er-Jahren in einer Vielzahl von Studien zur schulischen Zensurgebungspraxis zum Untersuchungsgegenstand erhoben worden[11]. Jürgens (1997) kommt auf der Grundlage eines detaillierten Überblicks verschiedenster Forschungsergebnisse zu folgendem Resümee:

> „Zusammenfassend kann auf der Grundlage der genannten und weiteren Untersuchungen gesagt werden, daß die Zensuren weder im außreichendem Maße Objektivität [...] noch Reliabilität und Validität beanspruchen können[.] [...] Der Prozeß der Notengebung muß damit als Schätzverfahren und die Zensur als subjektives Schätzurteil auf vorwissenschaftlichem Niveau bezeichnet werden [...][.]“ (ebd., S. 59)

Ähnlich wie auch viele andere Autor_innen (z. B. Sacher, 1996, S. 31 u. f.; Holmeier, 2013, S. 113 u. f.) kommt Jürgens zu einem vernichtenden Gesamturteil über die schulische Zensurgebungspraxis. In diesem Sinne sind auch Beutel & Vollstädt (2000) zu verstehen, wenn sie von einer „hinlänglich bekannten Problem[lage]“ sprechen, die im erziehungswissenschaftlichen Diskurs bereits mehrfach einen breit angelegten Meinungsstreit provoziert hat (vgl. ebd., S. 13). Dieser Meinungsstreit soll anhand einer Auswahl von Beispielpositionen skizziert werden:

So stellt sich für Wagenschein (1970b) lediglich schulische Zensurenvergabe als „grotesk“ dar, da sich hinter ihr die „Wahnidee“ verbirgt, „alles müsse sich in Zahlen einfangen lassen“ (ebd., S. 264), wohingegen seiner Ansicht nach Leistungsfeststellungen und -beurteilungen an sich „nicht zur Debatte [stehen, da] [j]eder Lernende [...] ein Anrecht auf ein Urteil des Lehrers [hat]“ (ebd., S. 263). Wagenscheins Aussage steht damit im Einklang mit einer reformpädagogisch orientierten Kritik an Zensuren, die „den meßtheoretischen Anspruch und die Standardisierung der Leitungsüberprüfung und -bewertung ab[lehnt]“ (Tillmann & Vollstädt, 2000, S. 32) und stattdessen individuelle Lernberichte befürwortet, die „nur innerhalb des jeweiligen sozialen Kontexts verstanden werden können [...] [und damit] für «Außenstehende» unverständlich bleiben müssen“ (ebd.). Eine solche Positi-

[11] Für einen aktuellen Gesamtüberblick der Forschung zur Zensurgebung der letzten 100 Jahre siehe Brookhart et al. (2016).

on ist allerdings selbst wiederum kritikwürdig, da sie die sozialen Funktionen schulischer Leistungsfeststellungen und -beurteilungen (vgl. Unterkapitel 1.2), die eine gewisse Nachvollziehbarkeit von Schülerleistungen auch für Außenstehende erforderlich machen, „nicht in den Blick [nimmt] [...] oder gar als illegitim zurück[weist]“ (ebd., S. 35).

Im Gegensatz dazu findet sich bei Holmeier (2013) keine derartige Trennung der Zensurengebung vom „Urteil der Lehrkraft“ (Wagenschein, 1970b, S. 264). Für die Autorin stellt sich die „Fragwürdigkeit der Zensurengebung“ (Ingenkamp, 1995) eher als Anlass dar, sich mit der „sehr komplexen Natur“ (Holmeier, 2013, S. 126) schulischer Leistungsfeststellungen und -beurteilungen und den sich hieraus ergebenden Problemlagen auseinanderzusetzen.

Eine dritte Position findet sich bei Terhart (2000), der das Urteil einer Mangelhaftigkeit schulischer Leistungsfeststellungen und -beurteilungen bezogen auf ihre Güte an sich infrage stellt:

> „Schulische Beurteilungssituationen sind mithin Problemlagen, in denen es eben nicht die *eine* richtige Lösung gibt; zumindest ist diese eine richtige Lösung nicht wirklich identifizierbar. Nur in der Perspektive von Testtheorie bzw. Pädagogischer Diagnostik erscheint diese gelebte Praxis u. U. als skandalös defizitär. [...] Lehrkräfte wissen um den ‹weichen›, kontextbezogenen Charakter der Notengebung und haben mit ihm zu leben gelernt, ja bestätigen und verlängern diese Praxis täglich neu; zum Teil wird dieser Charakter auch offensiv verteidigt.“ (ebd., S. 42, Hervorhebungen im Original)

Was an diesen drei Beispielen deutlich wird, ist, dass die Meinungen verschiedener Autor_innen zum Teil weit auseinandergehen, inwiefern eine Kritik an schulischen Leistungsfeststellungen und -beurteilungen (inklusive der Zensurengebung) überhaupt adäquat ist und inwieweit testtheoretische Gütemängel von Zensuren eine allgemeine Fragwürdigkeit von schulischen Leistungsfeststellungen und -beurteilungen impliziert. Dementsprechend gibt es auch unterschiedliche Positionen, wenn es darum geht aus der genannten Befundlage Konsequenzen abzuleiten. Im erziehungswissenschaftlichen Diskurs lassen sich hierzu drei Strömungen ausfindig machen:

1.3.2.1. Konsequenz einer stärkeren Vereinheitlichung schulischer Leistungsfeststellungen und -beurteilungen

Zum einen gibt es den Vorschlag schulische Leistungsfeststellungen und -beurteilungen zu vereinheitlichen bzw. stärker zu standardisieren, um so den drei Hauptgütekriterien der Testtheorie besser zu genügen (vgl. Tillmann & Vollstädt, 2000, S. 32). Als Vertreterin eines solchen Ansatzes kann Holmeier (2013) gelten. Ihre Überlegungen sind in Tabelle 1.2 zusammengefasst dargestellt. Für sie steht im Gegensatz zu Wagenschein (1970b) keine Generalkritik, sondern die Verbesserung der Notenvergabe, die sie als Teil schulischer Leistungsfeststellungs- und -beurteilungspraxis konzipiert, im Vordergrund. Hierzu benennt die Autorin mit Verweis auf die Überlegungen von Sacher (2009, S. 85) komplexitätserzeugende Aspekte, von denen anzunehmen ist, dass diese zu einem gewissen Anteil bestimmen, „dass Noten den Gütekriterien nur in einem unbefriedigenden Maß gerecht

Mögliche Quellen mangelnder Güte schulischer Leistungsbegutachtung	Mutmaßlich komplexitätsverringende Aspekte durch Vereinheitlichung von Leistungs-begutachtungen (z. B. zentrale Prüfungen)
• Die Lehrer_innen sind an der Herstellung der zu messenden Größen beteiligt. • Die Lehrer_innen bestimmen, was gemessen wird. • Die Lehrer_innen entwickeln Instrumente zur Leistungsbegutachtung selbst. • Die Leistungsfeststellung wird von den Lehrer_innen selbst durchgeführt. • Die Leistungsbeurteilung wird von den Lehrer_innen selbst durchgeführt.	• Inhalte sind abgestimmt. • Die Instrumente zur Leistungsbegutachtung sind abgestimmt. • Vorgaben zur Erhebung von Schülerleistungen sind abgestimmt. • Vorgaben zur Leistungsfestsstellung und –beurteilung sind abgestimmt.

Tabelle 1.2.: Mögliche Quellen mangelnder Güte schulischer Leistungsfeststellung und -beurteilung und mutmaßlich komplexitätsverringernde Aspekte durch Vereinheitlichung von Leistungsfeststellung und -beurteilung. Sinngemäß übernommen und abgewandelt aus Holmeier (2013, S. 129).

werden“ (Holmeier, 2013, S. 126). Dementsprechend besteht ihr Ansatz für eine mögliche Verbesserung der Güte schulischer Leistungsfeststellungen und -beurteilungen darin, diese stärker zu vereinheitlichen und damit deren Komplexität zu reduzieren.

1.3.2.2. Konsequenz einer stärkeren Gewichtung anderer Gütekriterien

Einen anderen Ansatz verfolgen Leisen & Höttecke (2011), sowie Höttecke & Wodzinski (2015, S. 6 u. f.). Die Autor_innen plädieren dafür, im Kontext schulischer Leistungsfeststellungen und -beurteilungen nicht primär die Gütekriterien der Testtheorie zurückzugreifen. Stattdessen bedarf es vordergründig eines Katalogs alternativer Gütemerkmale, die sowohl theoretisch abgesichert, als auch an der Praxis bewährt sind (Leisen & Höttecke, 2011). Leisen & Höttecke (2011) betonen dabei die Bedeutung einer Förderorientierung von Lehrkräften und sehen daher in einer pädagogisch günstigen Voreingenommenheit ein alternatives Gütemaß für schulische Leistungsfeststellungen und -beurteilungen. Was hierunter zu verstehen ist, wird an dem folgenden Zitat von Weinert & Schrader (1986) deutlich:

> „Lehrerdiagnosen müssen sich nicht durch neutrale Objektivität sondern durch pädagogisch günstige Voreingenommenheit auszeichnen. [...] Unter den Belastungen des Unterrichts sind nämlich bei Lehrern situationsabhängige Erlebnisse, Urteile über andere und die Regulation eigener Handlungen keineswegs analytisch getrennt, sondern aufs engste miteinander verknüpft. Wenn dem aber so ist, dann erscheint es unter praktischen Gesichtspunkten günstig, wenn der Unterrichtende im Vergleich zu den „wahren Werten“ die Leistungsunterschiede zwischen den Schülern einer Klasse *mäßig* unterschätzt, die Leistungsfähigkeit der einzelnen Schüler *leicht* überschätzt, ihre Erfolge subjektiv durch Begabung und ihre Mißerfolge durch mangelnde Anstrengung oder ineffektiven Unterricht erklärt und Handlungsanreize erschließt. Der Lehrer wird sich unter diesen Voraussetzungen auch dann noch um Lernfortschritte bei den Schülern intensiv bemühen, wenn er aufgrund objektiver Diagnosen vielleicht längst resigniert hätte. Pädagogische Erfolge werden sich dadurch natürlich nicht immer, aber häufig einstellen, weil – ausreichende didaktische Kompetenz bei den Lehrern

> unterstellt – wahrscheinlich nichts so motivierend und erfolgreich ist wie eine leicht optimistische Erfolgserwartung. Als pädagogisch ungünstig müssen demgegenüber diagnostische Voreingenommenheit von Lehrern angesehen werden, die häufig zu einer Überschätzung der Leistungsdifferenzen in einer Klasse, zu einer Unterschätzung der individuellen Lernmöglichkeiten und zu einer subjektiven Erklärung durch Zufall oder besonderer Anstrengung führen." (ebd., S. 19-20, Hervorhebungen im Original)

Daneben benennen Leisen & Höttecke (2011) weitere Qualitätsmerkmale für schulische Leistungsfeststellung und -beurteilung, die allerdings nur grob umrissen werden (z. B. Kohärenz, Transparenz und Entflechtung von Lehr- und Leistungssituationen). Insgesamt steht die Optimierung von Lehr-Lern-Prozessen unter Berücksichtigung der Lernbiographie der einzelnen Schüler_innen im Vordergrund des Ansatzes von Leisen & Höttecke (2011), sowie Höttecke & Wodzinski (2015). Die Autor_innen betonen damit im besonderen Maße die pädagogischen Funktionen schulischer Leistungsfeststellungen und -beurteilungen (vgl. Unterkapitel 1.2). Zudem orientiert sich dieser Ansatz an der bereits skizzierten Position von Terhart (2000), was insbesondere daran deutlich wird, dass die Autor_innen besonders hervorheben, dass schulische Leistungsfeststellung und -beurteilung (inklusive Schulnoten) weder objektiv, reliabel noch valide sein müssen (vgl. Leisen & Höttecke, 2011, S. 63; Höttecke & Wodzinski, 2015, S. 7).

1.3.2.3. Konsequenz einer Classroommetric Measurement Theory

Als ein dritter Vorschlag können die Überlegungen von Brookhart (2003), Moss (2003) und J. K. Smith (2003) gelten, die im Special Issue „Changing the Way Measurement Theorists Think About Classroom Assessment" der Zeitschrift *Educational Measurement* erschienen sind. Zunächst merken die Autor_innen ähnlich wie Leisen & Höttecke (2011) und Höttecke & Wodzinski (2015, S. 6 u. f.) an, dass die klassischen Gütekriterien der Testtheorie auch im Kontext schulischer Leistungsfeststellungen und -beurteilungen bis zu einem gewissen Grad ihre Berechtigung haben (vgl. Brookhart, 2003, S. 11). Jedoch ist es ihrer Ansicht nach ungenügend, die in der Testtheorie anerkannte Überlegungen und Konzepte auf diesen Kontext lediglich anzuwenden bzw. zu übertragen (vgl. ebd.). Stattdessen gilt es eine „classroommetric measurement theory" zu entwickeln (vgl. ebd.), die sich an den Intentionen und Funktionen schulischer Leistungsfeststellungen und -beurteilungen orientieren (vgl. Abschnitt 1.1.2 und Unterkapitel 1.2). Das Fundament einer solchen Entwicklungsarbeit sehen die Autor_innen in einer Neukonzeption der Begriffe Reliabilität und Validität.

So werden beim Reliabilitätskonzept der Testtheorie oftmals Personengruppen betrachtet und das Ziel der Stabilisierung von Rangreihensortierung oder Kategorisierungen von Personen einer Gruppe steht im Vordergrund (vgl. Brookhart, 2003, S. 9). Bei alltäglichen schulischen Leistungsfeststellung und -beurteilung liegt der Fokus allerdings weniger auf Schülergruppen, sondern einzelnen Schüler_innen und deren individuellen Leistungen. Die üblichen testtheoretischen Reliablitätskonzepte sind damit also kaum gegenstandsadäquat. Des Weiteren sind Paralleltestreliablität, Re-Test-Reliabilität, usw. für den schulischen Kontext keine geeigneten Reliabilitätskonzepte, da hier Leistungsfeststellungen und

-beurteilungen oftmals einmalige und meist nicht wiederholbare Ereignisse im Unterrichts- bzw. Schuljahrsverlauf sind (vgl. J. K. Smith, 2003, S. 26 u. f.). J. K. Smith (2003) schlägt daher vor, Reliabilität auf Grundlage der pädagogischen Funktionen schulischer Leistungsfeststellungen und -beurteilungen aus Sicht der Lehrkräfte neu zu konzipieren. Aus dieser Perspektive ist die entscheidende Frage, die eine Reliabilitätsüberprüfung zu beantworten hat, die, ob die Menge an Informationen, die dem_der Lehrer_in zur Verfügung steht, hinreichend groß ist, um Entscheidungen, welche die Optimierung von Lehr-Lern-Prozessen betreffen, begründen und rechtfertigen zu können (vgl. Brookhart, 2003, S. 9; J. K. Smith, 2003, S. 30).

Für eine Rekonzeption des Begriffs der Validität schlägt Moss (2003) vor, eine soziokuturell-hermeneutische Perspektive komplementär zu jener der Testtheorie einzunehmen. Zwecks dessen lässt sich auch hier zunächst den an pädagogischen Funktionen schulischer Leistungsfeststellungen und -beurteilungen orientieren (vgl. Brookhart, 2003, S. 9). Gemäß diesen gilt es bei Leistungsfeststellungen und -beurteilungen zu einem Verständnis über den Ist-Stand der Schüler_innen in Relation zu gesetzten Lernzielen zu gelangen und die gewonnenen Informationen dazu zu nutzen, zukünftiges Lernen effektiver zu gestalten (vgl. ebd.). Schulische Leistungsfeststellungen und -beurteilungen sind demnach valide, wenn sie derartige Informationen liefern bzw. wenn sie Informationen bereitstellen, die für die Optimierung von Lehr-Lern-Prozessen dienlich sind. Hierzu gehört aber auch, inwieweit Schüler_innen bewusst ist, dass Leistungsfeststellungen und -beurteilungen eine für ihr individuelles Lernen nützliche Informationsquelle darstellen und inwieweit sie tatsächlich hieraus Nutzen zeihen (vgl. ebd.). Folglich kann erst dann von einem validen Verfahren zur schulischen Leistungsfeststellung und -beurteilung gesprochen werden, wenn dieses auch Informationen über die eben beschriebene Perspektive der Schüler_innen liefert. Ein weiterer Aspekt, der für einen sich am Kontext der Schule orientiert Validitätsbegriff zu beachten ist, ist, dass sich Leistungsfeststellung und -beurteilung, sowie deren nachfolgenden Konsequenzen hier, im Gegensatz zu beispielsweise einer Erhebung im Rahmen eines Large-Scale-Assessments, nur bedingt voneinander trennen lassen (vgl. auch Abschnitt 1.1.2). Dies hängt wiederum damit zusammen, dass...

... schulische Leistungsfeststellungen und -beurteilungen Teil eines laufenden Lehr-Lern-Prozess sind (vgl. ebd.).

... spezifischen Inhalte nicht nur eine Domäne repräsentieren können, sondern auch einen Lehr-Lern-Prozess selbst und die im Laufe dieses Prozesses stattfindenden Leistungsfeststellungen und -beurteilungen (vgl. ebd.).

... Leistungsfeststellungen und -beurteilungen zu einem erheblichen Anteil von den subjektiven Vorstellungen, Überzeugungen und Handlungen von Lehrkräfte bezogen auf Fachinhalte und auf Schüler_innen bestimmt bzw. beeinflusst werden (vgl. ebd.).

Brookhart (2003) fasst dies als Konstruktrelevanz des Erhebungskontextes zusammen. Folglich ist ein Verfahren zur schulischen Leistungsfeststellung und -beurteilung erst dann valide, wenn bei ihm die eben genannten Kontextaspekte ebenfalls mitberücksichtigt sind.

1.3.3. Zwischenfazit

Das vorangegangenen Unterkapitel hat sich auf Aspekte der Güte von Leistungsfeststellung und -beurteilung konzentriert. Hierbei wurde deutlich, dass der testtheoretischen Perspektive in der Literatur eine bedeutende Rolle zukommt. Einen Verweis auf die drei Hauptgütekriterien Objektivität, Reliabilität und Validität in adaptierter Form findet sich in de facto jeder Abhandlung über Güte schulischer Leistungsfeststellungen und -beurteilungen. Am Beispiel des Befunds Zensuren seien nur mangelhaft objektiv, reliabel und valide wurde ferner deutlich, dass sich mit der Güte schulischer Leistungsfeststellungen und -beurteilungen im erziehungswissenschaftlichen Diskurs bereits seit Langem auseinandergesetzt wird und zum Teil unterschiedliche Meinungen vertreten werden. Letzteres betrifft nicht nur die Frage danach, inwieweit die genannte Befundlage generalisierbar ist, sondern insbesondere auch welche Konsequenzen sich hieraus ableiten lassen. Das Meinungsspektrum reicht dabei von einer Forderung nach stärkerer Vereinheitlichung schulischer Leistungsfeststellungen und -beurteilungen, um eine Verbesserung im Sinne der klassischen Gütekriterien zu erreichen, bis hin zu der eine am am Alltagskontext der Schule orientierten Messtheorie zu entwickeln, bei der eine Orientierung an der Testtheorie – überspitzt ausgedrückt – als inadäquat abgelehnt wird. Die Forderung Objektivität, Reliabilität und Validität zugunsten alternativen Gütekriterien weniger stark in der Vordergrund zu rücken kann dabei als Mittelweg zwischen diesen beiden „Extrempositionen“ verstanden werden.

1.4. Zusammenfassung

In diesem Kapitel wurden Gedankengänge zu schulischen Leistungsfeststellungen und -beurteilungen unternommen, die für den weiteren Verlauf der vorliegenden Arbeit grundlegend sind.

Zu Beginn von Unterkapitel 1.1 wurde zunächst der Wortsinn des Begriffs Leistung im erziehungswissenschaftlichen Diskurs näher bestimmt (vgl. Abschnitt 1.1.1). Hierbei wurde deutlich, dass dieser Begriff im Allgemeinen sehr weit gefasst werden muss und sich ferner drei wesentliche Bestimmungsmerkmale identifizieren lassen, mit Hilfe derer er enger bzw. weiter gefasst werden kann. Hierauf aufbauend wurde in Abschnitt 1.1.2 die in der Literatur verwandte Terminologie zum Themenkomplex der schulischen Leistungsfeststellung und -beurteilung erörtert. Hierbei wurde sichtbar, dass sich diese Terminologie überaus heterogen gestaltet, sich aber in einem groben Raster bezüglich der zeitlichen Stellung einer Leistungsfeststellung und -beurteilung im Lehr-Lern-Prozess und bezogen auf den Feststellungs- und Beurteilungsprozess selbst ordnen lässt. Hierauf aufbauend konnte schließlich eine Begriffsfestlegung unternommen werden, die Missverständnisse im weiteren Verlauf der vorliegenden Arbeit vorbeugen wird.

In Unterkapitel 1.2 wurde sich der Frage gewidmet, warum in der Schule Leistungsfeststellungen und -beurteilungen überhaupt vorgenommen werden. Hierzu wurde eine um-

fangreiche Aufarbeitung verschiedener Funktionen schulischer Leistungsfeststellung und -beurteilung unternommen, die sich in Anlehnung an Füller (1975) in vier Funktionsbereichen gruppiert. Dabei zeigte sich, dass diese zum Teil sehr unterschiedlichen Funktionen sich nicht nur gegenseitig ergänzen, sondern dass viele von ihnen auch miteinander konkurrieren. Die Konsequenzen, die sich hieraus ergeben, werden in Kapitel 2 noch erörtert.

Im letzten Unterkapitel – Unterkapitel 1.3 – wurde sich schließlich dem Thema gewidmet, wie sich die Güte von schulischen Leistungsfeststellungen und -beurteilungen fassen lässt. Dabei zeigte sich, dass den drei Hauptgütekriterien der Testtheorie (Objektivität, Reliabilität und Validität) ein beachtliches Gewicht zukommt, da sie in nahezu jeder, Güteaspekte schulischer Leistungsfeststellung und -beurteilung behandelnden Ausarbeitung erwähnt werden. Allerdings gehen im erziehungswissenschaftlichen Diskus, wie insbesondere in Abschnitt 1.3.2 dargelegt, die Meinungen deutlich auseinander, inwieweit die Gütekriterien der Testtheorie dem Alltagskontext von schulischer Leistungsfeststellung und -beurteilung gerecht werden. In diesem Spektrum unterschiedlicher Meinungen findet sich jedoch die Gemeinsamkeit, dass sie alle von einer tragenden Rolle von Lehrkräften bei der schulischer Leistungsfeststellung und -beurteilung ausgehen. Auch dieser Gedanke wird im sich nun anschließenden Kapitel 2 vertieft.

2. Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen

Nachdem in Kapitel 1 grundlegende Begrifflichkeiten geklärt und Überlegungen unternommen wurden, sollen nun Lehrerwissen und -können[12] zu schulischen Leistungsfeststellungen und -beurteilungen in den Vordergrund rücken. Dies begründet sich aus dem Gegenstand, mit dem sich im empirischen Teil der vorliegenden Arbeit auseinander gesetzt werden soll. In diesem wird die Genese von Lehrerleitungsurteilen über Schülertexte aus einer Leistungssituation im Physikunterricht untersucht. Ein Ziel dieses Kapitels ist daher, sich dem Erkenntnisinteresse des empirische Teils der vorliegenden Arbeit anzunähern, da in Kapitel 1 die Frage nach Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen noch unzureichend thematisiert wurde. Dementsprechend gilt es in der ersten Hälfte dieses Kapitels zunächst das weitläufige Feld der Forschung zu Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen zu ordnen und zu systematisieren. Die Darstellung beschränkt sich dabei auf die im (deutschsprachigen) erziehungswissenschaftlichen Diskurs am prominentesten vertretenen Forschungstradition, die jedoch möglichst detailreich erörtert werden.

Anschließend wird in Unterkapitel 2.2 die zuvor vorgenommenen Sichtung und Ordnung der Forschung über Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen im bisher ausschließlich im internationalen Diskurs thematisierten Rahmenkonzept einer *Assessment Literacy von Lehrkräften* zusammengeführt und dabei in den naturwissenschaftsdidaktischen Diskurs, sowie die von Terhart (2011) vorgeschlagenen drei Bestimmungsansätze von Professionalität im Lehrerberuf eingeordnet. Kurz charakterisiert[13] gehen diese Bestimmungsansätze von drei verschiedenen Betrachtungsweisen zum Begriff der Professionalität im Lehrerberuf aus:

[12] Im Rahmen der vorliegenden Arbeit ist der Begriff „Lehrerwissen und -können“ in Anlehnung an die hierzu vorgenommen Überlegungen von Neuweg (2014) als Sammelbegriff zu verstehen, der sowohl das theoretisches (Ausbildungs-)Wissen, als auch das berufspraktische Können von Lehrkräften umfasst. Details diesbezüglich sind vor allem Unterkapitel 2.1 zu entnehmen. Für eine Überblick zu verschiedenen Modellvorstellungen, die das komplexe Verhältnis zwischen Lehrerwissen und -können zu beschreiben versuchen siehe zudem Neuweg (2004).

[13] Auf eine umfassende Darstellung dieser breit angelegten (meta-)theoretischen Konzepte wird im Rahmen der vorliegenden Arbeit verzichtet. Für einen Gesamtüberblick sei stattdessen direkt auf den erwähnten Beitrag von Terhart (2011) verwiesen, sowie auf die dort zu findende reichhaltige Auswahl an Literaturempfehlungen zu den einzelnen Bestimmungsansätzen.

1. Im strukturtheoretischen Ansatz wird Professionalität als Fähigkeit verstanden, „die vielfachen Spannungen und [...] Antinomien [des Lehrerberufs] sachgerecht handhaben zu können“ (ebd., S. 206).
2. Im kompetenztheoretischen Ansatz gilt ein_e Lehrer_in als professionell, wenn sein_ihr Lehrerwissen und -können in verschiedenen Anforderungsbereichen möglichst hoch (bezüglich operationalisierter Standards) entwickelt ist (vgl. ebd., S. 207).
3. „Die berufsbiographische Zugangsweise versteht Professionalität [...] als ein berufsbiographisches Entwicklungsproblem[,] [...] [bei dem] eine Unterscheidung zwischen gelingender und misslingender, problematischer oder gefährdeter Entwicklung [entscheident ist]“ (ebd., S. 208).

Das Ziel, dass mit dem eben beschriebenen Vorgehen im zweiten Teil dieses Kapitels verfolgt wird, ist, möglichst allumfassend und dennoch präzise zu klären, was eine im Kontext von schulischer Leistungsfeststellung und -beurteilung professionell handelnde Lehrkraft ausmacht. Hierdurch soll zu einem Referenzrahmen gelangt werden, der als Heuristik für die stichhaltige Interpretation der Befunde im empirischen Teil der vorliegenden Arbeit dienen wird. Ferner soll dieser Referenzrahmen auch als Vergleichsskizze eines Lehrerideals herangezogen werden, um (fach-)didaktische Konsequenzen aus den Befunden des empirischen Teils der vorliegenden Arbeit ableiten zu können.

2.1. Erkenntnisstand zu Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen

2.1.1. Forschung zu diagnostischen Kompetenzen von Lehrer_innen

Eine in der Forschung bisher häufig genutzte Möglichkeit der Auseinandersetzung mit Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen eröffnet sich, wenn man den Blick auf die von der Kultusministerkonferenz beschlossenen „[bildungswissenschaftlichen] Standards für die Lehrerbildung“ richtet (vgl. Kultusministerkonferenz , 2004c). Deren aktuelle Fassung benennt neben inhaltlichen Ausbildungsschwerpunkten vor allem „Kompetenzen [...] über die eine Lehrkraft zur Bewältigung der beruflichen Anforderungen verfügt“ (ebd., S. 4). Im Zusammenhang mit dem hier als Standard definierten Lehrerwissen und -können haben sich in den letzten Jahrzehnten unterschiedliche Forschungszugänge entwickelt. Wird hierbei über schulischen Leistungsfeststellungen und -beurteilungen gesprochen, wird meist der Begriff *diagnostische*

Kompetenz von Lehrkräften verwendet[14]. Ferner lassen sich diesem Themengebiet drei auf methodischer und theoretischer Ebene unterschiedliche Forschungsansätze zuschreiben:

1. Der Urteilsgenauigkeitsansatz
2. Der Kompetenzmodellierungsansatz
3. Der Selbstauskunftsansatz

Diese Ansätze gilt es in den folgenden Unterabschnitten näher zu bestimmten. Dabei wird wie folgt vorgegangen: Zunächst wird die Konzeption des jeweiligen Zugangs skizziert und jeweilige Unterschiede werden herausgearbeitet. Anschließend erfolgt eine Übersicht über die auf der Grundlage des entsprechenden Zugangs durchgeführten empirischen Studien. Deren zentrale Befunde und die hieraus gewonnenen Erkenntnisse stehen dabei im Zentrum der Darstellung.

2.1.1.1. Der Urteilsgenauigkeitsansatz

Der Urteilsgenauigkeitsansatz ist der wohl älteste Ansatz[15] zur Erforschung der diagnostischen Kompetenz von Lehrkräften. Er besteht darin, das Ausmaß, in dem eine Lehrerleistungsfeststellung und -beurteilung mit der „tatsächlichen“ Schülerleistung übereinstimmt, als Indikator für diese Kompetenz anzusehen (vgl. Schrader, 2014, S. 869). In der deutschsprachigen Literatur wird dieser Indikator meist als *Veridikalität* bezeichnet (vgl. ebd.).

2.1.1.1.1. Empirische Zugänge zur Urteilsgenauigkeit

In Forschungsarbeiten, die nach dem Urteilsgenauigkeitsansatz vorgehen, werden Lehrer_innen darum gebeten, im Voraus oder parallel zu einem standardisierten Schulleistungstest Einschätzungen über das Abschneiden ihrer Schüler_innen in diesem Test abzugeben. Hierbei werden sie dazu aufgefordert, eine klasseninterne Rangfolge zu bilden (z. B. Demaray & Elliot, 1998), die Anzahl der insgesamt gelösten Testitems für jede_n Schüler_in zu schätzen (z. B. Hosenfeld, Helmke, & Schrader, 2002) oder für eine Auswahl von (Multiple-Choice-)Aufgaben ein Ja-Nein-Urteil darüber abzugeben, ob ein_e Schüler_in diese wird lösen können oder nicht (z. B. Coladarci, 1986; Brunner, Anders, Hachfeld, & Krauss, 2011). Bei der Veridikalität handelt es sich damit im Grunde um eine Sonderform der Übereinstimmungsgültigkeit bzw. der konvergenten Validität (vgl. Abschnitt 1.3.1), da hier „der Prädiktor (das Lehrerurteil) mit einer möglichst guten (zumindest aber besseren) Messung des [...] zu beurteilenden Merkmals [in Beziehung gesetzt wird]“ (Helmke, Hosenfeld, & Schrader, 2004, S. 123).

[14] In den Standards der Kultusministerkonferenz wird vom „Beurteilen“ gesprochen, das einen von vier Kompetenzbereichen der Lehrerbildung ausmacht (vgl. Kultusministerkonferenz , 2004c, S. 7 u. f.).

[15] Die älteste zitierte Studie im hierzu vorgenommen Literaturreview von Hoge & Coladaraci (1989) stammt aus dem Jahr 1962.

2.1.1.1.2. Komponenten der Urteilsgenauigkeit

Eine naheliegende Idee zur Auswertung der Daten, die mit den eben skizzierten Erhebungsmethoden gewonnen werden können, ist z. B. die mittlere quadratische Abweichung zwischen der Lehrereinschätzungen und den Testergebnissen der Schüler_innen zu bilden. Ein solches Globalmaß ist allerdings problematisch, da hier verschiedene Faktoren, welche die Leistungseinschätzungen einer Lehrkraft (negativ) beeinflussen können[16], miteinander konfundiert werden (vgl. Cronbach, 1955; Schrader & Helmke, 1987). Anstatt eine solchen Globalmaßes schlagen Schrader & Helmke (1987) daher vor, die Urteilsgenauigkeit einer Lehrkraft anhand von drei verschiedenen Bestimmungsmaßen zu charakterisieren, die sie als *Niveau-*, *Differenzierungs-* und *Vergleichskomponente*, bezeichnen:

> „Die *Niveaukomponente* charakterisiert die Tendenz von Lehrern, das Leistungsniveau der eigenen Klasse im Vergleich zu den tatsächlichen Ergebnissen insgesamt eher zu über- oder eher zu unterschätzen. Sie läßt sich operationalisieren als Differenz zwischen dem Mittelwert aller Urteile eines Lehrers und dem Mittelwert der korrespondierenden Kriteriumswerte innerhalb der Klasse. [...] Die *Differenzierungskomponente* kennzeichnet die Tendenz von Lehrern, die Streuung der Schülerleistungen in der Klasse zu über oder zu unterschätzen. [...] [Sie lässt sich] definiere[n] als Quotient aus der Streuung der Urteile des Lehrers (Zähler) und der Streuung der entsprechenden Kriteriumswerte (Nenner)[.] [...] Die *Vergleichskomponente* schließlich bezieht sich auf die zutreffende Einschätzung der relativen Leistungsposition der einzelnen Schüler innerhalb der Klasse. [...] [Sie] ist operationalisiert als Produkt-Moment-Korrelation zwischen den vom Lehrer vorhergesagten und den in seiner Klasse tatsächlich erzielten Leistungen der Schüler." (ebd., S. 30-31, Hervorhebungen im Original)

Studien, die den Zusammenhang zwischen Vergleichs-, Niveau- und Differenzierungskomponente untersuchten, stellten übereinstimmend nur geringe Interkorrelationen zwischen diesen Komponenten der Urteilsgenauigkeit fest (z. B. Schrader & Helmke, 1987, S. 42 u. f.; Spinath, 2005, S. 92 u. f.). Es wird deshalb angenommen, dass diese Kennwerte Indikatoren für voneinander unabhängige Dimensionen der Diagnosekompetenz von Lehrkräften sind (vgl. Karing, Matthäi, & Artelt, 2011, S. 160). Die Erkenntnisse und Befunde, welche bisher mit Hilfe des Urteilsgenauigkeitsansatzes gewonnen worden sind, werden daher im Folgenden entlang dieser drei Komponenten geordnet:

2.1.1.1.3. Forschungsstand zur Vergleichskomponente

Zur Vergleichskomponente finden sich in der Literatur mit Abstand die meisten empirischen Untersuchungen. Bei diesen wird meist ein über die befragten Lehrkräfte hinweg gemittelter Korrelationskoeffizient zwischen Lehrerurteil und Schülerleistungen angegeben. Für Studien, die ein enges Leistungsbegriffsverständnis zugrunde legen (vgl. Abschnitt 1.1.1), sind bis dato zwei Überblicksartikel erschienen. Im Literaturreview von Hoge & Coladaraci (1989) sind 17 Studien der Jahre 1962 bis 1988 zusammengefasst, in jenem von Südkamp, Kaiser, & Möller (2012) sind für die Jahre 1989 bis 2010 insgesamt 75 Studien aufgeführt. Diese Übersichtsarbeiten kommen übereinstimmend zu dem Ergebnis, dass

[16] Hierzu gehört beispielsweise der sog. Halo-Effekt. Eine Übersicht zu diesen „Stolpersteinen der Diagnose" findet sich z. B. bei Höttecke (2015).

Lehrer_innen im Durchschnitt ein „relativ gutes Gespür“ für Leistungsunterschiede in ihren Klassen haben. So geben Hoge & Coladaraci (1989, S. 303) und Südkamp et al. (2012, S. 750) für die Vergleichskomponente eine über die unterschiedlichen Studien hinweg gemittelte Korrelation (Median) von r=.66 bzw. r=.63 an. Auf der anderen Seite zeigen sich allerdings zum Teil große Unterschiede zwischen den in den einzelnen Studien erhobenen Werten: Bei den von Hoge & Coladaraci (1989) angegeben Studien schwanken die Korrelationskoeffizienten zwischen r=.28 und r=.92 und bei Südkamp et al. (2012) zwischen r=-.03 und r=.84. Des Weiteren wird durch einen genaueren Blick in die Einzelstudien deutlich, dass teilweise erhebliche Lehrerunterschiede anzutreffen sind. So bewegen sich beispielsweise in der Untersuchung von Schrader & Helmke (1987, S. 40) die Vergleichskomponente der befragten Lehrer_innen zwischen r=.04 und r=.88.

Zur Erklärung dieser Schwankungen schlagen Südkamp et al. (2012, S. 756 u. f.) ein heuristisches Modell vor, das in Abbildung 2.1 dargestellt ist[17]. Nach diesem wird die Höhe der Vergleichskomponente von Merkmalen der Lehrkraft (z. B. Berufserfahrung), der Schüler_innen (z. B. bisher erworbene Kompetenzen), des Testverfahrens (z. B. Testlänge oder Domäne) und der Einschätzungsaufgabe (z. B. Instruktionen an die Lehrkraft in der Erhebungssituation) bestimmt (durchgezogenen Linien in Abbildung 2.1), da zu erwarten ist, dass diese die Einschätzungen der Lehrkraft bzw. die Leistungen der Schüler_innen im Test beeinflussen (vgl. ebd.), aus denen wiederum die Vergleichskomponente bestimmt wird (große Pfeile in Abbildung 2.1). Zudem ist davon auszugehen, dass diese Merkmale auch wechselseitig in Beziehung zueinander stehen (vgl. ebd.), was durch die gestrichelten Linien in Abbildung 2.1 dargestellt ist. Die Autor_innen merken allerdings an, dass zu den vier im Modell aufgeführten Merkmalen zum Teil noch kaum empirische Evidenz vorliegt, weswegen sie ihr Modell als in Teilen hochgradig spekulativ bezeichnen (vgl. ebd.).

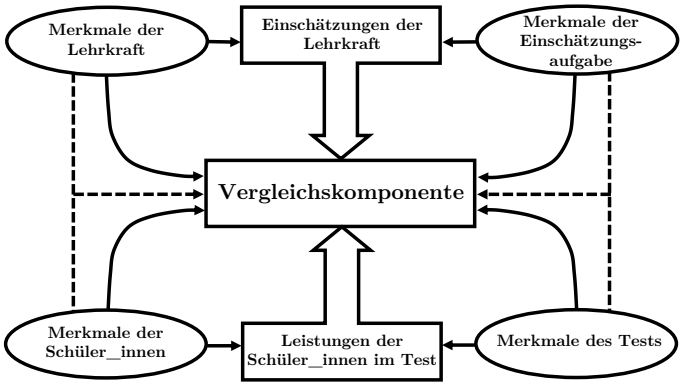


Abbildung 2.1.: Heuristisches Modell zu Moderatoren der Vergleichskomponente nach Südkamp et al. (2012, S. 756 u. f.).

Was aus den beiden erwähnten Übersichtsarbeiten allerdings auch deutlich wird, ist, dass sich die empirische Forschung zur Vergleichskomponente bislang vor allem auf sprachliche Schulfächer und Mathematik beschränkt hat (siehe Tabelle 2.1). Für den naturwissenschaftlichen Unterricht existieren dagegen kaum belastbare Daten. Ein Beispiel für die wenigen hierzu bekannten Studien ist die Untersuchung von Hopkins, George, & Williams (1985), in der die Vergleichskomponente von N = 42 Viert- und Fünftklasslehrkräften in

[17] Südkamp et al. (2012) merken an, dass sich dieses Modell auch auf die anderen Komponenten der Urteilsgenauigkeit übertragen lässt (vgl. ebd., S. 756).

	sprachliches Schulfach	Mathematik	weiteres Schulfach	sprachliches Schulfach & Mathematik	sprachliches & weiteres Schulfach	Mathematik & weiteres Schulfach	Mathematik, sprachliches & weiteres Schulfach
(Hoge & Coladarci, 1989)	6	1	---	7	---	---	3
(Südkamp et al., 2012)	32	8	3	25	2	---	5
Total	38	9	3	32	2	---	8

Tabelle 2.1.: Anzahl von Studien zur Vergleichskomponente in verschiedenen Domänen auf Basis der Literaturreviews von Hoge & Coladaraci (1989) und Südkamp, Kaiser, & Möller (2012).

fünf verschiedenen Inhaltsbereichen bestimmt wurde. Die mittlere Korrelationen zwischen den Einschätzungen der Lehrer_innen und den zeitgleich erhobenen Schülerleistungen lagen dabei zwischen r=.60 für die Domäne „Science“ und r=.74 für „Language Arts“ (vgl. ebd., S. 180). Mit Hilfe einer Varianzanalyse konnten die Autoren zudem zeigen, dass die von ihnen festgestellten Domänenunterschiede zwischen den Vergleichskomponenten statistisch signifikant sind (vgl. ebd.).

Weitere Untersuchungen liefern Hinweise zur Ausprägung der Vergleichskomponenten von Lehrereinschätzungen, die Schülerleistungen im erweiterten Sinne betreffen (vgl. Abschnitt 1.1.1). So bestimmte Spinath (2005, S. 91) die mittlere Vergleichskomponente bezogen auf die Schülermerkmale Intelligenz (r=.40), Fähigkeitenselbstwahrnehmung (r=.39), Lernmotivation (r=.20) und Leistungsängstlichkeit (r=.15), Karing (2009, S. 204) in Bezug auf das Interesse von Schüler_innen an den Fächern Mathematik (Gymnasium r=.32, Grundschule r=.37) und Deutsch (Gymnasium r=.21, Grundschule r=.30) und Praetorius, Karst, & Lipowsky (2011, S. 86) für das Fähigkeitenselbstkonzept von Schüler_innen in Mathematik (r=.55), Lesen (r=.52) und Schreiben (r=.25). Auch für Studien, zur Ausprägung der Vergleichskomponenten von Lehrereinschätzungen, die Schülerleistungen im erweiterten Sinne betreffen, sind bis dato zwei Überblicksarbeiten erschienen: Basierend auf 32 Artikel gibt Follman (1991) für Intelligenz eine mittlere Vergleichskomponente (Median) von r=.55 an. Machts, Kaiser, Schmidt, & Möller (2016) berichten in ihrer Metaanalyse, die auf 33 Studien basiert, mittlere Vergleichskomponenten für die Schülermerkmale Intelligenz (r=.50), Begabung (r=.36) und Kreativität (r=.34). Aus allen eben aufgeführten Befunden wird deutlich, dass hier die mittlere Vergleichskomponenten geringer ausfallen, als in Studien, deren Leistungsbegriff enger gefasst ist[18], „was auf die [hierbei] höheren Urteilsanforderungen [...], bei den Selbstberichtskalen zusätzlich aber

[18] Alles in allem bewegen sich die mittleren Vergleichskomponenten von Lehrereinschätzungen, die Schülerleistungen im erweiterten Sinne betreffen, aber auf einem akzeptablen Niveau (vgl. Machts et al., 2016, S. 98 u. f.). Allerdings zeigen sich auch hier zum Teil große Unterschiede zwischen den in den einzelnen Studien erhobenen Werten: Bei den von Follman (1991) zitierten Studien zum Merkmal Intelligenz schwanken die Korrelationskoeffizienten zwischen r=.25 und r=.88. Bei den von Machts

auch auf höhere Messungenauigkeiten Seitens des Kriteriums hindeuten könnte“ (Helmke et al., 2004, S. 127). In der Metaanalyse von Machts et al. (2016) konnte die Hypothese „[j]udgment accuracy is higher when the criterion is measured with tests of higher reliability“ (ebd., S. 91) empirisch jedoch nicht bestätigt werden (vgl. ebd., S. 99).

2.1.1.1.4. Forschungsstand zur Differenzierungs- und zur Niveaukomponente

Im Gegensatz zur Vergleichskomponente gibt es nur wenige Studien, die auch die Differenzierungs- und die Niveaukomponente erfasst haben (vgl. Helmke et al., 2004, S. 127). Nach der Studie von Spinath (2005, S. 91) zeichnet sich für ein weites Leistungsbegriffsverständnis kein einheitliches Bild über die Ausprägungen dieser beiden Komponenten der Urteilsgenauigkeit ab. Die Niveaus und die Streuungen der vier in dieser Studie untersuchten Schülermerkmale wurden von den befragten Lehrkräften im Mittel sowohl überschätzt (Niveau der Leistungsängstlichkeit; Streuung der Fähigkeitenselbstwahrnehmung und der Lernmotivation), als auch unterschätzt (Niveau der Fähigkeitenselbstwahrnehmung und der Lernmotivation; Streuung der Intelligenz und der Leistungsängstlichkeit) (vgl. ebd.). Lediglich das Niveau der Intelligenz wurde im Mittel relativ akkurat eingeschätzt (vgl. ebd.). Ferner kommen Praetorius et al. (2011, S. 86) zu dem Ergebnis, dass gemittelt Lehrkräfte das Niveau des Fähigkeitenselbstkonzept in Mathematik, Lesen und Schreiben unterschätzen und dessen Streuung überschätzen.

Für ein enges Leistungsbegriffsverständnis ist die Befundlage der wenigen Untersuchungen zur Differenzierungskomponte ähnlich diffus. In verschiedenen Studien wurde sowohl eine mittlere Überschätzung (z. B. Schrader & Helmke, 1987, S. 40), als auch eine Unterschätzung (z. B. Südkamp, Möller, & Pohlmann, 2008, S. 268) der Streuung der Schülerleistungen durch die befragten Lehrkräfte festgestellt. Bei den Untersuchungen zur Niveaukomponente deutet sich hingegen eine gewisser Trend ab. So kommen beispielsweise die Studien von Bates & Nettelback (2001, S. 182), Schrader & Helmke (1987, S. 38) und Kaiser & Möller (2017, S. 65) zu dem Ergebnis, dass (angehende) Lehrer_innen im Mittel das Niveau der Leistungen ihrer Schüler_innen leicht überschätzen. Schrader & Helmke (1987) sehen mögliche Ursachen hierfür darin, dass sich die befragten Lehrkräfte in der Feststellungs- und Beurteilungssituation, mit der sie konfrontiert wurden, zum einen deutlich an der kriterialen Bezugsnorm orientieren (vgl. Abschnitt 2.1.2) und zum anderen ihre Einschätzungen eher kompetenz- anstatt performanzorientiert vornehmen (vgl. ebd., S. 38 u. f.). Dieser Trend stellt allerdings nicht notwendigerweise einen Hinweis auf ein Feststellungs- und Beurteilungsdefizit Seitens der Lehrkräfte dar. Eine moderate Überschätzung von Schülerleistungen kann auch als Indikator für die Güte von Leistungsfeststellungen und -beurteilungen im Sinne einer pädagogisch günstigen Voreingenommenheit gelten (vgl. Abschnitt 1.3.2).

et al. (2016) zitierten Studien zu den Merkmalen Intelligenz, Begabung und Kreativität schwanken die Korrelationskoeffizienten zwischen r=-.18 und r=.71, r=.04 und r=.52, sowie r=.11 und r=.72.

2.1.1.1.5. Zwischenfazit

Zusammengefasst stehen beim Urteilsgenauigkeitsansatz Indikatoren, mit denen die Veridikalität empirische erfasst werden kann, im Zentrum der Überlegungen. Die Stärken dieses Ansatzes sind, dass durch die quantitative Erfassung Determinanten der Veridikalität und das Ausmaß ihres Einflusses identifiziert werden können und – für den naturwissenschaftlichen Unterricht zwar bislang noch nicht im befriedigenden Maße – konnten. Hinzu kommt, dass durch dieses Vorgehen der Vergleich und die Zusammenfassung verschiedener Studien begünstigt ist (vgl. Leuders, Leuders, & Philipp, 2014, S. 732).

2.1.1.2. Der Kompetenzmodellierungsansatz

Neben der eben erwähnten Stärke zeigte sich im vorherigen Unterabschnitt auch die zentrale Schwäche des Urteilsgenauigkeitsansatzes: Durch die deutliche Fokussierung auf empirisch erfassbare Indikatoren wird weitgehend die Frage übergangen, worum es sich bei der diagnostischen Kompetenz von Lehrkräften eigentlich handelt (vgl. Leuders et al., 2014, S. 732). Im Kompetenzmodellierungsansatz stellt diese Frage hingegen den Kern der Auseinandersetzung dar (vgl. von Aufschnaiter et al., 2015, S. 740). Bei diesem wird „[a]usgehend von einer möglichst genauen Aufgabenbeschreibung für den Lehrerberuf [...] [Wissen und Können in Form von Standards] definiert, [das] für die Bewältigung dieser Aufgaben wichtig bzw. notwendig [ist]“ (Terhart, 2011, S. 207). Sinn und Zweck dieser Operationalisierungen ist oftmals Papier-und-Bleistift-Tests zu entwickeln, mit denen versucht wird, die entsprechend modellierten Lehrerkompetenzen möglichst valide zu erfassen (vgl. Leuders et al., 2014, S. 731). Bei der Operationalisierung dieses Wissens und Könnens[19] orientiert sich die Forschung bis heute an der von Shulman (vgl. Shulman, 1986; Shulman, 1987) vorgeschlagenen Unterteilung in verschiedene sog. Wissensbereiche; im deutschsprachigen Raum zudem an der von Bromme (1992) (vgl. Neuweg, 2014, S. 586). Ferner hat sich inzwischen eine Dreiteilung in *fachliches Wissen* (*Content Knowledge*), *fachdidaktisches Wissen* (*Pedagogical Content Knowledge*) und *pädagogisches Wissen* (*Pedagogical Knowledge*) weitgehend durchgesetzt (vgl. von Aufschnaiter et al., 2015, S. 739). Wenn es allerdings um eine feinkörnigere Aufgliederung geht, bzw. darum welches Wissen und Können welchem der drei Wissensbereiche zuzuordnen ist, besteht jedoch noch kein Konsens (vgl. Cauet, 2016, S. 12 u. f.), zumal erschwerend hinzukommt, dass sich bereits Fachwissen, fachdidaktisches und pädagogisches Wissen nur bis zu einem bestimmten Grad theoretisch voneinander unterscheiden lassen (vgl. Neuweg, 2014, S. 588 u. f.). Hiermit hängt auch zusammen, dass verschiedene Studien bisweilen zu unterschiedlichen Ergebnissen kommen, wenn es um die Frage der empirischen Trennbarkeit dieser drei

[19] Auch wenn im Folgenden ein besonderer Fokus auf bestimmten Wissens- und Könnensaspekten liegt, darf damit (diagnostische) Kompetenz von Lehrkräften nicht Missverstanden werden als ein bloßes Sammelsurium dieser Einzelkomponenten. Vielmehr werden „[a]uf deren Basis [...] die entsprechenden Teilkompetenzen konstitutiv ausgebildet[.] [...] Dadurch wird eine Ebene erreicht, die über das reine Wissen [und Können] hinausgeht“ (Jäger, 2009, S. 108).

Wissensbereiche geht[20], „was das Problem von der Empirie- wieder auf die Theoriebene rückdelegiert“ (vgl. ebd., S. 592).

Aus dieser allgemeinen Übersicht über den Kompetenzmodellierungsansatz in der Lehrerbildungsforschung ergeben sich für die diagnostische Kompetenz von Lehrkräften folgende Fragen, die es im Folgenden zu klären gilt:

1. Wie wird die diagnostische Kompetenz von Lehrkräften im Rahmen des Kompetenzmodellierungsansatzes operationalisiert, in welche(n) der drei genannten Wissensbereiche lässt sich diese verorten und wie wird dies insbesondere in den Naturwissenschaftsdidaktiken diskutiert?
2. Welche Befundlagen liefert die bisherige Forschung, die auf dem Kompetenzmodellierungsansatz beruht, zum Lehrerwissen und -können bezogen auf Leistungsfeststellungen und -beurteilungen im schulischen Kontext?

2.1.1.2.1. Operationalisierung und Verortung der diagnostischen Kompetenz von Lehrkräften

Eine erste Vorstellung über eine mögliche Operationalisierung der diagnostischen Kompetenz von Lehrkräften, lässt sich aus den bereits erwähnten „[bildungswissenschaftlichen] Standards für die Lehrerbildung“ der Kultusministerkonferenz gewinnen (vgl. Kultusministerkonferenz , 2004c). Für den „Kompetenzbereich Beurteilen“ ist dort unter anderem folgendes Lehrerwissen und -können als Standard festgelegt:

> **„Standards für die theoretischen Ausbildungsabschnitte**
> Die Absolventinnen und Absolventen...
> [...]
> - kennen die Grundlagen der Lernprozessdiagnostik.
>
> [...]
> - kennen unterschiedliche Formen der Leistungsbeurteilung, ihre Funktionen und ihre Vor- und Nachteile.
> - kennen verschiedene Bezugssysteme der Leistungsbeurteilung und wägen sie gegeneinander ab.
>
> **Standards für die praktischen Ausbildungsabschnitte**
> Die Absolventinnen und Absolventen...
> [...]
> - stimmen Lernmöglichkeiten und Lernanforderungen aufeinander ab.
>
> [...]
> - wenden Bewertungsmodelle und Bewertungsmaßstäbe fach- und situationsgerecht an.
> - verständigen sich auf Beurteilungsgrundsätze mit Kolleginnen und Kollegen.“
>
> (ebd., S. 11)

Was aus diesen Beispielen erkennbar wird und auch für von anderen Autor_innen formulierte Operationalisierungen von Aspekten der diagnostischen Kompetenz von Lehrkräften gilt (z. B. Tamir, 1988, S. 100; Hashweh, 2005, S. 283 u. f.; von Aufschnaiter et al., 2015, S. 750), ist, dass sich die inhaltliche Ausgestaltung dieser Operationalisierungen vielfach auf die in Kapitel 1 unternommen grundlegenden Überlegungen zu Funktionen und Gü-

[20] Für einen Überblick zur hierzu aktuellen Befundlage siehe Cauet (2016, S. 14). Befunde zur Trennbarkeit von Fachwissen und fachdidaktischem Wissen finden sich zudem bei Neuweg (2014, S. 592).

teaspekte von Leistungsfeststellungen und -beurteilungen im schulischen Kontext (vgl. Unterkapitel 1.2 und 1.3) und/oder auf das Thema der Bezugsnormorientierung beziehen (siehe Abschnitt 2.1.2).

Daneben fällt auf, dass diese Standards von drei unterschiedlichen Auffassungen des Begriffs „Lehrerwissen“ ausgehen, die Neuweg (2014) als Wissen 1, 2 und 3 bezeichnet (vgl. ebd., S. 600). Sie können gleichsam verstanden werden als...

> „[...] das kodifizierte, mehr oder weniger systematische und insbesondere in der Ausbildung anzueignende [...] „Wissen im Buch“ [Wissen 1; M. S. F.], [...] [als] [d]ie *kognitiven Strukturen* von Lehrern [Wissen 2; M. S. F.] [...] [und als Zuschreibung einer] Verhaltensdisposition[,] [...] [dass der_die Lehrer_in] „weiß, wie es geht“[,] [...] [bei der] es sich aber nicht um das Wissen des Lehrers, sondern um das Wissen des Forschers [handelt], der die Logik des *Handelns* (!) von außen rekonstruiert [Wissen 3; M. S. F.].“ (ebd., S. 584-585, Hervorhebungen im Original)

Für ein besseres Verständnis ist anzumerken, dass Neuweg (2014) lediglich „Wissen 1“ explizit als Professionswissen bezeichnet und Lehrerkönnen mit dem Begriff „Wissen 3“ gleichsetzt (vgl. ebd., S. 584 u. f.). Des Weiteren ist zu erwähnen, dass bei diesen Wissensbegriffen sowohl ein expliziter wie auch ein impliziter Lernweg angenommen werden kann und dass diese auf mentaler Ebene explizit und/oder implizit repräsentiert sein können (vgl. ebd., S. 601).

Ein dritter Aspekt, der aus den eben zitierten Beispielstandards deutlich wird, ist, dass diese weitgehend fächerübergreifend formuliert sind. Dies legt zunächst den Gedanken nahe, Lehrerwissen und -können zum Thema schulische Leistungsfeststellungen und -beurteilungen dem pädagogischen und nicht dem fachdidaktischen Wissen zuzuordnen. Eine derartige Verortung findet sich beispielsweise im vielzitierten Modell professioneller Handlungskompetenz von Baumert & Kunter (2006), sowie Forschungsprojekten, die sich an stark diesem Modell orientieren (z. B. die Projekte *COACTIV* (vgl. Kunter et al., 2011), *BilWiss* (vgl. Linninger et al., 2015) und *FALKO* (vgl. Krauss et al., 2017)). Die Kultusministerkonferenz betont allerdings, „dass sich Erziehung und Unterricht an fachlichen Inhalten vollziehen“ (vgl. Kultusministerkonferenz , 2004c, S. 4) und stellt damit klar, dass die von ihr festgelegten Standards zwar fächerübergreifend formuliert, gleichzeitig aber aus der Perspektive des entsprechenden Fachunterrichts zu denken sind. Für das Themengebiet der schulischen Leistungsfeststellung und -beurteilung ist diese Festlegung in den später veröffentlichten „Ländergemeinsamen inhaltlichen Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung“ noch einmal bekräftigt worden (vgl. Kultusministerkonferenz , 2018). Dort heißt es, dass Lehrkräfte bei Abschluss ihrer Ausbildung über anschlussfähiges fachdidaktisches Wissen verfügen sollen, wobei diesem explizit Wissen und Könnerschaft zur fachspezifische Leistungsfeststellung und -beurteilung zugeordnet sind (vgl. ebd., S. 4). Nach Helmke et al. (2004, S. 121), sowie von Aufschnaiter et al. (2015, S. 739) umfasst dieses fachdidaktische Wissen Aspekte, wie sie die bildungswissenschaftlichen Standards für die Lehrerbildung festlegen, zusätzlich aber auch Wissen und Können bezogen auf „die Anforderungen in einem Lerngebiet, [...] Schwierigkeiten von Aufgaben [...] [und] mögliche Lösungsprozeduren, typische Vorgehens-

Referenz	Explizite Verortung von Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung	
	fachdidaktisches Wissen	**pädagogisches Wissen**
(Tamir, 1988)	×	×
(D. C. Smith & Neale, 1989)		
(Geddis, Onslow, Beynon, & Oesch, 1993)		
(van Driel, Verloop, & de Vos, 1998)		
(Magnusson, Krajcik, & Borko, 1999)	×	
(Hashweh, 2005)	×	×
(Lee & Luft, 2008)	×	
(Park & Oliver, 2008)	×	
(Rollnick, Bennett, Rhemtula, Dharsey, & Ndlovu, 2008)	×	
Paderborner Instrument (Riese, 2009; Vogelsang, 2014)	×	×
PLUS (Lange, 2010)	×	
(van Dijk & Kattmann, 2010)		
(Loughran, Berry, & Mulhall, 2012)	×	
ProwiN (Tepner et al., 2012; Kirschner, 2013; Cauet, 2016)		×
Profile-P (Gramzow, Riese, & Reinhold, 2013)	×	
KiL (Kröger, Neumann, & Petersen, 2013)	×	
QuiP (Ergönenç, Neumann, & Fischer, 2014)		
(Gess-Newsome, 2015)	×	×
FALKO-P (Schödl & Göhring, 2017)		×
(Förtsch et al., 2018)	×	×

Tabelle 2.2.: Übersicht zur expliziten Verortung von Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung in verschiedenen Operationalisierungen von fachdidaktischem Wissen und pädagogischem Wissen im naturwissenschaftsdidaktischen Diskurs. Zusammengestellt, erweitert und ergänzt aus Gramzow, Riese, & Reinhold (2013, S. 15), Kirschner (2013, S. 32) und Cauet (2016, S. 18).

weisen, Leistungen und Fehlern [sic!] bei Schülern unterschiedlichen Entwicklungsstands und unterschiedlicher Leistungsfähigkeit“ (Helmke et al., 2004, S. 121), wobei hier von einem weiten Leistungsbegriffsverständnis auszugehen ist (vgl. Abschnitt 1.1.1).

Alles in allem lässt sich damit diagnostische Kompetenz von Lehrkräften am ehesten als ein Mischform von pädagogischem und fachdidaktischem Wissen beschreiben (vgl. Neuweg, 2014, S. 594; Dori & Avargil, 2016, S. 1034), wobei die pädagogischen Anteile vor dem Hintergrund des entsprechenden Fachunterrichts zu denken sind. Ferner lässt sich in diesem Sinne auch das zunächst heterogen wirkende Meinungsspektrum im naturwissenschaftsdidaktischen Diskurs zu diesem Thema charakterisieren, das in Tabelle 2.2 zusammengefasst dargestellt ist. Acht der hier aufgelisteten Referenzen ordnen Lehrerwissen und -können zum Thema schulische Leistungsfeststellungen und -beurteilungen explizit dem fachdidaktischen Wissen zu, ohne sich dabei explizit zu anderen Wissensbereichen zu äußern. Fünf Referenzen nehmen dagegen eine explizite Verortung in fachdidaktische und pädagogische Wissensbereiche vor, weitere fünf Literaturverweise machen keinerlei

Aussage zum Lehrerwissen und -können zu schulische Leistungsfeststellungen und -beurteilungen und lediglich in den Projekten *FALKO* und *ProwiN* wird dieses Thema explizit ausschließlich pädagogischen Wissensbereichen zugeordnet.

2.1.1.2.2. Empirische Befunde zum Lehrerwissen und -können bezogen auf Leistungsfeststellungen und -beurteilungen im schulischen Kontext

Ohne Frage kann nicht davon ausgegangen werden, dass Lehrkräfte über bestimmte Wissensbestände bzw. Könnerschaften zur schulischen Leistungsfeststellung und -beurteilung verfügen, nur weil sie beispielsweise von der Kultusministerkonferenz als Standard beschlossen worden sind. Vielmehr stellt sich durch diese normative Setzung die Frage, inwiefern das Wissen und Können von Lehrer_innen diesen Güteanforderungen genügt. Interessanterweise gibt es allerdings sowohl national, als auch international kaum Untersuchungen, die Lehrerwissen und -können zur schulischen Leistungsfeststellung und -beurteilung direkt erfasst und davon berichtet haben (vgl. Marso & Pigge, 1993, S. 156; Helmke et al., 2004, S. 121; Neuweg, 2014, S. 594 u. f.; Schrader, 2014, S. 869). Die wenigen Studien, die sich von Mitte der 1960er bis Anfang der 1990er Jahre mit dieser Frage, bezogen auf das US-amerikanische Bildungssystem, beschäftigen, haben Marso & Pigge (1993) im Rahmen eines thematisch breiter angelegten Literaturreviews zusammengetragen und kommen aufgrund ihrer Recherche zu folgenden zentralen Befundlagen (vgl. ebd., S. 169 u. f.):

1. Insgesamt scheinen Lehrer_innen nur wenig über testtheoretisches Wissen und Können zu verfügen. Selbst elementare Grundkenntnisse besitzt offenbar nur eine Minderheit.
2. Ebenso verfügen Lehrer_innen scheinbar kaum über Wissen und Könnerschaft zu technischen Aspekten schulischer Leistungsfeststellung und -beurteilung, wie z. B. zum Umgang mit Bewertungsrastern oder zum Erkennen von Aufgabenfehlern.
3. Welche Art von Aufgaben Lehrer_innen zur Leistungsfeststellung und -beurteilung einsetzen, unterscheidet sich je nach Fach. Kognitiv anspruchsvollere Aufgaben, die über eine bloße Reproduktion deklarativer Kenntnisse hinaus gehen, finden sich eher im Fach Mathematik, sowie den Naturwissenschaften.
4. Lehrer_innen mit formaler Bildung zum Thema schulische Leistungsfeststellung und -beurteilung schneiden in Befragungen zu diesem Thema besser ab als Lehrer_innen ohne formale Bildung in diesem Themengebiet. Die Unterschiede sind allerdings gering.
5. Sowohl Lehrer_innen, als auch Schulleiter_innen und Seminarlehrer_innen scheinen über besonders wenig Wissen und Können bezogen auf einfache statistische Methoden zur Datenanalyse, informelle Verfahren zur Leistungsfeststellung und -beurteilung, sowie zur Erstellung kognitiv anspruchsvoller Leistungsaufgaben zu verfügen.

6. Lehrer_innen können nur schwer Fragen zur Interpretation der Ergebnisse von standardisierten Schulleistungstests und Vergleichsarbeiten beantworten.
7. Weder der Besuch von themenspezifischen Lehrerfortbildungen (sofern solche überhaupt angeboten werden) noch die Berufserfahrung per se scheinen das Wissen und die Könnerschaft zur schulischen Leistungsfeststellungen und -beurteilungen von Lehrkräften zu verbessern.

Hinweise über die gegenwärtige Situation in den Bildungssystemen im deutschsprachigen Raum liefern hingegen lediglich die Daten aus der Selbstauskunftsbefragungen, von der Oser (2001) berichtet und die im Rahmen des Projekts „Die Wirksamkeit des Lehrerbildungssystems in der Schweiz“ durchgeführt wurde (vgl. Criblez, 2001, S. 109). Bei dieser wurden N = 1286 angehende Lehrkräfte am Ende ihrer Ausbildung gebeten für eine Auswahl von Inhalten, die dem Themengebiet schulische Leistungsfeststellungen und -beurteilungen zugeordnet werden können, anzugeben, ob diese im Rahmen ihrer Ausbildung vermittelt wurden. Unter der Annahme, dass wenn eine solche Vermittlung stattgefunden hat auch Wissen und Könnerschaft erworben wurden, können hier zumindest vorsichtige Rückschlüsse auf das Lehrerwissen und -können der Befragten zum Zeitpunkt der Erhebung getroffen werden. Gemäß dieser Überlegung deuten die von Oser (2001) dargestellten Befunden darauf hin, dass nur eine Minderheit der befragten Lehramtsanwärter_innen am Ende ihrer Ausbildung über keinerlei Wissen und Können bezogen auf schulische Leistungsfeststellungen und -beurteilungen verfügen. So berichtet er, dass weniger[21] als 10 % der Befragten angeben in ihrer Ausbildung nichts darüber erfahren zu haben, „wie man einem Kind oder einem Jugendlichen entsprechend seiner Altersstufe ein Leistungswertsystem vermittelt“ (ebd., S. 281). Des Weiteren geben nur 25 % an in der Ausbildung nicht vermittelt bekommen zu haben, wie sich Leistungsfortschritte von Schüler_innen mit Hilfe unterschiedlicher Kriterien und Instrumente feststellen und beurteilen lassen (vgl. ebd., S. 281). Auf den „tatsächlichen“ Umfang des Lehrerwissens und -könnens der Befragten lässt sich aufgrund der in dieser Studie erhobenen Daten allerdings nicht schließen. Es ist daher zumindest denkbar, dass sich die durchweg negativen Befunde von Marso & Pigge (1993) auch auf das Lehrerbildungssystems in der Schweiz übertragen lassen.

2.1.1.2.3. Zwischenfazit

Zusammengefasst liefert der Kompetenzmodellierungsansatz vor allem Erkenntnisse darüber, welches Wissen und Können die diagnostische Kompetenz von Lehrkräften theoretisch umfasst und wie sich diese in verschiedene Wissensbereiche verordnen lassen. Empirisch Befunde sind allerdings rar, stammen vorrangig aus dem US-amerikanischen Raum und sind bislang vor allem für Lehrer_innen im Allgemeinen formuliert worden. Fachspezifische Befunde fehlen nahezu vollständig.

Am Ende dieses Abschnitts wurden zudem ausgewählte Befunde aus der Selbstauskunftsbefragung von Oser (2001) dargestellt. Wie bereits zu Beginn dieses Unterkapitels er-

[21] Die hier berichteten Prozent-Anteile sind in Ablesegenauigkeit aus den von Oser (2001) angegebenen Histogrammen entnommen.

wähnt, stellt die Erfassung der diagnostische Kompetenz von Lehrkräften mittels Selbstauskunft einen dritten methodisch-theoretischen Ansatz dar, um Einsichten in Lehrerwissen und -können im Kontext schulischer zu Leistungsfeststellungen und -beurteilungen zu gewinnen. Auf diesen Ansatz wird nun als letztes genauer eingegangen.

2.1.1.3. Der Selbstauskunftsansatz

2.1.1.3.1. Grundidee des Selbstauskunftsansatzes

Die methodische Grundidee hinter dem Selbstauskunftsansatz ist, wie die gewählte Bezeichnung bereits andeutet, die diagnostische Kompetenz von Lehrkräften durch Selbstauskunftsfragebögen zu erfassen. Auf der Hand liegt, dass dieser Ansatz, im Vergleich zum Urteilsgenauigkeitsansatz, deutlich ökonomischer ist und ferner die Erhebung größerer Datensätze im Rahmen einer einzelnen Untersuchung ermöglicht. Auf der anderen Seite sind selbsteingeschätzte Kompetenzen zu verstehen als „die Erfahrung, inwieweit Handlungsfähigkeit unter den Rahmenbedingungen des Lehrerhandelns vorliegt“ (Abs, 2006, S. 232). Insofern ist eine selbsteingeschätzte diagnostische Kompetenz ein Indikator für die subjektive Selbstwahrnehmung einer Lehrkraft und unterscheidet sich daher von den Maßen des Urteilsgenauigkeitsansatz bzw. des Kompetenzmodellierungsansatze, da diese als quantifizierte Fremdwahrnehmung einer Lehrkraft aufzufassen sind. Den Ansätzen liegt damit also ein unterschiedliches Verständnis diagnostischer Kompetenz zugrunde. Beim Selbstauskunftsansatz ist dieses die „Identifikation mit einem Beruf, wie er in der Praxis vorgefunden wird“ (ebd.), beim Urteilsgenauigkeitsansatz und dem Kompetenzmodellierungsansatz „die [fremdwahrgenommene] Befähigung zur Umsetzung eines [begründbaren] Ideals“ (ebd.). Ersterer kann damit auch als eine sturkturtheoretische und/oder berufsbiographische Sichtweise auf Lehrerwissen und -können verstanden werden.

2.1.1.3.2. Empirische Befunde des Selbstauskunftsansatzes

Nun zu zentrale Befunden aus Studien, in denen mit dem Selbstauskunftsansatz die diagnostische Kompetenz von Lehrkräften erfasst worden ist: Zunächst ist erneut die Sebstauskunftsbefragung aus dem Projekt „Die Wirksamkeit des Lehrerbildungssystems in der Schweiz“ zu beleuchten (siehe auch Unterabschnitt 2.1.1.2). Die hierbei befragten angehenden Lehrkräfte sollten nicht nur angeben, ob bestimmte, als Standards formulierte Aspekte schulischer Leistungsfeststellung und -beurteilung in ihrer Ausbildung thematisiert worden sind, sondern auch einschätzen, wie intensiv sie sich mit diesen auseinandergesetzt haben. Dies geschah auf einer fünf-stufigen Ordinalskala, deren Stufen von „Ich habe nichts von diesem Standard gehört“ bis „Ich habe Theorie, Übung und Praxis systematisch miteinander verbunden“ (vgl. Oser, 2001, S. 251 u. f.) und erfolgte damit im Sinne der in diesem Projekt vorgenommenen Festlegung, dass eine Kompetenz als erworben gilt, wenn „analytisch, theoretisch, nachahmend und praktisch selbsttätig gehandelt worden ist“ (Oelkers & Oser, 2000, S. 57). Die Befunde waren allerdings ernüchtern: So ergab sich für die beiden Aspekte, von denen Oser (2001) exemplarisch berichtet („Kennt-

nis der Erfolgskriterien vermitteln lernen“ und „Leistungsfortschritt messen können“; vgl. ebd., S. 280 u. f.), dass diese oftmals entweder nur auf rein theoretischer (ca. 35 % bzw. 40 % der Selbstauskünfte) oder auf rein praktischer Ebene vermittelt wurden (ca. 40 % bzw. 20 % der Selbstauskünfte). Für Oser (2001) sind dies Indikatoren dafür, dass diese Aspekte in der schweizerischen Lehramtsausbildung nicht intensiv behandelt werden und dementsprechend die selbstempfundene Diagnosekompetenz der überwiegenden Anzahl der befragten angehenden Lehrkräfte deutliche Defizite aufweist (vgl. ebd., S. 282). Zu einem ähnlichen Befunden kommen auch Jäger-Flor & Jäger (2008) durch Auswertung der Selbstauskunftsbefragung von N = 754 deutschen Lehrkräften aus dem „Bildungsbarometer zum Thema Förderung im Bildungssystem“[22], sowie Marso & Pigge (1993, S. 155) im Rahmen ihres Literaturreviews für den US-amerikanischen Raum.

Des Weiteren ist hier die Untersuchung von Abs (2006) zur „Bildung diagnostischer Kompetenz in der zweiten Phase der Lehrerbildung“ zu nennen. In dieser wurde auf die Selbstauskunftsdaten von N = 1102 Referendar_innen aus der ersten Erhebungswelle des vom Bundesland Hessen in Auftrag gegebenen Evaluationsprojekts „Pädagogische Bilanz an Studienseminaren“ zurückgegriffen (vgl. ebd., S. 223 u. f.). Die Befragten sollten dabei in Anlehnung an das Vorgehen von Oser (2001) zu bestimmten, als Standards formulierten Aspekten schulischer Leistungsfeststellung und -beurteilung eine Selbsteinschätzung ihrer entsprechenden Kompetenz auf einer fünf-stufigen Notenskala vornehmen. Diese Selbstauskünfte waren fachspezifisch zu erteilen und ferner war anzugeben, ob und wenn ja, welche Lernangebote im bisherigen Verlauf ihres Referendariats zu diesen Standards gemacht worden waren (vgl. ebd., S. 238). Die erhobenen Daten wurden anschließend mit Hilfe verschiedener quantitativer Methoden ausgewertet. Dabei ergaben sich folgende Befunde (vgl. ebd., S. 226 u. f.):

- Mittelwertvergleiche ergaben, dass Referendar_innen aus Studienseminaren für das gymnasiale Lehramt ihre diagnostische Kompetenz signifikant geringer einschätzen als Referendar_innen aus anderen Studienseminaren. Diese Unterschiede zeigten sich zudem durchgängig für Referendar_innen, die unterschiedlich weit in der zweiten Ausbildungsphase fortgeschritten waren.
- Varianzkomponentenanalysen zeigten, dass das Unterrichtsfach, für das die Selbsteinschätzungen der diagnostischen Kompetenz vorgenommen werden sollte, sowie die Interaktion dieses Personenmerkmals mit weiteren Gruppierungsmerkmalen, nur einen geringen Teil der Varianz in den Kompetenzselbsteinschätzungen aufklärt (Merkmal Fach: 3 %; Interaktion Fach × Person: 3 %; Interaktion Fach × Aspekt von schulischer Leistungsfeststellung und -beurteilung: 4 %). Allerdings zeigte sich, dass die entsprechenden Varianzaufklärungen im Verlauf des Referendariats zunehmen, wenn man Gruppen von Referendar_innen, die unterschiedlich weit in ihrer Ausbildung fortgeschritten waren, miteinander vergleicht und dass der Anteil der aufgeklärten Varianz hier höher ausfällt, als bei anderen selbsteingeschätzten Kompetenzen (z. B. zur Unterrichtsplanung).

[22] Zusammengefasst aufgearbeitet zu finden bei Jäger (2009, S. 106 u. f.).

- Regressionsanalysen mit den Daten von Befragten, die ihr Referendariat nahezu abgeschlossen haben, ergaben, dass sich die Prädiktion der Kompetenzselbsteinschätzungen unterscheidet, je nach dem, welches schulartspezifische Studienseminare besucht wurde. Bei angehenden Gymnasiallehrkräften ergab sich eine höhere Varianzaufklärung für die von ihnen angegebenen Lernangebote der entsprechenden Standards als für deren Beurteilung der Arbeitsqualität ihrer Lehrerbildner_innen[23] (18 % vs. 3 %). Bei angehenden Lehrkräften der Grund-, Haupt-, Real-, und Förderschule ergab sich ein umgekehrtes, wenn auch weniger stark ausgeprägtes Verhältnis (11 % vs. 18 %).

Abs (2006) schlussfolgert hieraus, dass es für zukünftige Forschung sinnvoll erscheint, die diagnostischen Kompetenzen von Lehrkräften schulartspezifisch zu erfassen und dass (vor allem für den gymnasialen Bereich) eine nähere Überprüfung der didaktischen Settings in der zweiten Ausbildungsphase angestrebt werden sollte (vgl. ebd., S. 229 u. f.). Ferner ist der Befund, dass angehende Gymnasiallehrkräfte durchgängig ihre diagnostische Kompetenz geringer einschätzen, ein möglicher Hinweis auf ein habituelles Grundverständnis, dass „Schülern innerhalb der Schulform Gymnasium eine höhere Eigenverantwortung für ihre Lernergebnisse [zuzugestehen ist] [...] [und] dass die zweite Phase [offenbar] kein entgegensteuerndes Moment [zu diesem Grundverständnis] darstellt“ (ebd., S 229-230).

2.1.1.4. Ergänzende Bemerkungen und Zwischenfazit

Alles in allem liefern die drei vorgestellten Forschungsansätze zur diagnostischen Kompetenz von Lehrkräften bereits einen weiten Einblick in Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen. Dies hängt vor allem damit zusammen, dass jeder dieser Ansätze durch seine methodisch-theoretische Eigenheiten bestimmte Vorzüge aber auch blinde Flecken aufweist, womit sich deren Befundlagen insgesamt gesehen gegenseitig ergänzen.

Deutlich wurde aber auch, dass im naturwissenschaftsdidaktischen Diskurs der Trend vorherrscht, sich der diagnostische Kompetenz von Lehrkräften mit Hilfe des Kompetenzmodellierungsansatzes anzunähern. Entsprechende Studien, die einem der beiden anderen vorgestellten Forschungsansätze zugeordnet werden können, finden sich bisher kaum. Nach diesen Ansätzen vorzugehen, könnten für Naturwissenschaftsdidaktiken in Zukunft allerdings durchaus ertragreich sein, was sich beispielsweise an den Erkenntnissen, die mit dem Urteilsgenauigkeitsansatz für die sprachlichen Fächer und die Mathematik gewonnen wurden, zeigt. Insbesondere ist festzustellen, dass die empirische Befundlage zur diagnostischen Kompetenz von Naturwissenschaftslehrkräften überaus beschränkt ist, weswegen diesbezüglich für den weiteren Verlauf der vorliegenden Arbeit primär auf die im vorangegangen Abschnitt vorgestellten theoretische Überlegungen zurückgegriffen wird.

[23] Die Arbeitsqualität ihrer Lehrerbildner_innen wurde von den angehenden Lehrkräften auf Likert-Skalen-Items, wie beispielsweise „[m]ein Ausbilder / meine Ausbilderin vermittelt Begeisterung für die Arbeit mit Kindern und Jugendlichen“ (Abs, Peter, Gerlach-Jahn, & Klieme, 2009, S. 58), eingeschätzt. Für Details hierzu siehe Abs et al. (2009, S. 57 u. f.).

Ferner wurden an der Darstellung des Selbstauskunftsansatzes Übergangsbereiche zum strukturtheoretischen und zum berufsbiographischen Bestimmungsansatz von Professionalität im Lehrerberuf angedeutet. Dies weist aber gleichzeitig darauf hin, dass sich der Forschungsstand zu Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen nicht allein auf das Konstrukt der diagnostische Kompetenz von Lehrkräften beschränkt. Weitere, noch nicht dargestellte Einblicke zeigen sich daher auch im nächsten Abschnitt, in dem die Forschung zu Bezugsnormen und Bezugsnormorientierung von Lehrkräften im Vordergrund steht.

2.1.2. Bezugsnormen und Bezugsnormorientierungen von Lehrkräften

Woran es eine Lehrkraft ausmacht, ob eine Schülerleistung als gut oder schlecht zu beurteilen ist, wurden in der bisherigen Darstellung nur am Rand angesprochen. Dies ist – wie Rheinberg (2001) feststellt – eine Frage, die auf den Begriff der Leistungsbeurteilung bezogen ist, da mit diesem „[der] Vergleich eines ermittelten Ergebnisses mit einem Standard [ausgedrückt wird]“ (ebd., S. 59). Dieser Standard bzw. das zum Vergleich herangezogene Gütemaß lässt sich wiederum auf verschiedene Normen zurückführen. In der Literatur wird meist von *Bezugsnormen* gesprochen und dementsprechend wird die habituelle Verwendung einer bestimmten Bezugsnorm durch Lehrkräfte als *Bezugsnormorientierung* bezeichnet (vgl. Rheinberg & Fries, 2010, S. 61 u. f.).

2.1.2.1. Die drei Bezugsnormen veranschaulicht an der „kleinen Bewertungsaufgabe“

Insgesamt werden drei Bezugsnormen unterschieden: die *individuelle*, die *kriteriale* und die *soziale Bezugsnorm* (vgl. ebd.). Anstatt dieses abstrakt herzuleiten, sollen diese anhand eines bewehrten Instruments zur Erfassung dieser Bezugsnormen[24] veranschaulicht werden, der sog. „kleinen Bewertungsaufgabe“ (vgl. Rheinberg, 1980; Rheinberg, 2001). Diese ist in Abbildung 2.2 dargestellt. Die zugehörige Aufgabenstellung lautet wie folgt:

> „Eine durchschnittliche Schulklasse macht in monatlichen Abständen Schulleistungstest, in denen jeweils der Unterrichtsstoff des letzten Monats abgefragt wird. In jedem Test kann man maximal 100 Punkte erreichen. Die Tests sind so aufgebaut, dass der Klassendurchschnitt bei ca. 50 Punkten liegt. Neun Schüler erreichten bei den letzten drei Tests die unten angeführten Punkte.
>
> Ihre Aufgabe besteht darin, *bei jedem der neun Schüler das letzte Testergebnis zu beurteilen*. Wenn Sie das Ergebnis eines Schülers für eine gute Leistung halten, so können Sie *einen bis fünf Pluspunkte* (++...) geben. Halten Sie dieses Ergebnis für eine schlechte Leistung, so können Sie *einen bis fünf Minuspunkte* (- -...) geben. Bitte geben Sie pro Ergebnis

[24] Genauer ist die „kleine Bewertungsaufgabe“ ein Instrument, mit dem sich lediglich eine Orientierung an der individuellen und/oder der sozialen Bezugsnorm valide feststellen lässt. Die kriteriale Bezugsnormorientierung lässt sich an ihr dennoch veranschaulichen.

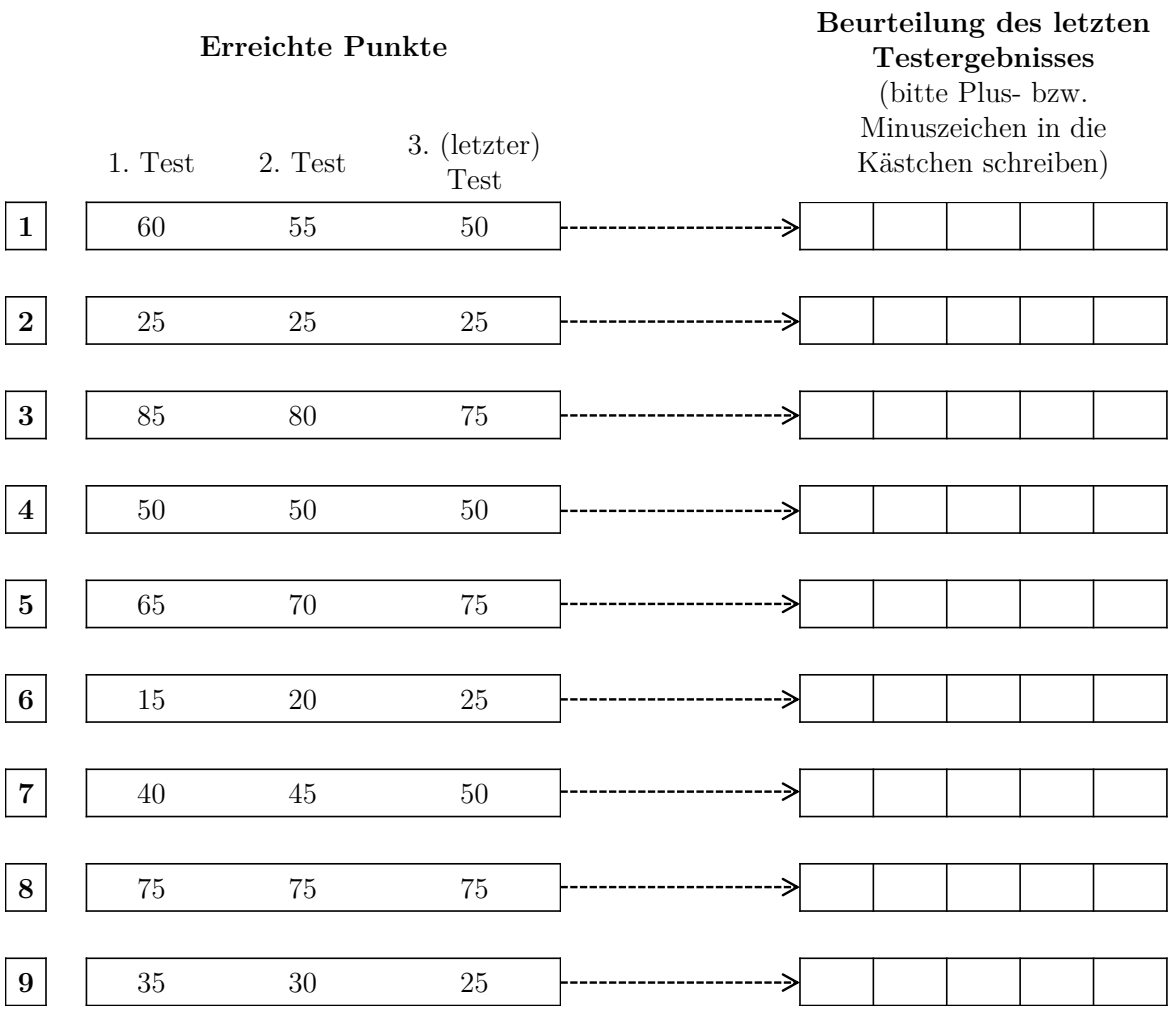


Abbildung 2.2.: Die kleine Bewertungsaufgabe. Übernommen aus Rheinberg (2001, S. 60).

> entweder *nur* Plus- oder *nur* Minuspunkte, also nicht beides gleichzeitig! Wenn sie in eine Zeile weder Plus- noch Minuszeichen schreiben, so bedeutet das, dass Sie das Ergebnis weder für eine gute noch für eine schlechte Leistung halten. Beziehen Sie sich bei Ihrer Beurteilung bitte auf eines Ihrer Unterrichtsfächer. [...] Es kann sein, dass Sie bei einigen Schülern sich unsicher über die „richtige“ Beurteilungsweise sind. Entscheiden Sie sich dann bitte so, wie Sie persönlich das für angemessen halten.“ (Rheinberg, 2001, S. 60, Hervorhebungen im Original)

Spätestens wenn man die „kleine Bewertungsaufgabe“ selbst durchführt, fällt auf, dass es verschiedenste Möglichkeiten gibt, eine Leistungsbeurteilung bei den einzelnen Schüler_innen vorzunehmen. Bei drei dieser Möglichkeiten werden die eben benannten Bezugsnormen in besondere Art und Weise sichtbar und damit unmittelbar verständlich:

1. Plus- und Minuspunkte können vergeben werden, je nach dem, ob vom ersten bis zum dritten Test die Punktzahl der Schüler_innen fällt, steigt, oder gleich bleibt. Gemäß dieser Norm erhalten Schüler_in 1, 3 und 9 Fünf Minuspunkte, Schüler_in 5 bis 7 fünf Pluspunkte und die übrigen drei Schüler_innen (2, 4 und 8) keinen Eintrag. Dies entspricht einer individuellen Bezugsnormorientierung, bei der der

Lern- bzw. Leistungsfortschritt der einzelnen Schüler_innen ausschlaggebend für die Beurteilung ist. Bei dieser wird eine Leistung als gut bzw. schlecht bewertet, wenn sich ein_e Schüler_in im Vergleich zu einem früheren Zeitpunkt verbessert bzw. verschlechtert hat (vgl. Sacher, 1996, S. 45).

2. Die Plus- und Minuspunkte können auch anhand der im dritten Test erreichten Punktzahl zugeordnet werden. Die maximal erreichbare Punktzahl (100 Punkte) wird dabei in 11 gleichgroße Punktintervalle untergliedert, da maximal fünf Minus- bzw. Pluspunkte, sowie kein Eintrag möglich sind. Gemäß dieser Beurteilungslogik erhalten Schüler_innen 2, 6 und 9 drei Minuspunkte, Schüler_innen 3, 5 und 8 drei Pluspunkte und Schüler_in 1, 4 und 7 keinen Eintrag. Die Beurteilung erfolgt hier also anhand der kriterialen Anforderung des dritten Tests (kriteriale Bezugsnorm), wobei eine gute Leistung genau dann vorliegt, wenn die Performanz eines_einer Schülers_Schülerin diesen Anforderung vollständig genügt (vgl. ebd.).

3. Eine weitere Möglichkeit der Vergabe von Plus- und Minuspunkte besteht darin, diese auf Grundlage der Durchschnittspunktzahl, welche die neun Schüler_innen im dritten Test erreicht haben (50 Punkte), zu verteilen. Die Leistungsbeurteilung erfolgt hierbei also anhand des Vergleichs einer Schülerleistung mit der Leistung einer Gruppe, was als soziale Bezugsnormorientierung bezeichnet wird (vgl. ebd.). Dementsprechen wird eine Schülerleistung, wenn sie die Leistung der Gruppe übertrifft (unterschreitet), als gut (schlecht) beurteilt. Bei der „kleinen Bewertungsaufgabe“ liegen die Punktzahlen der Schüler_in 3, 5 und 8 oberhalb des Klassendurchschnitts. Sie erhalten daher fünf Pluspunkte. Schüler_in 1, 4, und 7 haben genau die Durchschnittspunktzahl erreicht und bekommen keinen Eintrag, da ihre Leistung weder als gut noch schlecht zu beurteilen ist. Die übrigen drei Schüler_innen (2, 6 und 9) haben im dritten Test eine im Klassenvergleich unterdurchschnittliche Punktzahl erreicht und werden mit fünf Minuspunkten bewertet.

2.1.2.2. Das Verhältnis von Bezugsnormen zu Funktionen schulischer Leistungsfeststellungen und -beurteilungen

Das Beispiel der „kleinen Bewertungsaufgabe“ zeigt, dass gleiche Schülerleistungen, je nachdem welche Bezugsnorm angewandt wird, sehr unterschiedliche beurteilt werden können. Bei den drei Bezugsnormen handelt es sich aber nicht nur um mögliche Verfahrensweisen, die einer Lehrkraft zur Beurteilung von Schülerleistungen theoretisch zur Verfügung stehen. Vielmehr kommt hinzu, dass institutionelle Rahmenbedingungen von Lehrkräften verlangen, dass sie bei ihrer täglichen Arbeit auf jede dieser Bezugsnormen zurückgreifen, wie ein Blick in Richtung Zensurgebungspraxis vergegenständlicht:

- So geht nach Sacher (1996) die bis heute gültige Wortbedeutung der Noten „sehr gut“ bis „ungenügend“ auf den Beschluss der Kultusministerkonferenz vom 3. Oktober 1968 zurück (vgl. ebd., S. 51 u. f.). In diesem wurde beispielsweise festgelegt, dass „[d]ie Note „befriedigend“ [...] erteilt werden [soll], wenn die Leistung im

Allgemeinen den Anforderungen entspricht“ (Kultusministerkonferenz , 1968). Nach Meinung Sachers geht aus dem Begriff „Anforderungen“ eindeutig hervor, dass „hier letztendlich die kriteriale Bezugsnorm intendiert ist“ (Sacher, 1996, S. 52). Wenn es um die Vergabe von Noten geht, sollten Lehrkräfte dementsprechend auf die kriteriale Bezugsnorm festgelegt sein.

- Auf der anderen Seite wird von Lehrkräften verlangt sog. „pädagogische Zensuren“ zu verteilen, also ungünstige Lernvoraussetzungen der Schüler_innen durch eine „mildere“ Notenvergabe auszugleichen und/oder herausragende Anstrengungen durch „gute Noten“ zu honorieren (vgl. Jürgens, 1997, S. 41). Pädagogische Zensurenvergabe ist damit also gleichzusetzen mit der Vergabe von Noten unter Berücksichtigung der individuellen Bezugsnorm.
- Des Weiteren gilt in vielen Bundesländern ein sog. „Drittelerlass“, gemäß dem „bei Klassenarbeiten nicht mehr als ein Drittel der Zensuren «mangelhaft» und «ungenügend» sein soll“ (Mischo & Rheinberg, 1995, S. 149). Diese Erlasse zwingen Lehrkräfte also dazu, bei der Zensurenvergabe auch die soziale Bezugsnorm mit zu berücksichtigen (vgl. ebd.).

Der eben verdeutlichte Umstand lässt sich damit erklären, dass, isoliert betrachtet, keine der drei Bezugsnormen mit jeder Funktion, die schulische Leistungsfeststellungen und -beurteilungen erfüllen sollen, in Einklang zu bringen ist (vgl. Unterkapitel 1.2). So lassen sich die pädagogische und psychologische Funktionen am ehesten durch Anwenden der individuellen Bezugsnorm realisieren. Dies gilt jedoch weniger für die Repräsentationsfunktionen. Bei den sozialen Funktionen schulischer Leistungsfeststellungen und -beurteilungen kann sogar davon gesprochen werden, dass diese mit der individuellen Bezugsnorm im deutlichen Widerspruch stehen. Analoges gilt für die kriteriale und die soziale Bezugsnorm. Bei diesen lässt sich am ehesten mit die Repräsentationsfunktionen bzw. den sozialen Funktionen ein Zusammenhang herstellen, mit den jeweils übrigen Funktionen schulischer Leistungsfeststellungen und -beurteilungen jedoch eher kaum. Rheinberg (2001) spricht daher davon, dass jede der drei Bezugsnormen bestimmte Vorzüge, aber auch „blinde Flecken“ aufweist (vgl. ebd., S. 64 u. f.). Was hierdurch jedoch vor allem deutlich wird, ist, dass sich das antinomische Verhältnis der verschiedenen Funktionen schulischer Leistungsfeststellung und -beurteilung bei einer Betrachtung auf Ebene von Bezugsnormen bzw. Bezugsnormorientierungen auf die konkrete Situation, die Leistung eines_einer Schülers_Schülerin zu beurteilen, übersetzt. Anders ausgedrückt: Die Tendenz einer Lehrkraft, sich bei der Leistungsbeurteilung an einer bestimmten Bezugsnorm zu orientieren, kann verstanden werden als eine Fokussierung auf einen Teil der institutionelle Rahmenbedingungen, mit denen sie sich konfrontiert sieht. Ferner dient diese Fokussierung dem Zweck, durch das eigene Handeln einer bestimmten Auswahl von Funktionen, die schulische Leistungsfeststellungen und -beurteilungen erfüllen sollen, gerecht zu werden.

2.1.2.3. Bezugsnormorientierung und Lehrerunterschiede

Die Vermutung ist daher naheliegend, dass die Bezugsnormorientierungen von Lehrkräften „in Zusammenhang mit den Erziehungszielen [...] [stehen], von denen sich [...] [diese] im Unterricht leiten [...] [lassen]“ (Rheinberg & Fries, 2010, S. 65). Mischo & Rheinberg (1995) sind dieser Frage im Rahmen einer Studie mit N = 51 Gymnasiallehrer_innen nachgegangen. Mit Hilfe eines Selbstauskunftsfragebogens und anschließende Faktorenanalyse konnten die Autoren bei den befragten Lehrkräften zunächst vier handlungsleitende Erziehungsziele identifizieren, namentlich das „Bemühen um Förderung von Persönlichkeit und Sozialverhalten“, das „Bemühen um soziale Anerkennung und Zuneigung“, die „Ausrichtung an eher konservativen Bildungs- und Erziehungszielen“ (z. B. erhoben durch die Likert-Skalen-Items „Ehrgeiz des Schülers herausfordern“ oder „Beim Schüler Idealismus und Engagement für höhere Ziele vermitteln“) und das „Bemühen um disziplinierten Unterrichtsverlauf“ (vgl. ebd., S. 146). Eine anschließende multiple Regressionsanalyse zeigte allerdings, dass diese Ziele nur 18 % der Varianz in der individuellen Bezugsnormorientierung der befragten Lehrkräfte – erfasst mit Hilfe der „kleinen Bewertungsaufgabe“ – aufklären konnten (vgl. ebd., S. 148). Die Autoren vermuten daher, dass sich die Bezugsnormorientierung nur bis zu einem gewissen Grad als Strategie zur Erreichung von Erziehungszielen verstehen lässt (vgl. ebd., S. 139) und dass dieser...

> „[...] ein gewisser Eigenwert zu[kommt] und zwar als Konkretisierung der Überzeugung, daß man jemanden immer nur an dem messen kann, was ihm möglich ist (individuelle BnO [Bezugsnormorientierung; M. S. F.]) vs. an dem, was den meisten anderen Menschen gelingt, und deshalb – im Sinne gerechter Gleichbehandlung – auch von dieser Person erwartet werden darf (soziale BnO).“ (ebd., S. 149)

Anzumerken ist, dass das Ergebnis dieser Regressionsanalyse mit einem p-Wert von .068 als knapp nicht signifikant gelten muss und diese damit lediglich einen wenig belastbaren Befund liefert (vgl. ebd., S. 148). Dies gilt auch für den Forschungsstand im Allgemeinen: Worauf Unterschiede in der Bezugsnormorientierung zurückzuführen sind, konnte empirische Forschung bis dato nur unzureichend aufklären (vgl. Rheinberg & Fries, 2010, S. 66).

Ähnliches gilt auch für die Frage nach der Entwicklung der Bezugsnormorientierung von Lehrkräften. Bei diesem Entwicklungsprozess ist zumindest davon auszugehen, dass dieser nur langsam vonstatten geht und sich daher Änderungen in der Bezugsnormorientierung einer Lehrkraft wenn überhaupt nur langfristig zeigen (vgl. Rheinberg & Fries, 2010, S. 64). Eine der wenigen Studien, die hierzu Hinweise liefert, ist die Längsschnittuntersuchung von Rheinberg (1982). Im Rahmen dieser wurden N = 87 angehende Realschullehrer_innen zu Beginn, in der Mitte und am Ende ihres 1,5-jährigen Referendariats gebeten die „kleine Bewertungsaufgabe“ zu bearbeiten. Dabei zeigte sich, dass sich die – bezogen auf die Gesamtstichprobe – mittlere Bezugsnormorientierung über die gesamte Ausbildung hinweg kaum veränderte (vgl. ebd., S. 241 u. f.). Etwas deutlichere Ergebnisse zeigten sich allerdings in der Subgruppe von Lehramtsanwärter_innen, die zum ersten Messzeitpunkt nur eine leichte Tendenz zu einer bestimmten Bezugsnorm aufwiesen. Bei lediglich 13

Lehrer_innen mit dominanter individueller Bezugsnormorientierung...	**Lehrer_innen mit dominanter sozialer Bezugsnormorientierung...**
... tendieren dazu, Schülerleistungen auf situative, zeitlich variierende Ursachen, wie Fleiß oder Anstrengung, zurückzuführen. Dementsprechend gehen sie weniger von einer längerfristigen Vorhersagbarkeit von Schülerleistungen aus.	... tendieren dazu, Schülerleistungen auf zeitlich stabile Faktoren, wie Intelligenz oder Begabung, zurückzuführen. Sie erwarten dementsprechend, dass Schülerleistungen zeitlich gut vorhersagbar sind.
... orientieren sich bei Lob und Tadel von Leistungen eher an den Entwicklungstendenzen der_des entsprechenden Schüler_in.	... neigen dazu, die Leistungen von „fähigeren“ Schüler_innen zu loben und von „weniger fähigeren“ zu tadeln.
... bemühen sich um die Individualisierung von Anforderungen, sofern dies möglich ist.	... zeigen eine starke Tendenz ihren Unterricht angebotsgleich zu gestalten.

Tabelle 2.3.: Unterschiede zwischen Lehrkräften, die bei der Leistungsbeurteilung zur individuellen bzw. sozialen Bezugsnormen tendieren. Zusammengestellt aus den Darstellungen von Rheinberg & Fries (2010, S. 62 u. f.), sowie Sacher (1996, S. 53 u. f.).

dieser 46 Proband_innen zeigte sich im Vergleich zwischen Ausbildungsanfang und -ende keine Veränderung in der Bezugsnormorientierungen. Bei den übrigen Proband_innen zeigte sich hingegen entweder eine Verstärkung der anfänglichen Bezugsnormorientierungen (N = 20) oder ein deutlicher Umschlag zur im Vergleich zum Ausbildungsbeginn anderen Bezugsnorm (N = 13) (vgl. ebd, S. 245).

Neben dem bereits genannten Zusammenhang mit unterschiedlichen Erziehungszielen sind weitere Unterschiede zwischen Lehrkräften denkbar, die bei der Leistungsbeurteilung verschiedene Bezugsnormtendenzen zeigen. Empirisch ist dem vor allem für die individuelle und die soziale Bezugsnorm nachgegangen worden. In Tabelle 2.3 sind die hierzu, aufgrund des bisherigen Forschungsstandes als gesichert geltenden Erkenntnisse zusammengetragen. Zur kriterialen Bezugsnorm liegen hingegen bis dato nur unzureichend Forschungsergebnisse vor (vgl. Holmeier, 2013, S. 142). Sacher (1996) stellt diesbezüglich allerdings einige Mutmaßungen an. Er geht davon aus, dass eine kriteriale Bezugsnorm einer „weitgehend an der Struktur der Unterrichtsinhalte orientierten [Überzeugung entspricht]“ (ebd., S. 55). Ausgehend von dieser Vermutung schlussfolgert er, dass Lehrkräfte mit deutlicher Tendenz zu dieser Bezugsnorm...

... ihren Unterricht eher angebotsgleich gestalten (vgl. ebd.),

... wünschenswerte Anforderungen „um der Sache willen“ als nötig und damit für erfüllbar halten (vgl. ebd.) und

... dass sie dementsprechend davon ausgehen, dass der mehr oder weniger vorhandene „gute Willen“ der Schüler_innen die Ursache für deren Leistung ist, weswegen sie diesbezüglich durchaus auch mit kurzfristigen Änderungen rechnen (vgl. ebd.).

Aus Forschungsgesichtspunkten gilt es allerdings noch zu überprüfen, inwieweit sich diese Mutmaßungen erhärten lassen (vgl. ebd.).

2.1.2.4. Ergänzende Bemerkungen und Zwischenfazit

Aus der in diesem Abschnitt vorgenommenen Aufarbeitung des Erkenntnisstandes zu Bezugsnormen und Bezugsnormorientierungen von Lehrkräften ist deutlich geworden, dass hier vor allem aus Perspektive des strukturtheoretischen und berufsbiographischen Professionalitätsbestimmungsansatzes weiterhin Forschungsbedarf besteht. Insbesondere liegt zu den Ursprüngen und die Entwicklung unterschiedlicher Bezugsnormorientierungen, sowie zu Lehrerunterschieden, bezogen auf eine Tendenz zur kriterialen Bezugsnorm, bis dato noch zu wenig empirische Evidenz vor, um von einer gesicherten Befundlage zu sprechen (vgl. Unterabschnitt 2.1.2.3).

Ferner wurden Bezugsnorm und Bezugsnormorientierung ausschließlich im Zusammenhang mit Leistungsfeststellungen und -beurteilungen im schulischen Kontext im Allgemeinen diskutiert. Dies hängt wiederum damit zusammen, dass kaum erforscht ist, ob sich bei Lehrkräften unterschiedliche Bezugsnormtendenzen zeigen, je nachdem, welches Fach sie unterrichten. Auch dies muss als Desiderat gelten. Erste Hinweise diesbezüglich liefert allerdings die Studie von M. Maier (2001). Auf Basis einer Inhaltsanalyse von N = 468 Dritt- und Viertklassjahreszeugnissen (vgl. ebd., S. 138 u. f.) konnte er für Grundschullehrkräfte nachweisen, dass deren verbalen Zeugnisbeurteilungen in den Fächern Deutsch und Mathematik am häufigsten an der kriterialen Bezugsnorm orientiert sind (vgl. ebd., S. 181 u. f.). In den Fächern Kunst/Textiles Gestalten/Werken, Sachunterricht, Sport und Religion dominierte hingegen die individuelle Bezugsnorm (vgl. ebd.). Dass Lehrkräfte ihre Verbalbeurteilungen auch an der sozialen Bezugsnorm orientieren, konnte M. Maier (2001) zwar ebenfalls nachweisen, diese war im Vergleich zu den anderen beiden Bezugsnormen aber deutlich geringer ausgeprägt (vgl. ebd.).

Trotz der vielen, aufgrund mangelnder empirischer Evidenz, offen gebliebenen Fragen sprechen die in diesem Abschnitt vorgenommenen Kontrastierungen und die Darstellung sowohl theoretischer, wie auch empirischer Aspekte insgesamt jedoch eindeutig dagegen, dass Lehrkräfte bei schulischen Leistungsfeststellungen und -beurteilungen eine bestimmte Bezugsnorm generell bevorzugen sollten. „[Es] stellt sich [somit] [...] nicht die Frage nach einem Entweder-Oder, viel mehr geht es um ein sowohl als auch“ (Rheinberg & Fries, 2010, S. 66). Im Rahmen der vorliegenden Arbeit wird sich daher dem in der Literatur vorherrschenden Einvernehmen angeschlossen, dass sich ein wünschenswertes Lehrerverhalten durch eine Bezugsnormvielfalt auszeichnet (vgl. z. B. Lißman & Paetzhold, 1982, S. 213 u. f.; Blömeke, Herzig, & Tulodziecki, 2007, S. 208; Sacher, 2009, S. 99; Gläser-Zikuda, 2010, S. 373). Hierbei handelt es sich aber nicht um ein bloßes Wunschdenken, das sich auf den strukturtheoretischen Professionalitätsbegriff zurückführen lässt. In empirischen Befunden deuten sich an, dass es durchaus Lehrkräfte gibt, deren Handeln dem eben beschriebenen Ideal entspricht. So konnte Rheinberg (1980) nachweisen, dass es zwar durchaus Lehrer_innen gibt, „die es für richtig, gerecht und erstrebenswert halten, Leistungen ausschließlich im sozialen Vergleich zu bewerten“ (Holmeier, 2013, S. 135). Gleichzeitig zeigte sich aber, dass Lehrkräfte, die sich bei der Leistungsbeurteilung eher an der individuellen Bezugsnorm orientieren, nicht vollständig auf die soziale Bezugsnorm

verzichten, sondern je nach beabsichtigten Zweck zwischen diesen beiden Bezugsnormen hin und her wechseln (vgl. Rheinberg, 1987).

2.1.3. Berufsbezogene Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen

Im vorangegangenen Abschnitt ist am Beispiel von Bezugsnormen bzw. Bezugsnormorientierungen bereits angeklungen, dass subjektiven Überzeugungen von Lehrkräften eine bedeutende Rolle zukommt, wenn man ihr Wissen und Können zu schulischen Leistungsbegutachten in voller Breite erfassen möchte. In der Literatur finden sich diesbezüglich, neben dem Begriff der *berufsbezogenen Überzeugungen*, unterschiedlichste Termini (vgl. Pajares, 1992, S. 309). Nach Reusser, Pauli, & Elmer (2011, S. 479) zeichnet sich im deutschsprachigen Raum allerdings der Trend ab, mit Hilfe dieser Bezeichnung den in der internationalen Literatur am weitesten verbreiteten Begriff der *Teachers' Beliefs* zu übersetzen. In der Physikdidaktik wird jedoch vor allem der Terminus *Lehrervorstellung* verwendet (vgl. Klinghammer, Rabe, & Krey, 2016, S. 182). Die Begriffe „berufsbezogene Überzeugungen", „Teachers' Beliefs" und „Lehrerüberzeugungen" bzw. „-vorstellungen" werden daher im Folgenden synonym verwendet.

2.1.3.1. Der Begriff berufsbezogene Überzeugungen von Lehrkräften

Zunächst ist festzustellen, dass obwohl bereits mehrfach Bestimmungsanstrengungen unternommen worden sind (z. B. Pajares, 1992; Woolfolk Hoy, Davis, & Pape, 2006; Fives & Buehl, 2012), im erziehungswissenschaftlichen Diskrus bis dato kein allumfassender Konsens darüber besteht, was unter berufsbezogenen Überzeugungen von Lehrkräften zu verstehen ist[25]. Reusser et al. (2011, S. 479) führt dies auf die Unterschiedlichkeit der Annahmen und Methoden verschiedener Forschungstraditionen zurück, die sich mit diesem Themengebiet auseinandersetzen[26]. Nach Pajares (1992, S. 309) wiederum ergibt sich dieser Umstand im Wesentlichen daraus, dass verschiedene Autor_innen berufsbezogene Überzeugungen in unterschiedlicher Art und Weise von „Wissen" abgrenzen. Diese Abgrenzungsversuche lassen sich mit dem Problem der Unterscheidbarkeit von „Wissen 1" und „Wissen 2" gemäß der Modellierung des Lehrerwissens nach Neuweg (2014) gleichsetzen (vgl. Unterabschnitt 2.1.1.2). Neuweg (2014) selbst merkt hierzu folgendes an:

> „[Wissen 2] ist ein Begriff mit sehr unscharfen Rändern[,] [...] [unter anderem] weil die an „Wissen 1" anzulegenden Standards (z.B. Wahrheit, Wahrhaftigkeit, Begründbarkeit, Systematik) keine relevanten Einschlusskriterien für Mentales sind. Beispielsweise umgreifen kognitive Strukturen auch subjektive Theorien, Denkstile[,] [...] Überzeugungen [...] [und] Werthaltungen und entstehen über komplexe Prozesse der Transformation und Vernetzung

[25] Für eine Übersicht verschiedener Definitionen des Begriffs „Teachers' Beliefs" siehe z. B. Fives & Buehl (2012, S. 473).

[26] Eine Gesamtschau dieser Forschungstradition ist z. B. im historischen Überblick zur Forschung über „Teachers' Beliefs" von Ashton (2015) zu finden.

> von Informationen aus unterschiedlichen Wissensquellen. [...] Schon die Beziehung zwischen „Wissen 1“ und „Wissen 2“ ist [daher] komplex[.]“ (ebd., S. 584-586)

In diesem Zitat deutet sich insbesondere am metaphorischen Ausdruck „unscharfe Ränder“ an, dass das Verhältnis zwischen den Begriffen berufsbezogene Überzeugungen von Lehrkräften und „Wissen“ nicht als eine strikte Dichotomie zu verstehen ist, sondern eher für ein Kontinuum von Konzeptionsmöglichkeiten steht (vgl. Pajares, 1992, S. 311). Dementsprechend ist bezüglich der Unterscheidung dieser beiden Begriffe auch in Zukunft mit einer Vielzahl von Denkfiguren in der Forschung zu rechnen, die im unmittelbaren Vergleich miteinander unterschiedlich große Ähnlichkeit aufweisen können (vgl. Reusser et al., 2011, S. 479). Trotz alledem lässt sich der Literatur zumindest ein Leitkonzept entnehmen, dass den unterschiedlichen Konzeptionen des Begriffs „Teachers' Beliefs“ gemein ist (ebd.). Nach Skott (2015) umfasst dieses den folgenden „defining core“[27]:

> „The term [Teachers' Beliefs; M. S. F.] is used to designate individual, subjectively true, value-laden mental constructs that are the relatively stable results of substantial social experiences and that have significant impact on one's interpretations of and contribuations to classroom practice (Skott, 2013).“ (ebd., S. 19)

Vor dem Hintergrund dieses Leitkonzepts lässt sich nun der spärliche Forschungsstand zu berufsbezogenen Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen sinnvoll aufarbeiten.

2.1.3.2. Zusammenschau typischer berufsbezogener Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen

Einen ersten Einblick in die Forschung zu berufsbezogenen Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen liefert das bereits mehrfach erwähnte Literaturreview von Marso & Pigge (1993). Neben dem Befund, dass Sekundarstufenlehrkräfte für die Feststellung und Beurteilung von Schülerleistungen tendenziell vor allem auf schriftliche Klassenarbeiten zurückgreifen und auf diese als Informationsquelle vertrauen (vgl. ebd., S. 149 u. f.), liefern die Autoren in ihrem Literaturreview einen ausführlichen Überblick zu typischen berufsbezogenen Überzeugungen von Lehrkräften in im Kontext schulischer Leistungsfeststellung und -beurteilung, die in insgesamt 12 Forschungsarbeiten der 1980er Jahren identifiziert wurden. Sinngemäß übersetzt lauten diese wie folgt (vgl. ebd., S. 154 u. f.):

Lehrer_innen besitzen die berufsbezogene Überzeugung, dass...

- ... sie im Kontext schulischer Leistungsfeststellungen und -beurteilungen über geringere Kompetenzen verfügen, als in anderen berufsbezogenen Bereichen.
- ... schulische Leistungsfeststellungen und -beurteilungen eine anspruchsvolle aber auch undankbare Aufgabe ihres Berufsalltags darstellen.

[27] Analoge Beschreibungen dieses „defining cores“ finden sich beispielsweise auch bei Reusser et al. (2011, S. 480 u. f.), sowie Klinghammer et al. (2016, S. 182).

... schulische Leistungsfeststellungen und -beurteilungen Zeit- und Ressourceneffizient zu gestalten sind.

... schulische Leistungsfeststellungen und -beurteilungen das Lehren und Lernen im Klassenraum erleichtern.

... sich schulische Leistungsfeststellungen und -beurteilungen möglichst nah an den bisherigen Lerngelegenheiten der Schüler_innen orientieren sollten.

... ihre alltäglichen Erfahrungen und selbst durchgeführten Leistungsfeststellungen und -beurteilungen im Vergleich zu standardisierten Verfahren zuverlässigere Daten liefern, um unterrichts- und schülerbezogene Entscheidungen zu treffen.

... schulische Leistungsfeststellungen und -beurteilungen für sie nur dann von nutzen sind, wenn sie den Bedürfnissen ihres Unterrichts entsprechen, wenn sie einen „praktischen Wert“ haben und die entsprechenden Instrumente zur Leistungsfeststellungen und -beurteilungen auch einen kurzfristigen Einsatz ermöglichen.

... selbstentwickelte Instrumente zur Leistungsfeststellungen und -beurteilungen (z. B. eigene Klassenarbeiten) passgenauer zu den Bedürfnissen ihrer Klassen sind, als vorgfertigte (z. B. aus Praxiszeitschriften).

... die konkreten Unterrichtsinhalte und das in einer Klasse vorhandene Notenspektrum bestimmen, welche Instrumente zur Leistungsfeststellungen und -beurteilungen angemessen sind.

... in den Bereichen, in denen sich das Lernen von Schüler_innen eher in deren Handeln zeigt, Papier-und-Bleistift-Tests weniger vertrauenswürdig sind.

... sie in ihrer Ausbildung adäquates Hintergrundwissen zur schulischen Leistungsfeststellungen und -beurteilungen erworben haben, nicht aber zur erfolgreichen Integration von Leistungsfeststellung und -beurteilung in das tägliche Unterrichtsgeschehen.

... Verfahrensweisen, mit denen die Güte in standardisierte Testverfahren sichergestellt wird (z. B. Bestimmung eines Reliabilitätskoeffizienten) für sie nur geringen praktischen Wert besitzen.

... zur adäquaten Generierung von Noten und zum Fällen von Entscheidungen über Schüler_innen nicht nur die Ergebnisse von Klassenarbeiten, sondern auch weitere Informationen mit zu berücksichtigen sind (z. B. Schülerbeobachtungen).

... auch Schulverwaltung und Schüler_innen der Überzeugung sind, dass es lernförderlich ist, regelmäßig Klassenarbeiten durchzuführen.

... das Durchführen von Klassenarbeiten einen positiven Effekt auf Schüler_innen und ihrer Lernanstrengungen hat.

... das Durchführen von Klassenarbeiten, sowie die Besprechung der entsprechenden Ergebnisse sinnvoll genutzte Unterrichtszeit darstellt.

... Klassenarbeiten nützlich dafür sind, den Lernfortschritt von Schüler_innen zu diagnostizieren und zu dokumentieren, Noten zu generieren und um Schülergruppen einzuteilen.

... Klassenarbeiten ihnen helfen ihre Notengebung zu rechtfertigen.

... die Ergebnisse von Klassenarbeiten auch ohne mathematische Hilfsmittel interpretierbar und Schüler_innen vermittelbar sind.

... eine Klassenarbeit unterschiedliche Aufgabenformate enthalten sollte, damit sie fairer und passgenauer bezüglich der bisherigen Lerngelegenheiten der Schüler_innen sind.

... Klassenarbeiten auch kognitiv anspruchsvolle Aufgaben enthalten sollten.

... Leistungsaufgaben, in denen die Schüler_innen einen Text produzieren sollen, im Vergleich zu geschlosseneren Aufgabenformaten unpraktikabel sind, bei Schüler_innen weniger beliebt, größere Lernanstrengungen erfordern und auf einem höheren kognitives Anspruchsniveau anzusiedeln sind.

... bestimmte Formate von Leistungsaufgaben generell zweckdienlicher und ökonomischer sind als andere (eher zweckdienlich und ökonomisch sind: Zuordnungs-, Kurzantwort-, Vervollständigungs- und Multiple-Choice-Aufgabe; weniger zweckdienlich und ökonomisch sind Textproduktions- und Richtig-Falsch-Aufgaben).

Diese und eine immense Anzahl weiterer berufsbezogener Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen finden sich auch in Literaturübersichten neueren Datums (z. B. Brown, 2004, S. 304 u. f.; Barnes, Fives, & Dacey, 2015). Um einen besseren Gesamtüberblick zu erhalten sind in aktuelleren Arbeiten verschiedener Autor_innen Klassifizierungen dieser Vielzahl an berufsbezogenen Überzeugungen von Lehrkräften vorgenommen worden, zu denen Tabelle 2.4 einen Überblick gibt. Was aus dieser Übersicht unmittelbar deutlich wird, ist, dass die dort aufgeführten Klassifizierungen trotz gewisser Unterschiedlichkeiten, insgesamt große inhaltliche Ähnlichkeit aufweisen. Diese Ähnlichkeit ergibt sich zum einen aufgrund inhaltlicher Bezüge auf die verschiedenen Funktionen schulischer Leistungsfeststellungen und -beurteilungen (vgl. Unterkapitel 1.2), die sich in den Kategorien jeder der in Tabelle 2.4 aufgeführten Klassifizierungen identifizieren lassen (am deutlichsten in jener von Barnes et al., 2015), zum anderen aber auch durch in den Kategorien zu findende Hinweise auf den Umgang von Lehrkräften mit den Spannungen und Antinomien, mit denen sie im Kontext schulischer Leistungsfeststellung und -beurteilung konfrontiert sind (z. B. Verweise auf Rechenschaftspflichten oder auf die Relevanz schulischer Leistungsfeststellungen und -beurteilungen).

Ferner lohnt sich insbesondere ein Vergleich von Studien, in denen berufsbezogene Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen identifiziert wurden und die sich hinsichtlich ihrer schulpolitischen und/oder schulkulturellen Rahmenbedingungen voneinander unterscheiden. Barnes et al. (2015, S. 293 u. f.) stellen Zwecks eines solchen Vergleichs insgesamt sieben Studien einander gegenüber, die in Australien, der VR China, Hongkong, dem Iran, Neuseeland, den Niederlanden und Spanien durchgeführt wurden und denen gemeinsam ist, dass sie mit Hilfe des von Brown

Referenz	Kategorien für typische berufsbezogene Überzeugungen von Lehrkräften
(Brown, 2004)	Lehrer_innen besitzen berufsbezogene Überzeugungen... ... über schulische Leistungsfeststellungen und -beurteilungen als Beitrag zur Optimierung von Lehr-Lern-Prozessen. ... über Rechenschaftspflichten von Schüler_innen bezogen auf schulischer Leistungsfeststellung und -beurteilung. ... über Rechenschaftspflichten von Lehrer_innen und des Systems Schule bezogen auf schulische Leistungsfeststellung und -beurteilung. ... zur Irrelevanz schulischer Leistungsfeststellungen und -beurteilungen.
(Remesal, 2007)	Lehrer_innen besitzen berufsbezogene Überzeugungen über die Bedeutung schulischer Leistungsfeststellungen und -beurteilungen... ... für das Lernen von Schüler_innen. ... für die Unterrichtsgestaltung. ... für die Zertifikation institutionalisierten Lernens. ... als Indikator für die eigene Professionalität, aufgrund der sich aus ihnen ergebenden Rechenschaftspflicht über Schülerleistungen.
(Barnes, Fives, & Dacey, 2015)	Lehrer_innen besitzen berufsbezogene Überzeugungen... ... über die pädagogischen und psychologischen Funktionen schulischer Leistungsfeststellungen und -beurteilungen. ... über die sozialen Funktionen schulischer Leistungsfeststellungen und -beurteilungen. ... über die Repräsentationsfunktionen schulischer Leistungsfeststellungen und -beurteilungen und solche, die Mischformen der ersten beiden Kategorien sind. ... zur Irrelevanz schulischer Leistungsfeststellungen und -beurteilungen.

Tabelle 2.4.: Klassifikation berufsbezogener Überzeugungen von Lehrkräften zu schulischer Leistungsfeststellung und -beurteilung verschiedener Autor_innen.

(2004) entwickelten Likert-Skalen-Fragebogeninstruments und durch explorative und konfirmatorische Faktorenanalysen Rückschlüsse auf die berufsbezogenen Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen vorgenommen haben. Hierbei zeigten sich Unterschiede in den mit berufsbezogenen Überzeugungen von Lehrkräften verbundenen Werthaltungen:

1. Ein Vergleich der empirischen Befunde aus Ländern, deren Schulkultur vor allem Wert auf summatives Assessment legt, sowie deutlich vom Gedanken einer zentrale Überprüfung und Steuerung der Qualität des Bildungswesens geprägt ist (VR China, Hongkong, Iran) lässt vermuten, dass Lehrkräfte in diesen Ländern insbesondere Rechenschaftspflichten, die sich aus der Praxis schulischer Leistungsfeststellungen und -beurteilungen ergeben, ihrer Überzeugung nach gutheißen (vgl. Barnes et al., 2015, S. 294 u. f.).

2. In Ländern, in denen Schulautonomie betont wird und deren Schulkultur sowohl auf Assessment of Learning als auch auf Assessment for Learning Wert legt (Australien, Neuseeland, die Niederlande und Spanien) zeigt sich auf empirischer Ebene ein andersartiger Trend: Hier scheinen Lehrkräfte ihrer Überzeugung nach schulische Leistungsfeststellungen und -beurteilungen vor allem als positiven Beitrag zur Optimierung von Lehr-Lern-Prozessen zu sehen (vgl. ebd., S. 295 u. f.).

2.1.3.3. Ergänzende Bemerkungen und Zwischenfazit

Aus der Zusammenschau typischer berufsbezogener Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen deutet sich an, dass sich der Beitrag der vom Umfang her bisher stark begrenzten Forschung zu berufsbezogenen Überzeugungen von Lehrkräften für das weite Feld von Lehrerwissen und -können zu schulischer Leistungsfeststellungen und -beurteilungen in erster Linie auf strukturtheoretischer Ebene bewegt. Insbesondere die von Barnes et al. (2015) vorgenommene Gegenüberstellung von Studien, die auf das quantitative Erhebungsinstrument von Brown (2004) zurückgegriffen haben lässt vermuten, dass berufsbezogene Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen von schulpolitischen und/oder schulkulturellen Rahmenbedingungen und den dabei auftretenden Spannungen und Antinomien geformt werden. Die Autor_innen merken jedoch an, dass hierzu noch weitere Untersuchungen notwendig sind, vor allem zu den noch wenig erforschten Unterschieden zwischen Lehrkräften, die unter denselben schulischen Rahmenbedingungen agieren (Barnes et al., 2015, S. 298).

Ferner liefert der bisherige Forschungsstand über berufsbezogene Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen noch kaum Erkenntnisse aus einer speziell naturwissenschaftsdidaktischen, sowie einer berufsbiographischen Perspektive. Eine der wenigen zu diesen beiden Aspekten bisher veröffentlichte Studie stellt die explorative Untersuchung von M. A. Siegel & Wissehr (2011) dar. In der zugehörigen Publikation stellen die Autor_innen ein von ihnen konzipiertes Seminar für Lehramtsstudierende des Fachs Naturwissenschaften vor, dessen inhaltlicher Schwerpunkt das Thema formative schulische Leistungsfeststellungen und -beurteilungen ausmachte und berichten über Erkenntnisse, die sie durch qualitative Analysen von Theorie-Essays, Kurstagebüchern und Unterrichtsentwürfen der Seminarteilnehmer_innen gewonnen haben (vgl. ebd., S. 376). Die Analysen offenbarten dabei, neben einem Lernfortschritt der angehenden Naturwissenschaftslehrkräfte bezogen auf die Inhalte des Seminars, dass diese bereits berufsbezogene Überzeugungen aufweisen, wie sie Marso & Pigge (1993) in ihrem Literaturreview für praktizierende Lehrkräfte im Allgemeinen berichten (vgl. M. A. Siegel & Wissehr, 2011, S. 380 u. f.). Dieser Befund zeigt sich auch in den Ergebnissen der Mixed-Methods-Untersuchung von Ogan-Bekiroglu (2009), in der die berufsbezogenen Überzeugungen angehender Physiklehrkräfte in der Türkei untersucht wurden, sowie in der phänomenografischen Untersuchung von Wang, Kao, & Lin (2010), bei der die „conceptions about assessment of science learning“ angehende taiwanesische Grundschul-

lehrkräfte exploriert wurden. Bei den von M. A. Siegel & Wissehr (2011) untersuchten Unterrichtsentwürfen, sowie dem von Wang et al. (2010) analysierten Essays und Interviewtranskripten deutet sich zudem an, dass sich die berufsbezogenen Überzeugungen der befragten Studierenden deutlich an „traditional assessment goals“ orientieren (vgl. M. A. Siegel & Wissehr, 2011, S. 386), „[which] might trap [...] [them] into a teaching style committed to memorization of facts and focusing on what is in the textbook and text“ (Wang et al., 2010, S. 528). Beide eben genannten Befunde deuten mutmaßlich darauf hin, dass sich angehende (Naturwissensschafts-)Lehrkräfte „an den Unterrichtsausprägungen orientieren, die sie selbst als Schülerinnen und Schüler erlebt haben[,] [...] [die als] [t]ief verankerte Vorstellungen vom Lehren und Lernen in den Fächern [zu verstehen sind]“ (Klinghammer et al., 2016, S. 182). Ob sich dieser noch sehr vage Verdacht erhärtet, gilt es allerdings in zukünftiger (fachdidaktischer) Forschung zu klären.

2.1.4. Schulische Leistungsfeststellungen und -beurteilungen von Lehrer_innen aus Sicht der psychologischen Urteilsforschung

Neben den bereits vorgestellten Ansätzen lässt sich Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung auch aus Sicht der psychologischen Urteilsforschung betrachten, die den Prozess der Genese von Leistungsurteilen in den Vordergrund stellt (vgl. Kleber, 1992, S. 103; Leuders et al., 2014, S. 732) und damit „das Lehrerurteil nicht als Blackbox betrachtet, sondern gezielt [...] dem Mechanismus der Urteilsbildung [nachgeht]“ (Karst & Förster, 2017, S. 20). Einer derartigen Perspektive auf schulische Leistungsfeststellungen und -beurteilungen wird eine herausragende Bedeutung zugesprochen, da diese ermöglicht analytische und heuristische Urteilsbildung beschreiben und voneinander unterscheiden zu können (vgl. M. Böhmer, Englich, & Böhmer, 2017, S. 50), sowie Faktoren und Effekte zu identifizieren, welche die Genese von Leistungsurteilen moderieren bzw. beeinflussen (vgl. Förster & Böhmer, 2017, S. 46).

Im Rahmen dieser Forschung sind zum Teil höchst unterschiedlichen Modelle entstanden. Dabei finden sich in der Literatur...

... Modelle, die das Zustandekommen von Lehrerleistungsurteilen lediglich oberflächlich beschreiben und deshalb bezogen auf während der Urteilsbildung ablaufende Teilprozesse nur wenig erklärungsmächtig sind, wie beispielsweise das „naive Modell des Beurteilens“ von Jäger (2007, S. 62 u. f.).

... stark auf das jeweilige Erkenntnisinteresse der entsprechenden Autor_innen zugeschnittene Prozessmodelle, die sich kaum auf andere Kontexte übertragen und/oder verallgemeinern lassen (z. B. das Modell „der Beurteilung schriftlicher Leistungen in der Fremdsprache am Beispiel der Prüfung *Test Deutsch als Fremdsprache* (TestDaF)“ von Arras (2007, S. 441 u. f.) oder das Prozessmodell von Klug (2017) „zur Diagnose und Förderung von selbstreguliertem Lernen“).

... Modellierungen, deren Ausgangspunkt breiter gefasste Modelle der psychologischen Grundlagenforschung zur Urteilsbildung sind. Die prominentesten und gleichzeitig in der Forschung zur Genese von Lehrerleistungsurteilen bereits adaptierten Modelle sind das „Linsenmodell“ (vgl. Brunswik, 1956, S. 48. u. f.) und das Prozessmodell von Nickerson (1999) zur „Genese von Wissen über das Wissen anderer“.

In diesem letzten Abschnitt zum Erkenntnisstand zu Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen soll lediglich auf die zuletzt genannte Gruppe von Modellen näher eingegangen werden. Grund hierfür ist, dass sich bereits aus den eben aufgeführten Kurzcharakteristika abzeichnet, dass lediglich diese Modellgruppe gewinnbringend für das Ziel dieses Kapitels ist, die Frage zu klären, was eine im Kontext von schulischer Leistungsfeststellung und -beurteilung professionell handelnde (Physik-) Lehrkraft auszeichnet, sowie hierzu ein geeignetes Rahmenkonzept vorzustellen und zu erörtern (vgl. Kapitel 2; Einleitung). Überraschend ist allerdings, dass sich nur vereinzelt Arbeiten finden lassen, die explizit auf Basis derartiger Modelle dem Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen nachgegangen sind (vgl. Jäger, 2009, S. 113 u. f.; Leuders et al., 2014, S. 733). Hinzu kommt, dass sich diese Arbeiten überwiegend auf theoretischer Ebene bewegen und sich die bisherigen empirischen Befundlagen nahezu durchgängig in einem nicht speziell naturwissenschaftsdidaktischen Kontext ansiedeln. Im Folgenden werden daher die beiden genannten Modelle (das Linsenmodell und das Modell von Nickerson) vor allem theoretisch dargestellt, sowie Stärken der jeweiligen Modellierungen diskutiert. Abschließend erfolgt jeweils eine kurze Übersicht der empirischen Forschungsarbeiten, die in einem erziehungswissenschaftlichen Kontext und auf Grundlage dieser Modelle durchgeführt wurden.

2.1.4.1. Die Genese von Lehrerleistungsurteilen aus Perspektive des Linsenmodells

Wie in Unterkapitel 1.1 bereits dargelegt wurde, adressiert das Begriffspaar Leistungsfeststellung und -beurteilung schulische Verfahren, die dazu dienen das Lernen von Schüler_innen in Relation zu einem Gütemaß zu setzen. Je nach der hinter einer Leistungsfeststellung und -beurteilung liegenden Intention stehen hierbei bestimmte Merkmale von Schüler_innen im Fokus (vgl. Abschnitt 1.1.2). Oftmals handelt es sich dabei um nicht direkt beobachtbare Merkmale in Form eines Konstrukts (vgl. Kleber, 1976, S. 58; Kleber, 1992, S. 129 u. f.), wie z. B. die Schülerkompetenz im Bereich physikalisches Fachwissen. Ein allgemeines Modell zur Erklärung, wie Einschätzungen von derartigen Schülermerkmalen durch Lehrkräfte zustande kommen, ist das auf Brunswik (1956, S. 48. u. f.) zurückgehende Linsenmodell, das in Abbildung 2.3 dargestellt ist.

In diesem Modell wird der Prozess der Leistungsurteilsgenese in Analogie zum Verlauf von Lichtstrahlen durch zwei konvexe Linsen dargestellt, wobei sich das nicht direkt beobachtbare Schülermerkmal und die Einschätzung der Lehrkraft bildhaft gesprochen in den beiden Brennpunkten der Apparatur befinden (vgl. Kleber, 1976, S. 59). Dabei wird angenommen, dass ein_e Lehrer_in auf ein nicht direkt beobachtbares Schülermerkmal

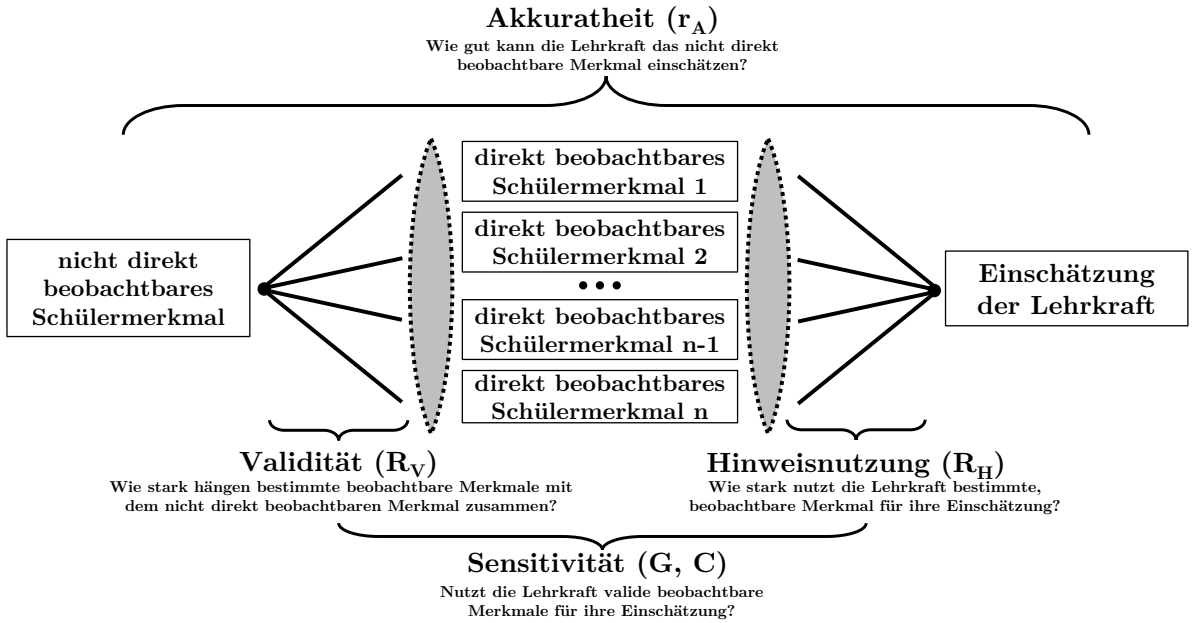


Abbildung 2.3.: Das Linsenmodell nach Brunswik (1956, S. 48 u. f.). Adaptiert aus Kleber (1976, S. 58 u. f.), sowie Förster & Böhmer (2017, S. 47 u. f.).

schließt, indem er_sie sich auf eine Reihe von direkt beobachtbaren Merkmalen[28], die ihm_ihr als „Hinweisreize oder Signale“ (Förster & Böhmer, 2017, S. 47) dienen, fokussiert. Diese *Hinweisnutzung* beschreibt aber lediglich, wie stark eine Lehrkraft bestimmte, beobachtbare Merkmale in ihrer Einschätzung berücksichtigt. Unbeantwortet bleibt, ob und wenn ja wie stark ein tatsächlicher Zusammenhang zwischen diesen Hinweisreizen und dem einzuschätzenden, nicht direkt beobachtbaren Schülermerkmal besteht (*Validität*). Die Hinweisnutzung wird daher von der *Sensitivität* unterschieden, die angibt, inwieweit eine Lehrkraft im Prozess der Leistungsurteilsgenese auf beobachtbare Merkmale zurückgreift, die auch mit dem nicht direkt beobachtbaren Schülermerkmal zusammenhängen (vgl. ebd., S. 48). Von dieser kann wiederum plausibel angenommen werden, dass sie die *Akkuratheit* einer konkreten Leistungsfeststellung und -beurteilung dieser Lehrkraft erklärt (vgl. ebd.).

Diese Überlegungen und Begrifflichkeiten lassen sich in Regressionsmodelle übersetzen, die wiederum zueinander in Beziehung stehen (vgl. M. Böhmer et al., 2017, S. 50). Hierbei wird zunächst die Akkuratheit als Produkt-Moment-Korrelation r_A zwischen der Einschätzung der Lehrkraft und der „tatsächlichen“ Ausprägung des entsprechenden Schülermerkmals verstanden und entspricht damit also der Vergleichskomponente des Urteilsgenauigkeitsansatzes (vgl. Unterabschnitt 2.1.1.1). Hinweisnutzung und Validität sind wiederum jeweils durch eine multiple Regression der Einschätzung der Lehrkraft bzw. der „tatsächlichen“ Ausprägung des nicht beobachtbaren Schülermerkmals mit den beobachtbaren Schülermerkmalen modellierbar und die zugehörigen multiplen Regressionkoeffizienten R_H und R_V damit Indikatoren für diese beiden Komponenten des Linsenmodells (vgl.

[28]Kleber (1976, S. 58 u. f.) bezeichnet diese direkt beobachtbaren Merkmale zusammenfassend als „proximale Merkmals-Linse“ einer Lehrkraft.

Abbildung 2.3). Die Beziehung zwischen diesen drei Regressionsmodellen lässt sich nun mit Hilfe einer sog. Linsenmodellgleichung beschreiben. In der Literatur wird dabei meist auf die folgende von Tucker (1964) vorgeschlagene Formel zurückgegriffen:

$$r_A = G \cdot R_H \cdot R_V + C \cdot \sqrt{1 - R_H^2} \cdot \sqrt{1 - R_V^2}$$

Die Faktoren G und C lassen sich mit Hilfe der zuvor aufgestellten Regressionsmodelle direkt ermitteln und beschreiben die „lineare“ und die „nicht lineare Passung“ zwischen den beiden multiplen Regressionsmodellen, die zur Bestimmung der Hinweisnutzung und der Validität herangezogen wurden (vgl. Hammond, Stewart, Brehmer, & Steinmann, 1975, S. 288 u. f.; Cooksey, Freebody, & Davidson, 1986, S. 47 u. f.). Sie können daher interpretiert werden als Maße für die Sensitivität einer Lehrkraft (vgl. ebd.).

Aus dieser Übersetzung des Linsenmodells in eine mathematische Gleichung, die sich wiederum aus empirisch erfassbaren Indikatoren aufbaut, zeigt sich die Stärke einer derartigen Modellierung des Zustandekommens von Lehrerleistungsurteilen: Im Unterschied zum Urteilsgenauigkeitsansatz lässt sich hier nicht nur ein Indikator für das Ausmaß, in dem eine Lehrerleistungsfeststellung und -beurteilung mit dem „tatsächlichen“ Schülermerkmal übereinstimmt angeben, sondern dieses Ausmaß auch erklären durch die generelle Hinweisnutzung der Lehrkraft (R_H), die Prädiktionierbarkeit des nicht direkt beobachtbaren Schülermerkmals durch beobachtbare Merkmale (R_V) und die Nutzung valider bzw. die Nichtnutzung nicht valider beobachtbare Merkmale durch die Lehrkraft (G und C).

Aufbauend auf dieser Überlegung lassen sich nun plausible prototypische Handlungsweisen von Lehrer_innen postulieren, die wiederum in beliebig vielen Mischtypen gedacht werden können. Kleber (1992) unterscheidet dabei die drei folgenden:

> „- Der Beurteiler verwendet [...] durch empirische Forschungsergebnisse begründbare [...] [direkt beobachtbare] Merkmale. Er gibt sich vor seinen Beurteilungen darüber Rechenschaft und bezieht seine proximale Merkmals-Linse in den Beurteilungsvorgang als vorhanden mit ein. Er kann seine proximale Merkmals-Linse auf Befragung sofort mit Kommentaren angeben. Er kennt eine Vielzahl [...] [derartiger] Merkmale, die er durch Studium der Fachliteratur weiter kritisch hinterfragt und verbessert.
> - Der Beurteiler verwendet weniger fundierte [direkt beobachtbare] Merkmale. Seine proximale Merkmals-Linse besteht aus validen und nicht-validen Merkmalen. Er verwendet seine [...] Merkmale ohne Reflexion. Auf Befragung kann er zunächst keine näheren Angaben über seine proximale Merkmals-Linse machen. Nach einigem Nachdenken nennt er einige Glieder seiner proximalen Merkmals-Linse.
> - Der Beurteiler verwendet unreflektiert implizite Persönlichkeitskonzepte, er urteilt intuitiv. Auf Befragen kann er zunächst keine Glieder seiner proximalen Merkmals-Linse angeben. Er beruft sich auf die Erfahrungen und auf die Unterschiedlichkeit einer beobachtbaren Population.“
>
> (ebd., S. 130-131)

Trotz der genannten Vorzüge des Linsenmodells und auch tiefgründigen Diskussionen über dessen Potential[29] für die Untersuchung von Lehrerwissen und -können (z. B. Snow,

[29] Hierbei wurde sich auch kritisch zum Linsenmodell geäußert. Kritisiert wird dabei vor allem, dass die Annahme hochgradig spekulativ ist, „die bei der Urteilsbildung stattfindende Informationsintegration

Referenz	Eingeschätztes Schülermerkmal	r_A	R_H	R_V	G	C
(Byers & Evans, 1980)	Reading Preferences	.23	.68	.67	.31	.16
(Cooksey et al., 1986)	Word Knowledge	.58	.96	.72	.87	-.02
	Reading Comprehension	.56	.92	.66	.93	-.00
(Marksteiner et al., 2012)	Academic Dishonesty	.02	.89	.30	nicht angegeben	nicht angegeben

Tabelle 2.5.: Zusammenschau empirischer Studien, die die Genese von Lehrerleistungsurteilen mit Hilfe des Linsenmodells untersucht haben. Die hier angegebenen Linsenmodell-Parameter sind die von den jeweiligen Autor_innen angegebenen Mittelwerte bezogen auf das jeweilige Gesamtsample an (angehenden) Lehrkräften.

1968; Shulman & Elstein, 1975, S. 25 u. f.), ist dieses Modell in der empirischen erziehungswissenschaftlichen Forschung bisher kaum eingesetzt worden (vgl. Cooksey et al., 1986, S. 49; Förster & Böhmer, 2017, S. 49). Tatsächlich lassen sich bis dato lediglich drei Studien ausfindig machen, die sich explizit auf das Linsenmodell beziehen (siehe Tabelle 2.5). Aus den in Tabelle 2.5 aufgeführten mittleren Linsenmodell-Parametern lässt sich dabei zusammenfassend festhalten, dass diese wenigen Befunde darauf hindeuten, dass sich auch auf empirischer Ebene die benannte Stärke dieses Modells zeigt, nämlich „die Akkuratheit bestimmter Beurteiler bezüglich bestimmter Zielpersonen und Eigenschaften und in ausgewählten Situationen [...] untersuchen und [...] erklären [zu können]“ (Förster & Böhmer, 2017, S. 48). So lässt sich die geringe mittlere Akkuratheit in der Studie von Marksteiner, Reinhard, Dickhäuser, & Sporer (2012) durch eine geringe Validität der direkt beobachtbaren Schülermerkmale und bei Byers & Evans (1980) durch eine geringe Sensitivität der befragten Lehrer_innen erklären. Analog sind die im Vergleich hierzu hohen mittleren Akkuratheiten der Lehrereinschätzungen bei Cooksey et al. (1986) auf durchgehend hohe Linsenmodell-Parameter zurückführbar. Aussagen darüber, ob sich die von Kleber (1992) postulierten Handlungstypen auch empirisch bestätigen, lassen sich auf Grundlage dieser Studien jedoch nicht treffen. Dies hängt damit zusammen, dass die Autor_innen der benannten Studien vor allem auf quantitative Forschungsmethoden zurückgreifen, ein empirischer Zugang zu den Handlungstypen von Kleber aber eher in einer qualitativen Befragung von Lehrkräften liegt. Ergänzend zu erwähnen ist zudem, dass sich sowohl die Studie von Cooksey et al. (1986), welcher ein enges Leistungsbegriffsverständnis zugrunde liegt (vgl. Abschnitt 1.1.1), als auch jene von Marksteiner et al. (2012), in der das Erkennen des Schülermerkmals „Academic Dishonesty“ (Erkennen eines Täuschungsversuchs eines_einer Schüler_in in einer Leistungssituation) untersucht wurde, dem Kontext von schulischer Leistungsfeststellung und -beurteilung zweifelsfrei zugeord-

[liese sich] durch [...] algebraische Gleichungen darstellen“ (M. Böhmer et al., 2017, S. 50). Es ist daher weniger ein Modell des „tatsächlichen“ Prozess einer Urteilsbildung, sondern beschreibt die Urteilsgenese vielmehr „als ob“ sie algebraischen Regeln folgt (vgl. Shavelson & Stern, 1981, S. 458; Cooksey et al., 1986, S. 49).

net werden können. Byers & Evans (1980) untersuchen dagegen, wie akkurat Lehrkräfte die „Reading Prefrences“ (Lesevorlieben) von Schüler_innen einschätzen können, also ein Schülermerkmal, das nur einem weiten Leistungsbegriffsverständnis zugeordnet werden kann und dementsprechend den Kontext von schulischer Leistungsfeststellung und -beurteilung eher peripher betrifft.

2.1.4.2. Die Genese von Lehrerleistungsurteilen aus Perspektive des Modells von Nickerson

Nun zum Prozessmodell von Nickerson: Sehr allgemein gesprochen sind schulische Leistungsfeststellungen und -beurteilungen Wissensgeneseprozesse einer oder mehrerer Personen über das Wissen anderer Personen. Für die Findung derartiger Personenurteile hat Nickerson (1999) ein allgemeines Prozessmodell entwickelt (siehe Abbildung 2.4), zu dem er eine Vielzahl stützender Befunde aus verschiedenen Studien zusammenträgt (vgl. Ostermann, Leuders, & Nückles, 2015, S. 51). Dieses Modell geht, im Gegensatz zum Linsenmodell von Brunswik (1956), nicht von einer analytischen, sondern von einer heuristischen Informationsverarbeitung aus. Hier wird also nicht angenommen, dass die Informationsintegration während des Urteilsprozesses einer algebraischen Logik folgt, sondern im wesentlichen „durch verkürzte, einfach anwendbare, automatisierte und somit meist unbewusste Urteilsstrategien [bestimmt wird]“ (M. Böhmer et al., 2017, S. 51), zu denen in der psychologischen Grundlagenforschung eine breite Studienlage existiert[30].

Eine der bekanntesten solchen Urteilsheuristiken, die sog. „Anker- und Anpassungsheuristik“, liegt dabei dem Aufbau des Prozessmodells von Nickerson zugrunde (vgl. Nickerson, 1999, S. 740). Gemäß dieser Heuristik bilden sich Menschen ein Urteil, indem sie von einem Anfangs-Anker ausgehen und anschließend Adjustierungen dieser Ankers vornehmen (vgl. Tversky & Kahnemann, 1974, S. 1128 u. f.). Nickerson geht in seinem Modell nun davon aus, dass die urteilende Person ihr eigenes Wissen als Anker für das Wissen der zu beurteilenden anderen Person heranzieht (vgl. Nickerson, 1999, S. 740). Indem die urteilende Person die aus ihrer Sicht ungewöhnlichen Facetten ihres eigenen Wissens ausklammert, nimmt sie eine erste Anpassung dieses Ankers vor und gelangt so zunächst zu einem Modell über das Wissen einer beliebigen anderen Person (vgl. ebd., S. 741). In einem zweiten Schritt wird dieses Modell dann in ein Ausgangsmodell für das Wissen der zu beurteilenden anderen Person weiter adjustiert. Zum einen geschieht dies, indem bereits vorhandene Kenntnisse über die zu beurteilende Person unmittelbar berücksichtigt werden und zum anderen dadurch, dass auf das Wissen der zu beurteilenden Person aufgrund ihrer (mutmaßlichen) Gruppenzugehörigkeit geschlossen wird (z. B. Gymnasialschülerin der sechsten Jahrgangsstufe zu sein) (vgl. ebd., S. 741 u. f.). In einer dritten und letzten Phase wird dieses Ausgangsmodell schließlich fortlaufend durch die Informationen, die bei der momentanen Interaktion mit der zu beurteilenden Person erhalten werden, zu einem immer feineren Arbeitsmodell des Wissens dieser Person modifiziert (vgl. ebd., S. 742).

[30] Für eine Übersicht dieser Forschung siehe z. B. Fiedler & von Sydow (2015).

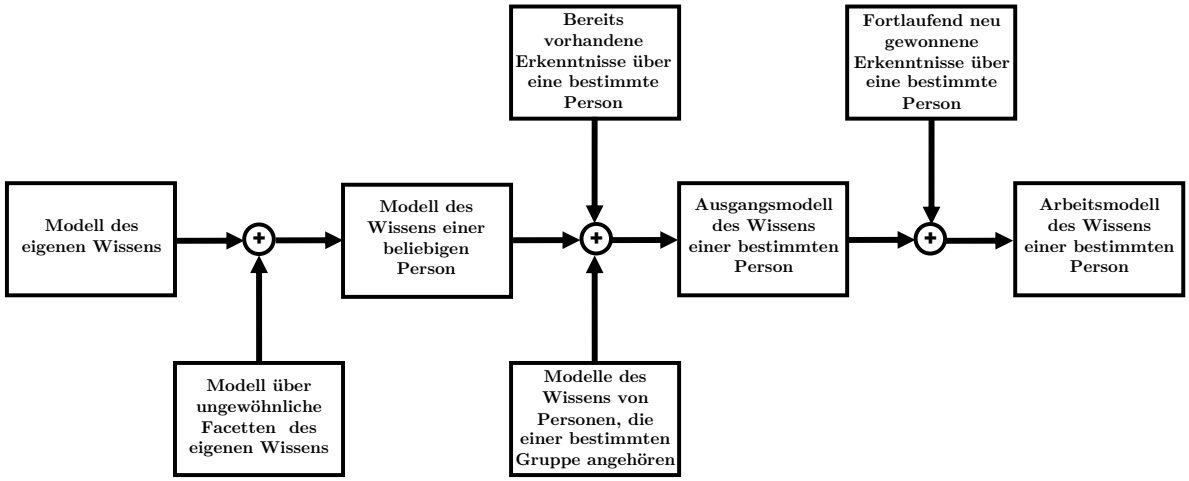


Abbildung 2.4.: Schematische Darstellung des Prozessmodells von Nickerson (1999) zur „Genese von Wissen über das Wissen anderer“.

Zu beachten ist, dass nicht bei jeder Einschätzung des Wissens eines_einer Anderen jede der drei eben beschriebenen Phasen durchlaufen werden muss (vgl. Nickerson, 1999, S. 740 u. f.). Wenn die zu beurteilende Person der urteilenden Person wohl vertraut ist, kann letztere oftmals direkt auf ein Ausgangsmodell des Wissens der zu beurteilenden Person zurückgreifen, da Menschen in einer länger andauernden Interaktion miteinander ein aus dem Gedächtnis abrufbares Modell davon entwickeln, welche Person über welche Wissensbestände verfügt (vgl. Wegner, 1995, S. 326 u. f.; Nickerson, 1999, S. 739). Ist die zu beurteilende Person jedoch unbekannt oder aber „[s]oll die Abschätzung auf einem Gebiet erfolgen, für das keine speziellen Annahmen über eine vertraute Person vorliegen, setzt der Prozess – wie geschildert – beim eigenen Wissen an“ (Jucks, 2001, S. 16).

Ergänzend zu der allgemeinen Beschreibung seines Modells hebt Nickerson (1999) hervor, dass einige aus der psychologischen Grundlagenforschung bekannte Effekte, die zu einer Verzerrung von Personenurteilen führen können, mit seinem Modell im Einklang stehen. Er benennt dabei explizit...

... den Effekt, dass Menschen den Anteil von Allgemeinwissen am eigenen Wissens oftmals überschätzen (vgl. ebd., S. 747 u. f.),

... den Effekt tendenziell zu überschätzen, inwieweit andere in derselben Art und Weise denken wie man selbst (*False Consensus Effect*) (vgl. ebd., S. 748 u. f.),

... die *Illusion der Einfachheit* (je vertrauter einer Person mit einem Sachverhalt ist, umso mehr neigt sie dazu diesen als „einfach“ einzuschätzen) (vgl. ebd., S. 750) und

... den Effekt, dass eigene Fachexpertise es einer Person eher schwieriger macht sich in das Leistungsvermögen eines Novizen und dessen Lehrnschwierigkeiten hineinversetzen zu können (der Effekt eines *Curse of Knowledge*) (vgl. ebd.).

Auch wenn Nickerson sich in seiner Darstellung nicht explizit auf schulische Leistungsfeststellung und -beurteilung bezieht, ist die Vermutung naheliegend, dass diese eben

aufgeführten allgemeinen Begleiterscheinung sozialer Urteilsbildung auch in diesem Kontext anzutreffen sind (z. B. als Verzerrungseffekte bei der Genese von Leistungsurteilen durch Lehrer_innen, die zur kriterialen Bezugsnorm tendieren; vgl. Abschnitt 2.1.2).

Insgesamt ist damit ein Lehrerleistungsurteil über eine_n Schüler_in nach dem Nickersonschen Prozessmodell als ein „Auszug“ des Arbeitsmodells der Lehrkraft über das Wissen der_des entsprechende_n Schüler_in zu verstehen. Inwieweit dieses Arbeitsmodell mit dem „tatsächlichen“ Wissen der_des Schülers_Schülerin übereinstimmt, bestimmt dementsprechend die Akkuratheit dieses Leistungsurteils. Da das Prozessmodell von Nickerson ferner eine in Phasen gegliederte Beschreibung angibt, wie eine Lehrkraft zu einem solchen Arbeitsmodell des Wissens eines_einer Schülers_Schülerin gelangt, liefert dieses – ähnlich wie auch das Brunswiksche Linsenmodell – Erklärungen für die Akkuratheit der Leistungsfeststellung und -beurteilung eines_einer Lehrers_Lehrerin: Im Allgemeinen hängt sie vom eigenen Wissen der Lehrkraft ab und der Einschätzung, welche Facetten ihres Wissens aus ihrer Sicht als ungewöhnlich gelten müssen. Des Weiteren hängt sie auch davon ab, wie gut die Lehrkraft den_die entsprechende Schüler_in bereits kennt, welche Gruppenzugehörigkeiten sie ihm_ihr zuschreibt und ob sie bei Menschen, die diesen Gruppen angehören, bestimmte Wissensbestände erwartet und falls ja welche. Zuletzt wird die Akkuratheit im Allgemeinen auch davon bestimmt, welche neuen Erkenntnisse die Lehrkraft über den_die zu beurteilenden_beurteilende Schüler_in während des Prozess der Leistungsfeststellung und -beurteilung gewonnen hat, da auch diese einen Einfluss auf ihr Arbeitsmodell über das Wissen des_der entsprechenden_entsprechende Schülers_Schülerin haben.

Allerdings gibt es auch einige wenige empirische Forschungsarbeiten, die das Prozessmodell von Nickerson als Grundlage wählen und die dem Kontext von schulischer Leistungsfeststellung und -beurteilung zugeordnet werden können. Aktuell lassen sich drei derartige Studien in der Literatur ausfindig machen:

Die Arbeiten von Krispenz, Dickhäuser, & Reinhard (2016) und Ostermann et al. (2015) widmen sich dabei dem Thema Aufgabenschwierigkeit, also der „Einschätzung von *Aufgabenanforderungen* vor der Bearbeitung durch Lernende[,] [...] [die] unabhängig vom Wissen um Lerngruppen und Individuen gefällt [wird]“ (Ostermann et al., 2015, S. 48, Hervorhebung im Original). Krispenz et al. (2016) legten hierzu N = 79 Lehramtsstudierenden Items aus den PISA-Studien von 2003 und 2006 vor und ließen diese die Aufgabenschwierigkeit auf einer fünf-stufigen Likert-Skala einmal bezogen auf sich selbst und zweites mal bezogen auf eine_eine typischen_typische Neuntklässler_Neuntklässlerin schätzen. Aus welcher Domäne die Items, die den Studierenden vorgelegt wurden, stammten, wurde dabei anhand ihrer jeweiligen Fächerkombination festgelegt. Zudem wurde das Gesamtsample in eine Experimental- (N = 39) und eine Kontrollgruppe (N = 40) geteilt. In der Experimentalgruppe erhielten die Studierenden einen zusätzlichen Impuls, der sie dazu anregte, sich bei der Einschätzung der Aufgabenschwierigkeiten verstärkt mit der Frage auseinanderzusetzen, welches Wissen ein_e Neuntklässler_in typischerweise aufweist (vgl. ebd., S. 872 u. f.). Dabei lies sich bei den Studierenden der Experimentalgruppe eine je-desto-Beziehung zwischen der geschätzten Aufgabenschwierigkeit bezogen auf sich

selbst und jener bezogen auf einen_eine typischen_typische Neuntklässler_Neuntklässlerin nachweisen, nicht jedoch in der Kontrollgruppe (vgl. ebd., S. 873 u. f.). Diesen Befund interpretieren die Autor_innen als eine Bestätigung ihrer allgemein formulierten Hypothese, „that inidividuals impute more of their own knowledge to others, the more they elaborate what these others might know“ (ebd., S. 865), was im Prozessmodell von Nickerson einer Dominanz des Ankers des eigenen Wissens entspricht.

Während Krispenz et al. (2016) ausschließlich Lehramtsstudierende befragten und die Frage eines möglichen fachbezogenen Charakters von Aufgabenschwierigkeitseinschätzungen außen vor ließen, untersuchten Ostermann et al. (2015) von welchen Faktoren die Schwierigkeitseinschätzungen von angehenden und erfahrenen Mathematiklehrkräften abhängen. In insgesamt drei Teilstudien[31], in denen unterschiedliche Gelegenheitsstichproben von Mathematik-Lehramtsstudierenden, Referendar_innen und Gymnasiallehrkräften erhoben wurden, konnten die Autoren vor allem die Hypothesen untermauern,...

... dass (angehende) Mathematiklehrkräfte die Schwierigkeit bestimmter Aufgabentypen stärker überschätzen andere (vgl. ebd., S. 61 u. f.),

... dass (angehende) Mathematiklehrkräfte ihren eigenen Bearbeitungsaufwand zur Lösung einer Aufgabe auf die Schwierigkeitsschätzung dieser Aufgabe für Schüler_innen projizieren (vgl. ebd., S. 63 u. f.),

... dass eine höhere Aufgabenschwierigkeitseinschätzung einer (angehenden) Mathematiklehrkraft einhergeht mit einer stärkeren Sensiblität für fachliche Hürden einer Aufgabe (vgl. ebd., S. 65 u. f.) und

... dass (angehende) Mathematiklehrkräfte mit wachsender Berufspraxis zu zunehmend akkurateren Aufgabenschwierigkeitsschätzungen gelangen (vgl. ebd. S. 67).

Wie Ostermann et al. (2015, S. 67 u. f.) zudem ausführlich erläutern, stehen auch diese Befunde im Einklang mit dem Modell von Nickerson.

Neben den beiden genannten Studien zum Thema Aufgabenschwierigkeitseinschätzung, findet sich in der Literatur aber auch zumindest eine Studie – jene von Herppich, Wittwer, Nückles, & Renkl (2011) –, deren Grundlage das Nickersonsche Prozessmodell ist, und die sich mit Lehrerurteilen über Schülerleistungen beschäftigt (vgl. Herppich, Wittwer, Nückles, & Renkl, 2013, S. 247). Die Autor_innen interessieren sich dabei für den Einfluss von Berufserfahrung auf die Akkuratheit von Lehrerleistungsurteilen, wobei sie Akkuratheit im Sinne des Urteilsgenauigkeitsansatzes verstehen (vgl. Unterabschnitt 2.1.1.1) (vgl. ebd., S. 251). Hierzu nehmen sie, im Setting einer unter Laborbedingungen stattfindenden Einzelnachhilfestunde zum Thema Herz-Kreislauf-System des Menschen, einen Experten-Novizen-Vergleich zwischen Biologie-Lehrkräften (N = 21) und -Studierenden (N = 25) vor (vgl. ebd., S. 247). Mit Hilfe von standardisierten Testverfahren, wurde dabei die Leistung der teilnehmenden Schüler_innen in der Mitte und zum Abschluss der Nachhilfestunde erfasst (vgl. ebd., S. 247 u. f.). Parallel hierzu wurden die teilnehmenden Lehrkräfte und Studierenden gebeten eine Einschätzungen über das Abschneiden ihrer

[31] Für eine detaillierte Darstellung der Teilstudien siehe Ostermann et al. (2015, S. 58 u. f.).

Nachhilfeschüler_innen in diesen Tests vorzunehmen (vgl. ebd.). Hierbei zeigte sich, dass sowohl die Biologie-Lehrkräften als auch die -Studierenden im Mittel die Leistung ihrer Nachhilfeschüler_innen zu beiden Einschätzungszeitpunkten überschätzen (vgl. Herppich et al., 2011, S. 81). Ferner zeigte sich mit Hilfe einer Varianzanalyse ein signifikanter Interaktionseffekt auf mittlerem Niveau ($\eta^2 = .09$; $p = .04$): Während die Überschätzung der Schülerleistung bei den Biologiestudierenden über den zeitlichen Verlauf der Nachhilfestunde zunahm, nahm diese bei den Lehrkräften leicht ab (vgl. ebd.). Zusammengefasst zeigt sich damit also in der Studie Herppich et al. (2011) ähnlich wie in der von Ostermann et al. (2015) ein „positiver Einfluss von Berufspraxis im Einklang mit dem Modell von Nickerson (1999)“ (ebd., S. 67).

2.1.4.3. Ergänzende Bemerkungen und Zwischenfazit

Welche Bilanz lässt sich aus diesem Abschnitt und insbesondere der hier vorgenommenen ausführlichen Darstellung des Linsenmodells und dem Prozessmodell von Nickerson ziehen?

Zunächst ist das Grundanliegen beider Ansätze kontextunabhängige, aber dennoch detallierte Beschreibungen sozialer Urteilsprozesse bereitzustellen. Wie neben theoretischen Überlegungen insbesondere auch die vorgestellten empirischen Forschungsarbeiten, die im Kontext von schulischer Leistungsfeststellung und -beurteilung auf Basis dieser Modelle durchgeführt wurden, deutlich erkennen lassen, sind sowohl das Linsenmodell, als auch das Prozessmodell von Nickerson in der Lage, Aufschluss über die Leistungsurteilsgenese von Lehrkräften zu liefern. Sie ermöglichen somit einen vertieften Einblick in die prozessbezogenen Aspekte von Lehrerwissen und -können in diesem Kontext, der mit den in diesen Unterkapitel bisher vorgestellten Herangehensweisen so nicht möglich ist.

Hinzukommt, dass sich bei beiden Ansätzen, aufgrund der hierbei vorgenommenen Überlegungen zur Akkuratheit von Lehrerurteilen, eine deutliche Querverbindung zum Urteilsgenauigkeitsansatz (vgl. Unterabschnitt 2.1.1.1) herstellen lässt. Es ist daher auch nicht verwunderlich, dass Prozessmodelle zur Genese von Lehrerleistungsurteilen von einigen Autor_innen der Forschungstradition um diagnostische Kompetenz von Lehrer_innen zuordnet werden (z. B. Leuders et al., 2014, S. 732 u. f.; von Aufschnaiter et al., 2015, S. 742 u. f.; Herppich et al., 2017, S. 85 u. f.). Allerdings ist Kompetenz im Linsenmodell und im Prozessmodell von Nickerson eher als eine Form von Performanz konzipiert. Die in Abschnitt 2.1.1 vorgestellte Forschung zu diagnostischen Kompetenzen von Lehrer_innen begreift dagegen Kompetenz eher als eine Form von Disposition. Anders ausgedrückt geht letztere von einem Kompetenzbegriff aus kognitivistischer Perspektive aus, die in diesem Abschnitt vorgestellten Ansätze zur Untersuchung von Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen aus Sicht der psychologischen Urteilsforschungs verstehen Kompetenz dagegen situationistisch (vgl. Blömeke, Gustafsson, & Shavelson, 2015, S. 5). Eine Distinktion beider Forschungstraditionen erscheint daher bis zu einem bestimmten Grad ebenso gerechtfertigt, wie eine Zusammenführung. In jedem Fall lassen sich aber die in diesem Abschnitt vorgestellten Ansätze zur Untersuchung

des Prozesses der Genese von Lehrerleistungsurteilen aufgrund der eben vorgenommenen Überlegung wohl am ehesten dem kompetenztheoretischen Bestimmungsansatz von Professionalität im Lehrerberuf zuordnen.

Des Weiteren ist in diesem Abschnitt ausführlich diskutiert worden, dass die beiden hier vorgestellten Prozessmodelle von zwei verschiedenen Ansätzen zur Beschreibung und Vorhersage von Urteilsgeneseprozessen ausgehen. Während das Linsenmodell von einer analytischen Informationsverarbeitung ausgeht, beschreibt das Prozessmodell von Nickerson den Prozess der Urteilsbildung auf Basis der „Anker- und Anpassungsheuristik". Da sich für beide Prozessmodell zumindest einige wenige Forschungsarbeiten finden lassen, die empirische Belege für das jeweilige Modell im Kontext von schulischer Leistungsfeststellung und -beurteilung liefern, kann dies auch so verstanden werden, dass zukünftige Forschung weniger von einem „entweder oder", sondern vielmehr von einem „sowohl als auch" beider Ansätze ausgehen sollte. Sogenannte duale Prozessmodelle sozialer Urteilsbildung, die analytische und heuristische Urteilsgenese miteinander verbinden, könnten hierfür eine fruchtbare Basis darstellen (vgl. M. Böhmer et al., 2017, S. 52). In neueren Arbeiten wird dieser dritte mögliche Modellierungsansatz von Urteilsgeneseprozessen zwar durchaus bereits theoretisch diskutiert (z. B. Leuders et al., 2014, S. 733; Ostermann et al., 2015, S. 56; M. Böhmer et al., 2017, S. 52 u. f.), überzeugende empirische Evidenz, die für eine solche Modellierung von Lehrerwissen und -können im Kontext von schulischer Leistungsfeststellung und -beurteilung spricht, liefern bisher allerdings nur die Studien von I. Böhmer, Gräsel, Krolak-Schwerdt, Höstermann, & Glock (2017) und Krolak-Schwerdt, Böhmer, & Gräsel (2012), die sich mit Schullaufbahnempfehlungen von Lehrkräften befasst. Inwieweit sich dieser Ansatz allerdings auch außerhalb des von beiden Autorengruppen gewählten Kontext als fruchtbar erweist, muss allerdings als Desiderat erziehungswissenschaftlicher Forschung gelten.

Für den weiteren Verlauf der vorliegenden Arbeit ist die bisherigen Forschung um Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen damit angemessen gesichtet. Im anschließenden Unterkapitel wird es darum gehen, die verschiedenen vorgestellten Forschungstraditionen und die aus ihnen gewonnenen Erkenntnisse zu einem Rahmenkonzept zusammenzuführen.

2.2. Das Konzept einer Assessment Literacy

Im vorangegangenen Unterkapitel wurden die folgenden vier Perspektiven auf Lehrerwissen und -können im Kontext von schulischer Leistungsfeststellung und -beurteilung erörtert:

1. Die Forschung zu diagnostischen Kompetenzen von Lehrer_innen
2. Die Forschung um Bezugsnormen und Bezugsnormorientierungen von Lehrkräften
3. Die Forschung zu berufsbezogenen Überzeugungen von Lehrkräften im Zusammenhang mit schulischen Leistungsfeststellungen und -beurteilungen

4. Schulische Leistungsfeststellungen und -beurteilungen aus Sicht der psychologischen Urteilsforschung

Hierbei wurde deutlich, dass jeder dieser Ansätze reichhaltige Einblicke auf unterschiedlichste Facetten von Lehrerwissen und -können in diesem Kontext ermöglichen. Isoliert betrachtet vermag jedoch keine der dargestellten Betrachtungsweisen eine allumfassende Antwort auf die Frage zu liefern, was eine professionell handelnde Lehrkraft bezogen auf den Themenkomplex der schulischen Leistungsfeststellung und -beurteilung auszeichnet, die für den empirischen Teil der vorliegenden Arbeit allerdings erforderlich ist (vgl. Kapitel 2; Einleitung). Mehr noch: Die ausführliche Diskussion der theoretischen Grundannahmen dieser vier Forschungstraditionen, sowie der mit deren Hilfe gewonnenen empirischen Befunde, fundiert, dass Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung derart vielschichtig ist, dass eine einzelne dieser Perspektive nicht ausreicht, um einen diesbezüglichen Referenzrahmen in befriedigender Art und Weise begründen zu können. Aus diesem Gedankengang folgt wiederum, dass die Begründung eines solchen Rahmens erst möglich ist, wenn man den bisherigen Forschungsstand über Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung in einem Metakonzept zusammenführt, da die theoretischen Konzepte einzelner Forschungstraditionen für dieses Anliegen „zu schmal“ ausgestaltet sind.

Zwecks dessen wurde im bisherigen Verlauf dieses Kapitels mit Hilfe der Bestimmungsansätze von Professionalität im Lehrerberuf im Sinne von Terhart (2011), sowie der von Neuweg (2014) vorgeschlagenen Unterteilung von Lehrerwissen und -können in Wissen 1, 2 und 3 bereits eine erste grobe Ordnung zwischen vorgestellten Arbeiten hergestellt. Jedoch sind diese Ansätze bzw. begrifflichen Unterscheidungen für das angestrebte Ziel, einen Referenzrahmen für professionelles Lehrerhandeln im Kontext von schulischer Leistungsfeststellung und -beurteilung möglichst allumfassend und dennoch präzise beschreiben zu können, zu weitläufig, was im wesentlichen mit dem diesen Konzepten innewohnenden fächerübergreifenden und allgemeindidaktischen Charakter zusammenhängt, der sich zudem bewusst nicht auf bestimmte Facetten des Lehrerberufs beschränkt.

Es soll nun ein Rahmenkonzept entworfen werden, dass zum einen den theoretischen Konzeptionen der vorgestellten Forschungstraditionen erhaben ist, zum anderen sich jedoch wiederum den metatheoretischen Überlegungen von Terhart (2011) und Neuweg (2014) unterordnen lässt. Teilweise sind im deutschsprachigen erziehungswissenschaftlichen Diskurs bereits Vorschläge für ein derartiges Rahmenkonzept unterbreitet worden. In den letzten Jahren geschah dies vor allem im Rahmen der Diskussion um die Modellierung diagnostischer Kompetenzen von Lehrer_innen (z. B. von Aufschnaiter et al., 2015; Herppich et al., 2017). Diese Rahmenkonzepte sind jedoch schon deswegen nicht unproblematisch, da hierdurch der Eindruck einer Ausschließlichkeit des kompetenztheoretischen Professionalitätsansatzes erweckt werden kann. Hinzu kommt, dass wie in den Abschnitten 2.1.1 und 2.1.4 bereits dargelegt, verschiedene Forschungsansätze zu Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen von zum Teil unterschiedlichen Kompetenzbegriffen ausgehen (z. B. die Unterscheidung zwischen einem kognitivistisch und einem situationistisch gedachten Kompetenzbegriff; vgl. Unterabschnitt 2.1.4.3), dies

jedoch bei den bislang in der Literatur vorgeschlagenen Rahmenkonzepten zur diagnostischer Kompetenzen von Lehrer_innen kaum mitbedacht wird. Zusammenführungen des bisherigen Forschungsstandes in ein gemeinsames Kompetenzmodell haben daher auch zur Folge, dass die Bedeutung des Begriffs „Kompetenz“ zunehmend an Schärfe verliert. Gerade dies widerspricht jedoch einer zentralen Leitidee der Kompetenzdebatte, nämlich sich von einer „inflationären“ Nutzung des Kompetenzbegriffs zu distanzieren und mit einem möglichst genau definierten und daher – eher aus pragmatischen, anstatt theoretischen Gründen – eng gefassten Verständnis dieses Begriffs zu operieren (vgl. Weinert, 2001a). Aus diesem Grund sollen die Überlegung in diesem Unterkapitel auch nicht unter dem begrifflichen Deckmantel der „Kompetenz“ unternommen werden. Stattdessen soll das Rahmenkonzept einer *Assessment Literacy von Lehrkräften* entwickelt werden, dem der angelsächsische Literacybegriff zugrunde liegt. Diesem Begriff lässt sich zwar, vor allem da er aufgrund der PISA-Studien an Popularität gewonnen hat (vgl. Tenorth & Tippelt, 2007, S. 486), eine gewisse Nähe zu einem engen Kompetenzbegriff zuschreiben (vgl. Kiel, 2009, S. 593), allerdings geht die angelsächsische Literacydebatte von einem Grundgedanken aus, der zur eben beschriebenen Leitidee der Kompetenzdebatte gegenläufig ist, nämlich den Begriff „Literacy“ nicht allein auf seine ursprüngliche Bedeutung als „Alphabetisierung“ (Tenorth & Tippelt, 2007, S. 486) zu verengen, sondern ihn „als Metapher für eine anwendungsorientierte Grundbildung[32] [zu begreifen,] [...] die sich auf sämtliche Bereiche beziehen kann“ (Nickel, 2007, S. 31). Anders ausgedrückt wird in der Literacydebatte im Gegensatz zur Kompetenzdebatte ein weitläufiges Begriffsverständnis angestrebt, weswegen der Literacybegriff, im Vergleich zum Kompetenzbegriff, für die nun folgenden Argumentation und Überlegungen deutlich geeigneter ist.

2.2.1. Ursprung des Begriffs „Assessment Literacy“ und frühe Überlegungen im Sinne des kompetenztheoretischen Bestimmungsansatzes

Zunächst ist festzuhalten, dass es sich bei „Assessment Literacy“ um einen Begriff handelt, der in der erziehungswissenschaftlichen Literatur nur wenig erwähnt wird. Hiermit hängt zusammen, dass sich bislang nur eine Hand voll von Definitionsversuchen zu diesem Begriff ausfindig machen lassen (vgl. Popham, 2011, S. 267). Zudem wurde die mit Assessment Literacy verbundene Idee im Laufe der Zeit von mehreren Autor_innen überarbeitet und weiterentwickelt. Zwecks einer ersten Annäherung soll daher zunächst die Entstehung des Begriffs „Assessment Literacy“, sowie frühe Überlegungen im Zusammenhang mit diesem Begriff grob nachgezeichnet werden, bevor anschließend aktuelle Begriffsverständnisse entfaltet werden:

[32] Der Begriff „Grundbildung“ ist eine inzwischen weitverbreitete Übersetzung des angelsächsischen Literacybegriffs und ist zu verstehen als eine funktionale Auslegung des Allgemeinbildungskonzepts (vgl. Tenorth & Tippelt, 2007, S. 486). Insgesamt ist der angelsächsische Literacybegriff allerdings in der deutschen Sprache schwer zu fassen, da es für ihn auf terminologischer Ebene keine Entsprechung gibt (vgl. Kühn, 2015, S. 7).

Seinen Ursprung hat der Begriff „Assessment Literacy“ in einem gleichnamigen Aufsatz Stiggins (1991). Anlass für seine Überlegungen war, dass sich im Verlauf der 1980er-Jahre der erziehungswissenschaftliche Diskurs im US-amerikanischen Raum vermehrt mit der Bedeutung von „educational outcomes“ auseinandersetzte und daher auch zunehmend die Frage gestellt wurde, wie gut das US-amerikanische Bildungssystem in der Lage ist, die Leistungen von Schüler_innen, angelehnt an output-orientierte Standards, festzustellen und zu beurteilen (vgl. ebd., S. 534). Bezogen auf diese Frage kommt Stiggins (1991) selbst zu dem vernichtenden Urteil, dass die USA eine „nation of assessment illiterates“ sind (vgl. ebd., S. 535), wobei er dies wie folgt spezifiziert:

> „Assessment illiterates don't understand what it takes to produce high-quality achievement data and so do not evaluate critically the data they use. Assessment illiterates accept achievement data at face value and can easily be intimidated by apparently technical information and by a complicated presentation of test scores.“ (ebd., S. 535)

Durch diese Charakterisierung eines Assessment Illiterate liefert Stiggins (1991) eine erste indirekte Definition von Assessment Literacy, nämlich als jene Grundbildung, die eine Person zu einem „kritischen Konsumenten“ von schulischen Leistungsfeststellungen und -beurteilungen werden lässt (vgl. ebd., S. 535). Konsumenten sind in diesem Zusammenhang alle Personen, die direkt oder indirekt an schulischen Leistungsfeststellungen und -beurteilungen beteiligt sind, sowie jene, denen durch diese schulischen Verfahren Informationen zur Verfügung gestellt werden. Assessment Literacy bezeichnet daher nicht nur eine wünschenswerte Grundbildung von Lehrer_innen, sondern auch von Schüler_innen, Erziehungsberechtigten, Schuldirektor_innen, (Fach-)Didaktiker_innen, usw. (vgl. Stiggins, 1991, S. 537 u. f.; Stiggins, 2014, S. 69 u. f.).

Da allerdings schulische Leistungsfeststellungen und -beurteilungen für jede dieser Personengruppen andere Funktionen erfüllen (vgl. Unterkapitel 1.2), unterscheidet sich diese Grundbildung auch von Gruppe zu Gruppe. In seiner ersten Konzeption unterscheidet Stiggins daher drei hierarchisch geordnete Level von Assessment Literacy, die er als „functional“, „practical“ und „advanced level of Assessment Literacy“ bezeichnet (vgl. Stiggins, 1991, S. 537). Das funktionale Level beschreibt dabei jene Grundbildung, die z. B. Schüler_innen oder Erziehungsberechtigte benötigen, um kritisch mit den Informationen umgehen zu können, die ihnen schulische Leistungsfeststellungen und -beurteilungen bereitstellen, wohingegen das erweiterte (advanced) Level der Grundbildung von Erziehungswissenschaftler_innen entspricht, deren Zuständigkeitsbereich standardisierte Schulleistungsstudien und Schulevaluation umfasst. Letzterem ist daher auch vertieftes testtheoretisches Wissen und Können zugeordnet (vgl. ebd.), womit Stiggins schlussendlich begründet, warum Lehrkräfte seiner Ansicht nach keine Personengruppe sind, die über ein „advanced level of Assessment Literacy“ verfügen müssen (vgl. ebd.). Stattdessen werden Lehrer_innen von ihm auf dem „pratical level of Assessment Literacy“ verortet, das die Aspekte des funktionalen Levels mit einschließt, jedoch die vertiefenden des erweiterten Levels ausspart (vgl. ebd.). Das „pratical level of Assessment Literacy“ beschränkt sich also auf eine Grundbildung, die notwendig ist, um als Lehrkraft in der

alltäglichen Praxis schulische Leistungsfeststellungen und -beurteilungen durchführen und die hierdurch gewonnenen Informationen handhaben zu können (vgl. ebd.).

Die Idee von Stiggins (1991), Assessment Literacy allgemein nicht nur als Grundbildung von Lehrkräfte aufzufassen, sondern als Oberbegriff für die Grundbildungen eines deutlich weiteren Personenkreises, hat sich bis heute durchgesetzt. Dies zeigt sich beispielsweise an der aktuellen Version der vom Michigan Assessment Consortium (2017) veröffentlichten „Assessment Literacy Standards", die für unterschiedliche Personengruppen differenziert formuliert worden sind. Ein deutlicher Schwerpunkt des gesamten Diskurses um Assessment Literacy liegt jedoch auf der Assessment Literacy von Lehrkräften, was damit zusammenhängt, dass schulische Leistungsfeststellungen und -beurteilungen für Lehrkräfte zum Kerngeschäft gehören (vgl. Kapitel 1; Einleitung) und daher der Anspruch geltend gemacht wird, dass sich insbesondere diese Personengruppe als bezüglich schulischer Leistungsfeststellung und -beurteilung grundgebildet auszeichnen sollte (vgl. Schafer, 1993, S. 124 u. f.). Diese Schwerpunktsetzung zeigt sich auch an zwei zwar veralteten, aber dennoch bis heute viel zitierten Definitionen von Assessment Literacy von Stiggins (1995) und Webb (2002), die vor allem die Grundbildung von Lehrkräften ansprechen:

> „[A]ssessment-literate educators [...] come to any assessment knowing what they are assessing, why they are doing so, how to generate sound samples of performance, what can go wrong, and how to prevent those problems before they occur." (Stiggins, 1995, S. 240)

> „[Assessment Literacy is] the knowledge about how to assess what students know and can do, interpret the results of these assessments, and apply these results to improve student learning and program effectiveness." (Webb, 2002, zitiert nach White, 2009, S. 7)

Solche Definitionsversuche wurden der Kritik unterzogen: Zum einen kann vor allem bei Stiggins (1995) Definition das Missverständnis aufkommen, dass Assessment Literacy gleich zu setzen wäre, mit dem Statistikwissen von Lehrkräften und insbesondere mit deren testtheoretischem Wissen um Güteaspekte schulischer Leistungsfeststellungen und -beurteilungen (vgl. Abschnitt 1.3.1), im Sinne eines in Wissen 2 transformierten Wissens 1 (vgl. Neuweg, 2014, S. 586). Tatsächlich wird aus den Ausführungen von Stiggins, wie bereits angedeutet, deutlich, dass er derartiges Wissen überwiegend nicht als Teil der Assessment Literacy von Lehrkräften versteht (vgl. Stiggins, 1991, S. 537; Stiggins, 1995, S. 242 u. f.). Seine Vorstellung einer diesbezüglichen Grundbildung von Lehrer_innen deckt sich in großen Teilen mit den in Abschnitt 1.3.2 vorgestellten Überlegungen von Leisen & Höttecke (2011), sowie Höttecke & Wodzinski (2015, S. 6 u. f.), die testtheoretischen Gütekriterien für die schulische Praxis eine eher nebengeordnete Rolle zukommen lassen (vgl. Stiggins, 1991, S. 537; Stiggins, 1995, S. 242 u. f.). Zum anderen wurde bei beiden Definitionen die Kritik laut, dass diese zu einseitig, aufgrund ihrer gewählte Formulierung im Sinne von output-orientierten Standards der Lehrerbildung, sowie ihrer deutlichen Fokussierung auf kognitive Facetten von Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen, wie sie in Abschnitt 2.1.1 und 2.1.4 erläutert

wurden, auf den kompetenztheoretischen Professionalitätsansatz verweisen[33] (z. B. White, 2009, S. 12 u. f.; Looney, Cumming, van Der Kleij, & Harris, 2018, S. 443 u. f.).

2.2.2. Überlegungen aktuellen Datums zum Begriff „Assessment Literacy“ im Sinne des strukturtheoretischen Bestimmungsansatzes

Überlegungen zum Begriff „Assessment Literacy“ aktuellen Datums versuchen daher unter anderem stärker weitere Aspekte im Sinne des strukturtheoretischen Professionalitätsansatzes mit zu berücksichtigen:

Eine vorsichtige Öffnung in diese Richtung nehmen die bereits erwähnten „Assessment Literacy Standards“, in denen die Assessment Literacy (von Lehrkräften) als „set of beliefs, knowledge and practices about assessment“ (vgl. Michigan Assessment Consortium , 2017, S. 2) definiert wird und damit auch berufsbezogene Überzeugungen von Lehrkräften zu schulischen Leistungsfeststellungen und -beurteilungen (siehe Abschnitt 2.1.3), sowie Bezugsnormorientierungen (siehe Abschnitt 2.1.2) in operationalisierter Form umfassen (vgl. ebd., S. 8).

Ähnliches gilt für die bisherigen speziell naturwissenschaftsdidaktischen Überlegungen zur Assessment Literacy von Lehrkräften. Hier ist zwar ebenfalls eine Schwerpunktsetzung in Richtung des kompetenztheoretischen Professionalitätsansatz festzustellen, zusätzlich wird dabei allerdings unter besonderer Berücksichtigung der Besonderheiten des naturwissenschaftlichen Fachunterrichts versucht die Verflechtungen von kompetenztheoretisch gedachtem Wissen und Können (vor allem in Wissen 2 und 3 transformiertes Wissen 1) mit berufsbezogenen Überzeugungen von Lehrkräften zur (fachspezifischen) Leistungsfeststellung und -beurteilung (vgl. Abschnitt 2.1.3) mit einfließen zu lassen[34]. Sowohl im Modell von Abell & Siegel (2011), als auch in jenem vom M. A. Siegel & Wissehr (2011)

[33] An dieser Stelle soll noch einmal erwähnt werden, dass der Begriff „Assessment Literacy“ bislang ausschließlich im internationalen erziehungswissenschaftlichen Diskurs thematisiert wurde, weswegen sich auch keiner der in diesem Unterkapitel zitierten Autor_innen explizit in den von Terhart (2011) und Neuweg (2014) vorgeschlagenen Trichotomien aus kompetenztheoretischem, strukturtheoretischem und berufsbiographischem Professionalitätsansatz bzw. Wissen 1, 2 und 3 verortet. Wenn daher im Folgenden davon gesprochen wird, dass sich die Überlegungen eines_einer Autors_Autorin in einer dieser Trichotomien einordnen lassen, ist dies keine von den entsprechenden Autor_innen selbst vorgenommene Zuschreibung, sondern eine Verortung, die sich aufgrund der Überlappung von zentralen Denkfiguren dieser Autor_innen und für den jeweiligen Bestimmungsansatz typischen Charakteristika und/oder Positionen ergibt.

[34] Dieselbe Aussage lässt sich auch für einige der speziell sprachdidaktischen Überlegungen zur Assessment Literacy von Lehrkräften treffen (z. B. Taylor, 2013; Crusan, Plakans, & Gebril, 2016). In Teilen sind die Gedankengänge in diesem Diskurs allerdings auch weiterführend (z. B. Fulcher, 2012; Scarino, 2013) und eher mit dem Assessment Literacy Modell von Xu & Brown (2016) vereinbar, das in den folgenden Absätzen vorgestellt wird. Alles in allem decken sich die sprachdidaktischen Überlegungen zur Assessment Literacy von Lehrkräften mit den Gedankengängen, die dieses Unterkapitel erörtert, weswegen sie in der vorliegenden Arbeit auch nicht weiter ausgeführt werden. Für einen aktuelle Gesamtschau dieses Diskurses sei aber auf den entsprechenden Überblick von Harding & Kremmel (2016, S. 413 u. f.) verwiesen.

zur Assessment Literacy von Naturwissenschaftslehrkräften wird diesbezüglich davon ausgegangen, dass berufsbezogene Überzeugungen als „mentale Strukturen einer Person [...] deren Wahrnehmung, Denken und Fühlen und somit bewusst oder unbewusst deren Handeln beeinflussen“ (Klinghammer et al., 2016, S. 182; vgl. auch Unterabschnitt 2.1.3.1). Des Weiteren ist bei genauer Betrachtung nicht nur anzunehmen, dass diese berufsbezogenen Überzeugungen wie ein „Wandler“ für neu anzueignendes Wissen im objektiven Sinn wirken (Transformation von Wissen 1 in Wissens 2, vgl. Neuweg, 2014, S. 586), sondern auch, dass aus dem (möglichen) Spannungsverhältnis zwischen eigenem theoretischen Wissen und eigenen berufsbezogenen Überzeugungen zur schulischen Leistungsfeststellung und -beurteilung eine subjektive Überzeugung darüber erwächst, wie schulische Leistungsfeststellungen und -beurteilungen vorgenommen werden sollten (vgl. Abell & Siegel, 2011, S. 210 u. f.; M. A. Siegel & Wissehr, 2011, S. 373 u. f.).

Daneben finden sich auch in den Überlegungen von Popham (2011) Aspekte, die sich als strukturtheoretisch interpretieren lassen. Besonders hervorzuheben ist seine Forderung, dass sich eine bezüglich schulischer Leistungsfeststellung und -beurteilung grundgebildete Lehrkraft dadurch auszeichnen sollte, dass es ihr in ihrer täglichen Arbeit gelingt, sich dem antinomischen Verhältnis von pädagogischen und sozialen Funktionen schulischer Leistungsfeststellungen und -beurteilungen (vgl. Unterkapitel 1.2) bewusst zu sein und dass sie mit diesem gemäß ihrer eigenen Zielsetzungen, aber auch je nach Kontext in dem sie agiert, adäquat umgehen kann (vgl. Popham, 2009; Popham, 2011, S. 268 u. f.).

Das gegenwärtig allerdings mit Abstand komplexeste und umfassendste Modell der Assessment Literacy von Lehrkräften, das sich mit dem strukturtheoretischen Professionalitätsansatz verbinden lässt, stammt von Xu & Brown (2016). In diesem modellieren die Autor_innen die Assessment Literacy von Lehrkräften als Praxis des bewusst selbst herbeiführenden Kompromissfindens zwischen drei Faktoren, die zueinander in schulische Leistungsfeststellung und -beurteilung betreffenden beruflichen Handlungsepisoden in einem Spannungsverhältnis stehen können (vgl. ebd., S. 157). Diese drei Faktoren sind:

1. Das kompetenztheoretisch gedachte Wissen und Können der Lehrkraft zur (fachspezifischen) Leistungsfeststellung und -beurteilung (vgl. ebd.).
2. Ihre berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung (vgl. ebd.).
3. Die makroskopischen (z. B. schulpolitische Vorgaben und schulkulturelle Gepflogenheiten bezogen auf die Umsetzung von schulischer Leistungsfeststellung und -beurteilung) und mikroskopischen Kontextbedingungen (z. B. Klassenzusammensetzung oder Stundenplan), die den tatsächlichen Handlungsspielraum der Lehrkraft in den entsprechenden Handlungsepisoden mehr oder minder begrenzen (vgl. ebd.).

Zum besseren Verständnis soll dies anhand eines hypothetischen Beispiels einer Handlungsepisode illustriert werden:

Einige Schüler_in der zwölften Jahrgangsstufe eines bayerischen Gymnasiums haben sich entschieden das Fach Physik als Abiturprüfungsfach zu wählen. Da sie allerdings nach

der sog. „Lehrplanalternative Biophysik“ unterrichtet wurden, können sie nach der aktuellen Schulordnung für die Gymnasien in Bayern in Physik nur eine mündliche, jedoch keine schriftliche Prüfung ablegen (vgl. Bayerische Staatsministerium für Unterricht und Kultus , 2007, §48 Abs. 1). Die Physiklehrkräfte ihrer Schule müssen daher einen Fachausschuss bilden, der die Themen und Fragen dieser mündlichen Prüfungen festlegt und diese nach einem vorgegebenen Verfahren durchführt (vgl. Bayerische Staatsministerium für Unterricht und Kultus , 2007, §46). Hierbei können allerdings aus Sicht der beteiligten Lehrkräfte Spannungen auftreten, zwischen denen es für sie gilt einen Vergleich zu finden:

Eine mögliche Spannungsquelle ist, inwieweit die beteiligten Lehrkräfte einer solchen Abiturprüfung ihren berufsbezogenen Überzeugungen nach skeptisch gegenüberstehen, da sie z. B. eine schriftliche Prüfung für angemessener halten. Auf eine andere Prüfungsform auszuweichen, die den Überzeugung der Mitglieder des Fachausschusses nach adäquatere wäre, stellt jedoch keine Handlungsoption dar, da dies der Gesetzgeber nicht vorsieht. Eine umsetzbare Lösung hingegen ist, dass die beteiligten Lehrkräfte einen Kompromiss finden, wie diese mündlichen Abiturprüfung im Fach Physik gestaltet und abgehalten werden sollte, sodass sie den gesetzlichen Regelungen entspricht und in dem ihre eigenen Vorbehalte bezüglich der vorgegebenen Prüfungsform mit berücksichtigt wurde (z. B. indem den Prüflingen gewährt wird, die Tafel als Unterstützung für ihre mündlichen Äußerungen zu verwenden). Dieses Kompromissfinden erfordert aber idealerweise, dass sich die beteiligten Lehrkräfte ihren eigenen berufsbezogenen Überzeugung zur Gestaltung und Umsetzung einer Abiturprüfung auch bewusst sind, da erst hierdurch ein fruchtbarer Austausch über evtl. Vorbehalte oder Bedenken möglich ist. Wenn dem so ist, würden Xu & Brown (2016) davon sprechen, dass die Mitglieder des Fachausschusses in dieser Handlungsepisode assessment literate sind.

Ein andersartige Spannung kann beim Erstellen der Prüfungsfragen durch die Mitglieder dieses Fachausschusses auftreten: Für Abituraufgaben im Fach Physik hat die Kultusministerkonferenz eine Liste von zu verwendenden Operatoren herausgegeben, an denen sich die Lehrkräfte bei der Erstellung von Prüfungsfragen orientieren sollen (vgl. Kultusministerkonferenz , 2004b, S. 14 u. f.). Diese Vorgabe kann aber zu dem (evtl. vorhandenen) Wissen der Mitglieder des Fachausschusses im Widerspruch stehen, dass für die Erstellung von Prüfungsfragen Operatoren gewählt werden sollten, die eindeutig mit bestimmten Denk- und Handlungsaufforderungen verknüpft sind (z. B. einen physikalischen Sachverhalt erklären oder eine physikalische Größe abschätzen) (vgl. Jatzwauk, Rumann, & Sandmann, 2008, S. 267 u. f.). Beispielsweise ist in der von der Kultusministerkonferenz vorgegeben Liste der Operator „bestimmen“ in zwei verschiedenen Bedeutungen angegeben (einmal als „aus einer Größengleichung physikalische Größe gewinnen“ (Kultusministerkonferenz , 2004b, S. 14) und ein weiteres mal als „einen Lösungsweg darstellen und das Ergebnis formulieren“ (ebd., S. 15)). Eine für die Lehrkräfte hier umsetzbare Lösung dieses Konflikts wäre, in ihren Prüfungsfragen den Operator „bestimmen“ zu meiden und stattdessen andere vorgegeben Operatoren zu verwenden, die eine ähnliche Bedeutung wie der Operator „bestimmen“ besitzen (z. B. „herleiten“ oder „berechnen“). Im Sinne des Assessment Literacy Modells von Xu & Brown (2016) erfolgt auch dieses Kompromissfinden

idealtypisch dadurch, dass sich hierbei die beteiligten Lehrkräfte ihrem Wissen zur Erstellung von Prüfungsaufgaben bewusst sind, da sie erst hierdurch ihre Operatorauswahl stichhaltig begründen können.

Aus den aufgeführten Überlegungen und dem eben dargestellten Beispiel wird deutlich, dass Xu & Brown (2016) die Assessment Literacy von Lehrkräften weniger als ein statisches Konzept oder als ein idealisiertes Sammelsurium bestimmter Wissens- und Könnensaspekte begreifen, sondern als weitgefasstes, dynamisches Konstrukt, das sowohl Wissens und Könnensaspekte bzgl. schulischer Leistungsfeststellung und beurteilung umfasst, als auch diesbezügliche berufsbezogene Überzeugungen und den Umgang von Lehrkräften mit Anforderungen und Zwängen schulischer Kontextbedingungen, unter denen Leistungsfeststellungen und beurteilungen stattfinden (ebd., S. 157). Ferner gehen Xu & Brown (2016) davon aus, dass nicht jede Lehrkraft zu einem bewusst selbst herbeigeführten Vergleich zwischen den eben aufgeführten, möglicherweise in einem Spannungsverhältnis stehenden Faktoren im Stande ist[35], sondern postulieren drei Qualitätsstufen („levels of mastery“, ebd., S. 159), auf denen sich die Assessment Literacy einer Lehrkraft bewegen kann:

> „[A teacher's Assessment Literacy] consists of three levels of mastery. First is a basic mastery of educational assessment knowledge, which includes the fundamental principles of the 'what', 'why', and 'how,' [sic!] without which teachers cannot engage with assessment at a deeper level. Second is an internalized set of understanding and skills of the interconnectedness of assessment, teaching, and learning. Unlike the 'should-do' kind of knowledge indicated by the first level, this is a more personal perception of how assessment should be, formed among the tensions between theoretical knowledge and teachers' own conceptions of assessment. Third is a self-directed awareness of assessment processes and one's own identity as an assessor[36]. Such awareness allows teachers to accommodate and translate assessment policies and principles into their classroom realities and institutional contexts while driving them to reflect on their assessment practices and to gain new insights.“ (ebd., S. 159)

Besonders hervorzuheben ist, dass das Modell von Xu & Brown (2016) der Assessment Literacy von Lehrkräften durch seine Fokussierung auf in der Praxis auftretenden Spannungsverhältnisse, die – sofern dies möglich ist – aufzulösen sind, nicht nur eine strukturtheoretische Komponente in sich birgt. Durch seine am Prozess der Genese von Lehrerleistungsurteilen orientierten Konzeption, bzw. als eine „Logik des *Handelns*“ (Neuweg, 2014, S. 585, Hervorhebungen im Original), im Sinne eines, sich unter anderem aus Wissen 2 speisenden Wissens 3, steht dieses Modell vor allem mit Ansätzen im Einklang, die schulische Leistungsfeststellung und -beurteilung aus Sicht der psychologischen Urteilsforschung betrachten (vgl. Unterabschnitt 2.1.4). Gleichzeitig unterscheidet sich das Modell von Xu & Brown (2016) damit erkennbar von den übrigen in diesem Unterabschnitt vorgestellten Überlegungen (mit Ausnahme derer von Popham (2011)), in denen die Assessment Li-

[35] Diese Annahme ist auch plausibel, da ein derartiger Vergleich „ein Bewusstsein [erfordert] [...], dass Handlungen und Ereignisse in Bezug auf multiple Perspektiven erklärbar sind oder auch, dass sie [...] in multiplen [...] Zusammenhängen angesiedelt sind und durch diese beeinflusst werden“ (Abels, 2011, S. 101), also eine Charakteristik eines_einer idealtypische_n Lehrers_Lehrerin als „Reflective Practitioner“ (vgl. ebd.).

[36] Auf den Begriff „identity as an assessor“ wird in Abschnitt 2.2.3 genauer eingegangen.

teracy von Lehrkräften eher als eine Form von Wissen 2 gedacht ist (vgl. Unterabschnitt 2.1.4.3).

2.2.3. Überlegungen aktuellen Datums zum Begriff „Assessment Literacy“ im Sinne des berufsbiographischen Bestimmungsansatzes

Neben den eben aufgeführten strukturtheoretischen Überlegungen, finden sich in der Literatur auch Ansätze, die dem Begriff „Assessment Literacy“ eher einen berufsbiographischen Wortsinn zuschreiben. Dies lässt sich mit Hilfe des folgenden Auszugs aus Pophams Artikel „Assessment Litercy Overlooked: A Teacher Educator's Confession“ motivieren:

> „[I]n the following paragraphs [...] I intend to make a forthright confession about a serious curricular sin I committed during my early years as a teacher educator. [...] Once I had finished my doctorate, [...] I was asked [...] to help would-be teachers learn how to teach. [...] I put all my academic energy [...] in promoting those teacher candiates' instructional smarts, that I paid no attention whatsoever to their assessment acumen. [...] And this, I confess, was my sin. [...] Later I realized, [...] that today's prospective teachers, in order to do their jobs properly, desperately need to become *assessment literate*.“ (Popham, 2011, S. 265-267, Hervorhebungen im Original)

In diesem Zitat deutet sich zunächst eine Denkfigur im Zusammenhang mit Assessment Litercy an, die sich in ähnlicher Art und Weise schon bei Stiggins (1991) finden lässt, nämlich dass Lehrkräfte sich ihre berufsbezogenen Grundbildungen zwar prozesshaft, vor allem im Rahmen der Ausbildung aneignen, ab einem bestimmten Punkt besitzen diese aber eine eher statische Natur. Während jedoch Stiggins (1991) beanstandet, dass das damalige US-amerikanische Lehrerbildungssystem (angehenden) Lehrkräften zu wenige Möglichkeiten eröffnet, sich – in Form von output-orientierte Standards formulierbares – Wissen und Können zur schulischen Leistungsfeststellung und -beurteilung anzueignen und sie deshalb „assessment illiterate“ verweilen, fasst Popham (2011) die Problemlage breitgefächerter, was sich im obigen Zitat in den weit auslegbaren Begriffen „smarts“ und „acumen“ andeutet. Der erste Eindruck ist daher irreführend, Popham (2011) würde ähnlich wie Stiggins (1991) von einer eher statische Idee von Assessment Literacy ausgehen. Dies bedarf jedoch einer genaueren Erläuterung, wozu sich die weiterführenden Überlegungen von Xu & Brown (2016), sowie Looney et al. (2018) eignen:

Beide Autorengruppen – explizit vor allem aber Xu & Brown (2016) – betonen, dass es für (angehende) Lehrkräfte für eine erfolgreiche berufliche Entwicklung nicht nur gilt eine Teilidentität als „instructor [of learning]“, sondern auch eine als „assessor of learning“ aufzubauen (vgl. Xu & Brown, 2016, S. 158). Mit diesem Gedankengang beanstanden sie ähnlich wie Popham (2011) und Stiggins (1991), dass Lehrerbildungssysteme oftmals den Bedürfnissen (angehender) Lehrkräfte zu wenig gerecht werden, verwenden aber ein anderes, zu Assessment Literacy in Beziehung stehendes Konstrukt, nämlich das einer beruflichen Teilidentität als Assessor of Learning. Die berufliche Teilidentität als Assessor of Learning einer Lehrkraft ist hierbei eine mentale Struktur, die kognitive, emotionale

und motivationale Facetten umfasst, und ist daher von der Assessment Literacy einer Lehrkraft als eine von außen rekonstruierten Logik des Handelns zu unterscheiden (vgl. Looney et al., 2018, S. 446 u. f.; siehe auch Abschnitt 2.2.2). Ansätze, die dem Begriff „Assessment Literacy“ einen auch berufsbiographischen Wortsinn zuschreiben, versuchen also den Zusammenhang zwischen dem beobachtbaren Handeln einer Lehrkraft im Kontext schulischer Leistungsbeurteilung und den mentalen Strukturen, die diese Handlungen hervorbringen, theoretisch aufzuklären (aus Wissen 2 speisendes Wissens 3; vgl. Neuweg, 2014, S. 584 u. f.).

Gemäß Keupp et al. (2002) lässt sich eine berufliche Teilidentität allgemein konzipieren als (für eine detaillierte theoretische Klärung siehe Abels, 2011, S. 34 u. f.)...

> „[...] [d]as Ergebnis der Integration selbstbezogener Erfahrungen [...] [als] ein Bild des Subjekts von sich selbst, in dem die vielen Facetten seines Tuns übersituative Konturen erhalten. [...] Solche [...] Teilidentitäten enthalten ein Mosaik an Erfahrungsbausteinen, die auf die Zukunft gerichtet sind[,] [...] sowie solche, die eher der Vergangenheit angehören[.] [...] [Ferner enthalten sie] ein Set von angewandten Bedeutungen, die Personen entwickeln, und die definieren, wer man glaubt zu sein (Burke, 1991, S. 837).“(Keupp et al., 2002, S. 218-219)

Das in diesen Zitat zuletzt genannte „Set von angewandten Bedeutungen“ umfasst dabei, im Fall der beruflichen Teilidentität als Assessor of Learning einer Lehrkraft, gemäß den Überlegungen von Looney et al. (2018, S. 455 u. f.) die folgenden fünf Facetten:

1. Was eine Lehrkraft glaubt über schulische Leistungsfeststellung und -beurteilung zu wissen
2. Welche Gefühle eine Lehrkraft mit schulischer Leistungsfeststellung und -beurteilung verbindet
3. Wie eine Lehrkraft ihre eigene Rolle im Kontext von schulischer Leistungsfeststellung und -beurteilung sieht
4. Welche berufsbezogenen Überzeugungen eine Lehrkraft glaubt zu schulischer Leistungsfeststellung und -beurteilung zu besitzen
5. Inwiefern sich eine Lehrkraft bezogen auf schulische Leistungfeststellung und -beurteilung für souverän hält

Besonders hervorzuheben ist, dass die Identitätstheorie von Keupp et al. (2002) davon ausgeht, dass eine Teilidentität fortlaufend Veränderungsprozessen unterliegt und daher nur mehr oder minder stabil ist (vgl. ebd., S. 217 u. f.). Diesem Identitätsverständnis entsprechend gehen Looney et al. (2018, S. 455) daher auch nicht von einer statischen, sondern vielmehr von einer dynamischen berufliche Teilidentität einer Lehrkraft als Assessor of Learning aus. Insbesondere die fünf eben aufgeführten Facetten, die diese berufliche Teilidentität bildhaft gesprochen rahmen, unterliegen also einem ständigen Wandel (vgl. ebd., S. 14 u. f.).

Die Modellvorstellung einer beruflichen Teilidentität als Assessor of Learning von Lehrkräften, lässt sich nun mit Hilfe des Vorschlags von Popham (2011), Assessment Literacy als „an individual's understandings of the fundamental assessment concepts and procedu-

res“ (ebd., S. 267) zu definieren, auf zweierlei Arten mit dem Konstrukt einer Assessment Literacy von Lehrkräften in Verbindung bringen:

1. Zum einen lässt sich der im aufgeführten Zitat von Keupp et al. (2002) beschriebene Prozess der gelungenen Entwicklung einer Teilidentität, wie Rehm (2012) feststellt, mit der Fähigkeit ein Verständnis über einen Sachverhalt aufbauen zu können gleichsetzen[37]. Folgt man diesem Gedankengang, dann ist obiger der Definitionsvorschlag von Popham (2011) nicht im Sinne von Stiggins (1991) Gedankengang aufzufassen, da hier mit „individuellem Verstehen“ das produktive Arbeiten an der eigenen sich stets weiterentwickelnden – und eben nicht statischen – beruflichen Teilidentität als Assessor of Learning gemeint ist (vgl. Looney et al., 2018, S. 456 u. f.).
2. Zum anderen sind, wie Keupp et al. (2002) ebenfalls theoretisch erörtern „[f]ür die Funktionalität in konkreten Handlungszusammenhängen [...] vor allem die [eigenen] Identitätsentwürfe [...] [zentral]. Sie bilden die Basis für Handlungsmotivation und Informationssteuerung, sie liefern Begründungen, die für Handeln unabdingbar sind“ (ebd., S. 238). Mit Hilfe dieses Gedankengangs lässt sich aber der Definitionsvorschlag von Popham (2011) nicht nur als produktives Arbeiten an einer eigenen berufsbezogenen Teilidentität lesen, sondern gleichzeitig kann der Begriff „an individual's *understandings*“ (ebd., S. 267, Hervorhebungen im Original) auch verstanden werden als eine Entfaltung der eigenen Teilidentität als Assessor of Learning im Rahmen verschiedener Handlungsepisoden (vgl. Xu & Brown, 2016, S. 158 u. f.). Auch dies stellt einen Unterschied zu den Überlegungen von Stiggins (1991) dar, da in diesen die Assessment Literacy von Lehrkräften kompetenztheoretisch konzeptioniert ist und dementsprechend dort von einer eher eng gefassten Disposition zu einer bestimmten Art von kontextspezifischer Performanz ausgegangen wird.

2.2.4. Zusammenführung verschiedener Konzeptionen der Assessment Literacy von Lehrkräften

Als erstes Zwischenfazit lässt sich festhalten, dass die in den vorherigen drei Abschnitten vorgenommenen Ausführungen mit Bezug auf prominente Veröffentlichungen zum Begriff „Assessment Literacy“ deutlich zeigen, dass die Gesamtheit der hierzu im erziehungswissenschaftlichen Diskurs bisher unternommenen Überlegungen ein Rahmenkonzept bildet, das allumfassend und dennoch präzise beschreibt, was eine im Kontext von schulischer Leistungsfeststellung und -beurteilung professionell handelnde Lehrkraft ausmacht. Zu-

[37] Rehm (2012) charakterisiert diese Fähigkeit wie folgt: „Das Ziel ist Selbstkompetenz in die/in seine [sic!] Welt. Wer von Grunde auf versteht, arbeitet produktiv an seinem Identitätsprojekt (Wagenschein 1970 [Wagenschein, 1970c; Wagenschein, 1970d; M. S. F.] und Keupp et al. 2002)“ (ebd., S. 128). Anzumerken ist, dass die Überlegungen von Rehm (2012) primär auf Verstehen im Zusammenhang mit dem Konzept einer „Scientific Literacy“ abziehlen. Sein allgemein gehaltener, weitgehend philosophischer Argumentationsgang legitimiert allerdings den von ihm herausgearbeiteten Zusammenhang zwischen dem Begriff „Verstehen“ und der Identitätstheorie von Keupp et al. (2002) auch auf die in diesem Unterkapitel thematisierte Grundbildung von Lehrkräften bezüglich schulischer Leistungsfeststellung und -beurteilung zu übertragen.

dem lässt sich die in Abschnitt 2.2.1 aufgeworfene These, die mit dem Begriff „Assessment Literacy von Lehrkräften“ verbundene Idee habe sich im Laufe des Diskurses zunehmend gewandelt, insofern präzisieren, als dass ältere Arbeiten, die sich hiermit auseinandersetzen, oftmals von einem kompetenztheoretischen Professionalitätsgedanken dominiert sind, aktuelle Ansätze aber auch im Sinne des strukturtheoretischen oder des berufsbiographischen Bestimmungsansatzes gelesen werden können.

Ferner zeigt sich, dass es verschiedene Konzeptionen der Assessment Literacy von Lehrkräfte gibt, die mit unterschiedlichen Wissensarten im Sinne der Trichotomie von Neuweg (2014) aus Wissen 1, 2 und 3 verknüpfbar sind und daher als bis zu einem bestimmten Grad unterschiedliche Auslegungen des Wortsinns dieses Begriffs nebeneinander stehen. Jedoch hat sich vor allem in den Abschnitten 2.2.2 und 2.2.3 gezeigt, dass insbesondere Ansätze neueren Datums zur Konzeption der Assessment Literacy von Lehrkräften stark miteinander verwoben sind. Die Gedankengänge unterschiedlicher Autor_innen lassen sich daher zu einer umfassenden Modellvorstellung der Assessment Literacy von Lehrkräften zusammenfassen. Aufgrund des zuerst gesagten ist es aber notwendig eine Setzung vorzunehmen, um welche Wissensart es sich bei der Assessment Literacy von Lehrkräften handelt, um so Klarheit herzustellen (siehe unten).

In Abbildung 2.5 ist diese umfassenden Modellvorstellung der Assessment Literacy von Lehrkräften schematisch dargestellt. Sie basiert auf zentralen Denkfiguren des Modells Xu & Brown (2016), sowie den Überlegungen zur beruflichen Teilidentität als Assessor of Learning einer Lehrkraft von Looney et al. (2018), da diese die zum gegenwärtigen Zeitpunkt komplexesten und umfassendsten Konzeptionen der Assessment Literacy von Lehrkräften bzw. eines Konstrukts, das mit dieser in Beziehung steht, gelten können.

Angelehnt an die Überlegungen von Xu & Brown (2016) (vgl. Unterabschnitt 2.2.2) ist in Abbildung 2.5 die Assessment Literacy einer Lehrkraft als eine Form von, sich unter anderem aus Wissen 2 speisenden Wissen 3 konzipiert, nämlich als ihre im Rahmen verschiedener Handlungsepisoden bewusst selbst herbeigeführte Realisierung eines Kompromisses zwischen...

... ihrem kompetenztheoretisch gedachten Wissen und Können zur (fachspezifischen) Leistungsfeststellung und -beurteilung (in Wissen 2 und 3 transformiertes Wissen 1),

... ihren berufsbezogenen Überzeugungen inklusive ihrer Bezugsnormorientierung (Wissen 2) und

... den makroskopischen und mikroskopischen Kontextbedingungen dieser Handlungsepisoden.

Diese drei Faktoren sind miteinander mehr oder weniger vereinbar und/oder können zueinander im Konflikt stehen. Folglich ist eine Lehrkraft, der es nicht oder nur teilweise gelingt bewusst einen Ausgleich zwischen diesen drei Faktoren herbeizuführen, eine Lehrkraft deren Grundbildung bezüglich schulischer Leistungsfeststellung und -beurteilung nicht oder nur unzureichend einer Idealvorstellung entspricht. Sie verfügt nur zu einem bestimmten Grad über Assessment Literacy, wobei sich dies auf theoretischer Ebene da-

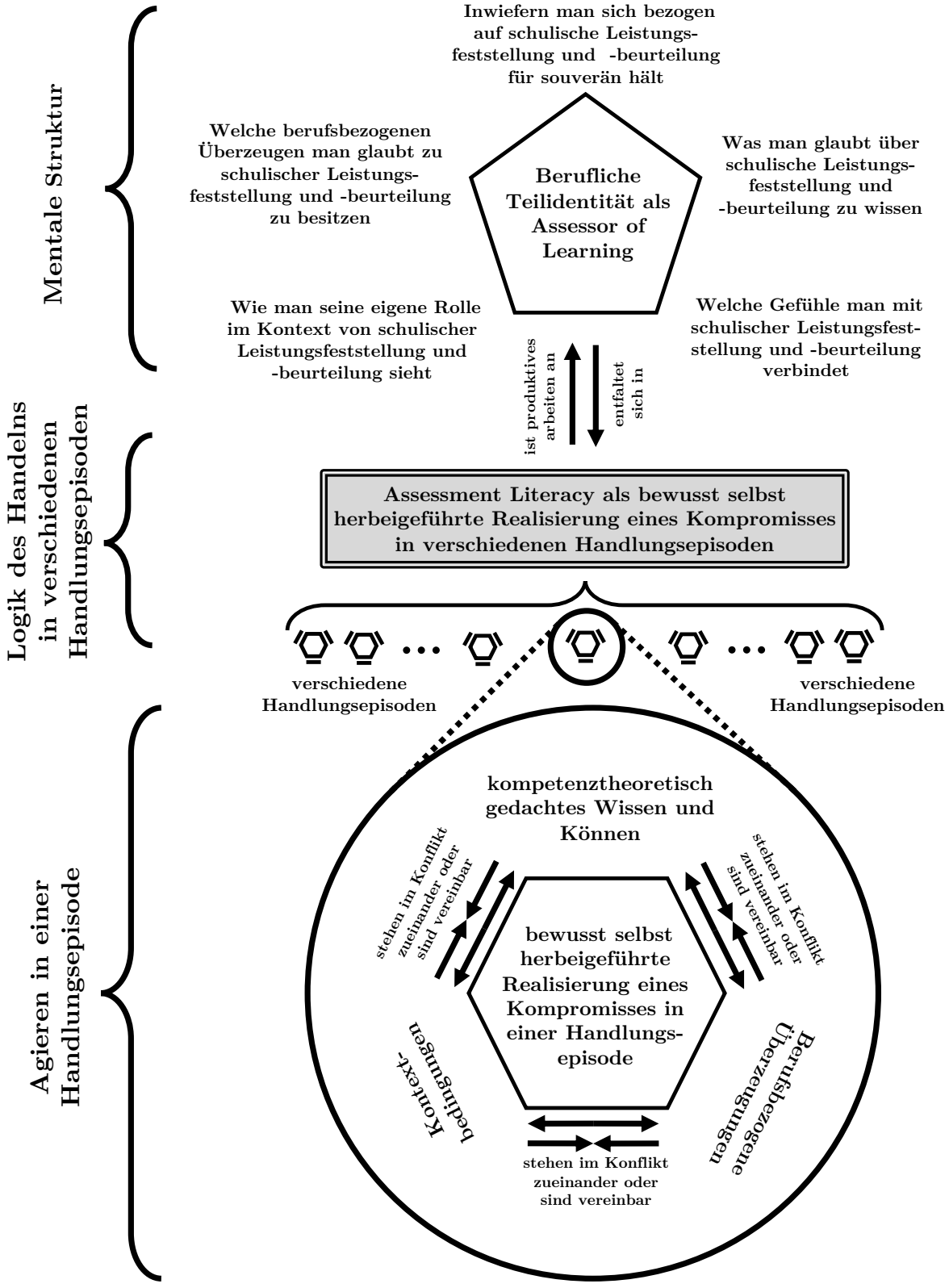


Abbildung 2.5.: Umfassende Konzeption der Assessment Literacy von Lehrkräften auf Grundlage eigener Überlegungen, sowie den Modellen von Xu & Brown (2016) und Looney, Cumming, van Der Kleij, & Harris (2018).

Ausprägungen	Die Lehrkraft agiert im Rahmen schulische Leistungsfeststellung und –beurteilung betreffenden Handlungsepisoden...	Hypothetisches Beispiel einer konkreten Handlungsepisode
Stufe 0	... ausschließlich auf Basis ihrer diesbezüglichen berufsbezogenen Überzeugungen. Sie besitzt kaum kompetenztheoretisch gedachtes Wissen und Können zur schulischen Leistungsfeststellung und –beurteilung, weswegen sie hierauf auch nicht zurückgreifen kann (man beachte hierzu insbesondere die ergänzenden Bemerkungen zu Unterabschnitt 2.1.3). Sie ist „assessment illiterate“.	Eine Lehrkraft hat in ihrer bisherigen Aus- und Weiterbildung kaum Wissen und Können zur adäquaten Erstellung von Leistungsaufgaben und zur Beurteilung von Schülerlösungen erworben. Sie erstellt Leistungsaufgaben und beurteilt Schülerlösungen im Rahmen einer Klassenarbeit daher so, wie ihr aufgrund ihrer berufsbezogenen Überzeugungen angemessen erscheint, ohne dabei auf theoretisches (Ausbilungs-)Wissen zurückgreifen zu können.
Stufe 1	... indem sie ihr kompetenztheoretisch gedachtes Wissen und Können (in Wissen 2 und 3 transformiertes Wissen 1) einsetzt, ohne dabei bewusst einen Ausgleich des mögliche Spannungsverhältnis zu ihren berufsbezogenen Überzeugungen und/oder zu den Kontextbedingungen, denen sie unterworfen ist, herzustellen. Sie agiert in dieser Art und Weise, „weil es so im Lehrbuch steht“.	Eine Lehrkraft hat im Rahmen einer Weiterbildung Wissen zum Einsatz von Operatoren in Leistungsaufgaben und zur Beurteilung von Schülerlösungen erworben. Sie konstruiert daher in ihrer nächsten Klassenarbeit Leistungsaufgaben mit Hilfe fest vorgegebener Operatoren und beurteilt Schülerlösungen anhand der wortwörtlich vorgegebenen Bedeutung dieser Operatoren. Dieses Vorgehen bringt sie aber weder mit ihren evtl. dabei auftretenden Bedenken bewusst in Einklang, noch berücksichtigt sie die tatsächlichen Lerngelegenheiten ihrer Schüler_innen.
Stufe 2	... auf Grundlage einer subjektiven Überzeugung darüber, wie schulische Leistungsfeststellungen und -beurteilungen unter optimalen Bedingungen sein sollten, losgelöst von den Kontextbedingungen, denen sie in diesen Handlungsepisoden unterworfen ist. Diese subjektive Überzeugung erwuchs dabei aus einem von ihr bewusst selbst herbeigeführten Vergleich ihres eigenen, kompetenztheoretisch gedachten Wissens und Könnens und ihren berufsbezogenen Überzeugungen zur schulischen Leistungsfeststellung und -beurteilung. Sie agiert in dieser Art und Weise, „weil es so im Lehrbuch steht“ und ihrer Überzeugung nach allgemein angemessen erscheint.	Eine Lehrkraft hat im Rahmen einer Weiterbildung Wissen zum Einsatz von Operatoren in Leistungsaufgaben und zur Beurteilung von Schülerlösungen erworben. Durch einen bewussten Vergleich dieses Wissens, mit ihren eigenen berufsbezogenen Überzeugungen, ist sie zudem zu einer subjektiven Überzeugung gelangt, wie unter optimalen Bedingungen mit Hilfe von Operatoren Leistungsaufgaben erstellt und Schülerlösungen beurteilt werden sollten. Auf Grundlage dieser Überzeugung konstruiert sie in ihrer nächsten Klassenarbeit Leistungsaufgaben und beurteilt die Lösungen ihrer Schüler_innen. Sie blendet hierbei jedoch Kontextbedingungen, wie z. B. die Vertrautheit ihrer Schüler_innen mit Operatoren weitgehend aus, beurteilt die Lösungen der Schüler_innen also so, als wären diese mit Operatoren in Leistungsaufgaben allumfassend vertraut.
Stufe 3	... idealtypisch. Sie realisiert bewusst selbst herbeiführend einen Kompromiss zwischen den drei Faktoren, die im Rahmen dieser Handlungsepisoden bis zu einem bestimmten Grad vereinbar sind und/oder einander konfligieren. Sie ist „assessment literate“.	Eine Lehrkraft hat im Rahmen einer Weiterbildung Wissen zum Einsatz von Operatoren in Leistungsaufgaben und zur Beurteilung von Schülerlösungen erworben. Durch einen bewussten Vergleich dieses Wissens, mit ihren eigenen berufsbezogenen Überzeugungen, ist sie zu einer subjektiven Überzeugung gelangt, wie Leistungsaufgaben erstellt und Schülerlösungen beurteilt werden sollten. Für die Durchführung ihrer nächsten Klassenarbeit vergleicht sie diese Vorstellungen zudem bewusst mit den Kontextbedingungen, denen sie hierbei unterworfen ist, wie z. B. mit der Vertrautheit ihrer Schüler_innen mit Operatoren in Leistungsaufgaben. Sie gelangt so zu einer subjektiven Überzeugung wie mit Hilfe von Operatoren und unter den Kontextbedingungen, denen sie unterworfen ist, Leistungsaufgaben erstellt und Schülerlösungen beurteilt werden sollten. Auf Grundlage dieser Überzeugung konstruiert sie in ihrer nächsten Klassenarbeit Leistungsaufgaben und beurteilt die Lösungen ihrer Schüler_innen.

Tabelle 2.6.: Theoretische Ausprägungen der Assessment Literacy einer Lehrkraft in Anlehnung an Xu & Brown (2016, S. 159), sowie Stiggins (1991).

nach stufen lässt, zwischen wie vielen der eben genannten drei Faktoren die Lehrkraft bewusst selbst herbeiführend einen Kompromiss realisiert (vgl. Tabelle 2.6). Ferner steht das eben beschriebene Finden eines Vergleichs im Rahmen von schulische Leistungsfeststellung und -beurteilung betreffenden Handlungsepisoden mit der beruflichen Teilidentität dieser Lehrkraft als Assessor of Learning in einer wechselseitigen Beziehung (↑↓-Pfeil in Abbildung 2.5):

Zum einen entfaltet sich diese berufliche Teilidentität, wie in Abschnitt 2.2.3 dargestellt, in Form der im Rahmen verschiedener Handlungsepisoden realisierten Kompromisse. Erkenntlich wird dies auch an den inhaltlichen Überlappungen zwischen dem in Abbildung 2.5 dargestellten Set angewandter Bedeutungen, das die beruflichen Teilidentität einer Lehrkraft als Assessor of Learning rahmt und den drei Faktoren zwischen denen es für eine Lehrkraft gilt im Rahmen einer Handlungsepisode einen Kompromiss zu finden.

Zum anderen ist ein in verschiedenen Handlungsepisoden wiederkehrendes Vergleichen zwischen miteinander vereinbaren und/oder zueinander im Konflikt stehenden Faktoren, eine andauernde selbstbezogene Erfahrung für eine Lehrkraft. Diese selbstbezogene Erfahrung führt im Idealfall zu „ein[em] temporäre[n] Zustand einer gelungenen Passung“ (Keupp et al., 2002, S. 276). Die Logik des Handelns einer bzgl. schulische Leistungsfeststellung und beurteilung grundgebildeten Lehrkraft (Assessment Literacy) ist damit Ausdruck ihres produktives Arbeiten an ihrer beruflichen Teilidentität als Assessor of Learning (vgl. Abschnitt 2.2.3).

2.2.5. Implikationen des Konzepts einer Assessment Literacy für die erziehungswissenschaftliche Forschung und für die Verbesserung der Lehrerbildung

Durch die in Abschnitt 2.2.4 entwickelte umfassende Modellvorstellung der Assessment Literacy von Lehrkräften verlieren die vier in den vorangegangen Unterkapiteln erörterten Forschungstraditionen um Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung nicht ihre jeweilige Daseinsberechtigung. Stattdessen ist die Forschung zu diagnostischen Kompetenzen, zu Bezugsnormen und Bezugsnormorientierng, zu berufsbezogenen Überzeugungen bzgl. schulischer Leistungsbeurteilung, sowie ferner die Forschung, die Leistungsfeststellung und -beurteilung als kognitiven Urteilsprozess auffassen, jeweils ein integraler Bestandteil der vorgestellten Konzeption der Assessment Literacy von Lehrkräften. Vor allem aber liegen diesen vier Forschungstraditionen – wie in den Abschnitten 2.1.1 bis 2.1.4 dargestellt – unterschiedliche Konzepte von Professionalität im Lehrerberuf zugrunde. Assessment Literacy als Rahmenkonzept dieser Forschungstraditionen ermöglicht deshalb eine (simultane) Auseinandersetzung mit Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung aus unterschiedlichen professionstheoretisch Perspektiven. Summa summarum zeichnet Assessment Literacy mit Bezug auf verschiedene Forschungstraditionen ein komplexes Gesamtbild davon, was eine bezüglich alltäglicher schulischer Leistungsfeststellung und -beurteilung grundgebildete

Lehrkraft auszeichnet. Das Konzept einer Assessment Literacy liefert damit eine reichhaltige Grundlage für Forschung, in der Erziehungwissenschaftler_innen aus den vier benannten Forschungstraditionen um Lehrer_innenwissen und können zu schulischer Leistungsfeststellung und -beurteilung interdisziplinär zusammenarbeiten. Durch derartiger Forschung ließen sich beispielsweise die offenen Fragen klären, mit welchen Testverfahren sich die Assessment Literacy von Lehrkräften valide erfassen lässt (vgl. DeLuca, LaPointe-McEwan, & Luhanga, 2016) und/oder inwieweit sich die in Tabelle 2.6 vorgestellte theoretischen Ausprägungen der Assessment Literacy von Lehrkräften empirisch bestätigen lassen.

Da die Assessment Literacy im gesamtes Berufsleben einer Lehrkraft einem Entwicklungsprozess unterworfen ist (berufsbiographischer Wortsinn des Begriffs; vgl. Abschnitt 2.2.3), lassen sich aus diesem Konzept unmittelbar Konsequenzen für alle Phasen der Lehrerbildung ableiten: Lehrkräften sollte in der Aus- und Weiterbildung adäquates Hintergrundwissen zu schulischer Leistungsfeststellung und -beurteilung vermittelt werden. Dieses Hintergrundwissen ist weniger an der Perspektive der Testtheorie, sondern vielmehr am täglichen Unterrichtgeschehen zu orientieren (z. B. Hintergrundwissen zur Herstellung von Kohärenz und Transparenz bei schulischer Leistungsfeststellung und -beurteilung oder zur Entflechtung von Lehr- und Leistungssituationen für Schüler_innen; vgl. Leisen & Höttecke, 2011). Lehreraus- und -weiterbildung sollte auch die Integration von Wissen und Können zu schulischer Leistungsfeststellung und -beurteilung unter verschiedenen Kontextbedingungen thematisieren. Hierzu zählt beispielsweise die Frage der Umsetzung schulischer Leistungsfeststellung und -beurteilung in sprachlich-kulturell heterogenen Lerngruppen (vgl. Lyon, 2013d) oder in inklusiven Lehr-Lern-Settings (vgl. von Bargen, 2017). Ferner müssen Lehrkräften in Aus- und Weiterbildung Möglichkeiten eröffnet werden, erworbenes Hintergrundwissen bezogen auf eigene berufsbezogene Überzeugungen zur Feststellung und Beurteilung von Schülerleistung zu reflektieren[38]. Erst ein solch breit gefächertes Aus- und Weiterbildungsangebot kann für sich in Anspruch nehmen, einen umfassenden Beitrag zur Grundbildung von Lehrkräften bzgl. schulischer Leistungsfeststellung und -beurteilung zu leisten.

2.3. Zusammenfassung

Zu Beginn dieses Kapitels wurde die Frage aufgeworfen, was eine im Kontext von schulischer Leistungsfeststellung und -beurteilung professionell handelnde Lehrkraft auszeichnet. Diese wurde in Unterkapitel 2.2 sowohl ausführlich, als auch präzise geklärt und somit das zentrale Ziel dieses Kapitels erreicht. Eine solche Lehrkraft zeichnet sich durch eine besondere Logik des Handelns, eine Form von Wissen 3 aus, die als Assessment Literacy bezeichnet wird und für die sich auf theoretischer Ebene unterschiedliche Ausprägungen beschreiben lassen (vgl. Abschnitt 2.2.4).

[38] Für eine Zusammenschau didaktisch-methodischer, theoretischer und empirischer Beiträge zur Gestaltung derartiger Aus- und Weiterbildungsangebote siehe Wang et al. (2010, S. 528).

Um diese Frage zu beantworten, aber auch um sich dem Erkenntnisinteresse des empirischen Teils der vorliegenden Arbeit anzunähern, galt es das weite Feld der Forschung um Lehrerwissen und -können zur schulischen Leistungsfeststellung und -beurteilung zu sichten. Hierzu wurden in Unterkapitel 2.1 die vier im (deutschsprachigen) erziehungswissenschaftlichen Diskurs diesbezüglichen am prominentesten vertretenen Forschungsperspektiven erörtert. Folgende Gedanken zu der vorgenommenen Sichtung dieser Forschungsstände sind besonders hervorzuheben:

1. Anzumerken ist, dass diese Sichtung entsprechender Forschungsstände, da sie anschließend mit dem bisher ausschließlich im internationalen Diskurs thematisierte Rahmenkonzept einer Assessment Literacy von Lehrkräften verbunden wurde, eine Form der Zusammenführung der Forschung um Lehrerwissen und -können zur schulischen Leistungsfeststellung und -beurteilung darstellt, wie sie in der deutschsprachigen erziehungswissenschaftlichen Literatur bislang fehlte.
2. Ferner büßen die vier vorgestellten Forschungstraditionen durch deren Unterordnung in ein gemeinsames Rahmenkonzept in keinerlei Hinsicht ihre jeweilige Daseinsberechtigung ein. Vielmehr wird hierdurch deutlich, dass diese vier Forschungstraditionen als unterschiedliche Herangehensweisen an ein und denselben Erkenntnisgegenstand (der Assessment Literacy von Lehrkräften) aufgefasst werden können, die vor allem erst durch ihre voneinander verschiedenen Perspektiven ermöglichen, ein Gesamtbild davon zu zeichnen, was eine im Kontext von schulischer Leistungsfeststellung und -beurteilung professionell handelnde Lehrkraft auszeichnet.
3. Des Weiteren zeigte sich bei jeder der vier vorgestellten Forschungsperspektiven, dass empirische Befunde zu Lehrerwissen und -können zur schulischen Leistungsfeststellung und -beurteilung im Allgemeinen sehr rar sind und zudem oftmals aus dem internationalen Raum stammen, weswegen ihre Übertragbarkeit auf Lehrkräfte, die im deutschen Schulsystem tätig sind, beschränkt ist. Die Kenntnis dieses Umstandes ist keineswegs neu, sondern wurde auch schon anderen Autor_innen beklagt (z. B. Marso & Pigge, 1993, S. 130 u. f.; Terhart, 2000, S. 40). Jedoch wird durch die hier vorgenommene aktuelle Aufarbeitung des Standes der Forschung aufgezeigt, dass diese Forschungslücke bis heute existiert, bzw. mutmaßlich noch zu wenige Anstrengungen unternommen wurden, diese zu verkleinern.
4. Ein besondere Schwerpunkt der Darstellung lag zudem darin, in Erfahrung zu bringen, welche speziell naturwissenschaftsdidaktischen Befunde und Überlegungen zu diesem Themenkomplex bisher existieren. Dabei zeigte sich, das derartige Erkenntnisse in einem Großteil der vier vorgestellten Forschungsperspektiven bzw. in vielen Teilströmungen dieser Perspektiven fehlen. Besonders bemerkenswert ist dabei eine gewisse Dominanz des Kompetenzmodellierungsansatzes im naturwissenschaftsdidaktischen Diskurs um Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung, obwohl selbst innerhalb des kompetenztheoretischen Bestimmungsansatzes von Professionalität im Lehrerberuf auch andersartige Herangehensweisen existieren bzw. diskutiert werden (vgl. Abschnitt 2.1.1 und 2.1.4). Die hier

vorgenommene Darstellung des Forschungsstandes um Lehrerwissen und -können zur schulischen Leistungsfeststellung und -beurteilung deckt damit auch eine Vielzahl der Desideraten auf, die zum Themenkomplex schulische Leistungsfeststellung und -beurteilung insbesondere in der Naturwissenschaftsdidaktik bestehen und denen zukünftige Forschung nachgehen sollte.

Vor allem wurden mit diesem Kapitel aber zentrale Grundlagen für den empirischen Teil der vorliegenden Arbeit gelegt. In diesem steht eine Untersuchung der Genese von Lehrerleistungsurteilen über Schülertexte aus einer Leistungssituation im Physikunterricht im Fokus des Erkenntnisinteresses. Der in Unterkapitel 2.1 dargestellten Stand der Forschung liefert allerdings darüber, wie Physiklehrkräfte bei der Beurteilung von Schülerleistungen in allgemeinen tatsächlich vorgehen, welchen Logiken sie dabei folgen und welche Maßstäbe sie hier für angemessen halten nahezu keine durch Empirie gestützten fachdidaktischen Erkenntnisse. Folglich erscheint es für den empirischen Teil der vorliegenden Arbeit zwar sinnvoll, die in Abschnitt 2.1.4 vorgestellte Perspektive der psychologischen Urteilsforschung auf schulische Leistungsfeststellung und -beurteilung besonders zu berücksichtigen, gleichzeitig aber eine eher explorativ geprägte Herangehenweise auf Grundlage des weiter gefassten Assessment-Literacy-Konzepts zu wählen. Das Rahmenkonzept einer Assessment Literacy von Lehrkräften, wie es in Abschnitt 2.2.4 entwickelt wurde, wird daher als heuristische Grundlage dienen, auf deren Basis die im Rahmen dieser Untersuchung gewonnenen Daten interpretiert werden.

Eine genauere Erörterung des Erkenntnisinteresses und der sich hieraus ableitenden Forschungsfragen der im empirischen Teil der vorliegenden Arbeit vorgestellten Studie, ist allerdings noch nicht möglich. Hierzu nötig ist die Aufarbeitung eines besonderen Teilaspekts der Assessment Literacy von Physiklehrkräften, der im bisherigen Verlauf der vorliegenden Arbeit noch nicht erfolgt ist: die Erörterung bisherigen Kenntnisse zum Umgang von Physklehrer_innen mit Sprache im Rahmen von schulischer Leistungsfeststellung und -beurteilung. Dementsprechend erfolgt im sich nun anschließenden dritten Kapitel ein gedanklicher Sprung zu diesem Themenkomplex, bevor anschließend auf den empirischen Teil der vorliegenden Arbeit übergeleitet wird.

3. Die Bedeutung von Sprache im Rahmen von Leistungsfeststellung und -beurteilung im Physikunterricht

Sprache stellt nicht nur Lernmedium, sondern auch einen Lerngegenstand im Physikunterricht dar. Aus Sicht der (deutschsprachigen) Naturwissenschaftsdidaktik ist dies keine neue Erkenntnis. Sowohl in der nahen als auch in der entfernten Vergangenheit hat sich kontinuierlich eine Vielzahl von Autor_innen mit diesem Gedanken intensiv auseinandergesetzt[39]. Unabhängig von der eigenen Verortung in den diesbezüglichen verschiedensten Perspektiven im naturwissenschaftsdidaktischen Diskurs ist es plausibel anzunehmen, dass Sprache auch eine Bedeutung bei der Leistungsfeststellung und -beurteilung im Rahmen des Physikunterrichts zukommt, da es sich bei Leistungsfeststellungen und -beurteilungen um schulische Verfahren handelt, die das Lernen von Schüler_innen in Relation zu einem Gütemaß stellen (vgl. Unterkapitel 1.1). Diese Bedeutung von Sprache wird im nun folgenden Kapitel mit Fokus auf die Perspektive von Physiklehrer_innen, also als Teil ihrer Assessment Literacy (vgl. Unterkapitel 2.2), erörtert.

3.1. Sprachgebrauch und Sprachgebrauchsnormen (in Leistungssituationen) im (Physik-)Unterricht

Will man sich den bisherigen Kenntnisstand zur Bedeutung von Sprache im Rahmen von Leistungsfeststellung und -beurteilung im Physikunterricht annähern, lohnt sich zunächst eine grobe Illustration der dort anzutreffenden sprachlichen Anforderungen an Schüler_innen. Hierfür eignen sich die folgenden drei Auszüge aus Veröffentlichungen der Kultusministerkonferenz, sowie aus einem aktuellen Lehrplan für das Fach Physik:

[39] Beispielsweise Neunhöffer (1967), Wagenschein (1970a), Mauermann (1976), Wagenschein (1976, S. 130 u. f.), Bach (1984), Cassels & Johnstone (1984), Wagenschein (1986), Steinmüller & Scharnhorst (1987), Diebold & Waldron (1988), Deppner (1989), Luchtenberg (1989), Lemke (1990), Leisen (1991), Bird & Welford (1995), Muckenfuß (1995, S. 245 u. f.), J. W. Rosenthal (1996), Ogborn, Kress, Martins, & McGillicuddy (1996), Fluck (1997), Demidow (1999), Wellington & Osborne (2001), Opitz (2002), Rincke (2007), Willems (2007), Bergeler (2009), Prophet & Babede (2009), Kulgemeyer (2010), Nitz, Nerdel, & Prechtl (2012), Özcan (2013), Heitzmann (2013), Riebling (2013b), Härtig, Bernholt, Prechtl, & Retelsdorf (2015), Krabbe & Beese (2015), Rincke & Leisen (2015), Lyon, Tolbert, Stoddart, Solís, & Bunch (2016), Höttecke, Ehmke, Krieger, & Kulik (2017), Höttecke (2017), Pineker-Fischer (2017), Tajmel (2017b), Höttecke, Feser, Heine, & Ehmke (2018), Rincke & Markic (2018).

Auszug aus den Bildungsstandards der Kultusministerkonferenz im Fach Physik für den mittleren Schulabschluss:
„Die Fähigkeit zu adressatengerechter und sachbezogener Kommunikation ist wesentlicher Bestandteil physikalischer Grundbildung. [...] Zur Kommunikation sind eine angemessene Sprech- und Schreibfähigkeit in der Alltags- und Fachsprache [...] erforderlich." (Kultusministerkonferenz , 2004a, S. 10)

Auszug aus dem Orientierungsrahmen für den Lernbereich globale Entwicklung:
„Als Bildungssprache ist sie [die deutsche Sprache; M. S. F.] auch das *Medium*, das Schulwissen und fachliches Wissen mit zum Teil spezifischen sprachlichen Mitteln transportiert und auch Elemente wissenschaftlichen Sprechens enthält, [...] und zwar über das Alltagssprachliche hinausgehende, situationsunabhängige Textsorten[.] [...] Deutsch ist Verständigungssprache im Fachunterricht (*Medium* und *Prinzip*)[.]" (Schreiber & Siege, 2016, S. 130-131, Hervorhebungen im Original)

Auszug aus dem Bildungsplan für das Fach Physik in der Sekundarstufe I des Gymnasiums der Freien und Hansestadt Hamburg:
„[Bildungssprachliche] Kompetenzen [...] [sind] die Grundvoraussetzung für erfolgreiches Lernen [...] [und] werden in der von der Alltagssprache dominierten Lebenswelt der Schülerinnen und Schüler nicht automatisch erworben, sondern ihr Aufbau ist Aufgabe aller Fächer[.] [...] Die Schülerinnen und Schüler werden an die besondere Struktur der Fachsprachen herangeführt, sodass sie erfolgreich am Unterricht teilnehmen können. Fachsprachen weisen verschiedene Merkmale auf, die in der Alltagssprache nicht üblich sind, aber in Fachtexten gehäuft auftreten[.]" (Behörde für Schule und Berufsbildung Hamburg , 2011, S. 13)

Was sich in diesen Auszügen andeutet, sind drei Facetten schulischer Sprachgebrauchsnormen, die im Folgenden aufgelistet werden und deren Verhältnis zueinander als kaskadenförmig geschachtelt zu denken ist. Aus diesen Normen leiten sich unmittelbar sprachliche Anforderungen ab, mit denen Schüler_innen (in Leistungssituationen) im Physikunterricht konfrontiert werden, da diese Sprachnormen als (Teil-)Antwort auf die Frage verstanden werden können, „[ü]ber welche Sprachfähigkeiten [...] schulisch erfolgreiche Schüler[_innen] [verfügen sollen]" (Feilke, 2012, S. 155):

- Erstens deutet sich vor allem in dem Zitat „Deutsch ist Verständigungssprache im Fachunterricht" (Schreiber & Siege, 2016, S. 131) an, dass in den oben aufgeführten Auszügen mit dem Wort „Sprache" sinngemäß Kenntnisse der deutschen Sprache gemeint sind. Hier wird also das monolinguale Selbstverständnis des deutschen Bildungswesens sichtbar, gemäß dem sich Schule deutschsprachig zu organisieren habe und damit Einsprachigkeit zur Norm erhoben wird (vgl. Gogolin, 2008, S. 3). Entsprechend dieser Norm müssen Schüler_innen dann aber zunächst über Kenntnisse der deutschen Sprache verfügen, um in schulischen Lern- und Leistungssituationen überhaupt erfolgreich sein zu können.
- Zweitens wird hier die „Schulnormthese" von Feilke (2012, S. 154 u. f.) deutlich, vor allem in der Aussage, dass Sprachfähigkeiten, über die schulisch erfolgreiche Schüler_innen verfügen müssen, „nicht automatisch erworben werden, sondern ihr Aufbau [...] Aufgabe aller Fächer [ist]" (Behörde für Schule und Berufsbildung Hamburg , 2011, S. 13). Diese These besagt, dass das Qualifikationsziel, welches mit

schulischen Sprachgebrauchsnormen verbunden ist, zunächst nicht die Vorbereitung auf den außerschulischen Sprachalltag, sondern primär der Spracherfolg im Sozialsystem Schule selbst ist. „Schulische Sprachnormen sind [...] [demnach] Normen sui generis“ (ebd., S. 154). Der Sprachgebrauch von (in Leistungssituationen) erfolgreichen Schüler_innen genügt also weniger den Normen des Gebrauchs der deutschen Sprache im Allgemeinen (vgl. ebd.), sondern in erster Linie den „spezifischen Sprach- und Sprachgebrauchsnormen des Systems Schule“ (ebd.).

- Drittens verweisen alle drei Auszüge darauf, dass ein Sprachgebrauch, der den Normen der Schule gerecht wird, nicht einer einzigen Art und Weise des Einsatzes von Sprache entspricht. Stattdessen bestimmen die unterschiedlichen situativen Bedingungen im (Physik-)Unterricht, welcher Sprachgebrauch als adäquat gilt (vgl. Tajmel, 2013, 242 u. f.). In den obigen Auszügen, wie auch generell im gegenwärtigen naturwissenschaftsdidaktischen Diskurs (z. B. Höttecke et al., 2017, S. 54 u. f.; Tajmel, 2017a), wird dabei vor allem auf die Varietäten[40] Fach- und Bildungssprache verwiesen, die sich von der Alltagssprache[41] der Schüler_innen unterscheiden. Damit werden (in Leistungssituationen) erfolgreiche Schüler_innen als solche umrissen, die den für verschiedene Unterrichtssituationen angemessenen bzw. erforderlichen Sprachgebrauch beherrschen.

Insgesamt zeichnet sich anhand dieser blitzlichtartigen Betrachtung ab, dass es für Schüler_innen gilt, nicht zu vernachlässigende sprachliche Anforderungen zu meistern, um (in Leistungssituationen) im Physikunterricht erfolgreich zu sein. Auch diese Feststellung ist der (deutschsprachigen) Naturwissenschaftsdidaktik keineswegs neu. Vielmehr ist sie bereits in den Prämissen zu erkennen, die sich in der Literatur zum Stellenwert von Sprache für das Lernen im Physikunterricht finden lassen. Das Meinungsspektrum schwankt hier zwischen den Thesen „[n]aturwissenschaftlicher Unterricht ist kein Sprachunterricht. Und doch ist Sprache für die Naturwissenschaften und das Lernen naturwissenschaftlicher Begriffe und Konzepte sehr bedeutsam“ (Höttecke, 2017, S. 107) und „[e]very science lesson is a language lesson [...] [and] [l]anguage is a major barrier (if not *the* major barrier) to most pupils in learning science“ (Wellington & Osborne, 2001, S. 2, Hervorhebungen im Original). Weitgehende Einigkeit besteht also darin, dass Sprache eine wesentliche Anforderung des Physikunterrichts darstellt (vgl. Tajmel, 2017b, S. 199 u. f.). Gerade aufgrund dieser Konsensmeinung scheint es sinnvoll und notwendig, die eben dargelegte

[40] Im Rahmen der vorliegenden Arbeit wird der Begriff „Varietät“ gemäß seiner geläufigsten Verwendung verstanden, nämlich als „allgemeiner Oberbegriff zur Erfassung der Heterogenität einer Einzelsprache. [...] Er umfaßt danach neben [...] Größen und Phänomenen wie *Dialekt*, *Jargon*, *Soziolekt* usw. auch Differenzierungen, die mit den Ausdrücken *Register*, *Funktionalstil* u.ä. [...] bezeichnet werden[.] [...] Der Ausdruck dient also in erster Linie als bequemer globaler Begriff, [...] [dem] kein spezifischer Forschungsansatz und keine einheitliche Konzeption zugeordnet werden kann. Immerhin impliziert er (tendenziell) wenigstens zweierlei: (a) Varietäten werden als Differenzierungen innerhalb einer Einzelsprache gefaßt; (b) sie erscheinen als in sich kohärente, diskrete systemartige Gebilde im Sinne von „Sprachen in der Sprache““ (Adamzik, 1998, S. 181-182, Hervorhebungen im Original).

[41] Für einen Vorschlag einer expliziten Bestimmung des Begriffs „Alltagssprache“ im Kontext von physikbezogenen Lehr-Lern-Settings siehe Tajmel (2017a, S. 254 u. f.). Zwecks einer Klärung der herausragenden Bedeutung, die Alltagssprache beim Lernen von Physik spielt, siehe neben Tajmel (2017a, S. 256 u. f.) zudem z. B. Wagenschein (1986), Muckenfuß (1995, S. 247 u. f.) oder Rincke (2010).

erste Annäherung an schulische Sprachgebrauchsnormen zu vertiefen, bevor sich genauer mit dem bisherigen Kenntnisstand zum Umgang von Physiklehrer_innen mit Sprache im Rahmen von schulischer Leistungsfeststellung und -beurteilung beschäftigt wird. Diese Vertiefung erfolgt dadurch, dass in den folgenden Abschnitten genauer geklärt wird, was unter dem Gebrauch von Fach- und Bildungssprache als sprachliche Anforderung an Schüler_innen im Physikunterricht zu verstehen ist und insbesondere in welchem Verhältnis Bildungssprache und Fachsprache zueinander stehen.

3.1.1. Die Fachsprache der Naturwissenschaft Physik im Physikunterricht

Als erstes soll die Fachsprache der Naturwissenschaft Physik, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird, näher bestimmt werden. Zu diesem Zweck eignet sich das Ordnungssystem, mit dessen Hilfe verschiedene Fachsprachen in der Fachsprachenforschung voneinander unterschieden werden. Am weitesten verbreitet ist hierbei eine Gliederung von Fachsprachen entlang zweier Dimensionen, die als *horizontale* und *vertikale Gliederung* bezeichnet werden (vgl. Roelcke, 2014):

- Verschiedene Fachsprachen horizontal zu gliedern bedeutet, diese „nach verschiedenen Fächern und Fachbereichen wie etwa Germanistik, Jursiprudenz oder Naturwissenschaft und Technik [zu unterscheiden]“ (ebd., S. 155).
- Verschiedene Fachsprachen vertikal zu gliedern bedeutet, diese „nach verschiedenen Abstraktionsebenen und Kommunikationsbereichen wie Theoriesprache, fachliche Umgangssprache oder Kommunikation zwischen Händlern und Verbrauchern [zu unterscheiden]“ (ebd.).

In der aktuellen Fachsprachenforschung werden zudem deutlich feinkörnigere Gliederungen anhand von Textsorten vorgenommen (z. B. Göpferich, 1995; Niederhaus, 2011), von denen angenommen wird, dass sie zukünftig „den Ansatz einer vertikalen (und ggf. horizontalen) Dimension obsolet machen [könnten]“ (Roelcke, 2014, S. 155). Dieser noch andauernde Diskurs soll allerdings im Rahmen der vorliegenden Arbeit nicht weiter aufgegriffen werden. Stattdessen wird sich auf den Ansatz einer horizontalen und vertikalen Gliederung verschiedener Fachsprachen beschränkt.

3.1.1.1. Horizontale Gliederung von Fachsprachen

Die horizontale Gliederung von Fachsprachen entlang unterschiedlicher Fächer bzw. Fachbereiche entspricht nach Auffassung von Adamzik (1998) jener Dimension, die eine Fachsprache am deutlichsten charakterisiert (vgl. ebd., S. 184). Dementsprechend besitzt jedes Fach seine eigene Fachsprache, die jeweils isoliert zu betrachten ist, weswegen auch Aussagen, die von „der“ Fachsprache im Allgemeinen sprechen, de facto wenig gegenstandsadäquat sind (vgl. Niederhaus, 2011, S. 43). Ferner ist der Begriff „Fach“ in diesem Zusammenhang weitläufig zu verstehen, nämlich als spezialisierter, menschlicher Tätig-

keitsbereich (vgl. Baumann, 1992, S. 145; Adamzik, 1998, S. 184; Roelcke, 2014, S. 155). Dieses weitläufige Verständnis begründet sich daraus, dass sich die Fachsprachenforschung bislang – wie Kalverkämper (1992) feststellt und nach Niederhaus (2011, S. 24) bis heute gilt – „noch gar nicht um den Begriff des Faches bemüht, bis auf wenige Einzelstimmen ihn als Problem noch nicht einmal erkannt hat“ (Kalverkämper, 1992, S. 32).

In jedem Fall lässt sich folgern, dass sich die Fachsprache der Naturwissenschaft Physik, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird, durch die in ihm thematisierten spezifischen Tätigkeitsbereiche bzw. Inhalte seiner Bezugsdisziplin (die Naturwissenschaft Physik) auszeichnet. Diese Inhaltsgebiete des Physikunterrichts lassen sich nun aus dem Diskurs darüber, welchen Beitrag der Physikunterricht zu einer *Scientific Literacy*, oder mit den Worten von Martin Wagenschein zur „ernsthafte[n] Wissenschaftsverständigkeit *aller* Bürger [leisten soll]“ (Wagenschein, 1983, S. 82, Hervorhebung im Original), ableiten. Wie J. Krüger (2017, S. 5 u. f.) ausführlich darlegt, existieren zum Begriff „Scientific Literacy“ im naturwissenschaftsdidaktischen Diskurs eine Vielzahl von zum Teil sehr unterschiedlichen Konzeptionen. Gemäß Härtig et al. (2015, S. 56) zeichnet sich aber im internationalen Raum der Trend ab, Scientific Literacy im Sinne der Überlegungen von Yore, Pimm, & Tuan (2007) als zwei miteinander verknüpfte Bereiche zu konzipieren, einer „interacting fundamental literacy“ (ebd., S. 561) und einem der „derived understandings“ (ebd.). Letzterer lässt sich dabei in Form von schlagwortartigen Wissensstandards umschreiben (vgl. Yore et al., 2007, S. 568 u. f.; Härtig et al., 2015, S. 56):

- Ein umfassendes Verständnis naturwissenschaftlicher Basiskonzepte
- Ein Verständnis der Natur der Naturwissenschaften
- Ein Verständnis über Methoden der naturwissenschaftlichen Erkenntnisgewinnung
- Ein Verständnis von und über Technik
- Ein Verständnis des Beziehungsgeflechts zwischen Naturwissenschaft, Technik, Gesellschaft und Umwelt

Diese fünf schlagwortartigen Wissensstandards umreisen auf globaler Ebene die Inhaltsgebiete naturwissenschaftlichen Unterrichts (also z. B. des Physikunterrichts). Sie charakterisieren also auch den fachlichen Sprachgebrauch in diesem spezialisierten menschlichen Tätigkeitsbereich und ermöglichen somit die Fachsprache der Naturwissenschaft Physik, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird, horizontal von anderen Fachsprachen zu unterscheiden.

3.1.1.2. Vertikale Gliederung von Fachsprachen

Neben einer horizontalen Gliederung von Fachsprachen finden sich in der Literatur auch vertikale Gliederungsvorschläge (z. B. Ischreyt, 1965, S. 38 u. f.; Hoffmann, 1985, S. 64 u. f.; Roelcke, 2014, S. 163 u. f.). Hierbei wird versucht „die zunehmende Präzisierung zu verfolgen, die die Sprache in der fachlichen Kommunikation erfährt, je weiter diese [...] als

Erkenntnis- und Kommunikationsinstrument vom Konkreten zum Abstrakten, vom Besonderen zum Allgemeinen, von der Erscheinung zum Wesen vordringt“ (Hoffmann, 1985, S. 64). Von unterschiedlichen Autor_innen wurden hierzu verschiedene Ansatzpunkte verfolgt:

So nimmt beispielsweise Ischreyt (1965) eine Unterteilung technischer Sprachen in wissenschaftliche Fachsprachen, Werkstatt- und Verkäufersprachen vor (vgl. ebd., S. 43 u. f.), wobei wissenschaftliche Fachsprachen (z. B. die der Naturwissenschaft Physik)...

> „[...] in ihrer erwünschten Abstraktion und Objektivität unmenschlich [sind.] [...] [E]ines der Mittel[,] [mit dem dies erreicht wird, ist] [...] die institutionelle Normierung[.] [...] [D]urch deren Verfahren [wird] eine Metapher unweigerlich getötet [...], indem man eine Definition für verbindlich erklärt [...] [oder] Metaphern [werden] durch die Terminologienormierung nach Möglichkeit ausgeschaltet und durch Zusammensetzung abgelöst.“ (ebd., S. 45)

In der von Ischreyt (1965) vorgeschlagenen vertikalen Gliederung von Fachsprachen stellt also das Abstraktionsniveau des Sprachgebrauchs das zentrale Unterscheidungsmerkmal dar. Hoffmann (1985) greift dies auf, nimmt aber eine Erweiterung vor, indem er die Merkmale „äußere Sprachform[,] [...] das Milieu in dem diese Sprachschichten gebraucht werden [und] [...] die Kommunikationsträger oder -teilnehmer [selbst mitberücksichtigt]“ (ebd., S. 65-66). Hierdurch gelangt er zu einer idealisierten vertikalen Gliederung von Fachsprachen in fünf sogenannte *Schichten*: der „Sprache der theoretischen Grundlagenwissenschaften“, der „Sprache der experimentellen Wissenschaften“, der „Sprache der angewandten Wissenschaften und der Technik“, „der Sprache der materiellen Produktion“ und der „Sprache der Konsumtion“ (ebd., S. 70). Dabei ist gemäß Steinmüller & Scharnhorst (1987) die „Sprache der Konsumtion“, bei der sich „die Sprachverwendung [...] durch einige Fachtermini und einige syntaktische Besonderheiten [auszeichnet[42],] [...] [ansonsten aber auf] allgemeinsprachlicher Basis gekennzeichnet [ist,] [...] das Niveau von Fachsprache, wie es im Unterricht der Sekundarstufe I angemessen ist und auch verwendet wird“ (ebd., S. 6).

An vertikalen Gliederungen, wie sie Ischreyt (1965) oder Hoffmann (1985) vorschlagen, lässt sich allerdings ihre Unvollständigkeit, ihre zu ausgeprägte Grobkörnigkeit, sowie ihre geringe Übertragbarkeit auf nicht naturwissenschaftlich-technische Fachsprachen bemängeln (vgl. Roelcke, 2014, S. 156 u. f.). Roelcke (2014) schlägt daher vor, Fachsprachen allgemeiner, nämlich anhand „der Unterscheidung zwischen Experten und Laien[,] [...] [sowie der] Kommunikation innerhalb eines Bereichs oder über dessen Grenzen hinaus [vertikal voneinander abzugrenzen]“ (ebd., S. 162). Durch diese Überlegung gelangt er ebenfalls zu einer fünfgliedrigen vertikalen Schichtung von Fachsprachen, die den eben genannten Kritikpunkten gerecht wird und wohl einen der aktuellsten Vorschläge zur vertikalen Gliederung von Fachsprachen darstellt (vgl. ebd., S. 163 u. f.):

[42] Diese Oberflächenmerkmale von fachlichem Sprachgebrauch im Rahmen von Physikunterricht sind vielzählig, wurden aber bereits von einer beträchtlichen Anzahl von Autor_innen detailliert und ausführlichst beschrieben (z. B. Deppner, 1989, S. 83 u. f.; Fluck, 1997, S. 35 u. f.; Leisen, 2005, S. 7; Eckhardt, 2008, S. 69 u. f.; Niederhaus, 2011, S. 48 u. f.; Härtig et al., 2015, S. 58 u. f.).

Schicht 1: Die Sprache von Expert_innen, die aus demselben Sachbereich eines Fachs stammen (z. B. beim Austausch zweier theoretischer Physiker_innen miteinander, die beide an einem bestimmten Themengebiet arbeiten)

Schicht 2: Die Sprache von Laien in einem bestimmten Sachbereich (z. B. beim Austausch von Schüler_innen im Rahmen des Physikunterrichts)

Schicht 3: Die Sprache von Expert_innen, die aus verschiedenen Sachbereichen desselben Fachs stammen (z. B. beim Austausch eines_einer theoretischen mit einem_einer Experimental-Physiker_in, die beide an einer bestimmten Forschungseinrichtung arbeiten)

Schicht 4: Die Sprache eines_einer Experten_Expertin eines Fachs und eines Laien im Austausch über einen bestimmten Sachbereich des entsprechenden Fachs (z. B. ein Lehrer_in-Schüler_innen-Gespräch im Rahmen des Physikunterrichts)

Schicht 5: Die Sprache eines_einer Experten_Expertin aus einem bestimmten Sachbereich eines Fachs und eines_einer Experten_Expertin aus einem bestimmten Sachbereich eines anderen Fachs (z. B. beim Austausch eines_einer Ornithologen_Ornithologin und eines_einer Astrophysikers_Astrophysikerin zum Thema nachhaltige Entwicklung)

Die aufgeführten Beispiele für Schicht 2 und 4 in obiger Auflistung illustrieren, dass diese vertikalen Schichten der Fachsprache der Naturwissenschaft Physik (in Leistungssituationen) im Physikunterricht auftreten können. Berücksichtigt man allerdings, dass Physikunterricht auch fächerübergreifend organisiert, im Rahmen von Exkursionen oder an außerschulischen Lernorten stattfinden kann, wird offensichtlich, dass Schüler_innen insgesamt mit jeder dieser fünf vertikalen Schichten der Fachsprache der Naturwissenschaft Physik im Rahmen des Physikunterrichts in Berührung kommen können. Ergänzend kommt hinzu, dass Schüler_innen nicht erst im Physikunterricht der Sekundarstufe mit der Fachsprache der Naturwissenschaft Physik konfrontiert werden, sondern bereits im Vor- und Grundschulalter über Erwachsene in Alltagskontexten immer wieder fachsprachliche Elemente aufgreifen können bzw. auf verschiedenste vertikale Schichten von Fachsprachen treffen (vgl. Luchtenberg, 1989).

3.1.1.3. Zwischenfazit

Insbesondere aus der vertikalen Gliederung der Fachsprache der Naturwissenschaft Physik lässt sich schlussfolgern, dass es (in Leistungssituationen) im Rahmen von Physikunterricht nicht „den“ Gebrauch von Fachsprache gibt, sondern dass im Physikunterricht unterschiedlichste Arten des fachlichen Sprachgebrauchs auftreten und diese eher in Ausnahmefällen mit dem Gebrauch von wissenschaftlicher Fachsprache im Sinne von Ischreyt (1965) gleichgesetzt werden können. Meist nimmt Fachsprache, wie sie im Physikunterricht verwendet wird, eine Zwischenstellung zwischen der Alltagssprache der Schüler_innen und wissenschaftlicher Fachsprache ein (vgl. Leisen, 1991; Leisen, 2005; Leisen,

2017) und lässt sich durch die Inhalte, die im Rahmen von Physikunterricht thematisiert werden charakterisieren (vgl. Unterabschnitt 3.1.1.1).

Ferner wurde in diesem Abschnitt darauf hingewiesen, dass sich fachlicher Sprachgebrauch im Physikunterricht „durch das häufige Auftreten von Fachbegriffen, Satz- und Textkonstruktionen sowie [durch besondere] morphologische und syntaktische Merkmale [auszeichnet]“ (Eckhardt, 2008, S. 69). Diese vielzähligen Oberflächenmerkmale von fachlichen Sprachgebrauch im Rahmen von Physikunterricht wurden bereits von einer großen Anzahl von Autor_innen detailliert und ausführlichst beschrieben, weswegen auf eine genaue Ausführung dieser Merkmale verzichtet wurde. Von zentraler Bedeutung ist vielmehr die sich hieraus ableitende Erkenntnis, nämlich dass die im Rahmen von Physikunterricht von Schüler_innen bis zu einem bestimmten Grad geforderte aktive und passive Beherrschung eines fachlichen Sprachgebrauchs durch bestimmte linguistische Merkmale charakterisierbar ist.

3.1.2. Bildungssprache und ihr Verhältnis zu Fachsprache, wie sie im Physikunterricht verwendet wird

Seit den frühen 2000er Jahren kann eine zunehmende Verbreitung des Begriffs „Bildungssprache“ im erziehungswissenschaftlichen Diskurs beobachtet werden (vgl. Morek & Heller, 2012, S. 67 u. f.). Der Begriff ist dabei keineswegs neu, sondern wurde schon im 19. und frühen 20. Jahrhundert in der erziehungswissenschaftlichen Literatur verwendet, „[um] eine ‚besonders wertvolle‘ Varietät [der deutschen Sprache] zu kennzeichnen“ (vgl. Gogolin & Duarte, 2016, S. 480). So ist beispielsweise für Drach (1928) Bildungssprache eine „der Hochsprache angenäherte, aber doch gelegentlich [...] von der Heimatmundart angehauchte [Art des Sprachgebrauchs]“ (ebd., S. 671). Nach Drach (1928) ist daher Bildungssprache durch sprachliche Merkmale charakterisiert, „die in jeder Provinz unter den Gebildeten durch Gebrauch beglaubigt [sind]“ (ebd., S. 671).

Das wissenschaftliche Verständnis des Begriffs „Bildungssprache“ hat sich seit dem 19. und frühen 20. Jahrhundert allerdings deutlich gewandelt. Gemäß der Heuristik von Morek & Heller (2012, S. 70) umfasst der Wortsinn von „Bildungssprache“ gegenwärtig drei Facetten[43], die diesen Begriff aus unterschiedlichen Perspektiven ausleuchten, aber dennoch miteinander verwoben sind:

1. Bildungssprache als Transfermedium für Wissen
2. Bildungssprache als Denkwerkzeug
3. Bildungssprache als Eintritts- und Visitenkarte

Diese drei Facetten werden im Folgenden beschrieben und dabei insbesondere das Verhältnis zwischen Bildungssprache und Fachsprache, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird, dargelegt.

[43] Eine analoge Unterscheidung verschiedener Bedeutungen des Begriffs „Bildungssprache“ findet sich zudem bei Gogolin & Duarte (2016, S. 483).

3.1.2.1. Bildungssprache als Transfermedium für Wissen

In erziehungswissenschaftlichen Publikationen herrscht aktuell der Trend vor, Bildungssprache (um sie als Transfermedium für Wissen beschreiben zu können) als Register auszulegen (z. B. Schleppegrell, 2004, S. 43 u. f.; Riebling, 2013a, S. 110 u. f.; Höttecke et al., 2017, S. 54 u. f.; Tajmel, 2017a, S. 254 u. f.). Was es bedeutet eine Sprachvarietät als Register aufzufassen, wird anhand der Definition von Reid (1956) deutlich, bei der es sich (mutmaßlich) um die erstmalige Verwendung des Begriffs „Register“ als linguistische Kategorie handelt (vgl. Meisel, 1975, S. 30; Hess-Lüttich, 1998, S. 209):

> „In Situationen, die den Anschein geben, linguistisch identischen Bedingungen zu unterliegen, wird er (der Sprecher, V.H.) bei verschiedenen Gelegenheiten gemäß dessen, was grob als unterschiedliche soziale Situation beschrieben werden kann, unterschiedlich sprechen (oder schreiben): Er wird eine Anzahl unterschiedlicher "Register" verwenden.“ (Reid, 1956, S. 32, Übersetzungen nach Hinnenkamp, 1982, S. 35)

Gemäß dieser frühen Definition spricht man bei einer Sprachvarietät also von einem Register, wenn es sich hierbei um eine Varietät handelt, die an bestimmte Charakteristika der Kommunikationssituation, in denen diese verwendet wird, gebunden ist (vgl. Hess-Lüttich, 1998, S. 209). Halliday, McIntosh, & Strevens (1970) greifen diesen Gedanken auf, indem sie den Begriff Register als Varietät einer Sprache „according to use“ (ebd., S. 87) definieren. Ferner schlagen sie drei Dimensionen vor („Field“, „Mode“ und „Style of Discourse“, vgl. ebd., S. 90 u. f.), mit Hilfe derer sich ein Register näher bestimmen bzw. von anderen Registern unterscheiden lässt und die sich – abseits terminologisch andersartiger Bezeichnungen – in den meisten Abhandlungen zum Registerbegriff finden lassen (vgl. Hess-Lüttich, 1998, S. 210 u. f.). Bei diesen Dimensionen gilt es allerdings zu beachten, dass sich verschiedene Register nicht notwendigerweise vollständig voneinander abgrenzen lassen. Vielmehr können verschiedene Register einander überlappen oder ein weit gefasstes Register kann eines oder mehrere eng gefasste Register umschließen (vgl. Biber, 2009, S. 823; Niederhaus, 2011, S. 41 u. f.):

- „Field of Discourse“: Diese Dimension „bezieht sich auf den *Redegegenstand*, den Inhalt der Verständigung, das Thema des Textes [...], den Texttyp, das Genre des Textes [...], das Sach-, Fach- und Arbeitsgebiet, in dem sprachlich gehandelt, über das sprachlich verhandelt wird“ (Hess-Lüttich, 1998, S. 210, Hervorhebungen im Original).
- „Mode of Discourse“: Mit der Dimension „Mode of Discourse“ werden Register danach unterschieden, wie stark (bzw. wie schwach) die sprachlichen Realisierungen dieses Registers konzeptionell am mündlichen bzw. am schriftlichen Sprachgebrauch orientiert sind, ohne dabei notwendigerweise auch tatsächlich in mündlicher oder schriftlicher Form vorliegen zu müssen (vgl. Koch & Oesterreicher, 1985, S. 17 u. f.; Riebling, 2013a, S. 113).
- „Style of Discourse“: Die Dimension „Style of Discourse“ charakterisiert ein Register durch die sprachlichen Nuancierungen, die sich dadurch ergeben, in welchem sozialen

Verhältnis die in der Situation agierenden Personen zueinander stehen (vgl. Halliday et al., 1970, S. 92 u. f.; Hess-Lüttich, 1998, S. 210).

Für das Register Bildungssprache lassen sich diese drei Dimensionen in Relation zu Fachsprachen, wie sie im Fachunterricht verwendet werden und die ebenfalls als Register beschreibbar sind (vgl. Hess-Lüttich, 1998), wie folgt auslegen:

- Das „Field of Discourse“ des Registers Bildungssprache ist „Wissen, das über das Alltagswissen hinausgeht[,] [...] das im Zuge einer verlängerten Ausbildung und/oder durch Teilnahme an öffentlichen Diskursen sowie aufgrund vertiefender Interessen und besonderer lebensweltlicher Erfahrungen erworben wird“ (Ortner, 2009, S. 2227; vgl. auch Habermas, 1978, S. 330). Das „Field of Discourse“ des Registers Bildungssprache umfasst damit den weiten Bezugsbereich schulischen aber auch außerschulischen Lernens. Bildungssprache fungiert damit domänenübergreifend (vgl. Ortner, 2009, S. 2229; Morek & Heller, 2012, S. 70) und unterscheidet sich hierdurch von Fachsprachen, die sich durch ihre Fachspezifität auszeichnen (vgl. Pineker-Fischer, 2017, S. 68). Gleichzeitig wird hier allerdings deutlich, dass sich Bildungssprache nicht trennscharf von Fachsprachen unterscheiden lässt (vgl. ebd.), da sich letztere vertikal auch über die Grenzen eines bestimmten Sach- oder Fachbereichs hinausgehend schichten lassen (vgl. Unterabschnitt 3.1.1.2). Verengt man den Blick auf einen bestimmten Fachunterricht, lässt sich daher das „Field of Discourse“ des Registers der dort anzutreffenden Bildungssprache kaum von jenem der Fachsprache, wie sie in diesem Fachunterricht verwendet wird, unterscheiden.
- Der „Mode of Discourse“ des Registers Bildungssprache wird in der Literatur übereinstimmend als konzeptionell schriftliche Realisierung des Sprachgebrauchs beschrieben (z. B. Habermas, 1978, S. 330; Gogolin, 2009, S. 270; Ortner, 2009, S. 2228; Neumann & Domenech, 2010, S. 5; Lengyel & Roth, 2012, S. 124; Riebling, 2013a, S. 118 u. f.). Bildungssprache ist demnach „unter anderem durch eine Entkopplung von Sprachproduktion und -rezeption (Monologizität) sowie eine weitgehende Unabhängigkeit von paralinguistischen Elementen des situativen Kontextes (hoher Grad der Versprachlichung, Explizitheit) gekennzeichnet“ (Riebling, 2013a, S. 70). Auch Fachsprachen weisen das Merkmal auf, konzeptionell schriftlich realisiert zu werden[44] (vgl. Hess-Lüttich, 1998, S. 211 u. f.; Niederhaus, 2011, S. 44 u. f.), weswegen sich bezüglich des „Mode of Discourse“ keine Unterschiede zwischen Bildungssprache und Fachsprachen feststellen lassen (vgl. Pineker-Fischer, 2017, S. 65 u. f.).
- Der „Style of Discourse“ des Registers Bildungssprache ist durch den öffentlich, institutionellen Rahmen, in dem bildungssprachliche Kommunikation stattfindet,

[44] Dies gilt jedoch nicht für *Fachjargon*, der von Fachsprache zu unterscheiden ist (vgl. Janich, 1998, S. 41 u. f.). In seiner geläufigsten Verwendung steht der Begriff „Fachjargon“ für eine „innerhalb eines Fachgebiets [herausgebildete,] quasi [...] fachbezogene (und meist nur gesprochene) Umgangssprache[,] [...] [die] neben der fachlichen Verständigung zusätzlich auch gruppenidentitätsstiftende, -demonstrierende oder -abgrenzende Funktion haben kann (z. B. Medizinersprache, Laborslang u.ä.) (vgl. Wright 1974, 4; Polenz 1981, 94 [Wright, 1974, S. 4; von Polenz, 1981, S. 94; M. S. F.])“ (Janich, 1998, S. 41). Fachjargon ist also im Gegensatz zu Fachsprache konzeptionell mündlich realisiert und besitzt einen persönlichen Tenor.

> bestimmt (vgl. Gogolin & Duarte, 2016, S. 486). Bildungssprache, aber ebenso auch Fachsprachen[44] (wie sie im Fachunterricht verwendet werden), ist dementsprechend durch emotionale Distanz, sowie eine weitgehende Fremdheit und offenkundige Hierarchie der Kommunikationspartner (z. B. in einem Lehrer_innen-Schüler_innen-Gespräch) gekennzeichnet (vgl. Schleppegrell, 2004, S. 58; Riebling, 2013a, S. 122 u. f.). „[Es wird] die Wahl eines unpersönlichen Tenors, die „Ich-Distanzierung“ (Ortner, 2009, S. 2228), erwartet“ (Riebling, 2013a, S. 122).

Bildungssprache als Transfermedium für Wissen ist damit also als die Varietät einer Sprache zu verstehen, „die für Bildungskontexte [im Generellen] angemessen und in diesen funktional ist“ (Gogolin & Duarte, 2016, S. 480). Im deutlich enger gefassten Kontext eines bestimmten Fachunterrichts lässt sie sich aber gerade deshalb kaum von den dort verwendeten Fachsprachen abgrenzen (vgl. Pineker-Fischer, 2017, S. 68). Dieses Abgrenzungsproblem wird auch nicht durch einen Blick auf sprachliche Oberflächenmerkmale gelöst, die von einer Vielzahl von Autor_innen als typische Merkmale des Registers Bildungssprache beschrieben werden (z. B. Lengyel, Heintze, Reich, Roth, & Scheinhardt-Stettner, 2009, S. 132 u. f.; Tajmel, 2011; Riebling, 2013a, S. 132 u. f.; Morek & Heller, 2012, S. 71 u. f.; Gogolin & Duarte, 2016, S. 487 u. f.). Diese sind weitestgehend deckungsgleich mit sprachlichen Oberflächenmerkmalen, die auch Fachsprachen zugeordnet werden (vgl. Unterabschnitt 3.1.1.2), zumal einige der eben aufgeführten Autor_innen für ihre Ausführungen zu sprachlichen Merkmalen von Bildungssprache explizit auch Merkmalslisten aus der Fachsprachenforschung heranziehen (z. B. Tajmel, 2011; Riebling, 2013a, S. 132 u. f.). Eine der wenigen Ausnahmen bildet die in Unterabschnitt 3.1.1.2 benannte Terminologienormierung, die für wissenschaftliche Fachsprachen typisch ist, als sprachliches Merkmal von Bildungssprache jedoch eher selten (z. B. bei Tajmel, 2011) benannt wird.

Alles in allem lassen sich damit auf Grundlage des bisherigen Forschungsstandes zwischen Fachsprache, wie sie in einem bestimmten Fachunterricht verwendet wird, und der für eine bestimmte Domäne spezifischen Ausprägung von Bildungssprache[45] (vgl. Riebling, 2013a, S. 128 u. f.; Riebling, 2013b, S. 47 u. f.) keine genuinen Unterschiede erkennen. Die Begriffe „Fachsprache der Naturwissenschaft Physik, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird“ und „domänenspezifische Ausprägung von Bildungssprache im Unterrichtsfach Physik“ werden daher im weiteren Verlauf der vorliegenden Arbeit synonym verwendet. Sehr wohl besteht allerdings ein Unterschied zwischen dem fachlichen Sprachgebrauch im Physikunterricht und Bildungssprache im Allgemeinen, da das Field

[45] In Anlehnung an den Ansatz der Fachsprachenforschung, verschiedene Fachsprachen durch horizontale und vertikale Gliederung voneinander zu unterscheiden (vgl. Unterkapitel 3.1.1), lässt sich auch Bildungssprache im Allgemeinen „in mindestens zwei Gliederungsdimensionen differenzier[en.] [...] In der horizontalen Gliederung lassen ich domänenspezifische Differenzierungen der allgemeinen Bildungssprache gegenüberstellen, in der vertikalen Gliederung Differenzierung nach den Stufen der Abstraktion von der Alltagssemantik“ (Riebling, 2013a, S. 125). Dementsprechend ist eine *domänenspezifische Bildungssprache* der Sprachgebrauch, wie er in einem bestimmten Unterrichtsfach (z. B. Physik) oder einer Fächergruppe (z. B. naturwissenschaftliche Schulfächer) aufgrund der dort thematisierten Unterrichtsgegenstände angemessen und funktional ist (vgl. Schleppegrell, 2004, S. 113 u. f.; Gogolin, 2007, S. 77) (man Vergleiche dies insbesondere mit der Darstellung in Unterabschnitt 3.1.1.1).

des Registers Bildungssprache im Generellen deutlich weiter bzw. nicht fachspezifisch ist (vgl. Pineker-Fischer, 2017, S. 68).

3.1.2.2. Bildungssprache als Denkwerkzeug

Die zweite Facette des Begriffs „Bildungssprache“ ist sie metaphorisch als „Werkzeug“ für die gedankliche Beschäftigung mit Begriffen, Erkenntnissen, Erinnerungen oder Vorstellungen zu verstehen. Hierbei wird angenommen, dass eine Person, die Bildungssprache als Transfermedium für Wissen adäquat nutzen kann, auch gedanklich dazu in der Lage ist, die von ihr sprachlich realisierbaren komplexen Operationen durchzuführen (z. B. Abstraktion, Konklusion oder Generalisierung) (vgl. Morek & Heller, 2012, S. 75).

Grundlegend für dieses Verständnis von Bildungssprache ist die von Cummins im Kontext von Zweitspracherwerbsforschung vorgeschlagene Unterscheidung von Sprachhandlungen in „Basic Interpersonal Communicative Skills“ (BICS) und einer „Cognitive Academic Language Proficiency“ (CALP) (vgl. Cummins, 1979; Cummins, 1981; Cummins, 2000), wobei BICS ins Deutsche als Alltagssprache und CALP als Bildungssprache übersetzt wird (vgl. Gogolin & Lange, 2011, S. 110 u. f.; Leisen, 2017, S. 59 u. f.; Pineker-Fischer, 2017, S. 52):

- Basic Interpersonal Communicative Skills sind Sprachfähigkeiten, die vor allem in Alltagssituationen eingesetzt werden und sich durch einen vergleichsweise geringen kognitiven Aufwand auszeichnen (vgl. Cummins, 2000). Dementsprechende Sprachhandlungen lassen sich charakterisieren als „face-to-face-Interaktion“ auf Basis eines den Kommunikationsteilnehmer_innen gemeinsamen Vorwissens über den Redegegenstand, bzw. dass ein bestimmter Kontext als selbstverständlich vorausgesetzt wird (vgl. Pineker-Fischer, 2017, S. 49 u. f.). Sprachhandlungen auf Grundlage von Basic Interpersonal Communicative Skills weisen typischerweise personelle, temporale oder lokale Bezüge auf, deren Bedeutung erst durch die Kenntnis des zugehörigen Kontextes deutlich wird (vgl. ebd.).
- Mit Cognitive Academic Language Proficiency können im Gegensatz zu Basic Interpersonal Communicative Skills auch komplexe Sinnzusammenhänge oder abstrakte Thematiken vermittelt werden, ohne hierzu z. B. auf kontextbezogene Referenzen zurück greifen zu müssen (vgl. Pineker-Fischer, 2017, S. 49 u. f.). Stattdessen werden Bezüge explizit versprachlicht. Personen, die Cognitive Academic Language Proficiency erworben haben, besitzen die Fertigkeit Sachzusammenhänge unabhängig von einer konkreten „face-to-face-Interaktion“ darstellen zu können (vgl. ebd.), was allerdings eines im Vergleich zu den Basic Interpersonal Communicative Skills höheren kognitiven Aufwandes bedarf (vgl. Cummins, 2000).

Cummins geht in seiner eben angeführten Unterscheidung also davon aus, dass es nicht nur eine Verzahnung zwischen Sprachhandlungen, Kontexteinbettung und Redegegenständen gibt, sondern dass durch das Einhergehen von erhöhten Denkanforderungen mit einem

bildungssprachlichen Sprachgebrauch auch ein enger Zusammenhang zwischen Sprachhandlungen und Denkprozessen besteht[46].

Neben der von Cummins vorgeschlagenen Unterscheidung von Sprachhandlungen, gelangt auch Halliday (1993) im Rahmen seiner registertheoretischen Überlegungen zum fachlichen Sprachgebrauch im naturwissenschaftlichen Unterricht zu dem Schluss, dass ein enger Zusammenhang zwischen Sprachhandlungen und Denkprozessen besteht. Er geht sogar noch weiter und postuliert: „'learning science' is the same thing as learning the language of science" (Halliday, 1993, S. 77). Morek & Heller (2012, S. 75) schreiben diese Überlegungen von Halliday (1993) ebenfalls der Facette des Begriffs „Bildungssprache" als Denkwerkzeug zu, obwohl sich diese, wie in dem eben aufgeführten Zitat deutlich zu erkennen ist, zunächst nur auf den Gebrauch von Fachsprache im naturwissenschaftlichen Unterricht beziehen.

Diesen Überlegungen sind jedoch jenen von Härtig et al. (2015, S. 59 u. f.) und Höttecke et al. (2017, S. 56 u. f.) gegenüberzustellen[47]: Beide Autorengruppen stellen übereinstimmend fest, dass für den naturwissenschaftlichen Unterricht „die Befundlage über den Zusammenhang aus fachsprachlichen[,] [allgemein-bildungssprachlichen] und fachlich-konzeptuellen Fähigkeiten noch nicht abschließend geklärt ist" (ebd., S. 56). Aus den empirischen Untersuchungen beider Autorengruppen zeichnet sich zudem ein Trend ab, der sowohl gegen eine gänzliche Ununterscheidbarkeit von fachlich-konzeptuellem und fachsprachlichem Lernen, wie Halliday (1993) annimmt, spricht, als auch eine eins-zu-eins-Übertragung von fachsprachlichen auf allgemein-bildungssprachliche Fähigkeiten fragwürdig erscheinen lässt, wie Morek & Heller (2012) sie vornehmen[47]: Insbesondere Kennwerte aus empirischen Modellrechnungen beider Autorengruppen weisen eher darauf hin, „dass sich sprachliche Fähigkeiten im Fachunterricht der Naturwissenschaften von unterrichtssprachlichen [allgemein-bildungssprachliche Fähigkeiten; M. S. F.] und rein fachlichen Fähigkeiten unterscheiden lassen, gleichzeitig müssen alle drei in einer gegenseitigen, sehr engen Beziehung stehen" (Härtig et al., 2015, S. 60).

3.1.2.3. Bildungssprache als Eintritts- und Visitenkarte

Die dritte Facette des Begriffs „Bildungssprache" betrifft nach Morek & Heller (2012, S. 76 u. f.) seine Bedeutung als Manifestation von Zwecken der Institution Schule (vgl. Fend, 1980, S. 13 u. f.; Jung, 1983, S. 26 u. f.) und damit insbesondere der mit diesen Zwecken im Zusammenhang stehenden sozialen und Repräsentationsfunktionen schulischer Leistungsfeststellungen und -beurteilungen (vgl. Unterkapitel 1.2). Dabei wird vor allem auf die Überlegungen von Bourdieu zurückgegriffen, gemäß denen...

[46] Eine Vielzahl von empirischen Studien untermauern diese Annahme. Für eine allgemeine Zusammenschau siehe z. B. Eckhardt (2008, S. 52 u. f.).

[47] Zu analogen Überlegungen und empirischen Befunden gelangen zudem Heitmann, Hecht, Schwanewedel, & Schipolowski (2014) in ihrer Untersuchung zu Gemeinsamkeiten und Unterschieden von Schülerfähigkeiten speziell zum argumentativen Schreiben im naturwissenschaftlichen Fachunterricht und im Unterricht in der Erstsprache.

> „[...] [d]er eigentliche soziale Wert der [...] Verwendung von Sprache [...] in ihrer Tendenz [liegt], Systeme von Unterschieden zu bilden, die das System der sozialen Unterschiede [...] widerspiegeln. [...] Diese Sprachstile, Systeme klassifizierter und klassifizierender, hierarchisch geordneter und hierarchisch ordnender Unterschiede, prägen diejenigen, die sie sich aneignen[.]“ (Bourdieu, 1990, S. 31-32)

Sprachgebrauch stellt gemäß Bourdieu (1990) also eine besondere Form des kulturellen Kapitals dar und ist im Individuum als „sprachlicher Habitus“ manifestiert (vgl. ebd.; Bourdieu, 1993, S. 91 u. f.). Da ferner das soziale Feld bestimmt, welche Sprache als „legitim“ gilt, ist Sprachgebrauch damit insbesondere sowohl ein Mittel sozialer Distinktion, als auch ein sozial-symbolisches Instrument, das der Identitätsstiftung dient (vgl. Bourdieu, 1990; Bourdieu, 1993, S. 91 u. f.). Unter „legitimer Sprache“ versteht Bourdieu dementsprechend folgendes:

> „Eine legitime Sprache ist eine Sprache[,] [...] die den üblichen Kriterien der Grammatikalität entspricht und neben dem, was sie sagt, ständig auch noch sagt, daß sie es gut sagt. Und dadurch glauben macht, daß das, was sie sagt, wahr ist[.]“ (Bourdieu, 1993, S. 100)

Bildungssprache ist dabei die „legitime Sprache“ im sozialen Feld institutionalisierter Bildung (vgl. Gogolin & Duarte, 2016, S. 481). Das Besondere ist nun, dass sowohl im Fach- wie auch im Sprachunterricht Schüler_innen nur selten explizit vermittelt wird welcher Sprachgebrauch als „legitim“ gilt (vgl. Steinmüller & Scharnhorst, 1987, S. 10; Schleppegrell, 2001, S. 433 u. f.). Vielmehr wird von schulisch erfolgreichen Schüler_innen selbstverständlich erwartet, dass sie diesen im Unterricht als adäquat geltenden Gebrauch von Sprache beherrschen (vgl. Pöhlmann-Lang, 2015, S. 106). Bildungssprache wird daher auch als Teil des „heimlichen Lehrplans“ verstanden (z. B. Schleppegrell, 2004, S. 2; Vollmer & Thürmann, 2010, S. 109), der „die *lautlosen Mechanismen* der Einübung in die Regeln und Rituale der Institution [Schule umfasst]“ (Meyer, 2016, S. 65, Hervorhebungen im Original). Sie ist damit also nicht nur ein Transfermedium für Wissen oder ein Denkwerkzeug (vgl. Unterabschnitt 3.1.2.1 und 3.1.2.2), sondern erfüllt (im Kontext von schulischer Leistungsfeststellung und -beurteilung) auch ungleichheitsreprodzuierende und identitätsstiftende Funktionen (vgl. Unterkapitel 1.2; Morek & Heller, 2012, S. 76 u. f.).

Diese Facette von Bildungssprache, die Morek & Heller (2012, S. 77) mit dem Begriffspaar „Eintritts- und Visitenkarte“ bezeichnen, lässt sich für den naturwissenschaftlichen Unterricht an den folgenden drei Transkriptausschnitten[48] aus der Untersuchung von Harren über sprachliche Anforderung an Schüler_innen in Unterrichtsgesprächen im Biologieunterricht veranschaulichen (vgl. Harren, 2011; Harren, 2015, S. 128 u. f.):

[48] Für die von der Autorin verwendeten Transkriptionsregeln siehe Harren (2015, S. 123 u. f.). Bei den Namen der Lehrer_innen und Schüler_innen handelt es sich um Pseudonyme (vgl. ebd.).

Transkriptausschnitt 1
(Jahrgangsstufe 7, Thema: Photosynthese; Harren, 2015, S. 132):

```
Frau Witt: fällt dir auch noch EIN? (- -)
           was für_n verSUCH wir mit dieser GLAS(.)glocke gemacht_ham?=

Fabian:    =ähh ja_wir_hAm die über eine brEnnende KERze
           geTAN-
           u[:n:

Timo:       [geSTÜLPT;

Frau Witt: [geSTÜLPT;

Fabian:    [geSTÜLPT;

Frau Witt: <<all>ja?>

Fabian:    und nach einiger ZEIT-
           is die kerze dann AUSgegangen;
```

Transkriptausschnitt 2
(Jahrgangsstufe 11, Thema: Enzymregulation; Harren, 2015, S. 142):

```
Herr Schäfer: das heißt- (-)
              wenn sie noch eben SAgen- (-)
              L:INKS von der membran,=
              =was ist DORT?
              und RECHTS von der membran- (.)
              was ist DORT.
              (.)

Timo:         =LINKS is halt- (-) äh
              ja fang_we mal mit RECHTS an-
              rechts is halt das INnere der zelle, =

Herr Schäfer: =ja.

Timo:         und LINKS is halt-
              äh sozuSAgn- (-)
              AUSSNwelt.
              (-)

Herr Schäfer: INterzellularraum.
              kann man SAgen.
              okee.
```

Transkriptausschnitt 3
(Jahrgangsstufe 12, Thema: Ökologie; Harren, 2015, S. 140):

```
Frau Witt: schmaROTzer.=
           =wie nenn wir die auf SCHLAU?
           (-)

Inge:      paraSItn?=

Frau Witt: =geNAU.
```

In Transkriptausschnitt 1 fällt auf, dass außer Frage steht, was Fabian mit dem Verb „geTAN-“ ausdrücken möchte. Trotzdem wird er von einem Mitschüler, sowie von seiner Lehrerin verbessert. Die Korrektur, stattdessen das Verb „geSTÜLPT“ zu verwenden, kann daher als Aufforderung gegenüber Fabian gelesen werden, die im Biologieunterricht „legitime Sprache“ zu verwenden, die sich „durch ein Beharren auf präzisere Wortwahl [auszeichnet]“ (Harren, 2011, S. 108), auch dann wenn die Kommunikationsteilnehmer_innen auf ein gemeinsames Vorwissens über den Redegegenstand zurückgreifen können und man deshalb auch bei einer weniger präzisen Wortwahl verstanden werden würde. Dass diese Korrekturaufforderung nicht nur von Frau Witt, sondern auch von einem Mitschüler unternommen wird und dass Fabian dieser Aufforderung auch entgegenkommt, untermauert dabei, „dass eine solche Orientierung an einer adäquaten Wortwahl nicht nur Anliegen der Lehrerin“ (ebd.), sondern aller am Unterrichtsgespräch beteiligten Personen ist.

In Transkriptausschnitt 2 macht der Schüler Timo in seiner Aussage durch das Adverb „sozuSAgn-“ deutlich, dass er bemüht ist die „legitime“ Sprache des Biologieunterrichts zu verwenden bzw. dass er das Nomen „AUSSNwelt“ für eine hier nicht adäquate Benennung hält, auch wenn hier ähnlich wie im Fall von Fabian verständlich ist, was er mit „AUSSNwelt“ zum Ausdruck bringen möchte (vgl. ebd., S. 109). Die Antwort seines Lehrers, die mit dem eingliedrigen Satz „okee.“ endet, kann daher nicht nur als Verbesserung von Fabians Wortwahl interpretiert werden, sondern auch als ein Gutheißen und Eingehen auf Fabians Anliegen sich in der „legitimen“ Sprache des Biologieunterrichts ausdrücken zu wollen.

In Transkriptausschnitt 3 erfragt die Lehrkraft Frau Witt eine fachliche Benennung für einen „schmaROTzer“, nachdem im vorherigen Unterrichtsgeschehen bereits inhaltlich geklärt wurde, was in der Ökologie hierunter verstanden werden kann (vgl. Harren, 2011, S. 112). Indem sie den Gebrauch von Fachlexik als „SCHLAU“ bezeichnet und die fragend betonte Antwort der Schülerin Inge mit „geNAU“ kommentiert, vermittelt sie ihren Schüler_innen daher auch, „dass das Verwenden von Fachbegriffen erwünscht ist und dass [...] Sprecher/innen sich durch den Gebrauch dieser Vokabeln als gebildet ausweisen können“ (Harren, 2011, S. 112-113).

Was sich also insgesamt in allen drei Transkriptausschnitten erkennen lässt, ist sowohl eine Orientierung des von Lehrkräften angestrebten Sprachgebrauchs ihrer Schüler_innen an der „legitimen Sprache“ des Biologieunterrichts (Einforderung von Bildungssprache als Eintrittskarte, um als erfolgreicher_erfolgreiche Lerner_Lernerin im Biologieunterricht gelten zu dürfen) als auch eine besondere Ausrichtung der von den Schüler_innen realisierten Sprache, die als ein Bemühen um eine dem „legitimen Sprachgebrauch“ entsprechende Ausdrucksweise interpretiert werden kann (Verwendung von Bildungssprache als Visitenkarte, um sich als ein_eine erfolgreicher_erfolgreiche Lerner_Lernerin im Biologieunterricht auszuweisen).

3.1.2.4. Zwischenfazit

Die in vorangegangenen Unterabschnitten erörterten drei Facetten des Begriffs „Bildungssprache“ lassen sich wie folgt zusammenfassen:

1. Bildungssprache ist die in Bildungskontexten im Generellen angemessene und funktionale Varietät einer Sprache, die sich durch die registertheoretischen Dimensionen „Field“, „Mode“ und „Style of Discourse“ beschreiben lässt und durch typische sprachliche Oberflächenmerkmale gekennzeichnet ist (vgl. Unterabschnitt 3.1.2.1). Die Fachsprache der Naturwissenschaft Physik, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird, kann kaum von einer für den Physikunterricht domänenspezifische Ausprägung von Bildungssprache unterschieden werden, weswegen im weiteren Verlauf der vorliegenden Arbeit beide Begriffe daher gleichbedeutend verwendet werden. Diese domänenspezifische Ausprägung von Bildungssprache bis zu einem bestimmten Grad aktiv und passiv zu beherrschen, stellt für Schüler_innen (in Leistungssituationen) im Physikunterricht eine sprachliche Anforderung dar.
2. Bildungssprache im Allgemeinen ist verschieden von ihrer domänenspezifischen Ausprägung im Physikunterricht. Beide Varietäten stehen miteinander aber in einer wechselseitigen und engen Beziehung. Ein solch enger Zusammenhang besteht auch zwischen bildungssprachlichen Sprachhandlungen und komplexen Denkprozessen (vgl. Unterabschnitt 3.1.2.2). Es ist daher gerechtfertigt anzunehmen, dass sich Anforderungen an Schüler_innen im Physikunterricht, die fachsprachliche bzw. domänenspezifisch-bildungssprachliche Fähigkeiten adressieren, von jenen, die fachlich-konzeptuelle Fähigkeiten ansprechen, unterscheiden. Andererseits müssen beide Anforderungsarten in einer sehr engen Beziehung zueinander stehen.
3. Ferner kann (domänenspezifische) Bildungssprache aus soziologischer Perspektive als die „legitime Sprache“ im (Fach-)Unterricht verstanden werden (vgl. Unterabschnitt 3.1.2.3). Sie ist (in ihrer domänenspezifischen Ausprägung) Teil des „heimlichen Lernplans“ und wird selten explizit vermittelt, sondern oftmals als ein selbstverständliches Merkmal von schulisch erfolgreichen Schüler_innen vorausgesetzt. Bildungssprache ist damit auch ein soziales Distinktionsmittel und ein sprecher- bzw. schreiberseitiges Instrument sozial-symbolischer Identifikation. Eine Anforderung an Schüler_innen im Physikunterricht besteht also auch darin, dass ihr „sprachlicher Habitus“ mit der „legitimen Sprache“ des Physikunterrichts im Einklang stehen muss, um sich als erfolgreicher_erfolgreiche Lerner_Lernerin im Physikunterricht ausweisen zu können bzw. als solche_r erkannt zu werden (vgl. Tajmel, 2017b, S. 251 u. f.).

Für den weiteren Verlauf der vorliegenden Arbeit sind damit Sprach- und Sprachgebrauchsnormen, mit denen Schüler_innen (in Leistungssituationen) im Physikunterricht konfrontiert werden, hinreichend aufgearbeitet. Das nun anschließende Unterkapitel widmet sich einer Bestandsaufnahme über die bisherige Forschung zum Umgang von Physik-

lehrer_innen mit Sprache im Rahmen von schulischer Leistungsfeststellung und -beurteilung.

3.2. Erkenntnisstand zum Umgang von Physiklehrer_innen mit Sprache im Rahmen von schulischer Leistungsfeststellung und -beurteilung

In diesem Unterkapitel wird erörtert, welche Befunde es bislang zum Umgang von Physiklehrer_innen mit Sprache in Leistungssituationen im Physikunterricht gibt. Es wird sich also explizit der Frage gewidmet, welche Erkenntnisse die bisherige erziehungswissenschaftliche Forschung über die Assessment Literacy von Lehrkräften (vgl. Unterkapitel 2.2) bezogen auf Sprache als Lernmedium und -gegenstand im Physikunterricht liefert.

Bisherige Studien aus dem internationalen Raum, die den Umgang von Lehrer_innen mit Sprache im naturwissenschaftlichen Fachunterricht untersuchten, konnten aufdecken, dass...

... Lehrkräfte Sprache oftmals nicht als Lerngegenstand des Physikunterrichts betrachten (z. B. Airey, 2012),

... Lehrkräften das (mögliche) Vorhandensein eines „cultural and language bias“ für sprachlich-kulturell heterogene Lerngruppen im Physikunterricht oft nicht bewusst ist (z. B. Luykx, Lee, Hart, & Deaktor, 2007) und

... Naturwissenschaftslehrkräfte beim Unterrichten sprachlich-kulturell heterogener Lerngruppen häufiger die sprachlich-kulturell bedingten Defizite von Schüler_innen im Blick haben, anstatt deren Fähigkeiten (z. B. Buxton, Salinas, Mahotiere, Lee, & Secada, 2013).

Die Vermutung ist daher naheliegend, dass Physiklehrkräfte im Rahmen von schulischer Leistungsfeststellung und -beurteilung sprachliche Anforderung an Schüler_innen und/oder sprachliche Leistungen von Schüler_innen weitgehend ausblenden, weil sie sich diesbezüglich als nicht zuständig fühlen oder aber dass Physiklehrkräfte den Sprachgebrauch ihrer Schüler_innen eher defizit- als fähigkeitsorientiert feststellen und -beurteilen.

Eine in Teilen ähnliche Vermutung lässt sich auch aus der Untersuchung von Riebling (2013b) zur Sprachbildung im naturwissenschaftlichen Unterricht ableiten: Riebling (2013b) führte im Jahr 2010 eine Selbstauskunftsbefragung mit insgesamt N = 229 Hamburger Naturwissenschaftslehrkräften durch, aus der sie mit Hilfe einer Clusteranalyse fünf empirisch und inhaltlich voneinander verschiedene Handlungstypen von Lehrkräften bezüglich ihres Umgangs mit sprachlicher Heterogenität im naturwissenschaftlichen Fachunterricht identifizieren konnte (vgl. ebd., S. 163 u. f.). Da diese Typen, deren Kurzcharakteristika in Tabelle 3.1 dargestellt sind, auf globaler Ebene das Handeln von Naturwissenschaftslehrkräften in sprachlich heterogenen Lerngruppen beschreiben, ist es plausibel

Typus	Unterstützung der Schüler_innen, um sprachlichen Fähigkeiten weiterzuentwickeln.	Art der sprachlichen Anforderungen an Schüler_innen im naturwissenschaftlichen Fachunterricht.
A	Unterstützung der Schüler_innen findet statt.	Hohe sprachliche Anforderungen an Schüler_innen. Sprachliche Entlastung der Schüler_innen findet kaum statt.
B1	Unterstützung der Schüler_innen findet statt.	Kaum sprachliche Anforderungen an Schüler_innen. Starke sprachliche Entlastung der Schüler_innen im Fachunterricht.
B2	Unterstützung der Schüler_innen findet statt.	Geringe sprachliche Anforderungen an Schüler_innen. Sprachliche Entlastung der Schüler_innen findet statt (deutlich weniger als bei Typ B1).
C	Kaum Unterstützung der Schüler_innen.	Geringe sprachliche Anforderungen an Schüler_innen. Entlastung der Schüler_innen findet statt (deutlich weniger als bei Typ B1).
D	Kaum Unterstützung der Schüler_innen.	Hohe sprachliche Anforderungen an Schüler_innen. Sprachliche Entlastung der Schüler_innen findet kaum statt.

Tabelle 3.1.: Empirische Handlungstypen bezüglich des Umgangs mit sprachlicher Heterogenität im naturwissenschaftlichen Fachunterricht nach Riebling (2013b, S. 163 u. f.).

anzunehmen, dass sich diese Handlungstypen auch im Umgang von Lehrer_innen mit Sprache im Rahmen von Leistungsfeststellung und -beurteilung im Physikunterricht zeigen. Dementsprechend ist damit zu rechnen, dass es sowohl Physiklehrkräfte gibt, die in Leistungssituationen hohe sprachliche Anforderungen an ihre Schüler_innen stellen (Typ A und D), als auch solche Lehrkräfte, bei denen sich derartige Anforderungen nur kaum finden lassen (Typ B1, B2 und C). Zusätzlich kann angenommen werden, dass es zum einen Physiklehrer_innen gibt, die im Rahmen von schulischen Leistungsfeststellungen und -beurteilungen Maßnahmen zur Unterstützung der sprachlichen Fähigkeiten ihrer Schüler_innen ergreifen (Typ A, B1 und B2), zum anderen aber auch solche, bei denen eine Unterstützung sprachlicher Fähigkeiten kaum stattfindet (Typ C und D).

Alle eben genannten Mutmaßungen über den Umgang von Physiklehrkräften mit Sprache im Rahmen von schulischer Leistungsfeststellung und -beurteilung müssen allerdings als hochgradig spekulativ gelten. Tatsächlich wurde sich der Frage, wie Physiklehrer_innen mit Sprache in Rahmen von schulischer Leistungsfeststellung und -beurteilung tatsächlich umgehen, bislang kaum gewidmet.

Zunächst ist hierbei die Explorationsstudie von Thieme & Mavruk (2018) zu nennen. In dieser wurden Gruppendiskussionen mit insgesamt N = 14 Biologielehrkräfte durchgeführt, um deren sprachbezogenen Kriterien zur Bewertung der Textprodukte von Schüler_innen im Fach Biologie identifizieren zu können (vgl. ebd., S. 292). Der zentrale Befund, der sich im Rahmen dieser Studie zeigte ist, dass die befragten Biologielehrkräfte

Textprodukte eines_einer Schülers_Schülerin danach bewerten, inwieweit „eine erwartete Angemessenheit erfüllt wird, die zwischen den beiden Polen »Ausführlichkeit« und »Knappheit« angesiedelt ist“ (ebd., S. 295). Beispielsweise wurde von den teilnehmenden Biologielehrkräften erwartet, dass Fachbegriffe von den Schüler_innen zu erläutern sind, gleichzeitig dürfen diese Erläuterungen aber nicht zu ausschweifend sein (vgl. ebd.).

Des Weiteren zeigte sich in der ethnographischen Untersuchung von Willems (2007) zum Prozess der Vergeschlechtlichung bei der Konstruktion schulischer Fachkulturen in den Fächern Physik und Deutsch ein bemerkenswerter Befund. Die Autorin stellt im Kontext von bilingualem Physikunterricht (mit Englisch als Unterrichtssprache)...

> „[...] aus Position der Lehrenden eine Ambivalenz der Bewertungspraxis [fest]: Auf der einen Seite formulieren die Lehrkräfte [...], dass die sprachliche Ebene nur dann in die Bewertungen von schriftlichen Tests eingeht, wenn der physikalische Inhalt dadurch falsch oder unvollständig wird. [...] [Gleichzeitig] lässt sich konstatieren, dass den Lernenden im bilingualen Physikunterricht nicht nur der sprachliche Bereich zur eigenen Einschätzung überlassen wird, sondern dieser auch weitgehend aus den Bewertungsrastern herausfällt.“ (ebd., S. 194-196).

Daneben finden sich zum gegenwärtigen Zeitpunkt zwei weitere (ebenfalls explorative) Studien, die sich allerdings mit dem Umgang von Physiklehrkräften mit Sprache im Rahmen von schulischer Leistungsfeststellung und -beurteilung tiefgreifender auseinandersetzten: die Untersuchung von Lyon zur Entwicklung der Expertise angehender Naturwissenschaftslehrkräften bezüglich des Umgangs mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen (vgl. Lyon, 2013a; Lyon, 2013b; Lyon, 2013c; Lyon, 2013d), sowie das Fallbeispiel von Tajmel zu Lehrerleistungsurteilen über die Originaltexte eines_einer Schülers_Schülerin mit Migrationshintergrund zum Thema Schwimmen und Sinken (vgl. Tajmel, 2010, S. 172 u. f.; Tajmel, 2011, S. 4 u. f.; Tajmel, 2017b, S. 251 u. f.). Beide Untersuchungen werden im Folgenden detailliert beschrieben und schließlich in Abschnitt 3.2.3 deren Quintessenz erörtert.

3.2.1. Lyons Untersuchung zur Entwicklung der Expertise angehender Naturwissenschaftslehrkräften im Umgang mit sprachlich-kultureller Heterogenität bei schulischer Leistungsfeststellung und -beurteilung

Das Erkenntnisinteresse der Untersuchung von Lyon lag darin zu explorieren, inwiefern bei angehenden Naturwissenschaftslehrkräften eine Entwicklung ihrer Expertise bezüglich schulischer Leistungsfeststellung und -beurteilung als „Assessment for Learning“ (vgl. Abschnitt 1.1.2) im Verlauf ihrer universitären Lehramtsausbildung stattfindet (vgl. Lyon, 2013c, S. 447). Ein Schwerpunkt lag hierbei auf der Untersuchung des Umgangs angehender Lehrkräfte mit sprachlich-kultureller Heterogenität in Leistungssituationen im naturwissenschaftlichen Fachunterricht (vgl. Lyon, 2013d).

Hierzu wurden N = 11 angehende Naturwissenschaftslehrkräfte während ihres 12 monatigen Lehramtsmasterstudiums wissenschaftlich begleitet (vgl. Lyon, 2013c, S. 447). Alle teilnehmenden Studenten_Studentinnen absolvierten dabei ihr Studium an einer kalifornischen Universität, in deren Lehramtscurriculum die Themen schulische Leistungsfeststellung und -beurteilung, sowie Umgang mit sprachlich-kultureller Heterogenität besonders berücksichtigt sind (vgl. ebd.).

Zwecks der Datengewinnung wurde von jedem_jeder teilnehmenden Studenten_Studentin an insgesamt drei Messzeitpunkten (zu Beginn, im Verlauf und am Ende ihres Masterstudiums) ein reichhaltiger Fundus verschiedenartigster Daten erhoben (vgl. Lyon, 2013c, S. 448 u. f.; Lyon, 2013d, S. 4 u. f.). Anschließend wurden die gewonnenen Daten im Sinne eines Mixed-Methods-Triangulationsdesigns ausgewertet (vgl. Creswell & Plano Clark, 2007, S. 62 u. f.), das heißt Teile der erhobenen Daten wurden mit Hilfe quantitativer Methoden ausgewertet, wohingegen Andere qualitativen Analysen unterzogen wurden, um so ein umfassenderes Bild des Forschungsgegenstandes zu erhalten.

3.2.1.1. Quantitative Datenanalyse und zentrale Befunde

Für die quantitative Analyse wurden die folgenden Produkte der teilnehmenden Studierenden als Datengrundlage herangezogen:

1. **Fiktive Assessmentpläne**: Drei schriftliche Planungen von Leistungsfeststellungen und -beurteilungen im Rahmen einer fiktiven Unterrichtseinheit, die von jedem_jeder Teilnehmer_in an jedem Messzeitpunkt einmal erhoben wurde (vgl. Lyon, 2013c, S. 448 u. f.).
2. **Fiktive Assessmentkritiken**: Drei schriftlich auszuarbeitende Kritiken an dem Vorgehen einer fiktiven Naturwissenschaftslehrerin in einer schulischen Leistungsfeststellungs- und -beurteilungsituation, das den Teilnehmer_innen in einer Textvignette beschrieben wurde (von jedem_jeder Teilnehmer_in an jedem Messzeitpunkte einmal erhoben) (vgl. ebd.).
3. **Unterrichtsplanung**: Die schriftliche Ausarbeitung der Planung eines eigenverantwortlich durchgeführten Unterrichts, die jeder_jede Teilnehmer_in als Teil der Abschlussprüfung seines_ihres Studiums anfertigen musste (vgl. ebd.).

Ausgangspunkt der quantitativen Analyse war ein eigenes für diese Studie entwickeltes Modell der Assessmentexpertise von angehenden (Naturwissenschafts-)Lehrkräften, das unter anderem auf Grundlage der bisherigen speziell naturwissenschaftsdidaktischen Überlegungen zur Assessment Literacy von Lehrkräften (vgl. Abschnitt 2.2.2) entwickelt wurde (vgl. Lyon, 2013b). Grob umschrieben umfasst dieses Modell die drei Dimensionen *Planung* und *Nutzung* von schulischen Leistungsfeststellungen und -beurteilungen im naturwissenschaftlichen Fachunterricht, sowie den *Umgang mit sprachlich-kultureller Heterogenität* bei schulischen Leistungsfeststellungen und -beurteilungen, wobei für jede dieser Dimensionen vier, aus der Literatur abgeleitete Entwicklungsstufen operationalisiert wurden (vgl. Lyon, 2013c, S. 445 u. f.).

	Berücksichtigung von Fairness	**Berücksichtigung von Zugänglichkeit**
Level 1 (Score=1)	Berücksichtigt nicht den Einfluss von sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen.	Berücksichtigt nicht die Partizipationsmöglichkeiten sprachlich-kulturell heterogener Lerngruppen bei schulischen Leistungsfeststellungen und -beurteilungen.
Level 2 (Score=2)	Berücksichtigt Faktoren, die einen Einfluss auf schulische Leistungsfeststellung und -beurteilung haben. Z. B. den Einfluss (a) des soziodemographischen Hintergrunds der Schüler_innen, (b) des sprachlich-kulturellen Hintergrunds der Schüler_innen, (c) der Durchführungsform einer Leistungsfeststellung und -beurteilung oder (d) der Einbettung einer Leistungsfeststellung und -beurteilung in eine Lehr-Lern-Sequenz.	Berücksichtigt (explizit) wie Leistungsfeststellungen und -beurteilungen als Assessment for Learning sprachlicher Anforderungen im naturwissenschaftlichen Unterricht gestaltet werden können (z. B. formatives Assessment des fachlichen Sprachgebrauchs von Schüler_innen), ohne besondere Bedürfnisse sprachlich-kulturell heterogener Lerngruppen mit zu bedenken.
Level 3 (Score=3)	Berücksichtigt mindestens eine Strategie zum Umgang mit Einflüssen von sprachlich-kultureller Heterogenität bei schulischer Leistungsfeststellung und -beurteilung (z. B. sprachliche Vereinfachung, Scaffolding, Binnendifferenzierung der Leistungsfeststellung und -beurteilung).	Berücksichtigt (explizit) wie Leistungsfeststellungen und -beurteilungen als Assessment for Learning sprachlicher Anforderungen im naturwissenschaftlichen Unterricht gestaltet werden können (z. B. formatives Assessment des fachlichen Sprachgebrauchs von Schüler_innen), wobei besondere Bedürfnisse sprachlich-kulturell heterogener Lerngruppen ebenfalls bedacht sind.
Level 4 (Score=4)	Berücksichtigt mindestens eine Strategie zum Umgang mit sprachlich-kultureller Heterogenität als Ressource bei schulischer Leistungsfeststellung und -beurteilung (z. B. Schüler_innen mit sprachlich-kulturell heterogenem Hintergrund gewähren sich in der Sprache auszudrücken, in der sie sich am besten verständlich machen können).	Berücksichtigt (explizit) wie Leistungsfeststellungen und -beurteilungen als Assessment for Learning sprachlicher Anforderungen im naturwissenschaftlichen Unterricht gestaltet werden können (z. B. formatives Assessment des fachlichen Sprachgebrauchs von Schüler_innen), wobei besondere Bedürfnisse sprachlich-kulturell heterogener Lerngruppen ebenfalls bedacht sind. Zusätzlich wird auch (explizit) berücksichtigt wie Leistungsfeedbacks zu gestalten sind, angepasst an die Bedürfnisse sprachlich-kulturell heterogener Lerngruppen.

Tabelle 3.2.: Sinngemäße Übersetzung des Rasters von Lyon (2013b) zur Operationalisierung der Expertise angehender Naturwissenschaftslehrkräfte im Umgang mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen.

Für die quantitative Datenanalyse diente dieses Modell als provisorische Grundlage, um in einem deduktiv-induktiven Vorgehen ein Kriterienraster zu entwickeln, mit dessen Hilfe die von jedem_jeder Teilnehmer_in angefertigten und zur Analyse herangezogenen Produkte einem holistischen Rating unterzogen werden konnten (vgl. Lyon, 2013c, S. 449). Die finale Version dieses Kriterienrasters[49] besteht dabei aus den drei Dimensionen des

[49]Für eine vollständige und ausführliche Beschreibung der finalen Version des Kriterienrasters siehe (Lyon, 2013b).

	Messzeitpunkt 1	Messzeitpunkt 2	Messzeitpunkt 3
Fiktive Assessmentpläne	3.09	3.30	3.64
Fiktive Assessmentkritiken	4.36	4.80	5.00
Unterrichtsplanung	- - -	- - -	5.46

Tabelle 3.3.: Arithmetische Mittel des Umgangs von angehenden Naturwissenschaftslehrkräften mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen (vgl. Lyon, 2013c, S. 453 u. f.).

theoretischen Ausgangsmodells, die jeweils in zwei Subdimensionen aufgefächert wurden (vgl. Lyon, 2013c, S. 449 u. f.). Jede dieser Subdimensionen wurde wiederum in vier Ausprägungen untergliedert, denen jeweils ein bestimmter Punktewert (Score) zugewiesen wurde (vgl. ebd.). Die Dimension „Umgang mit sprachlich-kultureller Heterogentität bei schulischen Leistungsfeststellungen und -beurteilungen“ des Ausgangsmodell wurde hierbei in die beiden Subdimensionen *Berücksichtigung von Fairness* und *Zugänglichkeit* schulischer Leistungsfeststellung und -beurteilung im Kontext sprachlich-kulturell Heterogentität untergliedert (siehe Tabelle 3.2). Bei der Anwendung des Kriterienrasters wurde jedem Produkt eines_einer teilnehmenden Studenten_Studentin zunächst eine Ausprägung für jede Subdimension zugeordnet und hierdurch mit insgesamt 6 Punktewerten versehen (vgl. ebd.). Anschließend wurde für jede der drei Dimensionen *Planung*, *Nutzung* und *Umgang mit sprachlich-kultureller Heterogenität* ein Gesamturteil gebildet, indem die Punktewerte der beiden zugehörigen Subdimensionen aufsummiert wurden (vgl. ebd.).

Die Einschätzung der Produkte der teilnehmenden Studenten_Studentinnen erfolgte durch insgesamt 3 geschulte Rater_innen, die in einem ersten Schritt unabhängig voneinander das gesamte Datenmaterial mit Hilfe des entwickelten Kriterienrasters auswerteten (vgl. Lyon, 2013c, S. 449). Da die Übereinstimmung der 3 Rater_innen allerdings vergleichsweise gering ausfiel ($\bar{\kappa} = .249$), wurden in einer anschließenden Diskussion Diskrepanzen besprochen, um so zu einem Konsens über die Zuordnung der Produkte der Teilnehmer_innen auf den Ausprägungen des Kriterienrasters zu gelangen (vgl. ebd.).

Im Anschluss an das eben beschriebene Rating wurden die durch dieses Vorgehen gewonnenen Summenscores für die fiktiven Assessmentpläne, Assessmentkritiken und die Unterrichtsplanungen der Teilnehmer_innen entlang der drei Messzeitpunkte deskriptiv analysiert (vgl. Lyon, 2013c, S. 449 u. f.). Dabei zeigten sich in den Mittelwerten der Summenscores für den Umgang mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen drei Muster (siehe Tabelle 3.3):

- Erstens deutete sich an, dass die Expertise der Teilnehmer_innen im Umgang mit sprachlich-kultureller Heterogenität bei schulischer Leistungsfeststellung und -beurteilung insgesamt moderat ausgeprägt war, da sich die arithmetischen Mittel der Summenscores eher in der gedachten Mitte zwischen dem minimal und maximal möglichen Werten (2.00 bis 8.00) bewegten (vgl. Lyon, 2013c, S. 449 u. f.; Lyon, 2013a, S. 285).

- Zweitens zeigte sich im zeitlichen Verlauf mutmaßlich zwar eine Zunahme der Expertise der Teilnehmer_innen im Umgang mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen, die Unterschiede zwischen den drei Messzeitpunkten sind allerdings eher gering und zudem (möglicherweise aufgrund der geringen Stichprobengröße) nicht signifikant (vgl. Lyon, 2013c, S. 449 u. f.).
- Drittens wurde aus einem direkten Vergleich der mittleren Summenscores für die fiktiven Assessmentpläne mit jenen der Assessmentkritiken deutlich, dass letztere durchgängig höher ausgeprägt waren als erstere. Dies lässt vermuten, dass die Teilnehmer_innen eine höhere Expertise darin aufweisen (bzw. erworben haben) den Umgang mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen zu reflektieren, als selbst zu praktizieren (vgl. ebd.).

3.2.1.2. Qualitative Datenanalyse und zentrale Befunde

Neben den aus der quantitativen Datenanalyse hervorgegangenen mittleren Tendenzen in der Ausprägung und der Entwicklung des Umgangs der Teilnehmer_innen mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen, wurde unter anderem durch einen genaueren Blick in die Punktevergabe bezüglich der einzelnen Subdimensionen zu den verschiedenen Messzeitpunkten deutlich, dass sich die Ausprägung und Entwicklung der Expertise der teilnehmenden angehenden Naturwissenschaftslehrkräfte zum Teil erheblich voneinander unterscheiden (vgl. Lyon, 2013c, S. 456 u. f.). Laut Lyon (2013d, S. 4) zeigt sich dabei, dass es sich bei den drei Teilnehmer_innen[50] Dean, Glenda und Lauren um besonders auffällige und stark unterschiedliche Fälle im Gesamtsamples handelte[51]. Lyon (2013d) unterzog daher die Daten, die er von diesen drei Studierenden erhoben hatte, einer zusätzlichen qualitativen Analyse, um hierdurch zu einem vertieften Einblick in die Entwicklung der Expertise dieser drei angehenden Naturwissenschaftslehrkräfte bezüglich des Umgangs mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen zu gelangen und insbesondere um weitere fallübergreifende Muster identifizieren zu können, die aus der quantitativen Analyse des Datenmaterials nicht hervorgingen.

Die qualitative Analyse der gesamten von Dean, Glenda und Lauren erhobenen Daten[52] erfolgte mit Hilfe der Grounded Theory (vgl. Corbin & Strauss, 1990). Dabei wurde in einem ersten Codierzyklus[53] das gesamte Material zunächst thematisch codiert (vgl.

[50] Bei den Namen der Studierenden handelt es sich um Pseudonyme (vgl. Lyon, 2013c, S. 447).

[51] Lyon (2013d, S. 4) berichtet lediglich, dass er zu diesem Urteil aufgrund einer Reanalyse der von den teilnehmenden Studierenden erhobenen Produkten gelangt ist, ohne diese jedoch genauer zu beschreiben.

[52] Neben den Daten, die bereits für die quantitative Datenanalyse verwendet wurden, umfasste der gesamte Datensatz mehrere leitfadengestützte Interviews, Beobachtungsprotokolle aus Unterrichtshospitationen und schriftliche Kurshausaufgaben der Teilnehmer_innen aus ihrem Studium (vgl. Lyon, 2013c, S. 448 u. f.; Lyon, 2013d, S. 4 u. f.).

[53] Für eine genaue Beschreibung des Codiervorgangs anhand von Beispielmaterial siehe Lyon (2013c, S. 450), sowie Lyon (2013a, S. 288 u. f.).

Saldaña, 2016, S. 102 u. f.). Daraufhin wurde eine zweite, longitudinale Codierung[53] vorgenommen (vgl. Saldaña, 2016, S. 260 u. f.), mit dem Ziel, im zuvor thematisch codierten Material „areas of stability, regression, or growth“ (Lyon, 2013c, S. 450) identifizieren zu können. Abschließend wurden in einem dritten Schritt, auf Grundlage der beiden Codierzyklen, ausführliche Fallnarrative über die Entwicklung der Expertise von Dean, Glenda und Lauren bezüglich des Umgangs mit sprachlich-kultureller Heterogenität bei schulischer Leistungsfeststellung und -beurteilung verfasst (vgl. Lyon, 2013d, S. 4), deren Validität durch ein Member-Checking mit den Teilnehmer_innen sichergestellt wurde (vgl. Birt, Scott, Cavers, Campbell, & Walter, 2016). Lyon (2013d) fasst diese Fallnarrative wie folgt zusammen[54]:

> „For Dean, equitable science assessment revolved around engaging individual students in dialog, thus integrating language and science [learning.] [...] Glenda's expertise in equitable science assessment, conveyed through her focus on multiple assessment forms and scientific explanations, evolved in that she considered explanations as not just a way to uncover what students know about science, but also a way to promote academic excellence for her students, regardless of second language proficiency. Finally, Lauren put most of her energy into particular language scaffolds[.] [...] The teachers all evolved by viewing assessment as providing a supportive role, in which it became important to incorporate some form of scientific discourse, such as through written predictions (Lauren), group discussion around prompts (Glenda), and scientific explanations (Dean). Unlike Lauren and Glenda, Dean also engaged each student in sustained dialog around the explanation. [...] Teachers recognized the mediating role of language and culture while assessing science, rather than viewing "learning styles" as the mediating factor.“ (ebd., S. 8)

Bei allen drei Teilnehmer_innen zeigte sich also insgesamt, im Einklang mit den Befunden auf quantitativer Ebene, ein Zuwachs ihrer Expertise bezüglich des Umgangs mit sprachlich-kultureller Heterogenität bei schulischer Leistungsfeststellung und -beurteilung. Allerdings wurde anhand der qualitativen Analyse der Daten von Dean, Glenda und Lauren auch deutlich, dass ihr Umgang mit sprachlich-kultureller Heterogenität bei schulischer Leistungsfeststellung und -beurteilung zwei besonders hervorstechende Auffälligkeiten aufweist:

Zum einen zeigte sich, dass Dean, Glenda und Lauren ein Verständnis dafür entwickelt haben, Sprache nicht nur als Lernmedium, sondern auch als Lerngegenstand im naturwissenschaftlichen Fachunterricht zu begreifen (vgl. Lyon, 2013d, S. 9). Bei schulischen Leistungsfeststellungen und -beurteilungen beschränkten sich allerdings alle drei Studierende bis zum Ende ihres Studiums weitgehend auf fachlich-konzeptuelles Wissen und Können von Schüler_innen (vgl. ebd.). Eine explizite Feststellung und Beurteilung der sprachlichen Fähigkeiten von Schüler_innen fand hingegen bei jedem der drei Teilnehmer_innen, wenn überhaupt, nur marginal statt. Dies deutet darauf hin, dass Dean, Glenda und Lauren mutmaßlich auch gegen Ende ihres Studium Sprache eher als Medium und weniger als Gegenstand schulischer Leistungsfeststellungen und -beurteilungen betrachten und/oder es ihnen an Wissen und Können mangelt, um die sprachlichen Fähigkeiten von Schüler_innen erfassen und von fachlich-konzeptuellen unterscheiden zu können (vgl. ebd.).

[54] Die vollständigen Fallnarrative von Dean, Glenda und Lauren sind bei Lyon (2013d, S. 4 u. f.) zu finden.

Zum anderen wurde bei allen drei angehenden Naturwissenschaftslehrkräften sichtbar, dass diese auf unterschiedliche Art und Weise handeln, wenn es um die Frage geht, inwieweit sprachliche Anforderungen an Schüler_innen in Leistungssituationen zu reduzieren sind und/oder im Rahmen von schulischen Leistungsfeststellungen und -beurteilungen Maßnahmen zur Unterstützung der sprachlichen Fähigkeiten ihrer Schüler_innen zu ergreifen sind (vgl. Lyon, 2013d, S. 9). Zusätzlich weisen die Fallnarrative von Dean, Glenda und Lauren darauf hin, dass sich deren diesbezüglicher Umgang mit sprachlicher Heterogenität bei schulischer Leistungsfeststellung und -beurteilung im Laufe ihres Studiums änderte (vgl. Lyon, 2013d, S. 4 u. f.). Dies kann interpretiert werden als Veränderungen des Handlungstypus der drei Studierenden im Umgang mit sprachlicher Heterogenität im naturwissenschaftlichen Unterricht im Sinne der von Riebling (2013b) theoretisch begründeten und empirisch nachgewiesenen Typologie (vgl. Abschnitt 3.2 Einleitung). Deans Handlungstyp scheint sich dabei im Verlauf des Studiums von Typ C hin zu Typ A verändert zu haben, bei Glenda ist eine Entwicklung von Typ D zu Typ A zu erkennen und bei Lauren scheint sich der Umgang mit sprachlicher Heterogenität im naturwissenschaftlichen Unterricht von Typ B1 hin zu Typ B2 gewandelt zu haben.

3.2.2. Das Fallbeispiel von Tajmel zu Lehrerleistungsurteilen zu Originaltexten eines_einer Schülers_Schülerin mit Migrationshintergrund zum Thema Schwimmen und Sinken

Nun zur explorativen Untersuchung von Tajmel (2017b): Das Erkenntnisinteresse dieser Studie lag darin zu erkunden, was es bedeutet den Gedanken „Naturwissenschaften sind ein relevanter Teil von Bildung“ in das normative Grundgerüst eines Rechts auf Bildung als absolutes und unveräußerliches Menschenrecht einzuordnen und wie sich die Begriffe „Fachkultur“, „Macht“, „Sprache im Fachunterricht“ und „Ungleichheit“ zu dem einer „naturwissenschaftlichen Bildung“ verhalten.

Hierzu widmet sich die Autorin in der ersten Hälfte ihrer zugehörigen Dissertationsschrift der detaillierten Skizze einer „Reflexiven Physikdidaktik“. Deren zentrale Charakteristika sind das Menschenrecht auf Bildung als Normbasis, die Erforschung von sich hieraus ergebenden Bildungsbarrieren, das Ziel Ansätze zur Überwindung dieser Barrieren zu entwickeln und die Konkretisierung des Begriffs „Zugang zu naturwissenschaftlicher Bildung“ auf Basis des Rahmenkonzepts von Tomasevški (2001) (vgl. Tajmel, 2017b, S. 125 u. f.).

Anschließend rückt Tajmel (2017b) die Rolle von Sprache als Zugangsbarriere für naturwissenschaftliche Bildung in den Vordergrund, zu deren Überwindung sie ein Konzept von Sprachbewusstheit als Teil reflexiver Professionalität von Physiklehrkräften entwickelt (vgl. ebd., S. 199 u. f.), dass „einerseits kognitive Aspekte im Sinne von *Sprachwissen* und andererseits reflexive Aspekte der Selektion aufgrund fachlicher und sprachlicher Normen im Sinne von *Machtwissen* [umfasst]“ (ebd., S. 267, Hervorhebung im Original). Ihren Argumentationsgang untermauert sie dabei anhand von drei Fallbeispielen, wobei sie diese

in Anlehnung an die Herangehensweise der Kasuistik[55] auswertet und darstellt (vgl. ebd. S. 202).

In einem dieser drei Fallbeispielen analysiert Tajmel (2017b) die Leistungsurteile von Naturwissenschaftslehrkräften zu Textprodukten von Schüler_innen (vgl. ebd., S. 251 u. f.). In diesem Fallbeispiel wurden zwei Aufgabenbearbeitungen eines_einer Schülers_Schülerin mit Migrationshintergrund[56] N = 73 Lehrkräften, die mindestens ein naturwissenschaftliches Fach unterrichten (vgl. Tajmel, 2017b, S. 254), jeweils zusammen mit folgender Arbeitsanweisung vorgelegt:

> „Bitte bewerten Sie die Antwort:
>
> 1. Ist die Antwort richtig oder falsch?
> 2. Wie viele von insgesamt 5 erreichbaren Punkten würden Sie geben?
> 3. Bitte begründen Sie Ihre Entscheidung!
>
> Zusatzinformation: Im Unterricht dieser Klasse wurde der Begriff „Dichte“ noch nicht eingeführt.“
>
> (Tajmel, 2010, S. 173)

In beiden Aufgaben, die der_die Schüler_in bearbeitete, geht es darum, durch Ankreuzen zu entscheiden und in Form eines kurzen Textes zu begründen, ob ein Baumstamm bzw. eine Metallplatte in Wasser schwimmt oder untergeht (vgl. Tajmel, 2010, S. 172; Tajmel, 2017b, S. 252). Beide Antworten des_der Schülers_Schülerin (siehe Tabelle 3.4) sind dabei zum einen fachlich-konzeptuell vollkommen korrekt[57] zum anderen zeigt sich aber auch, dass der_die Schüler_in im Erwerb des Deutschen als Zweitsprache weit fortgeschritten ist[58].

Die deskriptive Analyse der Punktevergaben der befragten Naturwissenschaftslehrkräfte offenbarte, dass diese im Mittel auf beide Schülertexte eine vergleichbare Punktzahl vergeben ($M_{Baumstamm} = 2.99$; $M_{Metallplatte} = 2.76$; ob der Unterschied beider Mittelwerte statistisch signifikant ist, berichtet Tajmel (2017b, S. 257 u. f.) nicht). Gleichzeitig ist diese mittlere Punktzahl deutlich geringer als die mögliche Maximalpunktzahl (5 Punkte) (vgl. Tajmel, 2010, S. 174). Dies deutet darauf hin, dass beide Schülertexte aus Perspektive der befragten Lehrkräfte im Sinne der an sie gerichteten Arbeitsanweisung nicht nur „Richtiges“ sondern auch „Falsches“ enthalten. Untermauert wird diese Interpretation, wenn man den Blick auf die Streuung der Punktevergabe richtet: Beide Schülertexte weisen eine

[55] Die Kasuistik kennzeichnet sich dadurch, dass es bei dieser „nicht primär um die Problemlösung [geht, sondern darum] [...] den Fall in seiner Besonderheit zu verstehen und darin das Allgemeine im Sinne von Theorien, gültigen Regeln, ethischen Grundsätzen u.v.m. zu entdecken“ (Kunz, 2015, S. 78). Die zentrale Frage ist hierbei: „Was ist der Fall?“ (vgl. Tajmel, 2017b, S. 202).

[56] Hierbei handelte es sich um einen_eine 13 jährigen_jährige Schüler_Schülerin der 7. Jahrgangsstufe mit russischer Herkunftssprache (vgl. Tajmel, 2011, S. 4; Tajmel, 2010, S. 172).

[57] In beiden Aufgabenbearbeitungen ist nicht nur die fachlich korrekte Antwortoption angekreuzt (siehe Tabelle 3.4), sondern auch die Schwimmfähigkeit eines Gegenstandes mit dem Material, aus dem er besteht begründet (Holz bzw. Metall) und nicht mit seinem Volumen oder Gewicht (vgl. Tajmel, 2010, S. 173).

[58] Z. B. werden Verbklammern („dan deht es unter“) und Präpositionalphrasen („aus Holz entschteht“) richtig gebildet (vgl. Tajmel, 2010, S. 173; Tajmel, 2011, S. 4 u. f.; Tajmel, 2017b, S. 254).

Aufgabenbearbeitung eines_einer Schüler_in der 7. Jahrgangsstufe (Tajmel 2011, S. 172; Tajmel 2017b; S. 253):	
Ein **Baumstamm...** ☐ ... geht unter, weil... ☒ ... schwimmt, weil... *„das ~~haub~~ baum aus Holz entschteht.“*	Eine **Metallplatte...** ☒ ... geht unter, weil... ☐ ... schwimmt, weil... *„der platte aus Metall entschteht und der Metall ist immer schwer egal ob es leicht oder schwer ist wen es ein Metall ist dan deht es unter!“*
Lehrerleistungsurteile über die Aufgabenbearbeitung des_der Schüler_in (Tajmel 2011, S. 173 u. f.; Tajmel 2017b; S. 258):	
Lehrerleistungsurteil A (5 von 5 Punkten): *„Das Kind hat den Zusammenhang das Material – Eigenschaft erkannt.“* **Lehrerleistungsurteil B (4 von 5 Punkten):** *„Die Antwort ist korrekt, weil die Schwimmfähigkeit des Baumstamms mit dem Material zusammenhängt. Für eine ausführlichere Antwort hätte es einen weiteren Punkt gegeben.“* **Lehrerleistungsurteil C (3 von 5 Punkten):** *„Die Antwort scheint mir grundsätzlich in Ordnung zu sein, die Begründung ist mir allerdings zu knapp. Schön wäre noch eine Erklärung, z.B. ''Holz ist meiner Erfahrung nach ziemlich leicht'' oder ähnliches.“* **Lehrerleistungsurteil D (3 von 5 Punkten):** *„Die Antwort ist falsch, weil eine wichtige Information fehlt, da ich positiv denkend diese Information hinter der Antwort vermute würde ich 3 Punkte geben.“* **Lehrerleistungsurteil E (3 von 5 Punkten):** *„Die Begründung ist nicht ausreichend.“* **Lehrerleistungsurteil F (3 von 5 Punkten):** *„Ich vermute, dass der Schüler bereits gesehen hat, dass Baumstämme mit Wasser (Flüsse) transportiert wird. Allerdings ist die Zuordnung Baum aus Holz falsch.“* **Lehrerleistungsurteil G (3 von 5 Punkten):** *„Grammatikfehler; Bezug fehlt (welches Holz?); Wasser? Erklärung fehlt“* **Lehrerleistungsurteil H (2 von 5 Punkten):** *„Indirekt hat der Schüler etwas richtig aufgeschnappt, kann es aber nicht in Worte fassen.“* **Lehrerleistungsurteil I (2 von 5 Punkten):** *„Die Schülerin kann vermutlich mit der Eigenschaft Holz etwas anfangen, kann aber die Erklärung nicht in richtige Worte fassen.“* **Lehrerleistungsurteil J (2 von 5 Punkten):** *„Ankreuzen ist richtig. Erklärung nicht ganz nachvollziehbar; scheint eine Erfahrungstatsache zu sein.“* **Lehrerleistungsurteil K (0 von 5 Punkten):** *„Völlig unsinnige Antwort auf eine unsinnige, nicht klar gestellte Aufgabe“*	**Lehrerleistungsurteil L (3 von 5 Punkten):** *„Die Antwort scheint grundsätzlich richtig zu sein. Schön ist auch die ausführliche Begründung. Leider widerspricht sich der/die Schüler/in in der Aufgabe selbst, so dass ich nicht die volle Punktzahl geben würde.“* **Lehrerleistungsurteil M (3 von 5 Punkten):** *„Kreuz ist richtig. Mit der Begründung versucht der Schüler zu erklären, dass Metall immer „schwerer“ ist als Wasser, egal ob Blei oder Alu.“* **Lehrerleistungsurteil N (3 von 5 Punkten):** *„Begründung nicht eindeutig.“* **Lehrerleistungsurteil O (3 von 5 Punkten):** *„Text z. T. unverständlich; Sprachl. Mängel“* **Lehrerleistungsurteil P (2 von 5 Punkten):** *„Dass die Platte aus Metall ist, verrät bereits die Bezeichnung ''Metallplatte''. Die Aussage, dass Metall immer untergehe, stimmt so nicht. Da das Material dennoch eine Rolle spielt, vergebe ich 2 von 5 Punkten.“* **Lehrerleistungsurteil Q (2 von 5 Punkten):** *„Metallplatte geht unter. Richtig. Begründung aber falsch. => metallischer Hohlköper – 1 Punkt Abzug wg. der sprachl. Fehler.“* **Lehrerleistungsurteil R (2 von 5 Punkten):** *„Vermutlich kennt der Schüler nur Metalle, die schwerer sind als Wasser“* **Lehrerleistungsurteil S (1 von 5 Punkten):** *„aus Erfahrung abgeleitet; richtige Gründe fehlen; Bewertung unabhängig von der sprachlichen Richtigkeit“* **Lehrerleistungsurteil T (1 von 5 Punkten):** *„Für die allg. Richtigkeit 1P, aber die Begründung ist tw. Widersprüchlich und nicht immer richtig (Oberflächenspannung)“* **Lehrerleistungsurteil U (1 von 5 Punkten):** *„Grammatikfehler; Bezug unklar (welches Metall?); Erklärung falsch; Oberfläche?“* **Lehrerleistungsurteil V (1 von 5 Punkten):** *„Antwort oberflächlich und widersprüchlich begründet, sprachlich unzureichend“* **Lehrerleistungsurteil W (0 von 5 Punkten):** *„Falsche Zuordnung von Begriffen“* **Lehrerleistungsurteil X (0 von 5 Punkten):** *„Metalle haben unterschiedliche Dichten“*

Tabelle 3.4.: Aufgabenbearbeitungen eines_einer Schülers_Schülerin der 7. Jahrgangsstufe und Beispiele von Leistungsurteilen verschiedener Naturwissenschaftslehrkräfte (vgl. Tajmel, 2010, S. 172 u. f.; Tajmel, 2017b, S. 251 u. f.). Der von dem_der Schüler_in produzierte Text, sowie die Begründungen der Lehrkräfte ihrer Leistungsurteile sind jeweils kursiv hervorgehoben.

Standabweichung von mehr als einem Punkt auf ($SD_{Baumstamm} = 1.30$; $SD_{Metallplatte} = 1.39$) auf. Zudem haben lediglich 12 (14) der 73 befragten Lehrkräfte[59] die Schülertexte zur Schwimmfähigkeit eines Baumstammes (einer Metallplatte) mit der Maximalpunktzahl bewertet (vgl. Tajmel, 2017b, S. 257).

[59] Die hier berichtete Anzahl sind in Ablesegenauigkeit aus den von Tajmel (2017b, S. 257) angegebenen Histogrammen entnommen.

Daneben haben sich die schriftlichen Begründungen der teilnehmenden Lehrer_innen als aufschlussreich erwiesen, von denen Tajmel insgesamt 24 wortwörtlich zitiert und die in Tabelle 3.4 ebenfalls aufgeführt sind. An diesen zeigen sich die folgenden vier Auffälligkeiten, welche vor allem die Sprachbewusstheit der befragten Lehrkräfte fragwürdig erscheinen lassen:

- Erstens erfolgen die Begründungen der Lehrkräfte anhand von Kriterien die fachlich-konzeptuelle Aspekte der Schülertexte adressieren und/oder deren sprachliche Realisierung (beides z. B. bei Lehrerleistungsurteil H, I, Q, S, V). Dies lässt vermuten, dass naturwissenschaftliche Fachlehrkräfte nicht nur fachlich-konzeptuelle Merkmale der Textprodukte von Schüler_innen feststellen und beurteilen, sondern stets auch „die sprachlichen Leistungen der Schülerinnen und Schüler mitbewerten“ (Tajmel, 2010, S. 174).
- Zweitens finden sich bezüglich fachlich-konzeptueller Aspekte zwar auch positive Feststellungen und Beurteilungen in den Begründungen (z. B. Lehrerleistungsurteil A, B, C, L, M, Q), die volle Punktzahl wird selbst dann jedoch kaum vergeben (Tajmel, 2017b, S. 259). Vielfach ist die Feststellung und Beurteilung fachlich-konzeptueller Aspekte der Schülertexte defizitorientiert, wobei die befragten Lehrkräfte dies mit Anforderungen an das Fachwissen begründen, über das der_die Schüler_in wahrscheinlich noch nicht verfügen kann (z. B. „Oberflächenspannung“ in Lehrerleistungsurteil T oder „Metalle haben unterschiedliche Dichten“ in Lehrerleistungsurteil X) (Tajmel, 2010, S. 173; Tajmel, 2017b, S. 257). Gerade da der_die Schüler_in vermutlich noch nicht über Fachwissen zur Lösung der beiden Aufgaben verfügen kann und die befragten Lehrkräfte hierüber auch informiert waren (siehe Arbeitsanweisung an die Lehrkräfte), ist bemerkenswert, dass anschlussfähige Denkfiguren, die sich in der Erklärung des_der Schülers_Schülerin finden lassen, in den Begründungen der Lehrkräfte zum Teil negativ konnotiert sind, z. B. als „Erfahrungstatsache“ der „richtige Gründe fehlen“ oder die „[i]ndirekt [...] aufgeschnappt“ sind (Lehrerleistungsurteil H, J, S).
- Drittens ist ebenfalls die Feststellung und Beurteilung der sprachlichen Realisierung der Schülertexte tendenziell defizitorientiert[60] (z. B. „- 1 Punkt wg. der sprachl. Fehler“ in Lehrerleistungsurteil Q oder „sprachlich unzureichend“ in Lehrerleistungsurteil V). Die Lehrer_innen begründen ihr Urteil hier allerdings oftmals mit eher vagen Kriterien (z. B. „sprachliche Richtigkeit“ in Lehrerleistungsurteil S oder „kann es [...] nicht in Worte fassen“ in Lehrerleistungsurteil H). Bemerkenswert ist diese defizitorientierte Grundhaltung vor allem deswegen, da sich bei dem_der Schüler_in – wie oben erwähnt – ein bereits weit fortgeschrittener Zweitsprachenerwerb konstatieren

[60] Dieser Trend zeigte sich zudem in Tajmels zusätzlicher Untersuchung mit N = 43 angehenden Naturwissenschaftslehrkräften im Rahmen einer universitären Lehrveranstaltung zu „Deutsch als Zweitsprache“ (vgl. Tajmel, 2017b, S. 260 u. f.). Die Studierenden sollten hierbei den Schülertext zur Schwimmfähigkeit einer Metallplatte in Form von Schulnoten bewerten (M = 3.40; SD = .76) und diese Benotung kurz begründen. Dabei zeigte eine Inhaltsanalyse der Begründungen, dass die angehenden Lehrkräfte vorrangig die sprachlichen Defizite dieses Schülertextes im Blick haben (vgl. ebd.).

lässt (vgl. Tajmel, 2010, S. 173; Tajmel, 2017b, S. 254). Die befragten Lehrkräfte scheinen daher also „hohe Ansprüche an die sprachliche Form [zu] stellen“ (Tajmel, 2010, S. 174).

- Viertens finden sich in einigen Begründungen die Auffälligkeit, dass positive (aber negativ konnotierte) Äußerungen zu fachlich-konzeptuellen Aspekten der Schülertexte mit negativen Anmerkungen bezüglich ihrer sprachlichen Realisierung gepaart sind. Zum Beispiel:
  - „Indirekt hat der Schüler etwas richtig aufgeschnappt, kann es aber nicht in Worte fassen“ (Lehrerleistungsurteil H; Hervorhebung M. S. F.).
  - „Die Schülerin kann vermutlich mit der Eigenschaft Holz etwas anfangen, kann aber die Erklärung nicht in richtige Worte fassen“ (Lehrerleistungsurteil I; Hervorhebung M. S. F.).
  - Metallplatte geht unter. Richtig. Begründung aber falsch => metallischer Hohlkörper - 1 Punkt Abzug wg. der sprachl. Fehler“ (Lehrerleistungsurteil Q; Hervorhebung M. S. F.).

  Dies lässt begründet vermuten, dass Physiklehrkräfte während der Genese der Leistungsbewertung ihre Beurteilung über fachlich-konzeptuelle und sprachliche Leistungen konfundieren. Ihre Bewertungslogik scheint mutmaßlich darin zu bestehen fachlich-konzeptuelle Merkmale eines Schülertextes mehr oder weniger zu relativieren, wenn ein Schülertext zu viele Mängel bezüglich einer sprachlichen Norm aufweist, deren Beherrschung von der Lehrkraft vorausgesetzt wird. Problematisch bezüglich einer adäquaten Leistungsfeststellung und -beurteilung ist vor allem, dass in den eben aufgeführten Beispielen selbst fachlich richtige oder zumindest anschlussfähige Denkfiguren von Schüler_innen eine solche Relativierung erfahren.

3.2.3. Quintessenz aus den Untersuchungen von Lyon und Tajmel

Was ist nun die Quintessenz, die sich aus den Studien von Lyon und Tajmel ergibt? – Beide Untersuchungen liefern alles in allem unterschiedliche Einblicke in den Umgang von (angehenden) Naturwissenschaftslehrkräften mit Sprache in schulischen Leistungssituationen, also in einen Teilaspekt ihrer Assessment Literacy:

Die Studie von Lyon vergegenwärtigt, dass nicht nur Schüler_innen im Physikunterricht mit nicht zu vernachlässigenden sprachlichen Anforderungen konfrontiert werden, sondern, dass auch für (angehende) Lehrkräfte der Umgang mit Sprache in Leistungssituationen im naturwissenschaftlichen Fachunterricht eine anspruchsvolle Herausforderung darstellt. Zum einen wird dies an der auf quantitativer Ebene insgesamt moderat ausgeprägten Expertise der von Lyon wissenschaftlich begleiteten angehenden Naturwissenschaftslehrkräften im Umgang mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen deutlich. Zum anderen zeigten sich sowohl auf quantitativer, vor allem aber auf qualitativer Ebene bei allen teilnehmenden Natur-

wissenschaftslehrkräften, eine eher geringfügige Entwicklung hin zu einem expertenhaften Umgang mit sprachlich-kultureller Heterogenität bei schulischer Leistungsfeststellung und -beurteilung. Bei den Teilnehmer_innen, deren Daten einer zusätzlichen qualitativen Analyse unterzogen wurden, deutet sich an, dass diese bis zum Ende ihres Studium Sprache eher als Medium und weniger als Gegenstand von Leistungsfeststellung und -beurteilung im naturwissenschaftlichen Fachunterricht betrachten und/oder es ihnen an Wissen und Können mangelt, sprachliche und fachlich-konzeptuelle Fähigkeiten von Schüler_innen voneinander zu unterscheiden. Dies lässt vermuten, dass Naturwissenschaftslehrkräfte wenn überhaupt nur langfristig eine derartige Expertise erwerben.

Auf der anderen Seite liefert die Studie von Tajmel Hinweise auf den von Lehrkräften in der Praxis umgesetzten Umgang mit Sprache in Leistungssituationen im naturwissenschaftlichen Fachunterricht. Zunächst erhärtet sich auf Grundlage des von Tajmel analysierten Fallbeispiels der Verdacht einer defizitorientierten Grundhaltung von Lehrkräften bezogen auf sprachliche Leistungen von Schüler_innen im naturwissenschaftlichen Fachunterricht, über die zu Beginn dieses Unterkapitels mit Bezug auf Befunde über den Umgang von Naturwissenschaftslehrkräften mit Sprache auf globaler Ebene bereits gemutmaßt wurde (vgl. Unterkapitel 3.2 Einleitung). Vor allem aber zeigen sich in einigen von der Autorin dargestellten Leistungsurteilsbegründungen der befragten Naturwissenschaftslehrkräfte deutliche Hinweise auf eine problematische Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile. Dies wirft die Vermutung auf, dass die von Tajmel befragten Lehrkräfte nur bis zu einem bestimmten Grad über Assessment Literacy bezogen auf Sprache als Lernmedium und -gegenstand im naturwissenschaftlichen Fachunterricht verfügen. Im Sinne der in Abschnitt 2.2.4 dargestellten umfassenden Konzeption der Assessment Literacy von Lehrkräften ist es denkbar, dass die befragten Naturwissenschaftslehrkräfte...

... entweder kaum über kompetenztheoretisch gedachtes Wissen und Können und/oder berufsbezogenen Überzeugungen darüber verfügen, wie bzw. dass fachlich-konzeptuelle und sprachliche Leistungen getrennt voneinander festgestellt und beurteilt werden können und/oder wie bzw. dass problematische Konfundierungen fachlich-konzeptueller und sprachlicher Leistungsurteile vermieden werden können,

... oder sie zwar über derartiges kompetenztheoretisch gedachtes Wissen und Können und/oder berufsbezogenen Überzeugungen verfügen, es ihnen aber nicht gelingt Kontextbedingungen, die ihren Handlungsspielraum begrenzen, mit zu berücksichtigen, weswegen es ihnen in ihrer Logik des Handelns nicht gelingt, problematische Konfundierungen fachlich-konzeptueller und sprachlicher Leistungsurteile zu vermeiden.

Es gilt allerdings festzuhalten, dass die Befunde beider Untersuchungen aufgrund ihres stark ausgeprägten Fallstudiencharakters nicht generalisierbar sind. Ferner lassen sich auf der Grundlage beider Studien keine sicheren Schlüsse darüber ziehen, wie Naturwissenschaftslehrkräfte tatsächlich mit Sprache umgehen, wenn sie in der täglichen Praxis damit konfrontiert sind, Leistungsurteile über ihre Schüler_innen zu genieren, welcher Logik sie dabei folgen und welche Maßstäbe sie dabei anwenden. Dies hängt damit zusammen, dass sich die Studie von Tajmel auf Begründungen von Naturwissenschaftslehrkräften stützt,

die erst nachdem sie bereits ein Leistungsurteil gebildet hatten, anzufertigen waren. Bei der Studie von Lyon begründet sich dies daraus, dass hier vornehmlich schriftliche Ausarbeitungen von bzw. Interviews mit angehenden Naturwissenschaftslehrkräften untersucht wurden, die von eigener tatsächlichen Unterrichtspraxis losgelöst waren.

3.3. Zusammenfassung

In diesem Kapitel galt es, die bisherigen Erkenntnisse zum Umgang von Lehrkräften mit Sprache in Leistungssituationen im Physikunterricht, also einen besonderen Aspekt der Assessment Literacy von Physiklehrkräften, aufzuarbeiten.

In Unterkapitel 3.1 erfolgte eine erste Annäherung an den Umgang von Physiklehrkräften mit Sprache im Rahmen schulischer Leistungsfeststellung und -beurteilung, dadurch dass die für Schüler_innen im Physikunterricht anzutreffenden sprachlichen Anforderungen illustriert wurden. In einer ersten blitzlichtartigen Betrachtung wurde dabei, neben dem monolingualen Selbstverständnis des deutschen Bildungswesens und der „Schulnormthese“ von Feilke (2012, S. 154 u. f.), vor allem auf die Sprachvarietäten Fach- und Bildungssprache, als im naturwissenschaftlichen Fachunterricht adäquat geltende Formen des Sprachgebrauchs aufmerksam gemacht. Anschließend wurde genauer geklärt, was unter dem Gebrauch von Fach- und Bildungssprache als sprachliche Anforderungen an Schüler_innen im Physikunterricht zu verstehen ist und insbesondere in welchem Verhältnis Bildungssprache und Fachsprache zueinander stehen:

- Der Gebrauch von Fachsprache (in Leistungssituationen) im Physikunterricht wurde dabei unter Zuhilfenahme der in der Fachsprachenforschung üblichen horizontalen und vertikalen Gliederung von Fachsprachen näher bestimmt (vgl. Abschnitt 3.1.1). Dabei wurde vor allem deutlich, dass es (in Leistungssituationen) im Physikunterricht nicht „den“ Gebrauch von Fachsprache gibt, sondern dass im Physikunterricht unterschiedlichste Arten des fachlichen Sprachgebrauchs auftreten, die sich durch eine Vielzahl linguistischer Oberflächenmerkmale charakterisieren lassen und dass diese Sprachgebrauchsarten eher in Ausnahmefällen mit dem Gebrauch von wissenschaftlicher Fachsprache gleichgesetzt werden können.
- In Abschnitt 3.1.2 wurden schließlich die drei zentralen Bedeutungen, die im gegenwärtigen erziehungswissenschaftlichen Diskurs dem Begriff „Bildungssprache“ zugewiesen werden, erörtert (Bildungssprache als Transfermedium für Wissen, als Denkwerkzeug und als Eintritts- und Visitenkarte). Dabei wurde insbesondere das Verhältnis zwischen Bildungssprache und Fachsprache, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird, dargelegt. Es zeigte sich,...

  ... dass die Fachsprache der Physik, wie sie (in Leistungssituationen) im Physikunterricht verwendet wird, kaum von einer für Physikunterricht domänenspezifischen Ausprägung von Bildungssprache unterschieden werden kann,

... dass Anforderungen an Schüler_innen im Physikunterricht, die domänenspezifisch-bildungssprachliche oder fachlich-konzeptuelle Fähigkeiten adressieren, in einer sehr engen Beziehung zueinander stehen, dennoch aber verschieden voneinander sind

... und dass aus soziologischer Perspektive domänenspezifische Bildungssprache als die „legitime Sprache“ (vgl. Bourdieu, 1993, S. 100) im Physikunterricht verstanden werden kann, die Schüler_innen beherrschen müssen, um sich als erfolgreicher_erfolgreiche Lerner_Lernerin ausweisen zu können und als solche_r von Lehrkräften auch erkannt zu werden.

Anschließend wurde sich in Unterkapitel 3.2 explizit der Frage gewidmet, welche Kenntnisse es bislang darüber gibt, wie Physiklehrer_innen mit Sprache in Leistungssituationen im Physikunterricht umgehen. Diese Forschungsstandsichtung erwies sich allerdings als ernüchternd. Zwar lassen sich aus Studien, die den Umgang von Lehrkräften mit Sprache im naturwissenschaftlichen Fachunterricht auf globaler Ebene untersuchten, durchaus plausible Mutmaßungen über den Umgang von Physiklehrkräften mit Sprache im Rahmen von schulischer Leistungsfeststellung und -beurteilung anstellen (vgl. Unterkapitel 3.2 Einleitung), diese müssen allerdings als spekulativ gelten.

Faktisch lassen sich zum gegenwärtigen Zeitpunkt in der Literatur lediglich zwei Studien ausfindig machen, in denen sich tiefgreifender mit dem Umgang von (angehenden) Lehrkräften mit Sprache (in Leistungssituationen) im naturwissenschaftlichen Fachunterricht auseinander gesetzt wurde. Die Befunde dieser beiden Studien – jene von Lyon und Tajmel – wurden in den Abschnitten 3.2.1 und 3.2.2 detailliert erläutert. Es zeigte sich, dass beide Untersuchungen alles in allem unterschiedliche Einblicke in den Umgang von (angehenden) Naturwissenschaftslehrkräften mit Sprache in schulischen Leistungssituationen liefern:

- Zum einen erhärten beide Studien Verdachtsmomente über den Umgang von Lehrkräften mit Sprache in Leistungssituationen im naturwissenschaftlichen Fachunterricht, die zu Beginn dieses Unterkapitels aufgeworfen wurden (z. B. das Auftreten verschiedener Handlungstypen im Umgang mit sprachlich-kultureller Heterogenität oder einer defizitorientierten Grundhaltung gegenüber den sprachlichen Fähigkeiten von Schüler_innen in Leistungssituationen im Fachunterricht).
- Zum anderen konnten aus diesen Untersuchungen besondere Auffälligkeiten aufgedeckt werden, wie die begründete Vermutung einer problematischen Konfundierung fachlich-konzeptueller und sprachlicher Beurteilungen bei der Genese von Leistungsurteilen durch naturwissenschaftliche Fachlehrkräfte.

Allerdings sind die Befunde sowohl der Studie von Lyon als auch jener von Tajmel aufgrund ihres ausgeprägten Fallstudiencharakters nicht generalisierbar. Insbesondere wurde deutlich, dass sich auf Grundlage der Erkenntnisse beider Studien keine sicheren Schlüsse darüber ziehen lassen, wie Naturwissenschaftslehrkräfte tatsächlich mit Sprache umgehen, wenn sie in der täglichen Praxis damit konfrontiert sind Leistungsurteile über ihre

Schüler_innen zu genieren, welcher Logik sie dabei folgen und welche Maßstäbe sie dabei anwenden.

In diesem Kapitel ist damit insbesondere deutlich geworden, dass die Fragen, welche Ressourcen Lehrkräfte zur Genese sprachlicher Leistungsurteile im naturwissenschaftlichen Fachunterricht einsetzen, sowie ob und falls ja inwiefern, Physiklehrkräfte bei der Genese von Leistungsurteilen fachlich-konzeptuelle und sprachlicher Beurteilungen problematisch miteinander konfundieren, aktuelle Desiderate fachdidaktischer Forschung darstellen. Diesen beiden Desideraten wird sich daher im nun folgenden empirischen Teil der vorliegenden Arbeit gewidmet.

Teil II:

Empirische Untersuchung

„[I]ch stelle fest, auch hier wieder, dass es nicht so ganz einfach ist [...] sozusagen mit dem gleichen Bewertungslevel oder Bewertungsschlüssel äh Texte zu korrigieren bzw. Texte zu bewerten. Weil's eben viele Kriterien gibt. Also sprachliche sind dann vielleicht eher nachgeordnet, wenn's um das Fach geht. Aber eben auch Benutzung von Fachbegriffen. Ich ganz persönlich [...] gehöre schon noch zu denjenigen die eigentlich von Schülern erwarten, dass sie zumindestens formal korrekte Sätze formulieren und nicht darum was sie hinklieren."

Herr Carboni (Transkiptsegment 489)

4. Erkenntnisinteresse und methodische Grundlegung der Untersuchung

4.1. Zielsetzung und Forschungsfragen

Teil I der vorliegenden Arbeit offenbarte einerseits, dass die Feststellung und Beurteilung von Schülerleistungen ein wesentlicher Bestandteil des Systems Schule ist, insbesondere dass sie einen bedeutenden Aspekt des Lehrerberufs darstellt. In der erziehungswissenschaftlichen Literatur wurde daher bereits viel diskutiert, welche Funktionen Leistungsfeststellungen und -beurteilungen im System Schule erfüllen sollen und welche sie tatsächliche erfüllen (vgl. Unterkapitel 1.2) sowie ferner, wie es um die Güte dieser schulischen Verfahren bestellt ist (vgl. Unterkapitel 1.3). Es verwundert daher auch nicht, dass sich in der Literatur ein weitläufiges Forschungsfeld zu Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen ausfindig machen lässt (vgl. Unterkapitel 2.1). Auf Grundlage dieser Forschung lässt sich auf theoretischer Ebene – wie in Unterkapitel 2.2 geschehen – eine sowohl detailreiche, wie auch umfassende Denkfigur davon entwickeln, was eine im Kontext von schulischer Leistungsfeststellung und -beurteilung professionell handelnde (Physik-)Lehrkraft auszeichnet (ihre Assessment Literacy). Des Weiteren wurde in Kapitel 3 deutlich, dass Sprache als Lernmedium und Lerngegenstand im Physikunterricht nicht nur für Schüler_innen eine nicht zu vernachlässigende Anforderung darstellt. Der Umgang mit Sprache in Leistungssituationen im naturwissenschaftlichen Fachunterricht ist für Lehrkräfte eine in gleicher Weise anspruchsvolle Herausforderung.

Andererseits ist in Teil I der vorliegenden Arbeit Folgendes deutlich geworden:

- Erstens ist auf genereller Ebene festzustellen, dass alltägliche Leistungsfeststellung und -beurteilung durch Lehrkräfte einen Gegenstand dargestellt, mit dem sich die erziehungswissenschaftliche und insbesondere physikdidaktische Forschung bislang nur sehr wenig auseinandergesetzt hat. Vor allem ist weitgehend unbekannt, wie (Physik-)Lehrkräfte in ihrer täglichen Berufspraxis bei der Feststellung und Beurteilung von Schülerleistungen tatsächlich vorgehen, welchen Logiken sie dabei folgen und welche Maßstäbe sie hierbei für angemessen halten (vgl. Abschnitt 2.1.4). Es erscheint daher relevant diese zum Teil höchst unterschiedlichen Facetten der täglichen Berufspraxis von Lehrkräften in einer empirischen Untersuchung zu explorieren. Der Einfachheit halber werden diese Facetten der täglichen Berufspraxis von Lehrkräften im Folgenden unter dem Begriff „Ressourcen“ zusammengefasst. Im Rahmen der vorliegenden Arbeit ist der Ressourcenbegriff also keiner ausformulierten Theo-

rie entlehnt, sondern dient – in Anlehnung an die Terminologie von Vogelsang & Reinhold (2013) – als möglichst neutraler Globalbegriff für Lehrerwissen und -können im Kontext schulischer Leistungsfeststellung und -beurteilung.

- Zweitens werden, wie bereits erläutert, sowohl Schüler_innen, als auch Lehrer_innen im Physikunterricht nicht nur mit anspruchsvollen fachlich-konzeptuellen, sondern auch mit hohen sprachlichen Anforderungen konfrontiert. Allerdings lassen sich auf Grundlage bisheriger Forschung in vielerlei Hinsicht lediglich Mutmaßungen darüber anstellen, inwieweit Lehrkräften in ihrer täglichen Berufspraxis ein adäquater Umgang mit sprachlichen Anforderungen in Leistungssituationen im Physikunterricht gelingt (vgl. Unterkapitel 3.2). Auf Grundlage des Fallbeispiels von Tajmel lässt sich jedoch die begründete Vermutung aufstellen, dass Physiklehrkräfte während des Prozesses zur Genese von Leistungsurteilen fachlich-konzeptuelle und sprachliche Leistungen miteinander konfundieren, insbesondere, dass selbst fachlich richtige oder zumindest anschlussfähige Denkfiguren von Schüler_innen eine Relativierung erfahren (vgl. Abschnitt 3.2.2). Neben der Frage nach den Ressourcen zur Genese fachlich-konzeptueller und sprachlicher Leistungurteile durch Physiklehrer_innen erscheint es daher zudem relevant sich der Frage zu widmen, inwieweit sich die eben benannte Mutmaßung einer Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile durch Lehrkräfte empirisch erhärten lässt.

Hinzu kommt, dass insbesondere die Forschung zu berufsbezogenen Überzeugungen von Lehrkräften zu schulischer Leistungsfeststellung und -beurteilung aufdecken konnte, dass Sekundarstufenlehrkräfte für die Feststellung und Beurteilung von Schülerleistungen tendenziell vor allem auf schriftliche Klassenarbeiten zurückgreifen und auf diese als Informationsquelle vertrauen (vgl. Marso & Pigge, 1993, S. 149 u. f.; siehe auch Abschnitt 2.1.3). Vor diesem Hintergrund wird die Angebrachtheit einer Fokussierung der eben aufgeworfenen Frage offenkundig: Es erscheint aus Sicht fachdidaktischer Forschung relevant, diese beiden Fragen insbesondere für die Feststellung und Beurteilung von schriftlicher Schülerleistung, z. B. bei der Korrektur von Klassenarbeiten, als eine typische, schulische Leistungsfeststellung und -beurteilung betreffende berufliche Handlungsepisode von (Physik-)Lehrkräften zu stellen. Der Begriff „Korrektur“ ist dabei eine in der schulischen Praxis gängige Bezeichnung für den Prozess und/oder das Produkt einer Leistungsfeststellung und -beurteilung durch Lehrkräfte im Rahmen einer Klassenarbeit (vgl. König, 2017, S. 15). In dieser Bedeutung wird der Korrekturbegriff in der vorliegenden Arbeit verwendet[61].

Ziel des empirischen Teils der vorliegenden Arbeit ist, sich den eben explizierten Desideraten physikdidaktischer Forschung zu widmen. Im weiteren Verlauf stehen daher folgende Forschungsfragen im Zentrum der Darstellung:

[61] Für eine detaillierte erziehungswissenschaftliche Bestimmung des Korrekturbegriffs siehe Mbaye (2018, S. 81 u. f.).

Forschungsfragen

(F1) Welche Ressourcen werden von Physiklehrkräften bei schriftlichen, aus einer Klassenarbeit stammenden Schülerleistungen zur Genese fachlich-konzeptueller und sprachlicher Leistungsurteile eingesetzt?

(F2) Inwieweit findet im Rahmen einer Klassenarbeit bei der Feststellung und Beurteilung von schriftlichen Schülerleistungen durch Physiklehrkräfte eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile statt?

4.2. Erste Vorüberlegungen zu Design und zum methodischen Vorgehen der Untersuchung

Die Forschungsfragen (F1) und (F2) sprechen für ein komplexes Studiendesign:

- Einerseits ist eine Studie mit explorativem Charakter erforderlich, weil der Sachverhalt, den Forschungsfrage (F1) adressiert, weitgehend unbekannt ist bzw. zu Frage (F2) lediglich vage Vermutungen angestellt werden können (vgl. Diekmann, 2013, S. 33 u. f.). Dies bedeutet allerdings nicht, dass die Forschungsfragen der vorliegenden Untersuchung losgelöst von jeglicher theoretischer Fundierung sind. Beiden liegt das Konstrukt einer Assessment Literacy von Lehrkräften als deren Logik des Handelns im Kontext von schulischer Leistungsfeststellung und -beurteilung als Referenzrahmen zugrunde (vgl. Abschnitt 2.2.4).
- Andererseits ist ein deskriptives Studiendesign anzustreben, da beide Forschungsfragen „weniger auf die Erforschung sozialer Zusammenhänge und Verhaltensursachen [zielen] als vielmehr auf die Schätzung von [...] Merkmalen der Verteilung sozialer Aktivitäten [...] in einer Bevölkerungsgruppe“ (ebd., S. 35), nämlich die eines besonderen Aspekts der vielschichtigen Alltagspraxis von Physiklehrkräften bei der Genese von schulischen Leistungsurteilen.

Ergo: Aufgrund der Komplexität der Forschungsfragen (F1) und (F2) ist ein Studiendesign erforderlich, dass sich zum einen als explorativ, zum anderen als deskriptiv charakterisieren lässt.

Während in rein explorativen Untersuchungen tendenziell eher qualitative Datenerhebungs- und Auswertungsmethoden verwendet werden, kommen in ausschließlich deskriptiv angelegten Untersuchungen vornehmlich quantitative Methoden zum Einsatz (vgl. Kuckartz, 2014, S. 61). Schon deshalb wurde zu Beginn des Forschungsprozesses deutlich, dass weder eine Festlegung auf ein rein qualitatives Vorgehen, noch auf eine ausschließlich quantitative Verfahrensweise, bezogen auf das Erkenntnisinteresse der geplanten Untersuchung, adäquat ist. Es scheint eher eine Kombination aus qualitativen und quantitativen Datenerhebungs- und Auswertungsmethoden der Komplexität der bei-

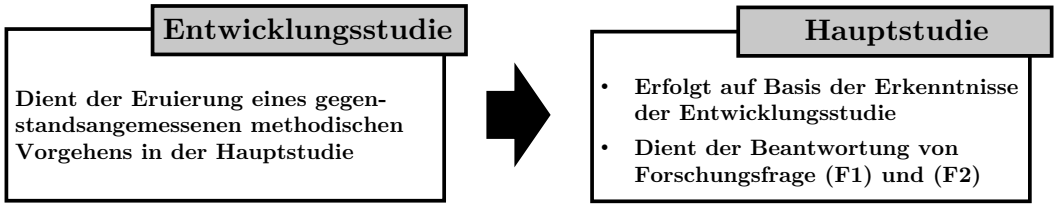


Abbildung 4.1.: Gliederung des empririschen Teils der vorliegenden Arbeit.

den Forschungsfragen gerecht zu werden. Allerdings gibt es, obwohl in der Praxis empirischer Sozialforschung bei bestimmten Studiendesigns auch ein bestimmtes methodisches Vorgehen in Teilen bevorzugt wird, „keine feste Verbindung zwischen Studientyp und Methodentyp“ (Kuckartz, 2014, S. 61). Die methodentheoretische Literatur empfiehlt daher inzwischen einhellig die präzise Wahl eines methodischen Vorgehens nicht ausschließlich auf Grundlage des angestrebten Studiendesigns vorzunehmen, sondern vor allem aufgrund seiner *Gegenstandsangemessenheit* (vgl. Stefer, 2013, S. 61 u. f.; Kuckartz, 2014, S. 50 u. f.). Der Begriff „Gegenstandsangemessenheit“ meint dabei, dass die gewählten Datenerhebungs- und Auswertungsmethoden jedwede für ein Forschungsvorhaben relevanten „Informationen erfassen, gleichzeitig die Besonderheiten von Forschungsgegenstand [...] und Forschungsfrage [...] sowie weitere beeinflussende Faktoren (z. B. Zeit- und Ressourcenbeschränkungen) berücksichtigen“ (Stefer, 2013, S. 61-62).

Alles in allem lässt sich also Folgendes feststellen: Aus den Forschungsfragen (F1) und (F2) lässt sich zwar unmittelbar die Notwendigkeit eines deskriptiv-explorativen Studiendesigns ableiten, jedoch keine genaue Beschreibung davon, wie die geplante Untersuchung gegenstandsangemessen anzulegen ist. Vielmehr offenbart sich, dass aufgrund der Komplexität der beiden Forschungsfragen der eigentlichen Hauptuntersuchung, Zwecks der Eruierung des methodischen Vorgehens, eine eigens hierfür angelegte Auseinandersetzung vorgeschaltet werden muss. Es ist also zunächst eine Entwicklungsstudie durchzuführen, die der folgenden Leitfrage folgt:

Leitfrage der Entwicklungsstudie

Welches methodische Vorgehen ist für die empirische Hauptstudie der vorliegenden Arbeit gegenstandsangemessen, ist also sowohl zu den Forschungsfragen (F1) und (F2), als auch zu den Eigenschaften des Untersuchungsgegenstands (die Alltagspraxis von Physiklehrkräften bei der Genese von schulischen Leistungsurteilen im Rahmen einer Klassenarbeit) passgenau und ermöglicht eine Verbindung zwischen Forschungsfragen und Untersuchungsgegenstand herstellen zu können?

Aufgrund des eben verdeutlichten Umstands gliedert sich der weitere Verlauf des empirischen Teils der vorliegenden Arbeit in zwei Teile (vgl. Abbildung 4.1): Als erstes wird in Kapitel 5 die Entwicklungsstudie, deren Notwendigkeit eben begründet wurde, ausführlich beschrieben. Anschließend erfolgt in Kapitel 6 die Darstellung der Hauptstudie, die auf Basis der Erkenntnisse der Entwicklungsstudie durchgeführt wurde.

5. Entwicklungsstudie

5.1. Gesamtüberblick über die Entwicklungsstudie

Der Ablauf der Entwicklungsstudie lässt sich in drei zeitlich aufeinanderfolgende Phasen gliedern (vgl. Abbildung 5.1): In der ersten Phase der Entwicklungsstudie galt es das methodische Vorgehen in der Hauptstudie durch die Wahl gegenstandsangemessener Erhebungsmethoden zu präzisieren. Zwecks dessen wurde eine erste Skizze einer für die Beantwortung der Forschungsfragen (F1) und (F2) geeigneten Laborsituation entwickelt. In dieser Laborsituation nehmen im Schuldienst aktive Physiklehrkräfte Leistungsfeststellungen und -beurteilungen von vier kontrastierenden Schülerlösungstexten zu einer Klassenarbeitsaufgabe vor. Die hierfür notwendigen Schülerlösungstexte wurden in Phase 2 der Entwicklungsstudie gewonnen. Schließlich wurde in der letzten Phase der Entwicklungsstudie die zuvor entwickelte Skizze der Laborsituation zu einem genauen Ablaufplan weiterentwickelt, sowie die in der Laborsituation (zur Datenerhebung) benötigten Materialien zusammengestellt. Dabei wurden der Ablaufplan und die Materialien mit angehenden Physiklehrkräften pilotiert. Ferner wurden in Phase 3 der Entwicklungsstudie Vorüberlegungen zu gegenstandsangemessenen Auswertungsmethoden für die in der Laborsituation erhobenen Daten unternommen.

In den folgenden Unterkapiteln werden die in den drei Phasen der Entwicklungsstudie vorgenommenen Gedankengänge bzw. das in ihnen gewählte Herangehen detailliert beschrieben. Anzumerken ist, dass die drei Phasen der Entwicklungsstudie, auch wenn Abbildung 5.1 dies zunächst suggeriert, nicht als scharf voneinander abgegrenzte Arbeitsschritte zu verstehen sind. Vielmehr kam es im Verlauf der Entwicklungsstudie zu einer zeitlichen Überlappung der einzelnen Phasen. Die in der vorliegenden Arbeit vorgenom-

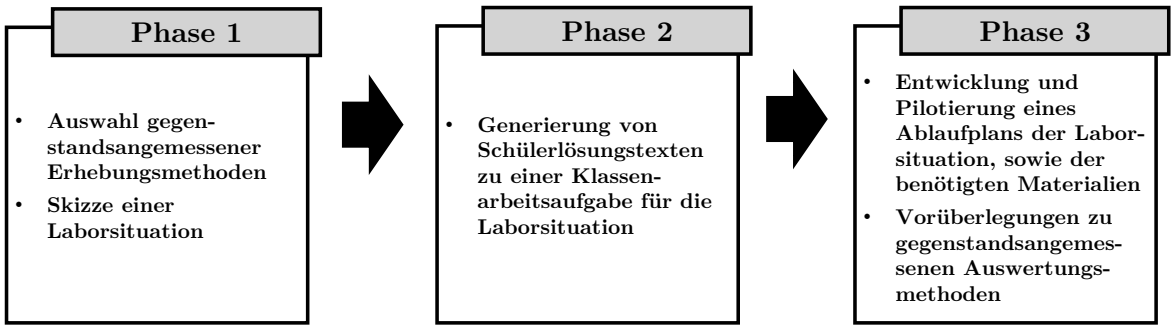


Abbildung 5.1.: Phasen der Entwicklungsstudie.

mene Gliederung in die Unterkapitel 5.2, 5.3 und 5.4 stellt also lediglich ein grobes Raster des Ablaufs der Entwicklungsstudie dar, das aus Gründen der besseren Lesbarkeit und Nachvollziehbarkeit gewählt wurde und gibt den tatsächlich Forschungsprozess nur bis zu einem gewissen Grad wirklichkeitsgetreu wieder.

5.2. Phase 1: Präzisierung des methodischen Vorgehens in der Hauptstudie durch die Wahl gegenstandsangemessener Erhebungsmethoden

5.2.1. Zur Gegenstandsangemessenheit einer authentischen Laborsituation

Die Forschungsfragen (F1) und (F2) adressieren die Genese fachlich-konzeptueller und sprachlicher Leistungsurteile durch Physiklehrkräfte im Rahmen einer Klassenarbeit und erfordern ein deskriptiv-exploratives Studiendesign (vgl. Unterkapitel 4.2). Ferner zeichnet sich bereits ab, dass erst eine Kombination aus qualitativen und quantitativen Forschungsmethoden der Komplexität beider Forschungsfragen gerecht wird. Hierdurch können sowohl vergleichbare, als auch unterschiedliche Aspekte des Gegenstandes beider Forschungsfragen ausgeleuchtet und (potenziell) konvergierende, komplementäre oder auch widersprechende Teilbefunde gewonnen werden (vgl. Kelle & Erzberger, 2015, S. 307 u. f.), woraus schließlich ein kaleidoskopartiges Gesamtbild der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile durch Physiklehrkräfte abgeleitet werden kann (vgl. H.-H. Krüger & Pfaff, 2004, S. 162). Bezogen auf die Datenerhebung sind für eine solche Kombination qualitativer und quantitativer Forschungsmethoden verschiedenste Realisierungsmöglichkeiten denkbar. Folgende Realisierungsmöglichkeiten erscheinen zunächst naheliegend, bei einer genaueren Betrachtung zeigt sich jedoch, dass diese den Forschungsfragen (F1) und (F2) nicht gegenstandsangemessenen sind:

- Auf den ersten Blick erscheint die Durchführung von qualitativen Interviews (vgl. Hopf, 2015) mit im Schuldienst aktiven Physiklehrkräften, in denen diese mit Fragen zu ihrer Alltagspraxis bei der Leistungsfeststellung und -beurteilung im Rahmen von Klassenarbeiten konfrontiert werden, eine naheliegende und gegenstandsangemessene Methode zur Datenerhebung darzustellen. Für ein solches Vorgehen spricht, dass die hierbei gewonnenen Antworten der befragten Lehrkräfte sowohl qualitativ, als auch quantitativ analysiert werden können. Beispielsweise ließe sich (im Sinne eines Mixed-Methods-Transferdesigns; vgl. Kuckartz, 2014, S. 87 u. f.) der Sinngehalt der Antworten durch eine geeignete qualitative Codierung aus dem Kommunikationsmaterial gewinnen und im Anschluss z. B. das Auftreten oder die Häufigkeit bestimmter Codierungen mit Hilfe adäquater statistischer Methoden analysieren (vgl. ebd.). Gegen ein derartiges Vorgehen spricht allerdings, dass davon auszugehen ist, dass das Vornehmen von Leistungsfeststellungen und -beurteilungen eine typischere Aktivität

für im Schuldienst aktive Physiklehrkräfte darstellt, als das Beschreiben des eigenen Handelns beim Feststellen und Beurteilen schriftlicher Schülerleistungen gegenüber einem_einer Dritten. Daher muss damit gerechnet werden, dass es im Schuldienst aktiven Physiklehrkräften schwer fallen könnte, Fragen zu beantworten, die sich auf ihr Vorgehen im Berufsalltag bei der Feststellung und Beurteilung schriftlicher, aus einer Klassenarbeit stammender Schülerleistungen beziehen und/oder ihren dabei verfolgten Logiken und/oder ihre zu diesem Zwecke angewandten Maßstäbe. Ferner spricht gegen eine Durchführung von Interviews mit im Schuldienst aktiven Physiklehrkräften, dass ein solches Vorgehen, ähnlich wie der Selbstauskunftsansatz der Forschung um diagnostische Kompetenzen von Lehrer_innen, vornehmlich Hinweise auf die subjektive Selbstwahrnehmung einer Lehrkraft liefern würde (vgl. Unterabschnitt 2.1.1.3). Die Forschungsfragen (F1) und (F2) adressieren allerdings weniger die Selbstwahrnehmung von Lehrkräften im Kontext schulischer Leistungsfeststellung und -beurteilung (als Teil ihrer berufliche Teilidentität als Assessor of Learning), sondern vielmehr die Assessment Literacy von Physiklehrkräften als eine Form von sich unter anderem aus Wissen 2 speisendem Wissen 3 (vgl. Abschnitt 2.2.4). Aus den eben aufgeführten Argumenten wird also deutlich, dass eine Durchführung von qualitativen Interviews mit im Schuldienst aktiven Physiklehrkräften alles in allem keine für die Forschungsfragen (F1) und (F2) gegenstandsangemessene Forschungsmethode darstellt.

- Eine andersartige, naheliegende Herangehensweise wäre eine ethnographische Beobachtung von im Schuldienst aktiven Physiklehrkräften, während sie die Feststellung und Beurteilung schriftlicher, aus einer Klassenarbeit stammender Schülerleistungen vornehmen (vgl. Lüders, 2015). Bei einem derartigen Vorgehen ist durchaus erwartbar reichhaltige und authentische Einblicke in die alltägliche Leistungsfeststellungs- und -beurteilungspraxis von Physiklehrkräften gewinnen zu können[62]. Für eine ethnographische Herangehensweise spricht zudem, dass...

  > „[...] Ethnografie kein einzelnes Verfahren [bezeichnet], sondern es handelt sich vielmehr um einen Sammelbegriff, der die Anwendung des ganzen Arsenals an Methoden unterstützt, welche die Sozialforschung zu bieten hat, unabhängig davon, ob diese dem qualitativen oder quantitativen Paradigma zuzuordnen sind. Durch das methodenplurale Vorgehen der Ethnografie wird der Anspruch nach Gegenstandsangemessenheit der Methodik am strengsten gewährleistet.“ (Thomas, 2010, S. 466-467)

  Allerdings würden bei einem solchen Vorgehen verschiedene Physiklehrkräfte in bezüglich ihrer makroskopischen und mikroskopischen Kontextbedingungen (mutmaßlich) höchst unterschiedlichen Handlungsepisoden beobachtet werden (vgl. Abschnitt 2.2.4). Ein unmittelbarer Vergleich zwischen den untersuchten Lehrkräften und damit eine Verallgemeinerbarkeit von Befunden würde hierdurch deutlich erschwert. Hinzu kommt, dass durch eine Beobachtung im Feld nur ein unvollständiges Bild des Geneseprozesses von Leistungsurteilen durch Physiklehrkräfte erfasst wer-

[62] Diese Annahme wird untermauert durch Einblicke, die z. B. die ethnographische Fallstudie von Kalthoff (1996) zur Leistungsfeststellung und -beurteilung von Lehrkräften im Kontext schriftlicher Klassenarbeiten und mündlicher Abiturprüfungen liefert.

den kann. Bei diesem Geneseprozess handelt es sich zu einem Großteil um eine gedankliche (und damit für Außenstehende durch bloße Beobachtung nicht zugängliche) Auseinandersetzung einer Lehrkraft mit dem Problem, das Lernen von Schüler_innen in Relation zu einem Gütemaß zu stellen (vgl. Abschnitt 2.1.4). Ähnlich wie eine Durchführung von qualitativen Interviews, entpuppt sich also auch eine ethnographische Beobachtung von im Schuldienst aktiven Physiklehrkräften als kein für die Forschungsfragen (F1) und (F2) gegenstandsangemessenes Vorgehen.

- Selbiges gilt auch für eine Adaption der überwiegend quantiativ orientierten Erhebungsmethoden von Studien, die Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen auf Grundlage des Linsenmodells oder des Prozessmodells von Nickerson untersuchten (vgl. Abschnitt 2.1.4). Das Vorgehen bisheriger Studien zur Genese von Lehrerleistungsurteilen auf Basis des Modells von Nickerson scheidet deswegen aus, da sich die dort angewandten Methoden am hypothesentestenden Erkenntnisinteresse dieser Untersuchungen orientierten. Forschungsfrage (F1) und (F2) verlangen aber ein methodischen Vorgehen, das vornehmlich die Exploration und Deskription der weitgehend unerforschten Alltagspraxis von Physiklehrkräften bei der Genese von schulischen Leistungsurteilen im Rahmen einer Klassenarbeit ermöglicht. Eine methodische Herangehensweise auf Grundlage des Linsensmodells kommt deshalb nicht infrage, da hierzu konkrete Annahmen über die „Gestalt“ der proximalen Merkmals-Linsen von Physiklehrkräften bei der Feststellung und -beurteilung schriftlicher, aus einer Klassenarbeit stammender Schülerleistungen getroffen werden müssten, um diese bei der Datenerhebung erfragen zu können. Solche Annahmen können aus der bisherigen Forschung zum Wissen und Können von Physiklehrkräften zu schulischer Leistungsfeststellung und -beurteilung jedoch nicht vorgenommen werden. Vorteilhaft an der Herangehensweise von Studien, deren theoretische Grundlage das Nickersonschen Prozessmodell oder das Linsenmodell darstellte, ist allerdings, dass diese Lehrerwissen und -können zu schulischen Leistungsfeststellungen und -beurteilungen unter Laborbedingungen untersuchten, die angelehnt an schulische Leistungsfeststellung und -beurteilung betreffende berufliche Handlungsepisoden von Lehrkräften ausgestaltet waren. Ein derartiges methodisches Vorgehen bei der Datenerhebung erleichtert einen unmittelbaren Vergleich zwischen den teilnehmenden Lehrkräften und begünstigt damit das Verallgemeinerbarkeitspotenzial von Befunden, die in der Datenauswertung gewonnen werden.

Die Darstellung verschiedener zunächst vielversprechender, bei einer genaueren Betrachtung jedoch den Forschungsfragen (F1) und (F2) nicht gegenstandsangemessener Erhebungsmethoden ist keineswegs vollständig. Aus der eben vorgenommenen beispielhaften Erörterung der Stärken und Schwächen unterschiedlicher Forschungsmethoden, bezogen auf das Erkenntnisinteresse des empirischen Teils der vorliegenden Arbeit, lassen sich aber bereits die wesentliche Charakteristika eines für die Hauptstudie gegenstandsangemessenen methodischen Vorgehens bei der Datenerhebung ableiten: Es bedarf eines Datenerhebungsverfahrens, das Rückschlüsse auf die Denkprozesse von Physiklehrkräften

bei der Feststellung und Beurteilung schriftlicher Schülerleistungen aus einer Klassenarbeit ermöglicht, um die Genese fachlich-konzeptueller und sprachlicher Leistungsurteile von Physiklehrkräften über schriftliche, aus einer Klassenarbeit stammende Schülerleistungen, sowie eine dabei eventuell auftretende Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile explorieren und deskriptiv beschreiben zu können. Ferner erscheint eine Untersuchung von im Schuldienst aktiven Physiklehrkräften in einer Laborsituation angemessener als eine Felduntersuchung, um das Verallgemeinerbarkeitspotenzial der in der Datenauswertung gewonnenen Befunde zu begünstigen. Zur Sicherstellung der ökologischen Validität sollte diese Laborsituation allerdings der Alltagspraxis von Physiklehrkräften bei der Leistungsfeststellung und -beurteilung im Rahmen einer Klassenarbeit möglichst ähnlich sein.

5.2.2. Erste Skizze einer für die Beantwortung der Forschungsfragen (F1) und (F2) geeigneten Laborsituation

Die am Ende des vorherigen Abschnitts aufgeführten Charakteristika dienten zu Beginn der ersten Phase der Entwicklungsstudie als Leitidee, aus der die folgende erste Skizze eines für die Hauptstudie gegenstandsangemessenen Datenerhebungsverfahrens abgeleitet wurde:

Man stelle sich folgende Situation vor (vgl. Abbildung 5.2): Im Rahmen einer Laborsituation werden verschiedenen, im Schuldienst aktiven Physiklehrkräften Schülerlösungen zu einer Klassenarbeitsaufgabe vorgelegt. Die teilnehmenden Lehrkräfte werden gebeten diese Schülerlösungen so zu korrigieren, wie sie dies in ihrem Berufsalltag unter normalen Umständen auch tun würden und für angemessen halten. Jeder Lehrkraft werden dabei dieselben vier Schülerlösungen vorgelegt, die zuvor anhand von zwei Kriterien ausgewählt wurden:

1. Bei den vier Schülerlösungen handelt es sich um Texte, die einen physikalischen Sachverhalt erklären[63] und zusätzlich weder zeichnerische, noch rechnerische Elemente enthalten. Hierdurch wird sichergestellt, dass sich die Schülerlösungen, die den Lehrkräften vorgelegt werden, kriterial vor allem bezüglich fachlich-konzeptueller und sprachlicher Merkmale voneinander unterscheiden lassen.
2. Bei den Texten handelt es sich um in zweierlei Hinsicht kontrastierende Schülerlösungen. Sie unterscheiden sich zum einen bezüglich ihrer fachlich-konzeptuellen Qualität (z. B. in ihrer fachwissenschaftlichen Richtigkeit). Zum anderen besteht ein Unterschied zwischen den Schülertexten bezogen auf ihre sprachliche Realisierung

[63] Erklären meint in diesem Zusammenhang „das Zurückführen eines Phänomens auf ein zugrundeliegendes Prinzip“ (Kulgemeyer & Tomczyszyn, 2015, S. 112). Im Kontext von naturwissenschaftlichen Fachunterricht ist Erklären vom Argumentieren zu unterscheiden (vgl. ebd., S. 114). „Bei Argumentationen werden andere Maßstäbe akzeptiert als bei Erklärungen. Während bei Erklärungen nur solche allgemeinen Gesetze als Begründungen akzeptiert werden, die empirisch überprüfbar und allgemein wissenschaftlich als wahr anerkannt sind [...], werden bei Argumentationen (insbesondere im Alltag) auch Plausibilitäten oder ad hoc konstruierte Gesetze verwendet“ (ebd.).

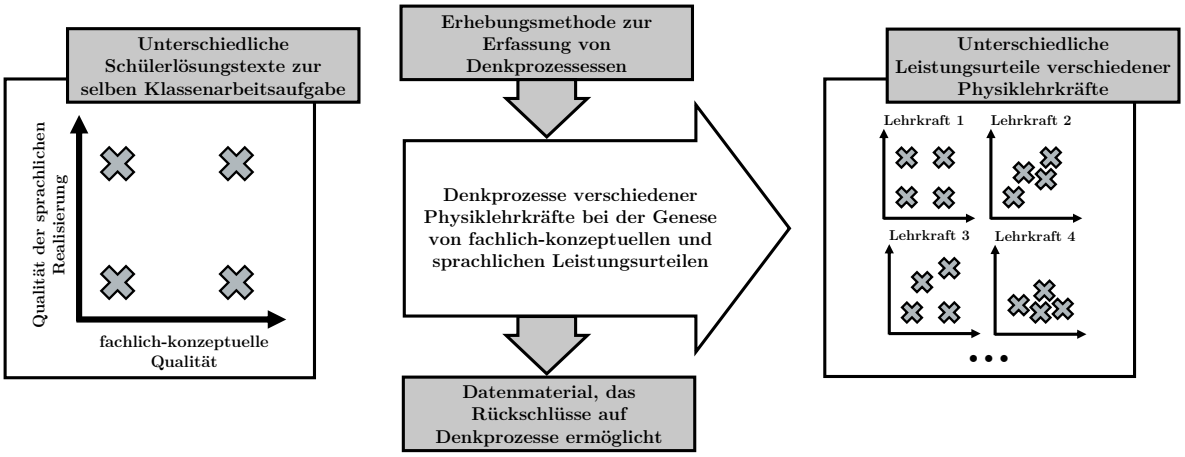


Abbildung 5.2.: Erste Skizze eines gegenstandsangemessenen methodischen Vorgehens in der Hauptstudie der vorliegenden Arbeit.

(entspricht die verwendete Sprache der Schüler_innen z. B. der „legitimen Sprache“ des Physikunterrichts; vgl. Unterabschnitt 3.1.2.3). Die vier Schülerlösungstexte sind dabei, wie im linken Kasten in Abbildung 5.2 dargestellt, in einem zweidimensionalen Koordinatensystem verortbar. Die horizontale Achse dieses Koordinatensystems beschreibt die fachlich-konzeptuelle Qualität eines Schülertextes und die vertikale Achse die Qualität der sprachlichen Realisierung eines Schülertextes. Die vier Schülerlösungstexte wurden so gewählt, dass sie in diesem Koordinatensystem die Eckpunkte eines Rechtecks bilden (Kreuze im linken Kasten in Abbildung 5.2): Zu jedem Schülerlösungstext gibt es jeweils genau einen weiteren Text mit vergleichbarer fachlich-konzeptueller Qualität, jedoch deutlich unterschiedlicher sprachlicher Realisierung (höhere bzw. niedrigere Qualität), einen Text der lediglich bezogen auf die Qualität seiner sprachlichen Realisierung mit besagtem Schülertext vergleichbar ist, sowie einen dritten, der sich sowohl auf fachlich-konzeptueller Ebene, als auch bezüglich seiner sprachlichen Realisierung deutlich vom entsprechenden Schülertext unterscheidet.

Aufgrund der kontrastierenden Auswahl der vier Schülerlösungstexte ist es plausibel anzunehmen, dass die im Rahmen der Laborsituation befragten Physiklehrkräfte bei der Genese ihrer Leistungsurteile auf unterschiedliche Ressourcen zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen zurückgreifen und hierdurch auch zu unterschiedlichen Leistungsurteilen (Korrekturen) der vier Schülerlösungstexte gelangen. Sofern die Physiklehrkräfte fachlich-konzeptuelle und sprachliche Leistungsurteile miteinander konfundieren, kann ferner damit gerechnet werden, dass die finale Leistungsfeststellung und -beurteilung der vier Schülerlösungstexte durch die Befragten – im Vergleich zur ursprünglichen Auswahl der Schülerlösungen – mehr oder minder „verzerrt“ ist. Bei beiden Annahmen ist zudem eine individuelle Ausprägung denkbar, weswegen nicht auszuschließen ist, dass die finalen Leistungsurteile der einzelnen Physiklehrkräfte im Rahmen der Laborsituation mehr oder minder verschieden voneinander sind

(exemplarisch veranschaulicht im rechten Kasten in Abbildung 5.2). Die Frage ob einander ähnlichen finalen Leistungsfeststellungen und -beurteilungen unterschiedlicher Lehrkräfte ein bis zu einem bestimmten Grad gleichartiger Geneseprozess fachlich-konzeptueller und sprachlicher Leistungsurteile zugrunde liegt, lässt sich beantworten, indem Rückschlüsse auf die Denkprozesse der befragten Physiklehrkräfte unternommen werden, während sie eine Feststellung und Beurteilung der vier Schülertexten vornehmen. Im Rahmen der Laborsituation sind daher Erhebungsmethoden zu wählen, mit deren Hilfe Daten über die befragten Lehrkräfte gewonnen werden können, die Rückschlüsse auf deren Denkprozesse bei der Feststellung und Beurteilung der vier Schülerlösungstexte ermöglichen (vertikale Pfeile in Abbildung 5.2).

5.2.3. Vergleich der Gegenstandsangemessenheit verschiedener instrospektiver Erhebungsmethoden

Erziehungs- und sozialwissenschaftliche Erhebungsmethoden, mit denen Daten gewonnen werden können, die Rückschlüsse auf die Denkprozesse von Untersuchungsteilnehmer_innen ermöglichen, werden in der methodentheoretischen Literatur unter den Oberbegriffen „Introspektion“ (z. B. Heine & Schramm, 2016), „introspektive Verfahren“ (z. B. Heine, 2013), „kognitive Verfahren“ (z. B. Häder, 2015, S. 402 u. f.), usw. zusammengefasst. Diesen Erhebungsverfahren ist gemeinsam, dass die Untersuchungsteilnehmer_innen...

> „[...] durch lautes Aussprechen Einblicke in ihre Gedanken und Emotionen gewähren, die der Beobachtung normalerweise unzugänglich sind. Nach einem weiten Begriffsverständnis zählen hierzu alle Formen von Interviews und Tagebuchdaten [...] [mit und] ohne Bezug auf eine konkrete Handlung[.] [...] [Nach einem engeren Begriffsverständnis werden jedoch nur] solche Verfahren als introspektiv bezeichnet[,] [...] bei denen gezielt Daten bezüglich einer bestimmten (mentalen oder interaktionalen) Tätigkeit [...] während bzw. direkt im Anschluss an eine zu untersuchende Tätigkeit [erhoben werden.]“ (Heine & Schramm, 2016, S. 173)

Die drei Erhebungsverfahren, die in einem engeren Begriffsverständnis als introspektiv bezeichnet werden, sind das *laute Denken*[64], das *laute Erinnern*[65] und die *retrospektive Befragung*[66] (vgl. Heine & Schramm, 2016, S. 173 u. f.). Auch wenn diese Erhebungsmethoden zum Teil unterschiedlich verfahren, weisen sie insgesamt große Ähnlichkeiten auf. Zum einen hängt dies damit zusammen, dass im engeren Begriffsverständnis introspektive Verfahren den „gemischten“ Methoden zugeordnet werden können. Sie lassen sich...

> „[...] per se weder einem qualitativen noch einem quantitativen Paradigma zuordnen; sie k[önnen] sowohl explorativ/deskriptiv als auch interpretativ und/oder hypothesentestend (Cohen 1996; Würffel 2001) [A. Cohen, 1996; Würffel, 2001; M. S. F.] eingesetzt werden

[64] In der Literatur unter anderem auch als „simultanes lautes Denken“ (z. B. Konrad, 2010, S. 481) oder „think aloud method“ (z. B. Heine, 2010, S. 84 u. f.) bezeichnet.

[65] In der Literatur unter anderem auch als „retrospective think aloud“ (z. B. Häder, 2015, S. 403) oder „stimulated recall“ (z. B. Heine & Schramm, 2016, S. 173) bezeichnet.

[66] In der Literatur unter anderem auch als „Nachfragetechnik“ (z. B. Häder, 2015, S. 403 u. f.) oder „retrospection“ (z. B. van Someren, Barnard, & Sandberg, 1994, S. 20 u. f.) bezeichnet.

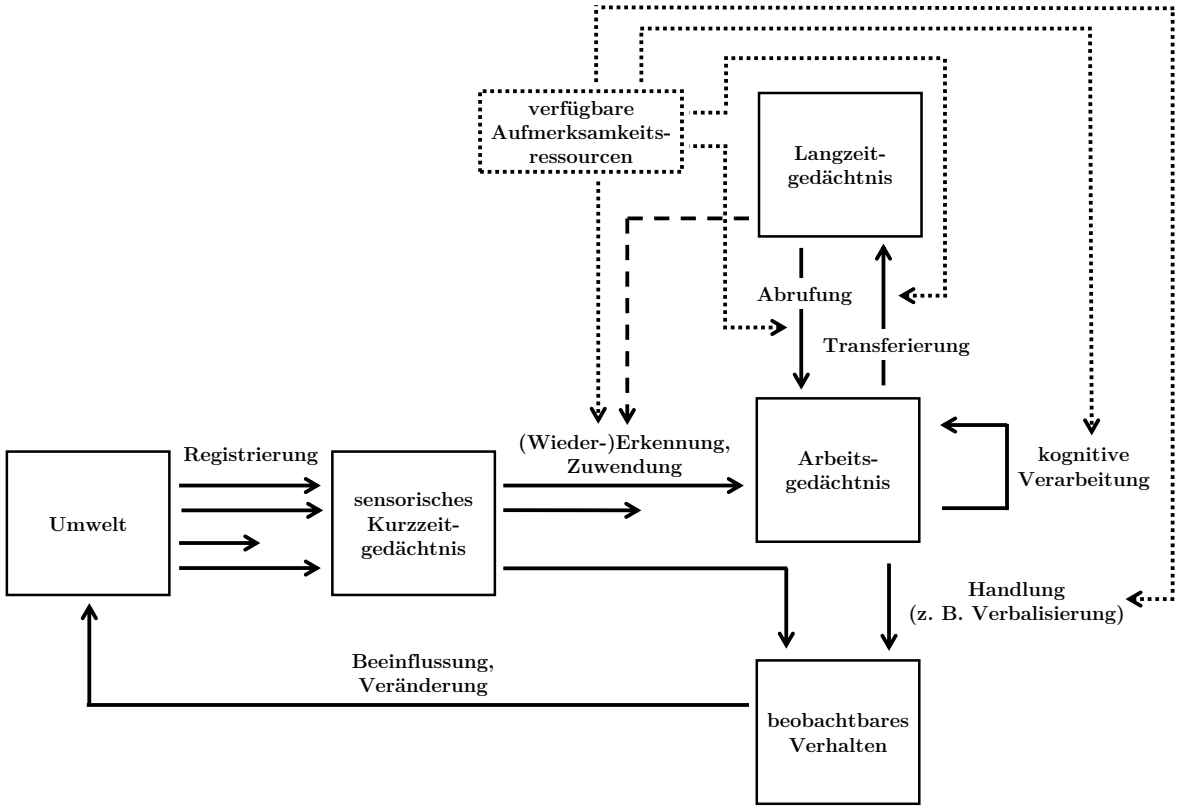


Abbildung 5.3.: Vereinfachte Version des Drei-Speicher-Modells von Wickens, Hollands, Banbury, & Parasuraman (2016, S. 4 u. f.).

> und qualitative wie auch quantitative Formen der Datensammlung und -analyse benutzen oder vereinen." (Konrad, 2010, S. 480)

Zum anderen begründet sich diese Ähnlichkeit daraus, dass alle im engeren Begriffsverständnis als introspektiv bezeichneten Erhebungsverfahren auf als gesichert geltenden Erkenntnissen der Psychologie über die Struktur des menschlichen Gedächtnisses und über den Ablauf kognitiver Informationsverarbeitung basieren (vgl. Ericsson & Simon, 1985, S. 10 u. f.; van Someren et al., 1994, S. 13 u. f.; Gilhooly & Green, 2002; Heine & Schramm, 2007, S. 168 u. f.; Liebl et al., 2012, S. 425 u. f.). Anstatt diese Erkenntnisse ausführlich theoretisch und auf Grundlage empirischer Befunde zu erörtern, sollen sie anhand einer vereinfachten Version des Drei-Speicher-Modells von Wickens, Hollands, Banbury, & Parasuraman (2016, S. 4 u. f.) veranschaulicht werden (vgl. Abbildung 5.3), da dieses Modell die Erkenntnisse der Gedächtnis- und Informationsverarbeitungspsychologie, auf denen introspektive Erhebungsmethoden beruhen, in sich vereint.

Die zentrale Annahme, auf der das Drei-Speicher-Modell von Wickens et al. (2016, S. 4 u. f.) basiert, ist das menschliche Gedächtnis in drei Teilgedächtnisse zu gliedern, die bei der Informationsverarbeitung und -speicherung unterschiedliche Aufgaben übernehmen[67]: das

[67] Den Modellcharakter dieser Annahme gilt es in besonderer Art und Weise hervorzuheben. Gemäß der ausführlichen Argumentation von Ericsson & Simon (1985) zu introspektiven Erhebungsverfahren haben „Personen nicht mehr Einblicke in ihre kognitiven Strukturen als außenstehende Beobachter

sensorische Kurzzeitgedächtnis, das *Arbeitsgedächtnis* und das *Langzeitgedächtnis* (vgl. ebd., S. 4). Das Langzeitgedächtnis stellt den dauerhaften Speicher von Gedächtnisinhalten dar, aus dem Inhalte abgerufen oder in den Informationen transferiert werden können (vgl. ebd.). Der Prozess des Abrufens und Transferierens von Informationen aus bzw. in das Langzeitgedächtnis wird dabei von den einer Person zur Verfügung stehenden Aufmerksamkeitsressourcen (die z. B. durch körperliche Erschöpfung, Gesundheitszustand, usw. bedingt sind; in Abbildung 5.3 als gepunktete Pfeile symbolisiert) moderiert (vgl. ebd., S. 5). Im sensorischen Kurzzeitgedächtnis werden alle von den Sinnen einer Person registrierten Informationen ihrer Umwelt kurzzeitig zwischengespeichert (vgl. ebd., S. 4). Nur ein Bruchteil dieser registrierten Umweltinformationen wird durch bewusste oder unbewusste Analyse ihrer Merkmale bewusst oder unbewusst (wieder-)erkannt (vgl. ebd.). Diese Merkmalsanalyse ist beeinflusst durch vergangene Erfahrungen der Person, die in ihrem Langzeitgedächtnis gespeichert sind (in Abbildung 5.3 als gestrichelter Pfeil symbolisiert), sowie durch ihre verfügbaren Aufmerksamkeitsressourcen (vgl. ebd., S. 4 u. f.). Erfolgt eine solche (Wieder-)Erkennung registrierter Umweltinformationen, kann dies entweder in einem intuitiven, unmittelbar beobachtbaren Verhalten der Person münden (z. B. das schnelle Ausweichen eines_einer Fahrradfahrers_Fahrradfahrerin in einer Gefahrensituation), oder – sofern die (Wieder-)Erkennung eines Umweltvorkommnisses dem Bewusstsein zugänglich ist – darin, dass sich die Person diesem Vorkommnis zuwendet (vgl. ebd., S. 4). Durch diese Zuwendung werden (wieder-)erkannte Informationen in das Arbeitsgedächtnis übertragen. Im Arbeitsgedächtnis findet die kognitive Verarbeitung der Informationen statt, die aus der Umwelt (wieder-)erkannt wurden (z. B. zum Zweck der Memorierung), aber auch von solchen, die aus dem Langzeitgedächtnis abgerufen worden sind (vgl. ebd., S. 4 u. f.). Beispielsweise wird eine als Frage erkannte Wortäußerung einer anderen Person inhaltlich erfasst, Antwortoptionen (aus Wissensbeständen aus dem Langzeitgedächtnis) generiert, diese gegeneinander abgewogen und die Entscheidung, eine bestimmte Antwort zu geben, getroffen. Auch dieser Prozess ist wiederum bedingt durch die einer Person zur Verfügung stehenden Aufmerksamkeitsressourcen (vgl. ebd., S. 5).

Auf Grundlage des Modells von Wickens et al. (2016, S. 4 u. f.) lassen sich zunächst das laute Denken und das laute Erinnern näher bestimmen und voneinander unterscheiden:

> „In Fällen, in denen Kognitionen direkt aus dem Arbeitsgedächtnis [...] verbalisiert werden[, spricht man vom lauten Denken,] [...] in Fällen in denen sie aus dem Langzeitgedächtnis aktiviert und nachträglich verbalisiert werden[, spricht man vom lauten Erinnern.] [...] [Lautes Denken] bezeichnet die aus dem Arbeits[gedächtnis] erfolgende simultane, ungefilterte Verbalisierung einer Person von Gedanken während einer (mentalen, interaktionalen oder aktionalen) Handlung. [...] [Lautes Erinnern] bezeichnet die aus dem Langzeitgedächtnis erfolgende nachträgliche, ungefilterte Verbalisierung einer Person von Gedanken während einer (mentalen, interaktionalen oder aktionalen) Handlung.“ (Knorr & Schramm, 2012, S. 185)

Lautes Denken und lautes Erinnern sind damit (als Erhebungsverfahren) besonders eng miteinander verwandt (vgl. Heine & Schramm, 2016, S. 173). Der wesentliche Unter-

[und können deshalb] Fragen zu ihrer kognitiven Organisation und ihren mentalen Abläufen [...] nicht valide beantwort[en]“ (Heine & Schramm, 2007, S. 171).

schied zwischen beiden Vorgehensweisen besteht darin, dass beim lauten Denken Untersuchungsteilnehmer_innen simultan zu einer Handlungssituation zur Verbalisierung ihrer Gedanken (also aus dem Arbeitsgedächtnis) aufgefordert werden, während beim lauten Erinnern die Aufforderung darin besteht, im Anschluss an eine (mentale, interaktionale oder aktionale) Handlung alle hierzu noch abrufbaren Gedächtnisinhalte als (während der Handlung im Arbeitsgedächtnis kognitiv verarbeite) Erinnerung aus dem Langzeitgedächtnis zu reaktivieren und zu verbalisieren. Mit lautem Denken werden daher Daten generiert, die Rückschlüsse auf die unreflektierte „innere Sprache[68]“ der Untersuchungsteilnehmer_innen ermöglichen (vgl. Heine & Schramm, 2007, S. 170 u. f.), wohingegen Laut-Erinnerungs-Daten Einblicke in ihre „reflection-in-action“ gewähren (vgl. Heine & Schramm, 2016, S. 178).

Eine besonders enge Verwandtschaft besteht zudem zwischen dem lauten Erinnern und der retrospektiven Befragung, da beide im Anschluss an eine (mentale, interaktionale oder aktionale) Handlung erfolgen. Im Unterschied zum lauten Erinnern werden bei einer retrospektiven Befragung Untersuchungsteilnehmer_innen allerdings nicht zur Reaktivierung und Verbalisierung all ihrer (noch abrufbaren) Erinnerungen zu einer Handlungssituation aufgefordert, sondern stattdessen gebeten Interviewfragen zu beantworten, die auf die Wiedergabe von Erinnerungen an konkrete, für das Erkenntnisinteresse der Untersuchung besonders relevante Aktivitäten in der Handlungssituation abzielen (vgl. Heine & Schramm, 2016, S. 173). Erwartbar ist, dass das gezielte Nachfragen im Rahmen einer retrospektiven Befragungen bei den den Untersuchungsteilnehmer_innen eine zusätzliche kognitive Verarbeitung ihrer Erinnerung an die Handlungssituation evoziert. Daten, die bei einer retrospektiven Befragung gewonnen werden, ermöglichen daher vor allem Rückschlüsse auf die „reflection-on-action“ der Untersuchungsteilnehmer_innen (vgl. ebd.).

Das Ziel dieses Abschnitts der vorliegenden Arbeit besteht darin, im engeren Sinn introspektive Erhebungsverfahren global, nämlich hinsichtlich ihrer Gegenstandsangemessenheit für die Forschungsfragen (F1) und (F2) zu bewerten und hierdurch eine überlegte Wahl von Erhebungsmethoden für die Hauptstudie vorzunehmen. Auf eine tiefgreifendere Erörterung und Gegenüberstellung der praktischen Umsetzung der Erhebungsmethoden lautes Denken, lautes Erinnern und retrospektive Befragung, die weitere, eher geringfügigere Unterschiede zwischen diesen Methoden aufdecken würde, wird daher an dieser Stelle der vorliegenden Arbeit verzichtet. Stattdessen folgt nun eine wie eben angekündigte Bewertung der Gegenstandsangemessenheit des lauten Denkens, des lauten Erinnerns und der retrospektiven Befragung auf Grundlage ihrer in den vorherigen Absätzen vorgenommen Kurzcharakterisierungen:

Forschungsfrage (F1) und (F2) verlangen aufgrund ihrer Komplexität ein methodenplurales Vorgehen aus qualitativen und quantiativen Forschungsmethoden (vgl. Abschnitt

[68] „Innere Sprache“ meint in diesem Zusammenhang, „dass viele Elemente einer [...] Gedankenfolge spontan und ohne zusätzlichen kognitiven Aufwand mit einer verbalen Form assoziiert werden, die normalerweise [...] unvokalisiert bleibt. Die Annahme ist nun, dass es prinzipiell möglich ist, diese verbalen Gedanken laut auszusprechen (wobei nicht zum Zweck der Kommunikation an ein Gegenüber oder an sich selbst gerichtete Sprache [...], sondern unreflektiertes lautes Mitvokalisieren ablaufender Gedanken gemeint ist)“ (Heine, 2013, S. 14).

5.2.1). Da lautes Denken, lautes Erinnern und retrospektive Befragung als „gemischte“ Methoden charakterisiert werden können (siehe oben), eignen sie sich daher auch besonders für eine Kombination aus qualitativen und quantitativen Datenerhebungs- und Auswertungsmethoden. Zudem ist bei keiner der drei im engeren Begriffsverständnis introspektiven Verfahren per se eine größere Eignung für ein methodenpluares Vorgehen erkennbar (vgl. Konrad, 2010, S. 480 u. f.). Lautes Denken, lautes Erinnern und retrospektive Befragung erscheinen daher zunächst, bezogen auf das Erkenntnisinteresse der Hauptstudie der vorliegenden Arbeit, in gleicher Weise gegenstandsangemessen zu sein.

Aus einer anderen Perspektive zeigen sich allerdings auch Unterschiede zwischen diesen drei Verfahren bezogen auf ihre Gegenstandsangemessenheit: Forschungfrage (F1) und (F2) adressieren die Genese fachlich-konzeptueller und sprachlicher Leistungsurteile von Physiklehrkräften im Rahmen einer Klassenarbeit. Sowohl eine Aufforderung zum lauten Erinnern, als auch eine retrospektive Befragung von Untersuchungsteilnehmer_innen erscheint diesbezüglich zunächst gegenstandsangemessen. Beide Erhebungsmethoden generieren Daten, die Einblicke in das (mentale, interaktionale oder aktionale) Handeln in einer konkreten Situation gewähren. Gleichzeitig erscheinen beide Methoden bezüglich ihrer Gegenstandsangemessenheit auch fragwürdig: Wie bereits dargestellt, ermöglichen lautes Erinnern und retrospektives Befragen vor allem Rückschlüsse auf die „reflection-in-action“ bzw. „reflection-on-action“ von Untersuchungsteilnehmer_innen. Durch lautes Erinnern und retrospektive Befragung gewonnene Daten enthalten daher in Teilen auch immer die subjektive Selbstwahrnehmung der Untersuchungsteilnehmer_innen, inwieweit sie in einer bestimmten Situation handlungsfähig sind. Diese subjektive Selbstwahrnehmung wird jedoch von Forschungsfrage (F1) und (F2) nicht primär adressiert (vgl. Abschnitt 5.2.1). Werden hingegen im Schuldienst aktive Physiklehrkräfte in einer Laborsituation gebeten, schriftliche Schülerlösungen zu einer Klassenarbeitsaufgabe ihren Gewohnheiten entsprechend zusätzlich allerdings laut denkend zu korrigieren, werden die Untersuchungsteilnehmer_innen ihre Gedankenabfolgen während ihrer Leistungsfeststellung und -beurteilung unreflektiert mitvokalisieren. Die so gewonnenen Daten ermöglichen dann unmittelbare Rückschlüsse auf die Ressourcen, auf die die teilnehmenden Physiklehrkräfte bei der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile zurückgreifen und inwieweit Physiklehrkräfte bei diesem Geneseprozess verschiedene Teilleistungsurteile miteinander konfundieren (z. B. fachlich-konzeptuelle oder sprachliche Leistungsurteile).

Alles in allem ist damit ein lautes Denken der teilnehmenden Physiklehrkräfte, während sie die vier Schülertexte korrigieren, den Forschungsfragen (F1) und (F2) im Vergleich zu den anderen beiden im engeren Begriffsverständnis introspektiven Erhebungsverfahren am gegenstandsangemessensten. Dennoch sollten die an der Hauptstudie teilnehmenden Physiklehrkräfte im Anschluss an das laut denkende Korrigieren darum gebeten werden, eine Einschätzung der fachlich-konzeptuellen Qualität, sowie der Qualität der sprachlichen Realisierung der vier Schülertexte vorzunehmen und dabei ihre eigene Einschätzung explizit zu begründen. Für eine derartige zusätzliche retrospektive Befragung sprechen zwei Reichhaltigkeitsargumente. Diese Reichhaltigkeitsargumente beziehen sich auf mögliche Befunde, die nur durch eine Triangulation von Daten (vgl. Denzin, 1970, S. 301 u. f.;

Flick, 2011, S. 36 u. f.), die durch lautes Denken und retrospektive Befragung erhoben wurden, gewonnen werden können:

1. **Steigerung der Reichhaltigkeit durch komplementäre Teilfunde**:
   Forschungsfragen (F1) und (F2) fokussieren die Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Leistungen durch Physiklehrkräfte. Auch wenn den Untersuchungsteilnehmer_innen hierzu Schülerlösungen aus einer Klassenarbeit vorgelegt werden, die sich kriterial vor allem bezüglich fachlich-konzeptueller und sprachlicher Merkmale voneinander unterscheiden lassen (vgl. Abschnitt 5.2.2), ist es zumindest denkbar, dass die Untersuchungsteilnehmer_innen bei der Mitvokalisierung ihrer Gedanken während der Genese ihrer Leistungsfeststellung und -beurteilung nur wenig zwischen fachlich-konzeptuellen und sprachlichen Merkmalen eines Schülertextes differenzieren. In einem solchen Fall ließen sich aus den Laut-Denk-Daten Ressourcen zur fachlichen und sprachlichen Leistungsfeststellung und -beurteilung nur wenig differenzieren, sowie kaum auf eine eventuelle Konfundierung fachlicher-konzeptueller und sprachlicher Leistungsurteile schließen. Forschungsfrage (F1) und (F2) könnten also nur schwer beantwortet werden. Das selbe Argument lässt sich zudem auch für eine Erhebung von Laut-Erinnerungs-Daten geltend machen. Ein ergänzendes lautes Erinnern würde daher keinen komplementären Beitrag zu den aus dem lauten Denken gewonnenen Daten leisten. Werden die Physiklehrkräfte allerdings um eine begründete Einschätzung der fachlich-konzeptuellen Qualität (bzw. der Qualität ihrer sprachlichen Realisierung) der vier Schülertexte auf Grundlage ihrer eigenen Korrekturen gebeten, werden zusätzliche Daten über die Leistungsfeststellung und -beurteilung der Untersuchungsteilnehmer_innen gewonnen, die in den Laut-Denk-Daten inhaltlich (möglicherweise) nur unzureichend abgebildet sind. Eine Ergänzung des lauten Denkens um eine retrospektive Befragung steigert damit (mutmaßlich) durch Komplementarität von Teilbefunden die Reichhaltigkeit der gewonnenen Erkenntnisse in der Hauptstudie.
2. **Steigerung der Reichhaltigkeit durch konvergierende Teilfunde**:
   Es ist allerdings auch denkbar, dass die befragten Physiklehrkräfte bei der laut denkenden Genese ihrer Leistungsfeststellungen und -beurteilungen deutlich zwischen fachlich-konzeptuellen und sprachlichen Merkmalen eines Schülerlösungstextes differenzieren und diese dabei bis zu einem bestimmten Grad miteinander konfundieren. Diese Differenzierung und Konfundierung sollten sich dann aber auch (zumindest in Teilen) in den Daten einer wie eben beschriebenen retrospektiven Befragung zeigen, da die Untersuchungsteilnehmer_innen hier Erinnerungen an die zuvor laut denkende Handlungsepisode aus dem Langzeitgedächtnis abrufen. Eine Kombination von lautem Denken und retrospektiver Befragung kann also zu konvergierenden Befundlagen führen. Auch dies steigert gegebenenfalls die Reichhaltigkeit der in der Hauptstudie gewonnenen Erkenntnis. Eine (mutmaßliche) Übereinstimmung von Teilbefunden spricht für die konvergente Validität der Forschungsergebnisse.

Nicht auszuschließen ist jedoch auch, dass sich Teilbefunde aus einer Kombination von lautem Denken und retrospektiver Befragung widersprechen. Es ist sogar erwartbar, dass sich die beim lauten Denken verbalisierte unreflektierte „innere Sprache“ der befragten Physiklehrkräfte inhaltlich von einer „reflection-on-action“ (retrospektive Befragung) aufgrund von Erinnerungsfehlern und zusätzlicher kognitiver Verarbeitung bis zu einem bestimmten Grad unterscheidet: Erstgenanntes ermöglicht Rückschlüsse auf die Logik des Handelns der Untersuchungsteilnehmer_innen und damit auf ihre Assessment Literacy. Letztgenanntes enthält (auch) die subjektive Selbstwahrnehmung der befragten Lehrkräfte, kann also (auch) Einblicke in die beruflichen Teilidentität als Assessor of Learning der Untersuchungsteilnehmer_innen liefern. Auf theoretischer Ebene ist aber davon auszugehen, dass zwischen der Assessment Literacy und der beruflichen Teilidentität als Assessor of Learning einer Lehrkraft eine ausgeprägte wechselseitige Beziehung besteht (vgl. Unterkapitel 2.2.4). Diese theoretische Annahme müsste bei einer vollständigen Divergenz von Teilbefunden aus einer Kombination von lautem Denken und retrospektiver Befragung kritisch hinterfragt werden[69]. Auch dies wäre aber ein Beitrag, die Frage zu klären, auf welche Ressourcen zur Leistungsurteilsgenese Physiklehrkräfte bei der Korrektur einer Klassenarbeit zurückgreifen (Forschungsfrage (F1)). Der Grad der Unterschiedlichkeit von Teilbefunden wäre dann ein Ausdruck dafür, welche Forschungsergebnisse aufgrund der theoretischen Fundierung der vorliegenden Arbeit erwartbar und welche überraschend sind. Ähnliches gilt für Forschungsfrage (F2). Divergierende Teilbefunde aus den Laut-Denk-Daten und den Daten der retrospektiven Befragung bezogen auf die Frage, inwieweit Physiklehrkräfte fachlich-konzeptuelle und sprachliche Leistungsurteile miteinander konfundieren, würden eine differenzierte Gesamtbefundlage darstellen. Ein solcher Fall würde empirische Evidenz dafür liefern, dass sich eine derartige Konfundierung von Teilleistungsurteilen entweder nur in der unreflektierten „innere Sprache“ (lautes Denken) der befragten Physiklehrkräfte oder nur in deren „reflection-on-action“ (retrospektive Befragung) zeigt.

5.2.4. Zwischenfazit

Zusammengefasst wurde in der Phase 1 der Entwicklungsstudie das methodische Vorgehen in der Hauptstudie wie folgt präzisiert:

1. Um das Verallgemeinerbarkeitspotenzial der Befunde der empirischen Hauptstudie zu begünstigen, ist eine Untersuchung von im Schuldienst aktiven Physiklehrkräften in einer Laborsituation einer Felduntersuchung vorzuziehen (vgl. Abschnitt 5.2.1). Aus ökologischen Validitätsgründen ist diese Laborsituation allerdings möglichst authentisch zu gestalten und sollte daher der Alltagspraxis von Physiklehrkräften bei

[69] Vollständig divergente Befunde könnten zudem auch – wie in jeder empirischen Untersuchung – auf Artefakte des methodischen Vorgehens oder eine unsaubere methodische Herangehensweise zurückzuführen sein. Auch aus diesem Grund erfolgt in Kapitel 6 unter anderem eine Diskussion der Grenzen der im Rahmen der vorliegenden Arbeit gewonnenen Befunde, sowie eine methodenkritische Auseinandersetzung.

der Leistungsfeststellung und -beurteilung im Rahmen einer Klassenarbeit möglichst ähnlich sein.

2. Hierzu wurde eine erste Skizze einer für die Beantwortung der Forschungsfragen (F1) und (F2) geeigneten Laborsituation entwickelt (vgl. Abschnitt 5.2.2). Diese Skizze lässt sich wie folgt zusammenfassen: In der Laborsituation werden die teilnehmenden Physiklehrkräfte aufgefordert vier Schülerlösungstexte zu einer Klassenarbeitsaufgabe ihren eigenen Gewohnheiten entsprechend zu korrigieren. Bei diesen Schülerlösungstexten handelt es sich um vier Kontrastfälle: Sie unterscheiden sich kriterial bezüglich ihrer fachlich-konzeptuellen Qualität und/oder der Qualität ihrer sprachlichen Realisierung.
3. Als Datenerhebungsmethode in der Laborsituation ist ein im engeren Begriffsverständnis introspektives Erhebungsverfahren zu wählen (vgl. Abschnitt 5.2.3). Hierdurch können von den teilnehmenden Physiklehrkräften Daten gewonnen werden, die Rückschlüsse auf deren Denkprozesse bei der Feststellung und Beurteilung der vier Schülerlösungstexte ermöglichen. Dabei ist die Methode des lauten Denkens den Forschungsfragen (F1) und (F2) im Vergleich zu anderen im engeren Begriffsverständnis introspektiven Erhebungsverfahren am gegenstandsangemessensten. Aus diesem Grund werden die teilnehmenden Lehrkräfte in der Laborsituation während sie die vier Schülerlösungen korrigieren zum lauten Denken aufgefordert. Im Anschluss an die Korrekturarbeit erfolgt eine zusätzliche retrospektive Befragung. Bei dieser nehmen die teilnehmenden Physiklehrkräfte eine explizite Einschätzung und Begründung der fachlich-konzeptuellen Qualität (bzw. der Qualität der sprachlichen Realisierung) der vier Schülerlösungstexte vor. Durch dieses Vorgehen wird es bei der Datenauswertung möglich sein, eine zusätzliche Datentriangulation vorzunehmen, die zur Bereicherung der Gesamtbefundlage beiträgt.

Das Herzstück des methodischen Vorgehens in der Hauptstudie sind die Schülerlösungstexte, die den teilnehmenden Physiklehrkräften zur Leistungsfeststellung und -beurteilung vorlegt werden sollen. Im Rahmen der Entwicklungsstudie galt es daher, die Auswahl einer Komposition von vier, bezüglich ihrer fachlich-konzeptuellen Qualität und/oder der Qualität ihrer sprachlichen Realisierung, kontrastierenden Schülerlösungstexten für die in Abschnitt 5.2.2 skizzierte Laborsituation so sorgfältig wie möglich vorzunehmen. Diese Auswahl wird im nun folgenden Unterkapitel beschrieben.

5.3. Phase 2: Generierung von Schülerlösungstexten für die Laborsituation der Hauptstudie[70]

Die Auswahl der vier Schülerlösungstexte, die den an der Hauptstudie teilnehmenden Physiklehrkräften zur Leistungsfeststellung und -beurteilung vorlegt werden sollen, erfolgte in

[70] Teile dieses Unterkapitels stellen eine überarbeitete und erweiterte Fassung von Feser, Höttecke, & Ehmke (2016), Feser & Höttecke (2017b), sowie Feser & Höttecke (2017c) dar.

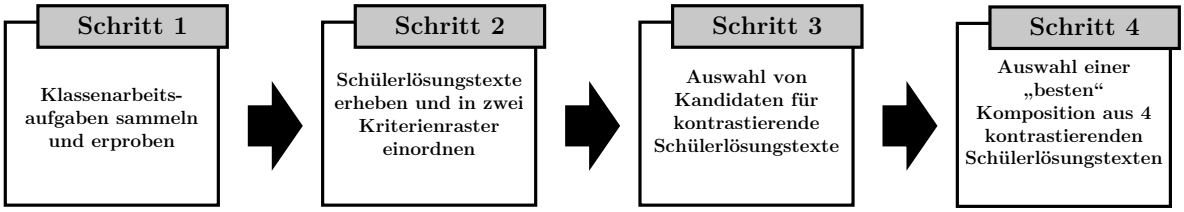


Abbildung 5.4.: Verfahrensschritte zu Auswahl vier kontrastierender Schülerlösungstexte.

einem mehrschrittigen Verfahren (vgl. Abbildung 5.4). Zu diesem Verfahren soll zunächst ein Gesamtüberblick gegeben werden, der in zeitlich umgekehrter Reihenfolge erfolgt:

Im vierten und letzten Schritt von Phase 2 der Entwicklungsstudie wurde eine für die Laborsituation der Hauptstudie am „besten“ geeignete Komposition aus 4 bezüglich ihrer fachlich-konzeptuellen Qualität und der Qualität ihrer sprachlichen Realisierung kontrastierenden Schülerlösungstexten ausgewählt. Die Auswahl dieser Komposition geschah durch 6 Physikdidaktiker_innen, die 8 zuvor ausgewählte Kandidaten für kontrastierende Schülerlösungstexte (Schritt 3) mit Hilfe von zuvor entwickelten Kriterienrastern (siehe unten) codierten. Die Kandidatenauswahl für kontrastierende Schülerlösungstexte erfolgte auf Grundlage einer Zuordnung von insgesamt N = 116 Schülerlösungstexten zu einer Klassenarbeitsaufgabe in zwei eigens hierfür entwickelte Kriterienraster (Schritt 2). Eines der beiden Kriterienraster ermöglichte die erhobenen Lösungstexte der Schüler_innen bezüglich ihrer fachlich-konzeptuellen Qualität voneinander zu unterscheiden. Das andere Kriterienraster ermöglichte eine Unterscheidung der Schülerlösungstexte hinsichtlich der Qualität ihrer sprachlichen Realisierung. Vor der Erhebung von Schülerlösungstexten galt es aber zunächst eine Klassenarbeitsaufgabe zu identifizieren, mit deren Hilfe für das weitere Vorgehen geeignete Lösungstexte von Schüler_innen gewonnen werden konnten. Im ersten Schritt von Phase 2 der Entwicklungsstudie wurden daher verschiedene Klassenarbeitsaufgaben gesammelt und erprobt.

Details zu den einzelnen Auswahlschritten werden in den nun folgenden Abschnitten dargestellt.

5.3.1. Klassenarbeitsaufgaben sammeln und erproben

Im ersten Schritt wurden verschiedene Klassenarbeitsaufgaben gesammelt und mit dem Ziel erprobt, eine für die Laborsituation der Hauptstudie geeignete Aufgabe zu finden. Das gesetzte ökologische Validitätskriterium, dass die Laborsituation einer realen Korrigiersituation möglichst ähnlich sein sollte, wurde hierbei besonders berücksichtigt. Deshalb wurden in Anlehnung an das Vorgehen von Thonhauser (2008) Physiklehrkräfte gebeten Aufgaben einzureichen, welche sie tatsächlich in Klassenarbeiten eingesetzt haben. Zusätzlich wurden publizierte Aufgabensammlungen von Physiklehrkräften nach geeigneten Aufgaben gesichtet (z. B. Meier, 2014; Strate, 2014; Wierzioch, 2018). In Orientierung an

die in Abschnitt 5.2.2 dargestellte Skizze der in der Hauptstudie geplanten Laborsituation, wurden folgende Auswahlkriterien für geeignete Klassenarbeitsaufgaben festgelegt:

1. Die Aufgaben fordern Schüler_innen dazu auf, einen physikalischen Sachverhalt zu erklären
2. Die Aufgaben fordern von Schüler_innen schriftliche Lösungen mit hohem Textanteil
3. Die Aufgaben fordern von Schüler_innen schriftliche Lösungen, die weder zeichnerische noch rechnerische Elemente enthalten
4. Die Aufgaben entsprechen den aktuellen Standards der Hamburger Bildungspläne bezüglich inhaltlicher Mindestanforderung an den Umgang mit Fachwissen von Schüler_innen am Ende der 8. Jahrgangsstufe im Fach Physik[71] (vgl. Behörde für Schule und Berufsbildung Hamburg , 2011, S. 19 u. f.; Behörde für Schule und Berufsbildung Hamburg , 2014, S. 20 u. f.)

Aus dem Aufgabenpool konnten 5 Klassenarbeitsaufgaben identifiziert werden, die den eben benannten Auswahlkriterien genügen (vgl. Tabelle 5.1). Diese 5 Aufgaben wurden am Ende des Schuljahres 2014/2015 in drei Teilerhebungen an einer Gelegenheitsstichprobe von insgesamt 23 Schüler_innen der 8. Jahrgangsstufe und 45 Schüler_innen der 9. Jahrgangsstufe verschiedener Hamburger Gymnasien und Stadtteilschulen[72] pilotiert. Den teilnehmenden Schüler_innen wurden dabei jeweils 4 der 5 Aufgaben zur Bearbeitung vorgelegt. Auf einem vorgefertigten Bearbeitungsbogen erhielten die Schüler_innen folgende Instruktionen für die Aufgabenbearbeitung, die ihnen zu Beginn der Erhebung zusätzlich mündlich mitgeteilt wurden:

> „**Bei der Bearbeitung der Aufgaben solltest du folgendes beachten:**
>
> - Lese die Aufgaben genau durch, bevor du beginnst.
> - Schreibe deine Antwort in ganzen Sätzen auf.
> - Schreibe deine Antwort auf, auch wenn du dir nicht sicher bist.“

Die ersten beiden Instruktionen dienten dazu, die teilnehmenden Schüler_innen zur genauen Auseinandersetzung mit den Aufgabenstellungen aufzufordern und keine stich-

[71] Der Zeitplan der Entwicklungsstudie erforderte, dass die Schülerlösungstexte, aus denen im weiteren Verlauf Kontrastfälle für die Laborsituation gewonnen wurden, in der ersten Hälfte des Schuljahres 2015/2016 erhoben werden mussten. Die hierfür notwendige Klassenarbeitsaufgabe war daher an Unterrichtsinhalten zu orientieren, die den teilnehmenden Schülern_Schülerinnen in vorherigen Schuljahren vermittelt wurden. In den Standards der Hamburger Bildungspläne ist allerdings lediglich für das Ende der 8. und 10. Jahrgangsstufe festgelegt, welche Inhalte Schüler_innen im Physikunterricht bis dahin zu vermitteln sind. Aus pragmatischen Gründen wurde daher entschieden, Lösungstexte von Schüler_innen der 9. Jahrgangsstufe zu erheben und die Auswahl von hierfür geeigneten Aufgaben an den Mindestanforderung an den Umgang mit Fachwissen von Schüler_innen am Ende der 8. Jahrgangsstufe im Fach Physik zu orientieren.

[72] Die Sekundarstufe im Hamburger Schulsystem ist seit der Schulreform von 2009 in die Schulformen Stadtteilschule und Gymnasium gegliedert (vgl. Hallwirth, 2015, S. 15 u. f.). Die Stadtteilschule ersetzte ab dem Schuljahr 2010/2011 die bisherigen Haupt-, Real- und Gesamtschulen, sowie seit dem Schuljahr 2013/2014 auch die Aufbaugymnasien (vgl. ebd., S. 7).

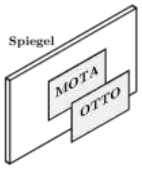
	Aufgabenstellung	exemplarische Schülerlösungstexte
Brennender Fluss	1952 stand plötzlich der Cuyahoga River (Cleveland, Amerika) in Flammen. Er war so stark mit Öl verschmutzt, dass er Feuer fing und brannte. Warum kann der Fluss brennen, wo doch Wasser nicht brennbar ist und es darüber hinaus zum Löschen von Feuer eingesetzt wird? Erkläre genau!	*a) „Das Wasser brennt nicht, sondern das Öl auf der Oberfläche bzw. die Gase."* *b) „Das liegt daran, dass das Öl auf dem Wasser schwimmt, weil es eine geringere Dichte hat."*
Energie-verbrauch	In den Medien (Nachrichten, Zeitungen, usw.) wird immer wieder der Begriff *„Energieverbrauch"* verwendet. Erkläre genau, warum dieser Begriff unter physikalischen Aspekten falsch ist und was damit eigentlich gemeint ist.	*a) „Energie wird nicht ~~um~~ verbraucht sondern umgewandelt."* *b) „Der Begriff ist falsch, weil Kraft in Energie umgewandelt wird. Die Energie wird nicht verbraucht, sondern in Strom ek. umgewandelt."*
Föhn	Beim Föhnen der Haare kannst du folgende Beobachtung machen: Solange die Haare noch feucht sind, ist der heiße Luftstrom des Föns wenig spürbar, er scheint sogar zu kühlen. Bei trockenen Haaren merkst du, dass der Luftstrom tatsächlich heiß ist. Wie sind die Unterschiede zu erklären?	*a) „Ich denke dadurch, dass das Wasser verdampft kühlt ~~das~~ die Luft ab."* *b) „weil die nassen Haare kühlen, und die trockenen nicht."*
Spiegelbild	Auf einem grauen Karton steht OTTO, im Spiegelbild jedoch MOTA (siehe Abbildung). Wie ist dies genau zu erklären? 	*a) „Es liegt an den Blickwinkel was man sieht"* *b) „Auf der anderen Seite des grauen Kartons steht Atom"*
Weltraum-spaziergang	Bei einem Weltraumspaziergang reißt zwischen zwei Astronauten die Funkverbindung ab. Obwohl der eine Astronaut aus Leibeskräften schreit, hört ihn sein Kamerad nicht. Der ältere Astronaut hält seinen in Panik geratenen Kollegen fest und presst seinen Helm an den des Kollegen. Plötzlich kann der jüngere den älteren leise hören. Erkläre beide Phänomene genau!	*a) „Da glas gegen glas ist, können sie sich besser hören. ~~weil~~"* *b) „~~Der~~ Dadurch, dass die Geräte wieder dich ~~beianande~~ ~~aunander~~ zusammen sind ~~trägt~~ ~~der~~ ist dss Signal Stäker."* *c) „Weil wenn die Schwerkraft zu stark ist kann man den Jungen nicht hören"*

Tabelle 5.1.: Für die Pilotierung ausgewählte Klassenarbeitsaufgaben mit exemplarischen Schülerlösungstexten aus der Aufgabenpilotierung.

punktartigen, sondern Lösungen in Form eines Textes zu produzieren. Die dritte Instruktion ermutigte die Schüler_innen dazu, auch Lösungen festzuhalten, die ihrer Ansicht nach möglicherweise Mängel aufweisen. Die dritte Instruktion dient also dazu, Aufgabennichtbearbeitungen zu minimieren und um einen Beitrag dazu zu leisten, die Heterogenität der erhobenen Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität und der Qualität ihrer sprachlichen Realisierung zu erhöhen.

Im Anschluss an die Piloterhebungen wurden die Aufgabenbearbeitungen der Schüler_innen einer Oberflächenanalyse unterzogen. Bei dieser Oberflächenanalyse wurden für jede Klassenarbeitsaufgabe und für jeden der drei Teilerhebungen die mittlere, die minimale und die maximale Wortzahl der erhobenen Schülerlösungstexte bestimmt, sowie die Häufigkeit von Aufgabennichtbearbeitungen ermittelt (absolute und prozentuale Häufigkeit). Die Ergebnisse dieser Oberflächenanalyse sind in Tabelle 5.2 zusammengetragen und zeigen die folgenden Auffälligkeiten[73]:

[73]Im Rahmen dieser Oberflächenanalyse wurde aus forschungsökonomischen Gründen darauf verzichtet, diese Auffälligkeiten mittels geeigneter statistischer Methoden zusätzlich auf Signifikanz zu überprüfen. Deshalb wird an dieser Stelle der vorliegenden Arbeit (lediglich) von Auffälligkeiten und nicht von Befunden gesprochen.

1. Teilerhebung (26.06.2015, Jahrgangsstufe 9, Stadtteilschule, N = 29)

Klassenarbeitsaufgabe	Wortzahl der Schülerlösungstexte (arithmetisches Mittel | Minimum | Maximum)	Aufgabe nicht bearbeitet (absolute Häufigkeit | prozentuale Häufigkeit)
Brennender Fluss	- - -	- - -
Energieverbrauch	18.1 | 4 | 45	4 | 13.8 %
Föhn	29.9 | 4 | 57	2 | 6.9 %
Spiegelbild	17.7 | 4 | 43	3 | 10.3 %
Weltraumspaziergang	25.8 | 5 | 51	3 | 10.3 %

2. Teilerhebung (07.07.2015, Jahrgangsstufe 8, Gymnasium, N = 23)

Klassenarbeitsaufgabe	Wortzahl der Schülerlösungstexte (arithmetisches Mittel | Minimum | Maximum)	Aufgabe nicht bearbeitet (absolute Häufigkeit | prozentuale Häufigkeit)
Brennender Fluss	- - -	- - -
Energieverbrauch	16.0 | 4 | 33	1 | 4.3 %
Föhn	22.8 | 6 | 52	5 | 21.7 %
Spiegelbild	19.4 | 8 | 45	2 | 8.7 %
Weltraumspaziergang	21.0 | 5 | 35	2 | 8.7 %

3. Teilerhebung (20.07.2015, Jahrgangsstufe 9, Gymnasium, N = 16)

Klassenarbeitsaufgabe	Wortzahl der Schülerlösungstexte (arithmetisches Mittel | Minimum | Maximum)	Aufgabe nicht bearbeitet (absolute Häufigkeit | prozentuale Häufigkeit)
Brennender Fluss	23.8 | 13 | 54	1 | 6.3 %
Energieverbrauch	25.6 | 12 | 56	0 | 0.0 %
Föhn	- - -	- - -
Spiegelbild	17.3 | 4 | 57	4 | 25.0 %
Weltraumspaziergang	32.1 | 12 | 57	1 | 6.3 %

Tabelle 5.2.: Oberflächenanalyse der Aufgabenbearbeitung von Gymnasial- und Stadtteilschulschüler_innen der 8. und 9. Jahrgangsstufe im Rahmen der Piloterhebung verschiedener Klassenarbeitsaufgaben.

1. Bei den Schülerlösungstexten zur Aufgabe *Spiegelbild* zeigt sich zu allen drei Teilerhebungen eine vergleichsweise niedrige mittlere Wortzahl. Ähnliches gilt für die Aufgabe *Energieverbrauch* bei der ersten und zweiten Teilerhebung, sowie für die Aufgabe *Brennender Fluss*, die lediglich bei der dritten Teilerhebung den teilnehmenden Schüler_innen vorgelegt wurde.
2. Die Schülerlösungstexte zur Aufgabe *Weltraumspaziergang* weisen hingegen zu allen drei Teilerhebungen eine vergleichsweise erhöhte mittlere Wortzahl auf. Gleiches gilt für die Aufgabe *Föhn*, die bei der ersten und zweiten Teilerhebung den teilnehmenden Schüler_innen vorgelegt wurde.
3. Die Aufgaben Föhn und Spiegelbild wiesen zu jeweils einer Teilerhebung eine auffällig erhöhte Anzahl an Nichtbearbeitungen auf (21.7 % bzw. 25.0 %). Die Aufgabe Energieverbrauch wurde hingegen bei der zweiten und dritten Teilerhebung im Vergleich zu den anderen Aufgaben seltener nicht bearbeitet[74].

[74]Mutmaßlich lässt sich dies darauf zurückführen, dass die Aufgabe Energieverbrauch stets die erste Aufgabe auf dem Bearbeitungsbogen der Schüler_innen war.

Eine zusätzliche grobe Sichtung der Schülerlösungstexte offenbarte zudem, dass sich die Lösungstexte zur Aufgabe Weltraumspaziergang inhaltlich und bezogen auf ihre fachliche Richtigkeit oftmals deutlich voneinander unterschieden. Hierzu findet sich in der rechten Spalte von Tabelle 5.1 eine exemplarische Auswahl besonders kontrastreicher Schülerlösungstexte zur Aufgabe Weltraumspaziergang aus der Aufgabenpilotierung. In Teilen Ähnliches zeigte sich auch bei den Schülerlösungstexten zu den Aufgaben Energieverbrauch und Föhn (vgl. ebenfalls die Beispiellösungstexte in Tabelle 5.1). Bei der Aufgabe Energieverbrauch wurden – wie bereits erwähnt – meist eher kurze Antworten gegeben, die entweder an memorierte Merksätze erinnern (z. B. „Energie wird nicht ~~um~~ verbraucht sondern umgewandelt“), oder in denen sich Hinweise auf in der Literatur vielfach beschriebene Alltagsvorstellungen von Schüler_innen finden lassen (z. B. eine Verwechslung des Energie- und des Kraftbegriffs; vgl. Duit, 1986a, S. 233 u. f.; Driver, Squires, Rushworth, & Wood-Robinson, 1994, S. 144 u. f.). Antworten zur Aufgabe Föhn, in denen sich ein fachliches Konzept von Verdunstungswärme andeutet, wie beispielsweise „[...] dadurch, dass das Wasser verdampft kühlt ~~das~~ die Luft ab“ wurden eher selten gegeben. Meist wurden Antworten wie „die nassen Haare kühlen“ gegeben, die eher an eine vage Alltagsvorstellung zum Begriff „Wärme“ erinnern (vgl. Erickson & Tiberghien, 1985, S. 55 u. f.; Duit, 1986b). Bei den Schülerlösungstexten zu den übrigen beiden Klassenarbeitsaufgaben zeigte sich hingegen ein anderes Bild. Die Lösungstexte zu den Aufgaben Brennender Fluss und Spiegelbild waren meist inhaltlich richtig aber unvollständig. In einem paarweisen Vergleich der Schülerlösungstexte zu diesen Aufgaben zeigte sich daher, dass diese zum Teil inhaltlich (sehr) ähnlich (siehe Beispiellösungstexte zur Aufgabe Spiegelbild in Tabelle 5.1), zum Teil aber auch inhaltlich komplementär zueinander waren (siehe Beispiellösungstexte zur Aufgabe Brennender Fluss in Tabelle 5.1).

Auf Grundlage der groben inhaltlichen Sichtung der Schülerlösungstexte, sowie den Auffälligkeiten, die sich in der Oberflächenanalyse der Aufgabenbearbeitungen zeigten, wurde letztlich die Aufgabe Weltraumspaziergang als Klassenarbeitsaufgabe ausgewählt, die in der Laborsituation der Hauptstudie eingesetzt werden soll. Diese Aufgabe zeigte bei keinem der drei Teilerhebungen eine auffällig erhöhte Anzahl an Nichtbearbeitungen, sie gehörte stets zu den Klassenarbeitsaufgaben, zu denen die teilnehmenden Schüler_innen vergleichsweise eher wortreiche Lösungstexte formulierten und ferner unterschieden sich die Schülerlösungstexte zur Aufgabe Weltraumspaziergang inhaltlich und bezogen auf ihre fachliche Richtigkeit oftmals deutlich voneinander.

5.3.2. Erhebung der Schülerlösungstexte und Kriterienrasterentwicklung

Ziel des zweiten Schritts von Phase 2 der Entwicklungsstudie war es möglichst unterschiedliche Schülerlösungstexte zur Aufgabe Weltraumspaziergang, aus denen Kontrastfälle ausgewählt werden sollten, zu erheben. Ferner galt es zwei Kriterienraster zu entwickeln, die es ermöglichen, die erhobenen Lösungstexte der Schüler_innen bezüglich ihrer fachlich-

Erhebungsklasse (fortlaufende Nummer)	1	2	3	4	5	6	7	**Total**
Erhebungsschule (fortlaufende Nummer)	1	1	2	3	3	4	4	
Schultyp (G: Gymnasium; StS: Stadtteilschule)	G	G	G	StS	StS	StS	StS	
Sozialindex der Schule	3	3	5	5	5	3	3	
Anzahl teilnehmender Schüler_innen	21	16	25	15	19	22	10	128
Anzahl an Aufgabenbearbeitungen	21	14	24	12	17	18	10	116
Anzahl an Aufgabennichtbearbeitungen	0	2	1	3	2	4	0	12

Tabelle 5.3.: Überblick über die systematische gezogene Stichprobe zur Erhebung kontrastierender Schülerlösungstexte zur Aufgabe Weltraumspaziergang.

konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung voneinander zu unterscheiden.

Da davon auszugehen war, dass Kontrastfälle eher selten vorzufinden sind, wurde die Aufgabe Weltraumspaziergang in der ersten Hälfte[75] des Schuljahres 2015/2016 einer größeren Anzahl von Schüler_innen der 9. Jahrgangsstufe im Bundesland Hamburg zur Bearbeitung vorlegt. Hierbei fand eine systematische Stichprobenziehung statt (vgl. Tabelle 5.3): Insgesamt wurden 128 Schüler_innen aus 7 Schulklassen, zweier Gymnasien und zweier Stadtteilschulen befragt. Beide Gymnasien bzw. die beiden Stadtteilschulen unterschieden sich voneinander bezüglich ihres Sozialindexes[76] (vgl. Bürgerschaft der Freien und Hansestadt Hamburg , 2013, S. 27 u. f.).

Von den 128 befragten Schüler_innen haben 116 (90.6 %) die Aufgabe Weltraumspaziergang bearbeitet. Mit Hilfe dieser 116 Schülerlösungstexte wurden in einer deduktiv-induktiven Verfahrensweise die beiden Kriterienraster zur Unterscheidung der Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung entwickelt. Diese deduktiv-induktive Verfahrensweise ist in Abbildung 5.5 schematisch veranschaulicht und ist dem Vorgehen der qualitativen Inhaltsanalyse zur deduktiv-induktiven Kategorienbildung entlehnt (vgl. Boyatzis, 1998, S. 37 u. f.; Schreier, 2012, S. 89 u. f.; Kuckartz, 2016, S. 95 u. f.; Saldaña, 2016, S. 74):

In einem ersten Schritt wurde die bisherige naturwissenschaftsdidaktische Forschungsliteratur nach Kriterienrastern durchsucht, die dazu eingesetzt wurden, die fachlich-konzeptuelle Qualität bzw. die Qualität der sprachlichen Realisierung von Schülertexten, in denen ein physikalischer Sachverhalt erklärt werden soll, zu erfassen. Bei dieser Literaturrecherche stellten sich die von Kang, Thompson, & Windschitl (2014) entwickelten „Criteria to Score Qualities of Explanations“ (vgl. ebd., S. 687 u. f.), die auf der theoretischen Vorarbeit von Braaten & Windschitl (2011) beruhen, sowie die „SOLO-Taxonomy“

[75] Der genau Erhebungszeitraum lag zwischen dem 07.10.2015 (erste Teilerhebung) und dem 09.12.2015 (letzter Teilerhebung).

[76] Allen staatlichen Grund- und weiterführenden Schulen der Freien und Hansestadt Hamburg wird ein Index zugewiesen, „der die soziale Situation der Schule widerspiegelt und damit eine Einschätzung über die soziale Belastung der Schulen erlaubt“ (Bos, Pietsch, Gröhlich, & Janke, 2006, S. 149). Die Werte des Sozialindex von 1 bis 6 bilden die Skala, auf deren Grundlage den Schulen Unterstützungsressourcen zugewiesen werden (vgl. ebd.; Schulte, Hartig, & Pietsch, 2014, S. 67): 1 – hohe soziale Belastung; 6 – geringe soziale Belastung.

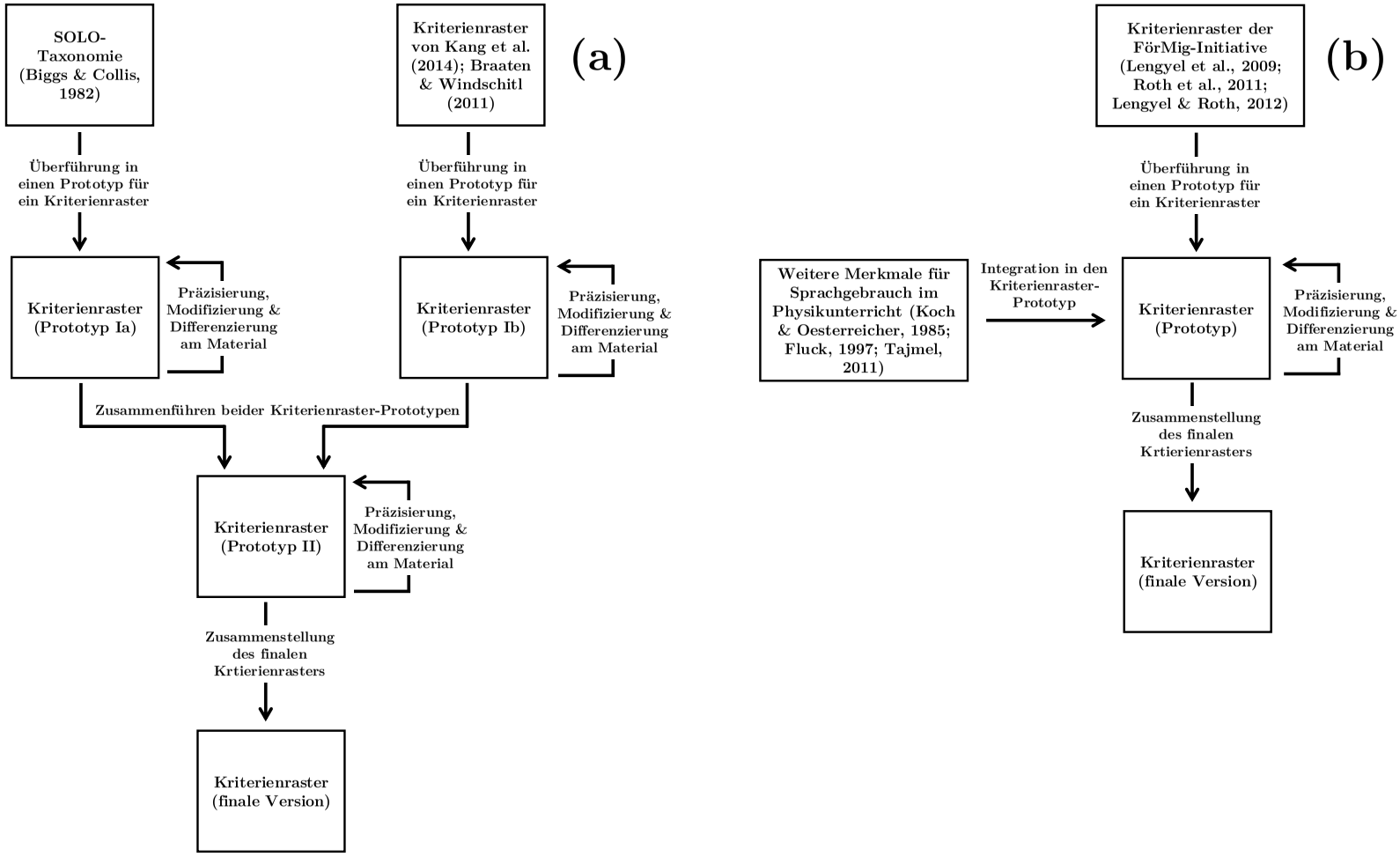


Abbildung 5.5.: Ablaufschema des deduktiv-induktiven Verfahrens zur Entwicklung zweier Kriterienraster zur Unterscheidung von Schülerlösungstexten zur Aufgabe Weltraumspaziergang bezüglich ihrer fachlich-konzeptuellen Qualität (Abbildung 5.5 (a)) bzw. der Qualität ihrer sprachlichen Realisierung (Abbildung 5.5 (b))).

von Biggs & Collis (1982) als besonders gut dokumentierte Kriterienraster für die fachlich-konzeptuelle Qualität von Schülertexten heraus. Für die Qualität der sprachlichen Realisierung von Schülertexten galt dies für das von der FörMig-Initiative (vgl. Salem, Neumann, & Dobutowitsch, 2013, S. 5) entwickelte Raster zur Beobachtung und Analyse bildungssprachlicher Fähigkeiten von Schüler_innen im natur- und sozialwissenschaftlichen Unterricht für die Sprachhandlung „Erklären“ (vgl. Lengyel et al., 2009, S. 135 u. f.; Roth, Reich, & Lengyel, 2011, S. 8 u. f.; Lengyel & Roth, 2012).

Aus den Dokumentationen dieser Raster wurden a-priori erste Prototypen von Kriterienrastern für die Unterscheidung von Schülerlösungstexten zur Aufgabe Weltraumspaziergang entwickelt[77]. Jedes A-Priori-Kriterienraster wurde dabei aus mehreren Kriterien aufgebaut, die jeweils ein anderes inhaltlich-konzeptuelles bzw. sprachliches Merkmal eines Schülerlösungstexts zur Aufgabe Weltraumspaziergang fokussieren. Ferner bestand jedes Kriterium jeweils aus mindestens 2 gestuften Ausprägungen. Da für das Raster zur Unterscheidung von Schülerlösungstexten bezüglich ihrer fachlich-konzeptuellen Qualität zwei gut dokumentierte Raster aus der Literatur zur verfügung standen, wurden hier zunächst zwei Kriterienraster-Prototypen entwickelt und erst im weiteren Verlauf der Entwicklungsarbeit zu einem integrierten Kriterienraster-Prototypen zusammengeführt (vgl. Abbildung 5.5 (a)). Für das Kriterienraster zur Unterscheidung von Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung wurde hingegen nur ein A-Priori-Prototypen basierend auf dem Raster der FörMig-Initative entwickelt. Da dieses allerdings bislang nur in einer Entwurfsfassung vorliegt (vgl. Lengyel et al., 2009, S. 130; Roth et al., 2011, S. 58) wurde das A-Priori-Kriterienraster um weitere linguistische Merkmale fachlichen Sprachgebrauchs im Physikunterricht vor allem aus den Arbeiten von Koch & Oesterreicher (1985), Fluck (1997, S. 35 u. f.) und Tajmel (2011) ergänzt, sowie Merkmale, die aus der Entwurfsfassung des FörMiG-Rasters übernommen wurden, weiter präzisiert[78].

Im nächsten Schritt wurden die zuvor a-priori entwickelten Kriterienraster in einem zirkulären Prozess am gesamten Datenmaterial (den N = 116 erhobenen Schülerlösungstexten zur Aufgabe Weltraumspaziergang) pilotiert und überarbeitet: In einem Zyklus dieses Prozesses wurden die Kriterienraster-Prototypen auf die Schülerlösungstexte aus zwei der sieben Erhebungsklassen angewendet. Dabei wurden Lösungstexte, denen bezüglich eines bestimmten Kriteriums keine der definierten Ausprägungen eindeutig zugeordnet werden konnten, in die zusätzliche Ausprägung „sonstige Lösungstexte“ eingeordnet. Nachdem

[77] Unterstützt wurde die Entwicklung erster Prototypen von Kriterienrastern durch die im Rahmen der vorliegenden Arbeit betreute Qualifikationsarbeit von Klemp (2015). In dieser wurde von der Autorin, in Anlehnung an das Vorgehen der vorliegenden Arbeit, ein inhaltsanalytisches Kategoriensystem zur Erfassung fachlicher und sprachlicher Qualitätsmerkmale von Schülerlösungen zu Leistungsaufgaben im Fach Physik entwickelt (vgl. ebd., S. 54 u. f.). Dieses Kategoriensystem diente als ergänzende Ausgangsbasis für die Entwicklung der A-priori-Kriterienraster.

[78] Ergänzt und präzisiert wurde beispielsweise das Konzept zwischen einer „Sprache der Nähe“ und einer „Sprache der Distanz“ zu unterscheiden (vgl. Koch & Oesterreicher, 1985), sowie sprachliche Oberflächenmerkmale, wie z. B. deiktische Elemente (vgl. Bußmann, 2008, S. 117 u. f.), Anaphern und Kataphern (vgl. Erfurt, 1996, S. 1390 u. f.), fachsprachliche Komposita (vgl. Fluck, 1997, S. 61 u. f.), fachsprachliche Kollokationen (vgl. Tajmel, 2011, S. 2 u. f.) und Funktionsverbgefüge (vgl. Fluck, 1997, S. 97 u. f.).

auf alle Schülerlösungstexte des gewählten Teildatensatzes die Kriterienraster-Prototypen angewandt wurden, wurden die Ausprägungen der einzelnen Kriterien so lange präzisiert, modifiziert und differenziert, bis alle Schülerlösungstexte, die zuvor der Ausprägung „sonstige Lösungstexte“ zugeordnet wurden, eindeutig in die Ausprägung eines Kriteriums eingeordnet werden konnten. Anschließend begann der Prozess von vorne, indem die überarbeiteten Kriterienraster auf die Schülerlösungstexte aus zwei anderen Erhebungsklassen angewandt wurden. Dieser zirkuläre Prozess wurde so lange durchlaufen, bis jedem Schülerlösungstexte aus dem gesamten Datenmaterial in jedem Kriterium eindeutig eine Ausprägung zugeordnet werden konnte.

Abschließend wurde finale Versionen der Kriterienraster zur Unterscheidung von Schülerlösungstexten zur Aufgabe Weltraumspaziergang bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung zusammengestellt. Diese finalen Versionen sind in Tabelle 5.4 und 5.5 überblicksartig dargestellt. Die vollständigen Kriterienraster sind in Anhang A zu finden. Jedes der beiden Kriterienraster ist – wie schon die A-priori-Prototypen – aus mehreren Kriterien aufgebaut. Im Kriterienraster zur Unterscheidung von Schülerlösungstexten bezüglich ihrer fachlich-konzeptuellen Qualität werden die Kriterien „Rahmenbau der Erklärung“, „Rolle von Evidenzbezügen in der Erklärung“, „Tiefe der Erklärung“ und „Konsinstenz der Erklärung“ unterschieden. Das Raster zur Unterscheidung von Schülerlösungstexten bezüglich der Qualität ihrer sprachlichen Realisierung ist in die Kriterien „Lexik/Semantik“, sowie „Syntax/Stilistik“ untergliedert. Ferner ist jedes Kriterium beider Kriterienraster aus mindestens 2 gestuften Ausprägungen aufgebaut (vgl. Tabelle 5.4 und 5.5). Entsprechend dieser Stufung ist jeder Ausprägung ein Score von 0 bis 2 Punkten zugewiesen. Bei der Anwendung[79] (Codierung) eines der beiden Kriterienraster wird jedem Schülerlösungstext für jedes Kriterium genau eine Ausprägung zugewiesen. Hierdurch erhält jeder Schülerlösungstext bezüglich jedes Kriteriums einen eindeutigen Score. Anschließend werden die Scores für jedes Kriterienraster (gewichtet) aufsummiert. Eine Gewichtung fand lediglich bei den Scores für das Kriterienraster zur Unterscheidung von Schülerlösungstexten bezüglich ihrer fachlich-konzeptuellen Qualität statt. Hier wurde der Score des Kriteriums „Konsistenz der Erklärung“ mit dem Faktor 3 gewichtet (= Anzahl der übrigen Kriterien im Kriterienraster), um bei der Bildung dieses Scores im besonderen Maße zu berücksichtigen, dass „bei [naturwissenschaftlichen] Erklärungen nur solche allgemeinen Gesetze als Begründungen akzeptiert werden, die empirisch überprüfbar und allgemein wissenschaftlich als wahr anerkannt sind“ (Kulgemeyer & Tomczyszyn, 2015, S. 112). Hierdurch erhält man für jeden Schülerlösungstext einen Summenscore für seine fachlich-konzeptuelle Qualität (0 bis 12 Punkte) und einen für die Qualität seiner sprachlichen Realisierung (0 bis 4 Punkte). Der Einfachheit halber werden diese Summenscores im Folgenden als *Fachscore* und *Sprachscore* bezeichnet. Beide Summenscores erlauben es, die Aufgabenlösung eines_einer Schülers_-Schülerin in einem Koordinatensystem mit den beiden Dimensionen „fachlich-konzeptuelle Qualität“ und „Qualität der sprachlichen Realisierung“ zu verorten.

[79] Details hierzu sind den Anwendungsregeln für die Kriterienraster in Anhang A zu entnehmen.

Kriterium	**Ausprägung 1** (Score=0)	**Ausprägung 2** (Score=1)	**Ausprägung 3** (Score=2)
Rahmenbau der Erklärung	Der Text ist eine bruchstückhafte Wiedergabe der Aufgabenstellung oder eine bloße Paraphrasierung von Merksätzen. Er ist nicht/kaum empirie- oder theoriegeleitet.	- - -	Argumentation durch empirische Daten, Erfahrungen oder Anwendung physikalischen Sachwissens. Der generalisierende Charakter des Textes dominiert.
Rolle von Evidenzbezügen	Der Text enthält keine Evidenzmittel zur Stützung von Aussagen.	Die verwendeten Evidenzmittel sind entweder nicht angemessen oder hinlänglich.	Die verwendeten Evidenzmittel sind sowohl angemessen als auch hinlänglich.
Tiefe der Erklärung	Erklärung erfolgt auf der „Was“-Ebene.	Erklärung erfolgt auf der „Wie“-Ebene.	Erklärung erfolgt auf der „Warum“-Ebene.
Konsistenz der Erklärung	Sachverhalte, die nicht im Fokus der Aufgabe stehen werden erklärt oder fachlich falsche Erklärungen werden gegeben.	Die Erklärung ist teilweise korrekt. Teile der Erklärung sind übergeneralisiert, irrelevant oder unvollständig.	Der Sachverhalt, der im Fokus der Aufgabe steht, wird allumfassend und fachlich richtig erklärt.

Tabelle 5.4.: Kurzfassung des Kriterienrasters zur Unterscheidung verschiedener Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität (basierend auf den Arbeiten von Kang, Thompson, & Windschitl (2014) und Braaten & Windschitl (2011), sowie der *SOLO-Taxonomie* von Biggs & Collis (1982)).

Kriterium	**Ausprägung 1** (Score=0)	**Ausprägung 2** (Score=1)	**Ausprägung 3** (Score=2)
Lexik/Semantik	Ein Alltagswortschatz mit dominant mündlichem Charakter wird verwendet.	Ein Alltagswortschatz, der sich deutlich von einer mündlichen Ausdrucksweise absetzt, wird verwendet	Ein unterrichtlicher Fachwortschatz wird verwendet.
Syntax/Stilistik	Es werden nur einfache/unvollständige Sätze oder einfache Satzreihen gebildet. Stilistisch entspricht der Text tendenziell einer Sprache der Nähe.	Unterschiedlichste satzverbindende Elemente werden eingesetzt. Sprachliche Verdichtung findet nicht/kaum statt. Stilistisch entspricht der Text tendenziell einer Sprache der Nähe.	Die Sätze sind sprachlich verdichtet. Stilistisch entspricht der Text tendenziell einer Sprache der Distanz.

Tabelle 5.5.: Kurzfassung des Kriterienrasters zur Unterscheidung verschiedener Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung (basierend auf den Arbeiten von Koch & Oesterreicher (1985), Fluck (1997, S. 35 u. f.), Lengyel, Heintze, Reich, Roth, & Scheinhardt-Stettner (2009, S. 135 u. f.), Roth, Reich, & Lengyel (2011, S. 8 u. f.), Tajmel (2011), sowie Lengyel & Roth (2012)).

5.3.3. Auswahl von Kandidaten für kontrastierende Schülerlösungstexte

Nach der Entwicklung beider Kriterienraster wurden deren finale Versionen auf das Gesamtmaterial angewendet und so jedem Schülerlösungstext zur Aufgabe Weltraumspaziergang ein Fach- und ein Sprachscore zugewiesen. Daraufhin erfolgte eine erste Vorauswahl von Schülerlösungstexten. Hierzu wurden der Wertebereich des Fach- und des Sprachscores in 3 ungefähr gleichgroße Intervalle zerlegt (Fachscore: 0 bis 4 Punkte, 5 bis 9 Punkte, 10 bis 12 Punkte; Sprachscore: 0 oder 1 Punkt, 2 oder 3 Punkte, 4 Punkte). Durch dieses Vorgehen wurde das gedachte Koordinatensystem mit den Dimensionen „fachlich-konzeptuelle Qualität“ und „Qualität der sprachlichen Realisierung“ in eine 3×3-Matrix überführt. Die Zellen dieser Matrix wurden nun soweit wie möglich mit jeweils 3 Schülerlösungstexten befüllt (vgl. Abbildung 5.6 (a)). Dabei galt die zusätzliche Auswahlregel, dass die Schülerlösungstexte in jeder Matrixspalte (ungefähr) den gleichen Fachscore bzw. in jeder Matrixzeile (ungefähr) den gleichen Sprachscore aufweisen sollten. Bei Zellen, für die es zunächst mehr als 3 passende Optionen gab, wurden diejenigen Schülerlösungstexte bevorzugt, die die größte Textlänge aufwiesen.

Wie in Abbildung 5.6 (a) schematisch dargestellt, konnte die 3×3-Matrix zunächst nur mit 20 Schülerlösungstexten aus der Datenerhebung befüllt werden, wodurch zunächst nicht alle 9 Zellen der Matrix mit jeweils 3 Texten besetzt waren. Um dennoch alle Matrixzellen vollständig befüllen zu können, wurden zusätzliche Lösungstexte zur Aufgabe Weltraumspaziergang „generiert“, indem der Inhalt und die sprachliche Realisierung von 8 bis dato noch nicht ausgewählten Schülerlösungstexten so lange systematisch abgewandelt wurde[80], bis jede Matrixzelle mit mindestens 3 Schülerlösungstexte zur Aufgabe Weltraumspaziergang besetzt war (vgl. Abbildung 5.6 (b); die „generierten“ Schülerlösungstexte sind hier durch schwarze Kreuze symbolisiert).

Die so gewonnene Vorauswahl von $20 + 8$ Schülerlösungstexten wurden einer Zweitcodierung unterzogen. Hierdurch sollte zum einen die Intercoderreliabilität der Kriterienraster zur Unterscheidung von Schülerlösungstexten zur Aufgabe Weltraumspaziergang bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung überprüft werden. Zum anderen sollten durch die Zweitcodierung der 28 vorausgewählten Texte, Kandidaten für kontrastierende Schülerlösungstexte für die Laborsituation der Hauptstudie identifiziert werden.

Vor der Zweitcodierung wurde der_die Zweitcodierer_in[81] im Umgang mit den Kriterienrastern an den Schülerlösungstexten zur Aufgabe Weltraumspaziergang aus der Piloterhebung (vgl. Abschnitt 5.3.1) geschult. Um dem_der Zweitcodierer_in zu verbergen, welche Schülerlösungstexte aus der Vorauswahl hervorgingen, wurden ihm_ihr die vorausgewähl-

[80] Beispielsweise wurde der Schülerlösungstext „da es im Weltraum keine Luft gibt kann sein Kolege ihn nicht verstehen. Und wenn sie ihre hälme an Einander drücken leitet das glas die Schallwellen weiter und Er kann ihn wieder hören“ abgewandelt zu: „Im All ist nichts durch das Ton geht, und er hört seinen Freund nicht. Dann kommt aber Ton durch die Helme, da Ton über Glas geht“.

[81] Bei dem_der Zweitcodierer_in handelte es sich um eine_n Doktorierende_n der Physikdidaktik.

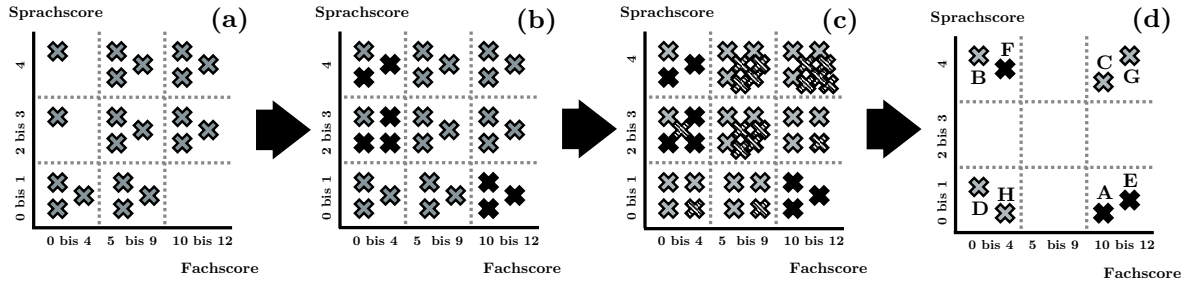


Abbildung 5.6.: Schematischer Ablauf der Auswahl von Kandidaten für kontrastierende Schülerlösungstexte.

ten 20 + 8 Schülerlösungstexte zusammen mit 17 weiteren, zufällig aus dem Gesamtmaterial ausgewählten Texten zur Zweitcodierung vorgelegt (vgl. Abbildung 5.6 (c); die zufällig ausgewählten Schülerlösungstexte sind hier durch schraffierte Kreuze symbolisiert). Insgesamt wurden hierdurch 31.9 % der Originalschülerlösungstexte einer Zweitcodierung unterzogen.

Als Maß für die Intercoderreliabilität wurde Krippendorffs α für ordinale Daten gewählt (vgl. Krippendorff, 2004, S. 233 u. f.). Bei diesem Koeffizient sind – als grobe Orientierung – Werte ab .667 zufriedenstellend (vgl. ebd., S. 241 u. f.). Wie aus Tabelle 5.6 und 5.7 hervorgeht, war im Rahmen der Zweitcodierung zunächst lediglich die Intercoderreliabilität für das Kriterium „Konsistenz der Erklärung“ und für den Fachscore zufriedenstellend. In einer anschließenden Diskussion zwischen Erst- und Zweitcodierer_in wurden daher Diskrepanzen besprochen, um zu einem Konsens über die Zuordnung der Schülerlösungstexte auf den Ausprägungen der Kriterienraster zu gelangen. Im Rahmen dieser Diskussion stellte sich heraus, dass die Diskrepanzen zwischen Erst- und Zweitcodierier_in vor allem auf missverständliche Formulierungen in beiden Kriterienrastern zurückzuführen waren (insbesondere beim Kriterium „Syntax/Stilstik“). Durch Aufklären dieser Missverständnisse und eine Überarbeitung der Formulierungen in beiden Kriterienraster[82,83] konnte die Intercoderreliabilität deutlich verbessert werden. Nachfolgend an die Diskussion zwischen Erst- und Zweitcodierer_in, war die Intercoderreliabilität für alle Kriterien in beiden Kriterienrastern, sowie für den Fach- und den Sprachscore zufriedenstellend (vgl. Tabelle 5.6 und 5.7).

Im Anschluss an die Zweitcodierung wurden aus den bis dahin 28 Textoptionen 8 Kandidaten kontrastierender Schülerlösungstexte für die Laborsituation der Hauptstudie ausgewählt. Dabei wurden, wie in Abbildung 5.6 (d) dargestellt, je zwei Schülerlösungstexte aus jeder „Ecke“ der 3×3-Matrix ausgewählt, bei denen Erst- und Zweitcodierer_in im Anschluss an die gemeinsame Diskussion allen Kriterien beider Kriterienraster identische

[82] Dabei wurden lediglich missverständliche Begriffe und Satzstrukturen in beiden Kriterienrastern durch solche ersetzt, die sich im Anschluss an die Diskussion von Erst- und Zweitcodier_in als besser geeignet herausstellten, um die Bedeutung der Ausprägungen der einzelnen Kriterien beschreiben zu können. Die ursprüngliche inhaltliche Ausgestaltung beider Kriterienraster wurde durch die Überarbeitung also nicht verändert.

[83] Bei den in Anhang A aufgeführten Kriterienrastern handelt es sich um die im Anschluss an die Diskussion von Erst- und Zweitcodier_in überarbeiteten Versionen.

Kriterium	**α vor der Diskussion der Codierer_innen**	**Interpretation** (vgl. Krippendorff, 2004, S. 241 u. f.)	**α nach der Diskussion der Codierer_innen**	**Interpretation** (vgl. Krippendorff, 2004, S. 241 u. f.)
Rahmenbau der Erklärung	.487	nicht akzeptabel	.897	zufriedenstellend
Rolle von Evidenzbezügen	.504	nicht akzeptabel	.884	zufriedenstellend
Tiefe der Erklärung	.588	nicht akzeptabel	.885	zufriedenstellend
Konsistenz der Erklärung	.731	zufriedenstellend	.924	zufriedenstellend
Fachscore	.818	zufriedenstellend	.946	zufriedenstellend

Tabelle 5.6.: Intercoderreliabilität (Krippendorffs α) des Kriterienrasters zur Unterscheidung verschiedener Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität.

Kriterium	**α vor der Diskussion der Codierer_innen**	**Interpretation** (vgl. Krippendorff, 2004, S. 241 u. f.)	**α nach der Diskussion der Codierer_innen**	**Interpretation** (vgl. Krippendorff, 2004, S. 241 u. f.)
Lexik/Semantik	.513	nicht akzeptabel	.915	zufriedenstellend
Syntax/Stilistik	.123	nicht akzeptabel	.915	zufriedenstellend
Sprachscore	.454	nicht akzeptabel	.907	zufriedenstellend

Tabelle 5.7.: Intercoderreliabilität (Krippendorffs α) des Kriterienrasters zur Unterscheidung verschiedener Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

Ausprägungen zugewiesen hatten[84] und bei denen es sich vorzugsweise nicht um „generierte“ Schülerlösungstexte handelte.

Um diese 8 Kandidaten kontrastierender Schülerlösungstexte im weiteren Verlauf einfacher voneinander unterscheiden zu können, werden diese im Folgenden als Schülerlösungstext A, B, C, D, E, F, G und H bezeichnet[85] (vgl. Abbildung 5.6 (d)). In Tabelle 5.8 sind diese 8 Schülerlösungstexte aufgeführt, zusammen mit dem Fach- bzw. dem Sprachscore, den Erst- und Zweitcodierer_in im Anschluss an die gemeinsame Diskussion den Schülerlösungstexten zugewiesen haben. Die Ausprägungen der einzelnen Kriterien beider Kriterienraster, die Erst- und Zweitcodierer_in diesen Schülerlösungstexten zugewiesen haben, finden sich zudem in Tabelle 5.13, die sich im nun folgenden Abschnitt der vorliegenden Arbeit befindet.

[84] Die einzige Ausnahme bildete Schülerlösungstext H, dem Erst- und Zweitcodierer_in auch im Anschluss an die gemeinsame Diskussion bzgl. des Kriteriums „Tiefe der Erklärung“ unterschiedliche Ausprägungen zugewiesen haben (Ausprägung 1 bzw. Ausprägung 2). Schülerlösungstext H wurde dennoch in die Kandidatenauswahl kontrastierender Schülerlösungstexte aufgenommen, da bei diesem Text die eben beschriebene Diskrepanz zwischen Erst- und Zweitcodierer_in nicht zur Folge hatte, dass dieser eventuell einer anderen Zelle der 3×3-Matrix hätte zugewiesen werden müssen.

[85] Diese Bezeichnung wurde den 8 Schülerlösungstexten in Phase 3 der Entwicklungsstudie gegeben (vgl. Unterkapitel 5.4). Die Texte A bis D sind dabei die „beste“ Komposition aus 4 kontrastierenden Schülerlösungstexten, die am Ende von Phase 2 der Entwicklungsstudie für die Laborsituation ausgewählt wurde. Ferner ergab sich die Bezeichnung der Texte A bis D bzw. E bis H daraus, dass in Phase 3 der Entwicklungsstudie entschieden wurde, vier kontrastierende Schülerlösungstexte den an der Hauptstudie teilnehmenden Physiklehrkräften in einer kontraintuitiven Reihenfolge vorzulegen.

Bezeichnung	fachlich-konzeptuelle Qualität (Fachscore)	Qualität der sprachlichen Realisierung (Sprachscore)	Schülerlösungstext
A	hoch (11 Punkte)	gering (0 Punkte)	*Im All ist nichts, durch das Ton geht, und er hört seinen Freund nicht. Dann kommt aber Ton durch die Helme, da Ton über Glas geht.*
B	gering (2 Punkte)	hoch (4 Punkte)	*Die beiden Astronauten können sich wieder hören, weil der geringe Abstand zwischen den beiden Funkgeräten eine bessere Funkverbindung herstellt. Deswegen kann der jüngere den älteren wieder leise hören.*
C	hoch (12 Punkte)	hoch (4 Punkte)	*1. Nicht hörbar: Im Weltall herrscht ein Vakuum, also können sich die Schwingungen nicht fortbewegen. Sie werden nämlich durch Luft geleitet.* *2. Noch hörbar: Die Helme bestehen aus Glas. Wenn man innerhalb des Helmes spricht, kann sich die Stimme ausbreiten, da es Luft gibt. Hält man zwei Helme aneinander, werden die Schallwellen durch die Schwingungen des Glases weitergegeben.*
D	gering (2 Punkte)	gering (1 Punkt)	*Sie haben sich nicht gehört, weil die Frequenz nicht gut genug war, haben sie sich nicht gehört.*
E	hoch (11 Punkte)	gering (1 Punkte)	*Er hört ihn nicht, weil draußen keine Luft ist, durch die der Schrei kommen kann. Wenn sie aneinander halten, können sie die Schreie spüren und leise etwas hören.*
F	gering (4 Punkte)	hoch (4 Punkte)	*Die beiden Astronauten können sich zunächst nicht hören, da die Funkverbindung umso besser wird, je näher man aneinander ist und im Weltraum ist die Funkverbindung sehr schlecht.*
G	hoch (12 Punkte)	hoch (4 Punkte)	*Im Weltraum herrscht ein Vakuum, das heißt, dass kein Sauerstoff, bzw. keine Luft vorhanden ist. Schwingungen können somit nicht übertragen werden, da sie die Luft dazu brauchen und aus diesem Grund hört der ältere Astronaut den jüngeren nicht. Jedoch ist es so, dass als die beiden Helme gegeneinander gepresst werden, dass die Schwingungen eine sogenannte "Brücke" haben um übertragen zu werden. Man kann es sich auch so vorstellen, dass wenn ein Zug über eine Brücke fährt und man den Sockel der Brücke oder irgendeinen anderen Teil von ihr berührt, man die Schwingungen spürt, die der Zug auf den Schiene ausübt. Sie werden übertragen und die Brücke ist der Leiter.*
H	gering (1 oder 2 Punkte)	gering (1 Punkt)	*Vielleicht hat der jüngere keine Luft mehr bekommen, weil er keinen Helm auf hatte.*

Tabelle 5.8.: Kandidaten kontrastierender Schülerlösungstexte zur Aufgabe Weltraumspaziergang. Die aufgeführten Fach- und Sprachscores ergeben sich aus den Ausprägungen, die Erst- und Zweitcodierer_in im Anschluss an die gemeinsame Diskussion den Schülerlösungstexten zugewiesen haben.

5.3.4. Auswahl einer „besten“ Komposition aus 4 kontrastierenden Schülerlösungstexten

Aus den 8 verbleibenden Schülerlösungstexten zur Aufgabe Weltraumspaziergang ließen sich insgesamt 16 verschiedene Kompositionen aus 4 kontrastierenden Schülerlösungstexten bilden. Des Weiteren konnten auf Grundlage der Fach- und Sprachscores aus der Zweitcodierung im dritten Schritt von Phase 2 der Entwicklungsstudie für jede dieser 16 Kompositionen die folgenden Erwartungen an die Rangfolge der jeweiligen 4 Schü-

lerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung formuliert werden:

1. **Komposition aus Schülerlösungstext A, B, C und D:**
   B=D<A=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   A=D<B=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

2. **Komposition aus Schülerlösungstext A, B, C und H:**
   B=H<A=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   A=H<B=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

3. **Komposition aus Schülerlösungstext A, B, G und D:**
   B=D<A=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   A=D<B=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

4. **Komposition aus Schülerlösungstext A, F, C und D:**
   F=D<A=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   A=D<F=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

5. **Komposition aus Schülerlösungstext E, B, C und D:**
   B=D<E=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   E=D<B=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

6. **Komposition aus Schülerlösungstext A, B, G und H:**
   B=H<A=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   A=H<B=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

7. **Komposition aus Schülerlösungstext A, F, G und D:**
   F=D<A=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   A=D<F=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

8. **Komposition aus Schülerlösungstext E, F, C und D:**
   F=D<E=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   E=D<F=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

9. **Komposition aus Schülerlösungstext E, B, C und H:**
   B=H<E=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
   E=H<B=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

10. **Komposition aus Schülerlösungstext A, F, C und H:**
    F=H<A=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
    A=H<F=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

11. **Komposition aus Schülerlösungstext E, B, G und D:**
B=D<E=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
E=D<B=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

12. **Komposition aus Schülerlösungstext A, F, G und H:**
F=H<A=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
A=H<F=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

13. **Komposition aus Schülerlösungstext E, B, G und H:**
B=H<E=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
E=H<B=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

14. **Komposition aus Schülerlösungstext E, F, C und H:**
F=H<E=C – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
E=H<F=C – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

15. **Komposition aus Schülerlösungstext E, F, G und D:**
F=D<E=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
E=D<F=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

16. **Komposition aus Schülerlösungstext E, F, G und H:**
F=H<E=G – Rangfolge der 4 Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität;
E=H<F=G – Rangfolge der 4 Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung.

Im letzten Schritt von Phase 2 der Entwicklungsstudie galt es daher die Frage zu klären, welche dieser 16 möglichen Kompositionen für die Laborsituation der Hauptstudie die geeignetste ist. Als Kriterien für diese Eignung wurden festgelegt, dass jene der 16 möglichen Kompositionen für die Laborsituation der Hauptstudie auszuwählen ist, bei der sich in einer dritten Codierung der Schülerlösungstexte A bis H...

... die eben aufgeführten Erwartungen an die Rangfolgen der jeweiligen 4 Schülerlösungstexte und

... die Ausprägungen der Kriterien beider Kriterienraster, die Erst- und Zweitcodierer_in den 8 Schülerlösungstexten zugewiesen haben (vgl. Tabelle 5.13), am besten reproduzieren ließen.

Des Weiteren sollte die dritte Codierung der 8 Kandidatentexte von einer größeren Anzahl Codierer_innen durchgeführt werden, um *Codierer-Effekte*[86] zu minimieren. Deshalb wurden insgesamt 6 Physikdidaktiker_innen[87], analog zum Vorgehen im dritten Schritt von Phase 2 der Entwicklungsstudie (vgl. Abschnitt 5.3.3), im Umgang mit beiden (überar-

[86] *Codierer-Effekte* lassen sich nach Degen (2015) definieren als „jene Effekte auf das spezifische Codier-Ergebnis, die auf den Codierer selbst oder auf dessen ›Interaktion‹ mit der konkreten Codier-Situation oder dem zu codierenden Medieninhalt zurückgeführt werden können“ (ebd., S. 81).

[87] Bei diesen 6 Physikdidaktiker_innen handelte es sich um eine_n Professor_Professorin, einen Postdoktorierende_n und 4 Doktorierende der Physikdidaktik.

beiteten) Kriterienrastern geschult. Anschließend wurden den geschulten Codierer_innen die Schülertexte A bis H zusammen mit 5 weiteren zufällig ausgewählten Schülerlösungstexte zur Einordnung in die Kriterienraster zur Unterscheidung von Schülerlösungstexten bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung vorgelegt[88].

Im Anschluss an die dritte Codierung der 8 Kandidatentexte wurden die hierdurch gewonnenen Daten mit folgenden parameterfreien Maßen bzw. Methoden ausgewertet[89]:

Zunächst wurde der Konkordanzkoeffizient W von Kendall bestimmt (vgl. Kendall & Gibbons, 1990, S. 117 u. f.). Dieser diente als Maß für die Intercoderreliabilität der Kriterienraster zur Unterscheidung von Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung bei der Einschätzung der 8 Kandidatentexte durch die 6 Physikdidaktiker_innen. Für Kendalls W gelten – als grobe Orientierung – signifikante[90] Werte ab .10 als „sehr schwach“, ab .30 als „schwach“, ab .50 als „moderat“, ab .70 als „stark“ und Werte ab .90 als „sehr stark“ (vgl. R. C. Schmidt, 1997, S. 767). Wie aus den Tabellen 5.9 und 5.10 hervorgeht, zeigte sich beim dritten Codiervorgang bei allen Kriterien beider Kriterienraster, sowie für den Fach- und den Sprachscore, eine starke bis sehr starke Konkordanz der Einschätzungen der 6 befragten Physikdidaktiker_innen ($.717 \leq W \leq .944$). Die Intercoderreliabilität der Codierungen der 6 befragten Physikdidaktiker_innen war also zufriedenstellend, weswegen im Anschluss an den dritten Codiervorgang – anders als im dritten Schritt von Phase 2 der Entwicklungsstudie (vgl. Abschnitt 5.3.3) – auf eine Diskussion unterschiedlicher Codierungen verzichtet werden konnte.

Anschließend wurde die Vereinbarkeit zwischen den oben benannten Erwartungen an die Rangfolgen der 4 Schülerlösungstexte in allen 16 möglichen Kompositionen (zu reproduzierende Kontrastierung) und dem Ergebnis der dritten Codierung der Schülerlösungstexte A bis H überprüft. Zu diesem Zwecke wurde auf Grundlage der Fach- und Sprachscores aus der dritten Codierung Trendtests nach Page (1963) durchgeführt. Beim Trendtest nach Page (1963) wird der Nullhypothese, dass für die Lageparameter $m_1, m_2, ..., m_k$ von k Untersuchungsobjekten $m_1 = m_2 = ... = m_k$ gilt, die Trendalternativhypothe-

[88] Dies erfolgte, um den 6 Physikdidaktiker_innen zu verbergen, bei welchen Schülerlösungstexte es sich um vorausgewählte Kontrastfälle handelte (vgl. Abschnitt 5.3.3). Zudem wurde jedem_jeder Physikdidaktiker_in jeweils andere 5 zufällig ausgewählte Schülerlösungstexte vorgelegt. Dieses Vorgehen, diente ebenfalls dazu, Codierer-Effekte bei der dritten Codierung der 8 Kandidatentexte zu minimieren (teilweise Randomisierung der zu Schülerlösungstexten, mit denen die einzelnen Codierer_in bei der Codierung in Interaktion treten) (vgl. Degen, 2015, S. 79 u. f.).

[89] Bei der Auswertung fand ein mutiples Testen in derselben (Teil-)Stichprobe statt. Da es sich bei der vorliegenden Arbeit um eine Studie mit explorativem Charakter handelt (vgl. Unterkapitel 4.2), wurde in dieser Untersuchung generell auf eine Adjustierung des globalen α-Niveaus verzichtet (vgl. Bender, St., & Ziegler, 2002, S. T6; Victor, Elsäßer, Hommel, & Blettner, 2010, S. 55). Wenn daher im empirischen Teil der vorliegenden Arbeit von „Signifikanz“ gesprochen wird, meint dies streng genommen also (lediglich) das Auftreten „von auffällig kleinen p-Werten, die als Motivation für eventuelle weitere (dann vielleicht konfirmatorische) Studien dienen“ (Victor et al., 2010, S. 55).

[90] Zur Signifikanzbewertung von Kendalls W wird mittels eines χ^2-Tests überprüft, ob zwischen den Einschätzungen der befragten Personen ein Zusammenhang besteht (vgl. S. Siegel, 1976, S. 233 u. f.). Im Rahmen der vorliegenden Arbeit wurde hierbei das Signifikanzniveau $\alpha = .05$ gewählt.

Kriterium	**Kendalls** W	**Interpretation** (vgl. R. C. Schmidt, 1997, S. 767)
Rahmenbau der Erklärung	.717****	starke Konkordanz
Rolle von Evidenzbezügen	.754****	starke Konkordanz
Tiefe der Erklärung	.736****	starke Konkordanz
Konsistenz der Erklärung	.944****	sehr starke Konkordanz
Fachscore	.943****	sehr starke Konkordanz

Tabelle 5.9.: Konkordanz der Einschätzung der befragten Physikdidaktiker_innen (Kendalls W) für alle Kandidaten kontrastierender Schülerlösungstexte für die Kriterien des Rasters zur Unterscheidung von Schülerlösungstexten hinsichtlich ihrer fachlich-konzeptuellen Qualität (****:$p \leq .0001$; χ^2-Test).

Kriterium	**Kendalls** W	**Interpretation** (vgl. R. C. Schmidt, 1997, S. 767)
Lexik/Semantik	.817****	starke Konkordanz
Syntax/Stilistik	.790****	starke Konkordanz
Sprachscore	.914****	sehr starke Konkordanz

Tabelle 5.10.: Konkordanz der Einschätzung der befragten Physikdidaktiker_innen (Kendalls W) für alle Kandidaten kontrastierender Schülerlösungstexte für die Kriterien des Rasters zur Unterscheidung von Schülerlösungstexten hinsichtlich der Qualität ihrer sprachlichen Realisierung (****:$p \leq .0001$; χ^2-Test).

se $m_1 \leq m_2 \leq ... \leq m_k$ gegenübergestellt[91]. Die oben benannten Erwartungen an die Rangfolge von 4 Kandidatentexten stellen allerdings strengere Trendhypothesen der Form $m_1 = m_2 < m_3 = m_4$ dar. Deshalb wurden für jede Erwartung an die Rangfolge von 4 Kandidatentexten jeweils 4 Trendtests nach Page (1963) mit den Trendalternativhypothesen $m_1 \leq m_2 \leq m_3 \leq m_4$, $m_2 \leq m_1 \leq m_3 \leq m_4$, $m_1 \leq m_2 \leq m_4 \leq m_3$ und $m_2 \leq m_1 \leq m_4 \leq m_3$ durchgeführt. Im Fall eines signifikanten Ausgangs aller 4 durchgeführten Trendtests, ist die Annahme gerechtfertigt, dass die Rangfolge der 4 Kandidatentexte, die sich empirisch, auf Basis der Fach- bzw. Sprachscores, die die befragten Physikdidaktiker_innen ihnen zugewiesen haben, ergibt, mit den oben beschriebenen theoretischen Erwartung an ihre Rangfolge vereinbar ist. Die Ergebnisse dieser insgesamt $2 \cdot 16 \cdot 4 = 128$ Trendtests nach Page (1963) sind in den Tabellen 5.11 und 5.12 zusammengetragen. In Tabelle 5.11 finden sich die Ergebnisse für die Trendtests auf Grundlage der Fachscores der 4 Schülerlösungstexte der 16 möglichen Kompositionen, in Tabelle 5.12 jene auf Grundlage der Sprachscores. Wie aus beiden Tabellen hervorgeht, hatten alle 128 durchgeführten Trendtests einen zum Signifikanzniveau $\alpha = .05$ signifikanten Ausgang. Im Rahmen des dritten Codiervorgangs konnten also alle oben benannten Erwartun-

[91] Page (1963) gibt in seiner ursprünglichen Veröffentlichung die Trendalternativhypothese $m_1 > m_2 > ... > m_k$ an. Spätere Veröffentlichungen (z. B. S. Siegel & Castellan, 1988, S. 184 u. f.; van de Wiel & Di Bucchianico, 2001, S. 276 u. f.; Hollander, Wolfe, & Chicken, 2014, S. 304 u. f.) stellten jedoch richtig, dass beim Trendtest nach Page (1963) die Trendalternativhypothese $m_1 \leq m_2 \leq ... \leq m_k$ lautet, wobei wobei mindestens eine strikte Ungleichung gilt.

Trendhypothese: B=D<A=C

Getestete Trendalternativhypothese	Pages L
D≤B≤A≤C	177.5***
B≤D≤A≤C	175.5***
D≤B≤C≤A	172.5***
B≤D≤C≤A	170.5**

Trendhypothese: B=H<A=C

Getestete Trendalternativhypothese	Pages L
H≤B≤A≤C	179.0***
B≤H≤A≤C	174.0***
H≤B≤C≤A	174.0***
B≤H≤C≤A	169.0**

Trendhypothese: B=D<A=G

Getestete Trendalternativhypothese	Pages L
D≤B≤A≤G	177.5***
B≤D≤A≤G	175.5***
D≤B≤G≤A	172.5***
B≤D≤G≤A	170.5**

Trendhypothese: F=D<A=C

Getestete Trendalternativhypothese	Pages L
D≤F≤A≤C	178.5***
F≤D≤A≤C	174.5***
D≤F≤C≤A	173.5***
F≤D≤C≤A	169.5**

Trendhypothese: B=D<E=C

Getestete Trendalternativhypothese	Pages L
D≤B≤E≤C	178.0***
B≤D≤E≤C	176.0***
D≤B≤C≤E	172.0***
B≤D≤C≤E	170.0**

Trendhypothese: B=H<A=G

Getestete Trendalternativhypothese	Pages L
H≤B≤A≤G	179.0***
B≤H≤A≤G	174.0***
H≤B≤G≤A	174.0***
B≤H≤G≤A	169.0**

Trendhypothese: F=D<A=G

Getestete Trendalternativhypothese	Pages L
D≤F≤A≤G	178.5***
F≤D≤A≤G	174.5***
D≤F≤G≤A	173.5***
F≤D≤G≤A	169.5**

Trendhypothese: F=D<E=C

Getestete Trendalternativhypothese	Pages L
D≤F≤E≤C	178.5***
F≤D≤E≤C	174.0***
D≤F≤C≤E	172.0***
F≤D≤C≤E	167.5**

Trendhypothese: B=H<E=C

Getestete Trendalternativhypothese	Pages L
D≤B≤E≤C	179.5***
B≤D≤E≤C	174.5***
D≤B≤C≤E	173.5***
B≤D≤C≤E	168.5**

Trendhypothese: F=H<A=C

Getestete Trendalternativhypothese	Pages L
H≤F≤A≤C	179.5***
F≤H≤A≤C	173.5***
H≤F≤C≤A	174.5***
F≤H≤C≤A	168.5**

Trendhypothese: B=D<E=G

Getestete Trendalternativhypothese	Pages L
D≤B≤E≤G	178.0***
B≤D≤E≤G	176.0***
D≤B≤G≤E	172.0***
B≤D≤G≤E	170.0**

Trendhypothese: F=H<A=G

Getestete Trendalternativhypothese	Pages L
H≤F≤A≤G	179.5***
F≤H≤A≤G	173.5***
H≤F≤G≤A	174.5***
F≤H≤G≤A	168.5**

Trendhypothese: B=H<E=G

Getestete Trendalternativhypothese	Pages L
H≤B≤E≤G	179.5***
B≤H≤E≤G	174.5***
H≤B≤G≤E	173.5***
B≤H≤G≤E	168.5**

Trendhypothese: F=H<E=C

Getestete Trendalternativhypothese	Pages L
H≤F≤E≤C	179.5***
F≤H≤E≤C	173.0***
H≤F≤C≤E	173.0***
F≤H≤C≤E	166.5*

Trendhypothese: F=D<E=G

Getestete Trendalternativhypothese	Pages L
D≤F≤E≤G	178.5***
F≤D≤E≤G	174.0***
D≤F≤G≤E	172.0***
F≤D≤G≤E	167.5**

Trendhypothese: F=H<E=G

Getestete Trendalternativhypothese	Pages L
H≤F≤E≤G	179.5***
F≤H≤E≤G	173.0***
H≤F≤G≤E	173.0***
F≤H≤G≤E	166.5*

Tabelle 5.11.: Trendtest nach Page (1963) für alle möglichen Kompositionen aus 4 kontrastierenden Schülerlösungstexten auf Grundlage der Fachscores aus dem dritten Codiervorgang (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$).

Trendhypothese: A=D<B=C

Getestete Trendalternativhypothese	Pages L
D≤A≤B≤C	175.5***
A≤D≤C≤B	176.0***
D≤A≤B≤C	170.0**
A≤D≤C≤B	170.5**

Trendhypothese: A=H<B=C

Getestete Trendalternativhypothese	Pages L
H≤A≤B≤C	174.5***
A≤H≤C≤B	177.0***
H≤A≤B≤C	169.0**
A≤H≤C≤B	171.5**

Trendhypothese: A=D<B=G

Getestete Trendalternativhypothese	Pages L
D≤A≤B≤G	175.5***
A≤D≤G≤B	176.6***
D≤A≤B≤G	170.0**
A≤D≤G≤B	170.5**

Trendhypothese: A=D<F=C

Getestete Trendalternativhypothese	Pages L
D≤A≤F≤C	177.0***
A≤D≤C≤F	177.0***
D≤A≤F≤C	171.0**
A≤D≤C≤F	171.0**

Trendhypothese: E=D<B=C

Getestete Trendalternativhypothese	Pages L
D≤E≤B≤C	175.5***
E≤D≤C≤B	176.0***
D≤E≤B≤C	170.0**
E≤D≤C≤B	170.5**

Trendhypothese: A=H<B=G

Getestete Trendalternativhypothese	Pages L
H≤A≤B≤G	174.5***
A≤H≤G≤B	177.0***
H≤A≤B≤G	169.0**
A≤H≤G≤B	171.5**

Trendhypothese: A=D<F=G

Getestete Trendalternativhypothese	Pages L
D≤A≤F≤G	177.0***
A≤D≤G≤F	177.0***
D≤A≤F≤G	171.0**
A≤D≤G≤F	171.0**

Trendhypothese: E=D<F=C

Getestete Trendalternativhypothese	Pages L
D≤E≤F≤C	177.0***
E≤D≤C≤F	177.0***
D≤E≤F≤C	171.0**
E≤D≤C≤F	171.0**

Trendhypothese: E=H<B=C

Getestete Trendalternativhypothese	Pages L
H≤E≤B≤C	174.5***
E≤H≤C≤B	177.0***
H≤E≤B≤C	169.0**
E≤H≤C≤B	171.5**

Trendhypothese: A=H<F=C

Getestete Trendalternativhypothese	Pages L
H≤A≤F≤C	175.0***
A≤H≤C≤F	177.5***
H≤A≤F≤C	168.5**
A≤H≤C≤F	171.0**

Trendhypothese: E=D<B=G

Getestete Trendalternativhypothese	Pages L
D≤E≤B≤G	175.5***
E≤D≤G≤B	176.0***
D≤E≤B≤G	170.0**
E≤D≤G≤B	170.5**

Trendhypothese: A=H<F=G

Getestete Trendalternativhypothese	Pages L
H≤A≤F≤G	175.0***
A≤H≤G≤F	177.5***
H≤A≤F≤G	168.5**
A≤H≤G≤F	171.0**

Trendhypothese: E=H<B=G

Getestete Trendalternativhypothese	Pages L
H≤E≤B≤G	174.5***
E≤H≤G≤B	177.0***
H≤E≤B≤G	169.0**
E≤H≤G≤B	171.5**

Trendhypothese: E=H<F=C

Getestete Trendalternativhypothese	Pages L
H≤E≤F≤C	175.0***
E≤H≤C≤F	177.5***
H≤E≤F≤C	168.5**
E≤H≤C≤F	171.0**

Trendhypothese: E=D<F=G

Getestete Trendalternativhypothese	Pages L
D≤E≤F≤G	177.0***
E≤D≤G≤F	177.0***
D≤E≤F≤G	171.0**
E≤D≤G≤F	171.0**

Trendhypothese: E=H<F=G

Getestete Trendalternativhypothese	Pages L
H≤E≤F≤G	175.0***
E≤H≤G≤F	177.5***
H≤E≤F≤G	168.5**
E≤H≤G≤F	171.0**

Tabelle 5.12.: Trendtest nach Page (1963) für alle möglichen Kompositionen aus 4 kontrastierenden Schülerlösungstexten auf Grundlage der Sprachscores aus dem dritten Codiervorgang (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$).

gen an die Rangfolgen der 4 Schülerlösungstexte in alle 16 möglichen Kompositionen empirisch bestätigt werden. Dies bedeutete allerdings auch, dass auf Grundlage der durchgeführten Trendtests nach Page (1963) keine Entscheidung darüber getroffen werden konnte, welche der 16 möglichen Kompositionen kontrastierender Schülerlösungstexte für die Laborsituation der Hauptstudie ausgewählt werden sollte.

Aus diesem Grund wurden die Ergebnisse der dritten Codierung einer genaueren Analyse unterzogen. Zwecks dessen wurde für alle 8 Kandidatentexte für alle Kriterien beider Kriterienraster der Konsenskoeffizient Ξ bestimmt. Dieser Koeffizient wurde eigens für die Entwicklungsstudie der vorliegenden Arbeit entwickelt und ist eine Verallgemeinerung des Einigkeitskoeffizienten η nach Einhaus (2007), der „die Beurteilungsübereinstimmung bei einem bestimmten Item hinsichtlich eines bestimmten Merkmals [beschreibt]“ (ebd., S. 33). Im an dieser Stelle der vorliegenden Arbeit beschriebenen Fall (die Einschätzung von 8 Schülerlösungstexten zur Aufgabe Weltraumspaziergang durch 6 geschulte Physikdidaktiker_innen) beschreibt der Konsenskoeffizient Ξ also, inwieweit beim dritten Codiervorgang zwischen den codierenden Physikdidaktiker_innen ein Konsens darüber besteht, ob einem bestimmten Schülerlösungstext eine bestimmte Ausprägung eines Kriteriums aus einem der beiden Kriterienraster zuzuweisen ist. In Anhang B ist das Verfahren zur Berechnung und Interpretation des Konsenskoeffizienten Ξ ausführlich erläutert. Die zentralen Merkmale und Schritte dieses Verfahrens lassen sich wie folgt zusammenfassen:

1. Der Konsenskoeffizient Ξ wird für jedes Merkmal eines Untersuchungsobjekts, das von mehreren Beurteilern_Beurteilerinnen eingeschätzt wurde, bestimmt (hier: Für jedes Kriterium beider Kriterienraster für jeden der 8 Kandidatentexte, die von 6 geschulten Physikdidaktiker_innen codiert wurden).
2. Für den Konsenskoeffizient gilt $0 \leq \Xi \leq 1$, wobei $\Xi = 0$ genau dann eintritt, wenn die Einschätzungen der Beurteiler_innen über die Merkmalsausprägung eines Untersuchungsobjekts maximal nicht übereinstimmen und $\Xi = 1$ genau dann, wenn alle Beurteiler_innen zu derselben Einschätzung über die Merkmalsausprägung eines Untersuchungsobjekts gelangt sind.
3. Nimmt Ξ mindestens einen kritischen Wert Ξ_{krit} an, so ist die Übereinstimmung der Beurteiler_innen groß genug, dass von einem hinreichenden Konsens unter den Befragten gesprochen werden kann. In diesem Fall ist der Modalwert der Beurteilereinschätzungen über die Merkmalsausprägung eines Untersuchungsobjekts als dessen Konsensmerkmalsausprägung zu interpretieren.
4. Ξ_{krit} wird maßgeblich von den Mehrheitsverhältnissen der Befragten untereinander bestimmt (vgl. Anhang B). Für den an dieser Stelle der vorliegenden Arbeit beschriebenen Fall wird ein besonders strenges „cut-off“-Kriterium gewählt, nämlich dass 4 der 6 befragten Physikdidaktiker_innen dieselbe Ausprägung eines Kriteriums zugewiesen haben müssen (Zweidrittelmehrheit).

Rahmenbau der Erklärung	**Schülerlösungstext:**							
	A	B	C	D	E	F	G	H
Ausprägung (Score) bei der Vorauswahl	2	0	2	0	2	2	2	0
Modale Ausprägung (Score) im 3. Codiervorgang d	2	0	2	0	2	2	2	0
Häufigkeit der modalen Ausprägung im 3. Codiervorgang k_d	5	4	6	6	6	4	6	6
Konensenskoeffizient im 3. Codiervorgang Ξ	.44	.11	1.00	1.00	1.00	.11	1.00	1.00
Kritischer Wert des Konsenskoeffizienten Ξ_{krit}	.11	.11	.11	.11	.11	.11	.11	.11
Obere kumulierte Binomialverteilung $P(6, k_d, \frac{1}{2})$	.094$^{\times}$	.234	.000*	.000*	.000*	.234	.000*	.000*

Rolle von Evidenzbezügen	**Schülerlösungstext:**							
	A	B	C	D	E	F	G	H
Ausprägung (Score) bei der Vorauswahl	2	1	2	1	2	1	2	1
Modale Ausprägung (Score) im 3. Codiervorgang d	1	1	2	1	1	1	2	0
Häufigkeit der modalen Ausprägung im 3. Codiervorgang k_d	4	6	6	5	4	5	6	4
Konensenskoeffizient im 3. Codiervorgang Ξ	.43	1.00	1.00	.65	.43	.65	1.00	.43
Kritischer Wert des Konsenskoeffizienten Ξ_{krit}	.43	.43	.43	.43	.43	.43	.43	.43
Obere kumulierte Binomialverteilung $P(6, k_d, \frac{1}{3})$	.082$^{\times}$	.000*	.000*	.016*	.082$^{\times}$	.016*	.000*	.082$^{\times}$

Tiefe der Erklärung	**Schülerlösungstext:**							
	A	B	C	D	E	F	G	H
Ausprägung (Score) bei der Vorauswahl	1	1	2	1	1	1	2	0/1
Modale Ausprägung (Score) im 3. Codiervorgang d	1	1	2	1	1	1	2	1
Häufigkeit der modalen Ausprägung im 3. Codiervorgang k_d	4	6	6	5	5	5	6	3
Konensenskoeffizient im 3. Codiervorgang Ξ	.43	1.00	1.00	.65	.65	.65	1.00	.35
Kritischer Wert des Konsenskoeffizienten Ξ_{krit}	.43	.43	.43	.43	.43	.43	.43	.43
Obere kumulierte Binomialverteilung $P(6, k_d, \frac{1}{3})$	.082$^{\times}$	.000*	.000*	.016*	.016*	.016*	.000*	.219

Konsistenz der Erklärung	**Schülerlösungstext:**							
	A	B	C	D	E	F	G	H
Ausprägung (Score) bei der Vorauswahl	2	0	2	0	2	0	2	0
Modale Ausprägung (Score) im 3. Codiervorgang d	1	0	2	0	1	0	2	0
Häufigkeit der modalen Ausprägung im 3. Codiervorgang k_d	4	6	6	6	5	6	5	6
Konensenskoeffizient im 3. Codiervorgang Ξ	.43	1.00	1.00	1.00	.65	1.00	.65	1.00
Kritischer Wert des Konsenskoeffizienten Ξ_{krit}	.43	.43	.43	.43	.43	.43	.43	.43
Obere kumulierte Binomialverteilung $P(6, k_d, \frac{1}{3})$	.082$^{\times}$	.000*	.000*	.000*	.016*	.000*	.016*	.000*

Lexik/Semantik	**Schülerlösungstext:**							
	A	B	C	D	E	F	G	H
Ausprägung (Score) bei der Vorauswahl	0	2	2	0	0	2	2	0
Modale Ausprägung (Score) im 3. Codiervorgang d	0	1	2	0	0	1	1	0
Häufigkeit der modalen Ausprägung im 3. Codiervorgang k_d	6	6	4	5	6	4	4	5
Konensenskoeffizient im 3. Codiervorgang Ξ	1.00	1.00	.43	.65	1.00	.43	.43	.65
Kritischer Wert des Konsenskoeffizienten Ξ_{krit}	.43	.43	.43	.43	.43	.43	.43	.43
Obere kumulierte Binomialverteilung $P(6, k_d, \frac{1}{3})$	.000*	.000*	.082$^{\times}$	.016*	.000*	.082$^{\times}$	.082$^{\times}$	.016*

Syntax/Stilistik	**Schülerlösungstext:**							
	A	B	C	D	E	F	G	H
Ausprägung (Score) bei der Vorauswahl	0	2	2	1	1	2	2	1
Modale Ausprägung (Score) im 3. Codiervorgang d	0	1	2	0	0	1	2	0
Häufigkeit der modalen Ausprägung im 3. Codiervorgang k_d	4	4	5	5	4	3	4	3
Konensenskoeffizient im 3. Codiervorgang Ξ	.43	.33	.65	.65	.43	.15	.65	.35
Kritischer Wert des Konsenskoeffizienten Ξ_{krit}	.43	.43	.43	.43	.43	.43	.43	.43
Obere kumulierte Binomialverteilung $P(6, k_d, \frac{1}{3})$	.082$^{\times}$	.082$^{\times}$	.016*	.016*	.082$^{\times}$	.219	.016*	.219

Tabelle 5.13.: Zusammenfassung deskriptiver Befunde des dritten Codiervorgangs für alle Kandidaten kontrastierender Schülerlösungstexte, geordnet nach den Kriterien beider Kriterienraster (*:$P(m, k_d, \frac{1}{w}) \leq .05$; $^{\times}$:$P(m, k_d, \frac{1}{w}) \leq .10$).

5. Um im Fall $\Xi \geq \Xi_{krit}$ eine zufällige Übereinstimmung der Beurteiler_innen auszuschließen, wird für die Häufigkeit des Auftretens der Konsensmerkmalsausprägung in der Befragung ein rechtsseitiger Binomialtest ($H_0 : p_d = \frac{1}{w}$; $H_1 : p_d > \frac{1}{w}$; w : Anzahl der möglichen Merkmalsausprägungen) als Signifikanztest durchgeführt. Da im an dieser Stelle der vorliegenden Arbeit beschriebenen Fall lediglich 6 Physikdidaktiker_innen befragt wurden, wurde für diese Binomialtests mit $\alpha = .10$ ein höheres Signifikanzniveau angesetzt, als für gewöhnlich üblich.

Die vorgenommenen Berechnungen der Konsenskoeffizienten für alle 8 Kandidatentexte sind in Tabelle 5.13 zusammengefasst. Es zeigte sich,...

... dass in 91.7 % der Fälle (44 von 48 Berechnungen) zwischen den 6 befragten Physikdidaktiker_innen ein hinreichenden Konsens ($\Xi \geq \Xi_{krit}$) darüber bestand, dass einem bestimmten Schülerlösungstext zur Aufgabe Weltraumspaziergang eine bestimmte Ausprägung eines Kriteriums zuzuweisen ist,

... dass in 89.6 % der Fälle (43 von 48 Berechnungen) der durchgeführte Binomialtest (zum Signifikanzniveau $\alpha = .10$) signifikant ausfiel und

... dass in 70.8 % der Fälle (34 von 48 Berechnungen) die Konsensmerkmalsausprägung mit jener der Vorauswahl (der Zweitcodierung aus dem dritten Schritt von Phase 2 der Entwicklungsstudie) übereinstimmte.

Mit Hilfe dieser Befunde lies sich aus den 16 möglichen eine „beste“ Komposition aus 4 kontrastierenden Schülerlösungstexten für Laborsituation der Hauptstudie identifizieren. Hierzu wurden aus allen 16 Kompositionsmöglichkeiten diejenigen identifiziert, in denen alle drei eben genannten Kriterien ($\Xi \geq \Xi_{krit}$, signifikanter Ausgang des Binomialtests und Übereinstimmung der Konsensmerkmalsausprägung mit jener der Vorauswahl) am häufigsten gleichzeitig erfüllt waren. Dies war für die Komposition aus den Schülerlösungstexten A, B, C und D, sowie für jene aus den Schülerlösungstexten A, F, C, D der Fall (jeweils in 19 von 24 Berechnungen). Von den 16 möglichen Kompositionen blieben also lediglich 2 übrig. Wie aus Tabelle 5.13 allerdings auch hervorgeht, führten die Berechnungen der Konsenskoeffizienten für Schülerlösungstext B zu leicht besseren Befunden, als im Fall von Schülerlösungstext F (höherer Konsenskoeffizient bezüglich des Kriteriums „Rolle von Evidenzbezügen“ und signifikanter Ausgang des Binomialtests bezüglich des Kriteriums „Syntax/Stilistik“). Aufgrund dessen wurde die Komposition aus den Schülerlösungstexten A, B, C und D als diejenige ausgewählt, die in der Laborsituation der Hauptstudie eingesetzt werden soll.

5.3.5. Zwischenfazit

In Phase 2 der Entwicklungsstudie galt es für die Laborsituation der Hauptstudie eine Komposition aus vier, bezüglich ihrer fachlich-konzeptuellen Qualität und/oder der Qualität ihrer sprachlichen Realisierung kontrastierende Schülerlösungstexte so sorgfältig wie möglich „zu generieren“. Deshalb wurden in einem mehrschrittigen Verfahren zunächst

Klassenarbeitsaufgaben gesammelt und erprobt und für das weitere Vorgehen die Aufgabe Weltraumspaziergang ausgewählt (vgl. Abschnitt 5.3.1). Anschließend wurden 116 Schülerlösungstexte zu dieser Aufgaben erhoben (vgl. Abschnitt 5.3.2). Durch insgesamt drei Codiervorgänge mit Hilfe von zwei eigens hierfür entwickelten Kriterienrastern (vgl. Abschnitt 5.3.2 und Anhang A) konnte aus den erhobenen Schülerlösungstexten eine „beste“ Komposition kontrastierender Schülerlösungstexte identifiziert werden (vgl. Abschnitt 5.3.3 bis 5.3.4).

Auf Grundlage der Ergebnisse der zweiten Phase der Entwicklungsstudie, wurde die in Phase 1 entwickelte Skizze der Laborsituation (vgl. Abschnitt 5.2.2) zu einem genauen Ablaufplan weiterentwickelt. Ferner konnten die in der Laborsituation (zur Datenerhebung) benötigten Materialien zusammengestellt und pilotiert werden. Dies wird im nun folgenden Unterkapitel, das sich der dritten Phase 3 der Entwicklungsstudie widmet, detailliert dargestellt.

5.4. Phase 3: Entwicklung und Pilotierung eines Ablaufplans der Laborsituation der Hauptstudie und erste Überlegung zu gegenstandsangemessenen Auswertungsmethoden

In Phase 3 der Entwicklungsstudie wurde die in Abschnitt 5.2.2 dargestellte Skizze einer für die Beantwortung der Forschungsfragen (F1) und (F2) geeigneten Laborsituation zu einem genauen Ablaufplan weiterentwickelt und die in der Laborsituation benötigten Materialien zusammengestellt. Das dabei gewählte Vorgehen lässt sich in drei Schritte gliedern:

Schritt 1: Auf Basis der Empfehlungen der methodentheoretischen Literatur zur praktischen Umsetzung von lautem Denken und retrospektiver Befragung als Erhebungsmethode[92] wurden (a) ein vorläufiges Manual erstellt, in dem die Durchführung der Erhebung detailliert beschrieben ist und (b) ein Aufgabenheft, sowie ein Fragebogen für die an der Hauptstudie teilnehmenden Physiklehrkräfte entwickelt. Das Manual stellt für den_die Forscher_in, der_die die Laborsituation leitet[93], das Regelwerk dar und dient dazu, den Ablauf und die Bedingungen der Laborsituation konstant zu halten. Das Aufgabenheft wurde so aufgebaut, dass die teilnehmenden Lehrkräfte dieses selbstständig und laut

[92] Hierzu wurden vor allem die Arbeiten von Ericsson & Simon (1985), van Someren et al. (1994), A. Green (1998), C. Green & Gilhooly (2002), Heine & Schramm (2007), Knorr & Schramm (2012), Arras (2013) und Heine & Schramm (2016) herangezogen. Die Instruktionen und Übungsaufgaben im Manual zur Durchführung der Erhebung sind zum Teil in gekürzter und/oder geänderter Form aus den Arbeiten von Arras (2007, S. 499) und van Someren et al. (1994, S. 174) übernommen.

[93] Der Einfachheit halber wird der_die Forscher_in, der_die die Laborsituation leitet im Folgenden kurz als *Leitung* bezeichnet.

denkend bearbeiten können, ohne hierbei unterbrochen werden zu müssen (z. B. um zusätzliche Instruktionen zu geben). Zweck des Lehrkräftefragebogens ist eine detaillierteren Beschreibung des Samples im Rahmen der Auswertung der Hauptstudie. Deshalb werden im Fragebogen von den Teilnehmer_innen Eckdaten zu ihrer Person erfasst, sowie Selbstauskunftsskalen, die aus der Dokumentation der Erhebungsinstrumente der *COACTIV*-Studie (vgl. Baumert et al., 2009) und der Untersuchung von Riebling (2013b) übernommen wurden, abgefragt (vgl. Anhang C).

Schritt 2: Der Ablauf der Laborsituation und die dabei benötigten Materialien (unter Anderem das Aufgabenheft und der Lehrkräftefragebogen) wurden anschließend mit zwei Referendaren_Referendarinnen und einem Lehramtsstudenten pilotiert[94]. Bei der Pilotierung wurde sich streng an die Ablaufschritte des vorläufigen Manuals gehalten. Im unmittelbaren Anschluss an die Pilotierung der Laborsituation wurde mit jedem_jeder Teilnehmer_in ein halboffenes Interview geführt. Dabei wurden die Teilnehmer_innen gebeten die Fragen zu beantworten, wie sie aus ihrer Sicht mit der Laborsituation zurechtgekommen sind, für wie verständlich und (weniger) hilfreich sie die einzelnen Materialien empfunden haben, sowie ob sie an der Laborsituation Änderungen vornehmen würden und falls ja, welche. Die halboffenen Interviews wurden dabei audiographiert.

Schritt 3: Auf Grundlage der in der Pilotierung erhobenen Daten (lautes Denken, retrospektive Befragung und halboffene Interviews) wurden das Manual zur Durchführung der Erhebung, das Aufgabenheft und der Lehrkräftefragebogen zu Endversionen überarbeitet. Diese Endversionen werden in Abschnitt 5.4.1 kommentiert dargestellt. Die Originale des Manuals zur Durchführung der Erhebung, des Aufgabenhefts und des Lehrkräftefragebogens sind in Anhang C zu finden.

Des Weiteren wurden in Phase 3 der Entwicklungsstudie Vorüberlegungen zur gegenstandsangemessenen Auswertung der in der Laborsituation gewonnenen Daten unternommen. Diese Überlegungen werden in Abschnitt 5.4.2 dargestellt.

[94] Die Pilotierungen fanden am 09.02.2016, 14.02.2016 und 16.02.2016 in den Privatwohnungen der Teilnehmenden statt. Bei den beiden Referendaren_Referendarinnen handelte es sich um einen Referendar für gymnasiales Lehramt für die Fächer Physik und Biologie und eine Referendarin für gymnasiales Lehramt für die Fächer Physik und Mathematik. Beide Referendare_Referendarinnen hatten kurz vor der Erhebung das erste Schulhalbjahr ihrer zweiten Ausbildungsphase abgeschlossen und gaben zum Erhebungszeit 3 bzw. 2 Stunden eigenverantwortlichen Physikunterricht in der Sekundarstufe I. Bei dem Studenten handelte es sich um einen Lehramtsstudenten für gymnasiales Lehramt für die Fächer Physik und Mathematik (Staatsexamen) im 9. Fachsemester, der laut eigenen Angaben – abgesehen den Pflichtpratika im Rahmen seines Studiums – bislang keine Lehrerfahrung besaß.

5.4.1. Geplanter Ablauf der Laborsituation der Hauptstudie

5.4.1.1. Gesamtüberblick über den Ablauf der Laborsituation

Zunächst soll ein Gesamtüberblick über den Ablauf der Laborsituation gegeben werden: Der zeitliche Umfang einer einzelnen Erhebung ist auf zirka 90 Minuten angesetzt (vgl. Anhang C.3 Durchführungsmanual, S. 2). Die Laborsituation lässt sich in drei thematisch unterschiedliche und chronologisch aufeinanderfolgende Teile gliedern (vgl. Abbildung 5.7.):

Teil 1: Eine Trainingsphase in der die teilnehmende Lehrkraft über die Erhebungsmethode des lauten Denken informiert wird und in der sie in das Anwenden des lauten Denkens bei einer (mentalen, interaktionalen oder aktionalen) Handlung geschult wird.

Teil 2: Anschließend erfolgt das laut denkende Korrigieren der 4 Schülerlösungstexte, die im Rahmen von Phase 2 der Entwicklungsstudie ausgewählt wurden, durch die teilnehmende Lehrkraft. Der_Die Teilnehmer_in erhält dabei (durch das Aufgabenheft) die Anweisung entsprechend ihrer eigenen Praxis im Berufsalltag und mit Hilfe eines eigenes zuvor erstellen Erwartungshorizontes vorzugehen.

Teil 3: Zuletzt erfolgt die retrospektive Befragung der teilnehmenden Physiklehrkraft. Bei diese Befragung nimmer der_die Teilnehmer_in Paarvergleiche der 4 von ihm_-ihr korrigierten Schülertexte vor. Bei diesen Paarvergleichen wird der_die Teilnehmer_in gebeten, (nachträglich) eine Einschätzungen über die fachlich-konzeptuelle Qualität bzw. die Qualität der sprachlichen Realisierung der Schülerlösungstexte vorzunehmen und ihre Einschätzungen zu begründen. Nach Abschluss der retrospektiven Befragung füllt der_die Teilnehmer_in den Lehrkräftefragebogen aus.

Abbildung 5.7.: Chronologischer Gesamtüberblick über den Ablauf der Laborsituation.

5.4.1.2. Vorbereitung und Beginn der Erhebung

Die Erhebung findet in einem eigens hierfür präparierten Raum statt (siehe unten; vgl. Anhang C.3 Durchführungsmanual, S. 2). Das Geschehen während der Durchführung wird audiographiert. An ausgewählten Stellen des lauten Denkens der teilnehmenden Lehrkraft und der retrospektiven Befragung werden von der Leitung Notizen im Durchführungsmanual vermerkt (siehe unten).

Vor dem Beginn der Durchführung wird für die teilnehmende Lehrkraft ein Code generiert, um deren Anonymität zu gewährleisten. Dieser besteht aus den ersten drei Buchstaben des Vornamens ihres Vaters, ihrer Körpergröße in Zentimetern und einer beliebigen Zahl zwischen 0 und 9 (z. B. BER1762). Die teilnehmende Lehrkraft wird in der Tonbandaufnahme ausschließlich mit diesem Code angesprochen. Ferner wird, um eine Zuordnung im Anschluss an die Erhebung zu gewährleisten, der Code auf den Materialien vermerkt, die der_die Teilnehmer_in während der Erhebung bearbeitet, sowie auf dem Durchführungsmanual der Leitung.

Nach dem Start der Tonbandaufnahme wird der teilnehmenden Lehrkraft die folgende Instruktion laut vorgelesen:

> „*Ziel dieser Studie ist es, möglichst viel darüber herauszufinden, wie Sie bei der Bewertung einer Klassenarbeit vorgehen. Es geht also* ***nicht*** *darum, Ihre Arbeit zu kontrollieren. Vielmehr geht es darum, mehr über die Strategien herauszufinden, die Sie bei der Beurteilung anwenden. Die Bewertungsarbeit sollte deshalb möglichst so ablaufen, wie Sie dies unter normalen Umständen auch tun würden. Alle hierfür notwendigen Materialien stehen Ihnen zur Verfügung.*“ (Anhang C.3 Durchführungsmanual, S. 3, Hervorhebungen im Original)

Der_Die Teilnehmer_in wird zu Beginn der Erhebung also über die allgemeinen Ziele und das Vorgehen während der Untersuchung informiert. Er_Sie wird dabei allerdings bewusst im Verborgenen gehalten, dass die Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Merkmale von Schülerlösungstexten zu einer Klassenarbeitsaufgabe im Vordergrund des Erkenntnisinteresses stehen (vgl. Forschungsfrage (F1) und (F2)). Hierdurch soll vermieden werden, dass es durch die einleitende Instruktion zu einer Verschiebung des Aufmerksamkeitsfokus der teilnehmenden Lehrkraft bezüglich fachlich-konzeptueller bzw. sprachlicher Merkmale von Schülerlösungstexten kommt, die von ihrer Alltagspraxis bei der Feststellung und Beurteilung von Schülerleistungen abweicht.

5.4.1.3. Teil 1: Trainingsphase

Nach der einleitenden Instruktion wird die teilnehmende Lehrkraft zunächst über die Erhebungsmethode des lauten Denken und den Ablauf der Trainingsphase informiert. Hierzu wird ihr folgende Instruktion laut vorgelesen und darauffolgend ihre dabei aufkommenden Fragen beantwortet:

> „*Wir verwenden die sog. Think-Aloud-Methode (Methode des lauten Denkens). Das bedeutet: Während Sie bewerten, sollen Sie all das laut äußern, was sie gerade denken und machen. Stellen Sie sich am besten vor, dass Sie alleine im Raum sind und mit sich selbst sprechen.*

> *Ein Diktiergerät wird dabei Ihre Äußerungen aufzeichnen. Das ist nicht ganz einfach, weil man sehr viel schneller denkt, als man verbalisieren kann. Deshalb werde ich Ihnen bevor wir mit der eigentlichen Untersuchung beginnen ein Lernvideo zur Think-Aloud-Methode zeigen. Anschließend werden wir eine kurze Übungssequenz machen, bei der Sie sich auf die Methode einstellen können. Bei dieser Übungsphase werde ich Ihnen ggf. noch ein paar Hinweise geben.*“ (Anhang C.3 Durchführungsmanual, S. 4, Hervorhebungen im Original)

Anschließend wird der teilnehmenden Lehrkraft das angekündigte Lernvideo[95] zur Methode des lauten Denkens gezeigt und danach ihre aufgekommenen Fragen beantwortet. In diesem wird an einem einfachen Beispiel das Anwenden des lauten Denkens bei einer (mentalen, interaktionalen oder aktionalen) Handlung beschrieben und illustriert (vgl. Trump, 2015, S. 85). Bei dem Beispiel handelt es um einen zirka 4 minütigen Videosequenz, die eine Person dabei zeigt, wie sie laut denkend das Problem löst, ob sich in einem von drei Überraschungseiern eine Figur befindet (vgl. ebd.; Abbildung 5.8).

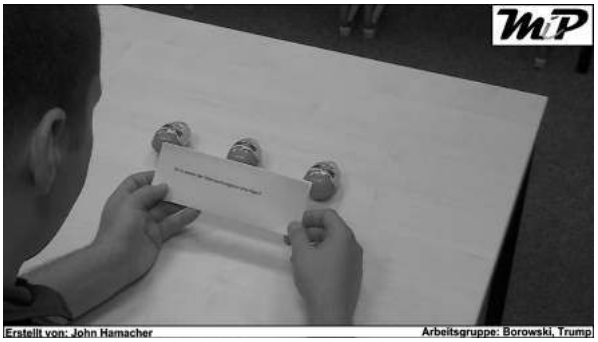


Abbildung 5.8.: Momentaufnahme aus dem Lernvideo zur Methode des lauten Denkens.

Zum Abschluss der Trainingsphase wird die teilnehmende Lehrkraft gebeten an einer einfachen Aufgabenstellung das Anwenden des lauten Denkens bei einer (mentalen, interaktionalen oder aktionalen) Handlung selbst zu üben. Zu diesem Zweck wird sie aufgefordert, ihr eigenes Handy griffbereit auf den vor ihnen platzierten Tisch zu legen. Anschließend wird dem_der Teilnehmer_in folgende Instruktion gegeben:

> „Ihre Aufgabe besteht darin ein technisches Gerät zu verbessern. Ich werde Ihnen ein technisches Gerät nennen und Ihre Aufgabe ist es, **fünf** Verbesserungen für dieses Gerät zu erfinden. Denken Sie dabei laut. Name des Gerätes: Ihr Handy“ (Anhang C.3 Durchführungsmanual, S. 5, Hervorhebungen im Original)

Während des lauten Denkens der teilnehmenden Lehrkraft nimmt die Leitung auf einem Stuhl im Raum Platz, der so positioniert wurde, dass er sich außerhalb des Blickfeldes der Lehrkraft befindet. Dies dient der Unterstützung der teilnehmende Lehrkraft während der Aufgabenbearbeitung ablaufende Gedanken unreflektiert mitvokalisieren zu können („innere Sprache“) und nicht (unbewusst) in ein lautes Aussprechen von Gedanken „zum Zweck der Kommunikation an ein Gegenüber [überzugehen]“ (Heine, 2013, S. 14; vgl. auch Knorr & Schramm, 2012, S. 189). Ferner notiert die Leitung ihr auffallende Schwierigkeiten der teilnehmende Lehrkraft mit der Anwendung des lauten Denkens. Diese Schwierigkeiten werden nach Abschluss der Übungssequenz mit dem_der Teilnehmer_in besprochen und dabei aufkommenden Fragen der Lehrkraft beantwortet.

[95] Dieses Lernvideo wurde von der Arbeitsgruppe von Prof. Dr. Andreas Borowski (Didaktik der Physik; Universität Potsdam) erstellt (vgl. Trump, 2015, S. 85) und für die empirische Untersuchung der vorliegenden Arbeit zur Verfügung gestellt.

5.4.1.4. Teil 2: lautes Denken der Teilnehmer_innen

Im Anschluss an die Trainingsphase erfolgt durch die teilnehmende Lehrkraft das laut denkende Korrigieren der 4 Schülerlösungstexte, die im Rahmen von Phase 2 der Entwicklungsstudie ausgewählt wurden. Hierzu wird dem_der Teilnehmer_in das Aufgabenheft, in dem unter Anderem die Schülertexte A, B, C und D abgedruckt sind, sowie ein rot und ein blau schreibender Stift zur Verfügung gestellt. Vor dem Beginn der Laut-Denk-Phase erfolgt eine Einweisung des_der Teilnehmers_Teilnehmerin in den Aufbau des Aufgabenhefts und den Umgang mit diesem. Während des lauten Denkens des_der Teilnehmers_Teilnehmerin verlässt die Leitung erneut das Blickfeld der Lehrkraft. Von ihrer Position aus notiert die Leitung besondere Vorkommnisse während der Datenerhebung. Insbesondere vermerkt sie im Durchführungsmanual (soweit dies von ihrer Position einsehbar ist) die Reihenfolge, in der die teilnehmende Lehrkraft die 4 Schülerlösungstexte im Aufgabenheft korrigiert, da hierzu die Arbeitsanweisungen den Teilnehmer_innen keine festen Vorgaben machen (vgl. Anhang C.3 Durchführungsmanual, S. 7).

Das Aufgabenheft besteht aus insgesamt zwei Aufgaben, die der_die Teilnehmer_in laut denkend bearbeiten soll (siehe unten). Diese Aufgaben dienen dazu, Rückschlüsse auf die Denkprozesse des_der Teilnehmers_Teilnehmerin bei der Feststellung und Beurteilung der Schülerlösungstexte A, B, C und D zur Aufgabe Weltraumspaziergang zu ermöglichen. Beide Aufgaben sind zudem kontextualisiert, um die Aufgabenbearbeitungen für den_die Teilnehmer_in, entsprechend ihrer Alltagspraxis bei der Leistungsfeststellung und -beurteilung im Rahmen einer Klassenarbeit, so authentisch wie möglich zu gestalten. Vor der eigentlichen Aufgabenbearbeitung wird daher der_die Teilnehmer_in zunächst gebeten, sich in folgenden Kontext hineinzuversetzen:

> „Stellen Sie sich folgende Situation vor:
>
> - Sie unterrichten eine 9. Klasse in Physik und haben eine Klassenarbeit geschrieben.
> - In dieser Klassenarbeit haben Sie die Aufgabe „Weltraumspaziergang“ als Grundwissensaufgabe eingesetzt (siehe Seite 3).
> - Sie haben Ihren Schülerinnen und Schülern zusätzlich folgende Anweisung gegeben:
>   1. „*Schreibt eure Antwort in ganzen Sätzen auf.*“
>   2. „*Skizzen oder Zeichnungen können bei dieser Aufgabe nicht gezählt werden.*“
> - Für die Aufgabe „Weltraumspaziergang“ möchten Sie 0 bis maximal 5 Punkte vergeben.“
>
> (Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 2, Hervorhebungen im Original)

Ferner enthält jede der beiden Aufgaben zusätzliche Instruktionen, die darauf hinweisen, welche Seiten des Aufgabenhefts für welche Aufgabe benötigt werden und was zu tun ist wenn ein bestimmter Arbeitsauftrag abgeschlossen ist[96]. Hierdurch kann die teilnehmende

[96] Beispielsweise finden sich im Aufgabenheft die Instruktionen „**Für Aufgabe 1 benötigen Sie nur die Seiten 2 und 3**“ (Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 2, Hervorhebungen im Original) oder „**Wenn Sie mit Aufgabe 1 fertig sind, blättern Sie weiter zu Seite 4**“ (ebd., Hervorhebungen im Original)

Lehrkraft das Aufgabenheft selbstständig und laut denkend bearbeiten, ohne durch zusätzliche Anweisungen seitens der Leitung unterbrochen werden zu müssen. Aus selbigem Grund erhält die teilnehmende Lehrkraft auf der ersten Seite des Aufgabenhefts folgenden allgemeinen Instruktionen, die unter Anderem noch einmal explizieren, was bei der laut denkenden Bearbeitung des Aufgabenhefts zu beachten ist:

„**Instruktion**

1. Ziel der folgenden Aufgaben ist es, möglichst viel darüber herauszufinden, wie Sie bei der Bewertung einer Klassenarbeit vorgehen.
2. Die Aufgaben haben eine feste Reihenfolge und sollen nur mit Hilfe bestimmter Materialien bearbeitet werden. Dies wird Ihnen in den Aufgaben genau beschrieben. Weichen Sie hiervon nicht ab.
3. Wenden Sie das sogenannte laute Denken an, während Sie die Aufgaben bearbeiten. Ein Diktiergerät wird dabei Ihre Äußerungen aufzeichnen.
4. Sie sollten die Texte auch laut vorlesen, damit deutlich wird, an welchen Stellen Sie ggf. Schwierigkeiten haben.
5. Da keine Videokamera mitläuft, ist es auch wichtig, dass Sie jeweils verbalisieren, was Sie gerade tun.
6. Während Sie die Aufgaben bearbeiten, werde ich anwesend sein. Ggf. werde ich etwas sagen, wenn Sie den Redefluss zu lange unterbrechen. Meine Anwesenheit sollte Sie jedoch nicht irritieren.“

(Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 1, Hervorhebungen im Original)

Auf Seite 2 des Aufgabenhefts findet sich die erste Aufgabe für die teilnehmende Lehrkraft. Bei dieser wird der_die Teilnehmer_in gebeten „für die [...] beschriebene Situation einen geeigneten Erwartungshorizont für die Aufgabe „Weltraumspaziergang“ [zu erstellen]“ (Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 2) und dabei so vorzugehen, „wie [s]ie dies unter normalen Umständen auch tun [würde]“ (ebd.). Die Verschriftlichung des Erwartungshorizonts nimmt die teilnehmende Lehrkraft auf Seite 3 des Aufgabenhefts vor. Diese besteht aus der Aufgabenstellung zu Aufgabe Weltraumspaziergang, sowie einem $17\,cm \times 19.5\,cm$ großen karierten Schreibfeld.

Nachdem der_die Teilnehmer_in einen aus ihrer Sicht für die Kontextsituation geeigneten Erwartungshorizont für die Aufgabe Weltraumspaziergang erstellt hat, bearbeiten sie die zweite Aufgabe im Aufgabenheft. Die Arbeitsanweisung an die teilnehmenden Lehrkraft lautet bei dieser wie folgt:

„**Aufgabe 2:**

Bewerten Sie die Antworten A, B, C, D mit 0 bis maximal 5 Punkten. Verwenden Sie hierzu Ihren Erwartungshorizont auf Seite 3 und einen Rotstift. Gehen Sie dabei so vor, wie Sie dies unter normalen Umständen auch tun würden.

Wenn Ihnen beim Korrigieren einer Antwort etwas auffällt, ist es sinnvoll, die Zeilennummer zu verbalisieren. Hierdurch wird in der Tonbandaufnahme klar, welche Stelle der Antwort Sie gemeint haben.“

(Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 4, Hervorhebungen im Original)

Die vier Schülerlösungstexte A, B, C und D befinden sich als photokopierte handschriftliche Abschrift jeweils auf einer eigenen Seite im Aufgabenheft (vgl. Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 5 u. f.). Von den Textmerkmalen Schriftbild und Rechtschreibung ist ein Einfluss auf die Leistungsfeststellung und -beurteilung durch Lehrkräfte im Allgemeinen bekannt (z. B. Briggs, 1970; Birkel & Birkel, 2002; Tajmel, 2017b, S. 260 u. f.). Da dieser Einfluss aber nicht im Erkenntnisinteresse von Forschungsfrage (F1) und (F2) liegt, wurde entschieden, diese Textmerkmale im Rahmen der Laborsituation zu kontrollieren. Schülerlösungstext A, B, C und D sind daher, worauf die teilnehmende Lehrkraft im Aufgabenheft hingewiesen wird, „in einer einheitlichen Handschrift geschrieben und von Rechtschreibfehlern bereinigt worden“ (Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 4).

In welcher Reihenfolge die teilnehmende Lehrkraft die Schülerlösungstexte A, B, C und D korrigiert und ob sie dabei z. B. linear oder zirkulär vorgeht, steht ihr offen (im Aufgabenheft werden hierzu keine Arbeitsanweisungen vorgegeben). Ferner werden ihr keine Zeitvorgaben gemacht, sondern sie beendet selbst das laut denkende Korrigieren, indem sie der Leitung ein Zeichen gibt, sobald sie die Korrektur aller vier Schülerlösungstexte ihrer Ansicht nach abgeschlossen hat.

5.4.1.5. Teil 3: retrospektive Befragung der Teilnehmer_innen

Zu Beginn der retrospektiven Befragung wird die teilnehmende Lehrkraft gebeten, eine mündliche Selbstauskunft darüber abzugeben, wie sie mit dem Anwenden des lauten Denkens bei der Bearbeitung des Aufgabenheftes zurecht gekommen ist. Hierzu werden ihr von der Leitung folgende Fragen gestellt: „∘ *Wie geht es Ihnen denn jetzt gerade?* ∘ *Wie sind Sie denn mit dem lauten Denken zurecht gekommen?* ∘ *Was ist Ihnen leicht bzw. was ist Ihnen schwer gefallen?*“ (Anhang C.3 Durchführungsmanual, S. 8, Hervorhebungen im Original). Dieses Vorgehen dient dazu bei der Datenauswertung einschätzen zu können, inwieweit die teilnehmende Lehrkraft das laute Denken während der Aufgabenbearbeitung als „[stark] unterbrechend und belastend [empfunden hat[97]] (Heine & Schramm, 2007, S. 175). Sollte dies der Fall gewesen sein, sind die Laut-Denk-Daten dieses_dieser Teilnehmers_Teilnehmerin aufgrund von Validitätsbedenken von der Datenanalyse auszuschließen (vgl. ebd; Heine & Schramm, 2016, S. 176 u. f.).

Anschließend wird die fachlich-konzeptuelle Qualität und die Qualität der sprachlichen Realisierung der zuvor von der teilnehmenden Lehrkraft korrigierten Schülerlösungstexte noch einmal beleuchtet. Dazu werden dem_der Teilnehmer_in die Schülerlösungstexte A, B, C und D in Paaren vorgelegt[98]. Bei vier verschiedenen Schülerlösungstexten gibt

[97] Allgemein lassen sich kaum Pauschalaussagen über die Reaktivität der Erhebungsmethode des lauten Denkens machen (vgl. Heine, 2010, S. 170), „da verschiedene Menschen ihre Gedanken offenbar unterschiedlich stark mit verbalen Formen verknüpfen“ (Heine & Schramm, 2016, S. 176). Für eine Zusammenschau bisheriger Untersuchungen zur Reaktivität des lauten Denkens als Erhebungsmethode siehe z. B. Heine (2010, S. 89).

[98] Dabei werden der teilnehmenden Lehrkraft die vier Schülerlösungstexte aus ihrem Aufgabenheft vorgelegt. Die Schülerlösungstexte enthalten daher auch die eigenen Korrekturanmerkungen des_der

es insgesamt sechs unterschiedliche Paare. Jedes dieser sechs Paare wird der teilnehmend Lehrkraft zweimal vorgelegt. Beim ersten bzw. zweiten Mal erhält der_die Teilnehmer_in folgende „Instruktion 1“ bzw. „Instruktion 2“:

> „**Instruktion 1:**
> *Beurteilen Sie, ob eine der beiden Antworten* ***fachlich besser ist, oder ob sie fachlich gleich gut sind****. Ob evtl. eine der beiden Antworten sprachlich besser ist, soll hierbei komplett unberücksichtigt bleiben.* ***Bitte begründen Sie Ihre Entscheidung****.*
>
> [...]
>
> **Instruktion 2:**
> *Beurteilen Sie, ob eine der beiden Antworten* ***sprachlich besser ist, oder ob sie sprachlich gleich gut sind****. Ob evtl. eine der beiden Antworten fachlich besser ist, soll hierbei komplett unberücksichtigt bleiben.* ***Bitte begründen Sie Ihre Entscheidung****.*“
>
> (Anhang C.3 Durchführungsmanual, S. 9, Hervorhebungen im Original)

Während der Befragung vermerkt die Leitung in einer Tabelle mit geschlossenen Ankreuzmöglichkeiten (vgl. Anhang C.3 Durchführungsmanual, S. 9) zu welchen Entscheidungen die teilnehmende Lehrkraft im Rahmen der Paarvergleiche gekommen ist (z. B. dass Schülerlösungstext B und D „fachlich gleich gut sind“).

Nachdem die teilnehmende Lehrkraft in der retrospektiven Befragung alle $2 \cdot 6$ Paarvergleichen vorgenommen hat, wird durch die Leitung das Ende der Laborsituation eingeleitet. Der_Die Teilnehmer_in wird abschließend darum gebeten, den Lehrkräftefragebogen (vgl. Anhang C.2 Lehrkräftefragebogen) auszufüllen. Zusätzlich wird ihr die Möglichkeit eröffnet Fragen über Erhebung beantwortet zu bekommen und/oder eigene Ergänzungen vorzunehmen (vgl. Anhang C.3 Durchführungsmanual, S. 10).

Teilnehmers_Teilnehmerin. Dieses Vorgehen dient als indirekte Impuls an den_die Teilnehmer_in, beim paarweisen Vergleichen der Schülerlösungstexte, (auch) retrospektiven Bezüge zur ihrer laut denkenden Korrekturarbeit vorzunehmen.

5.4.2. Vorüberlegungen zur gegenstandsangemessenen Auswertung der in der Laborsituation gewonnenen Daten

Aus Abschnitt 5.4.1 geht hervor, dass im Rahmen des geplanten Ablaufs der Laborsituation verschiedenste Daten von den an der Hauptstudie teilnehmenden Physiklehrkräften erhoben werden. Diese Daten lassen sich in einem groben Raster in „Verbaldaten“, „Teilnehmerprodukte“ und „Leitungsmitschriften“ unterteilen:

Verbaldaten
(lautes Denken;
retrospektive Befragung):

- Erwartungshorizonterstellung durch die Teilnehmer_innen
- Korrektur der 4 Schülerlösungstexte durch die Teilnehmer_innen
- Paarvergleiche zur fachlich-konzeptuellen Qualität der Schülerlösungstexte
- Paarvergleiche zur Qualität der sprachlichen Realisierung der Schülerlösungstexte

Teilnehmerprodukte
(Aufgabenheft):

- Erwartungshorizont der Teilnehmer_innen zur Aufgabe Weltraumspaziergang
- Korrekturanmerkungen der Teilnehmer_innen zu den 4 Schülerlösungstexten
- Punktevergabe der Teilnehmer_innen zu den 4 Schülerlösungstexten

Leitungsmitschriften
(Durchführungsmanual):

- Reihenfolge, in der die Teilnehmer_innen die 4 Schülerlösungstexte im Aufgabenheft korrigierten
- Einschätzungen der Teilnehmer_innen bezüglich der fachlich-konzeptuellen Qualität der Schülerlösungstexte in den Paarvergleichen
- Einschätzungen der Teilnehmer_innen bezüglich der Qualität der sprachlichen Realisierung der Schülerlösungstexte in den Paarvergleichen

Wie in Abschnitt 5.2.3 dargestellt, sollen die durch das laute Denken der Teilnehmer_innen und die retrospektive Befragung gewonnenen Daten im Rahmen der Auswertung einer Triangulation unterzogen werden, um so die Reichhaltigkeit der in der Hauptstudie gewonnenen Erkenntnisse zu maximieren (*Datensorten-Triangulation*; vgl. Denzin, 1970, S. 301 u. f.; Flick, 2004, S. 179; Flick, 2011, S. 12 u. f.; Kuckartz, 2014, S. 46). Ferner, wie unter Anderem in Abschnitt 5.2.1 erörtert, erfordern Forschungsfrage (F1) und (F2) aufgrund ihrer Komplexität ein methodenplurales Vorgehen aus qualitativen und quantiativen Forschungsmethoden. Diese Forderung betrifft nicht nur die Wahl gegenstandsangemessener Erhebungsmethoden (vgl. Unterkapitel 5.2), sondern auch die gegenstandsangemessner Auswertungsmethoden für die in der Laborsituation der Hauptstudie gewonnenen Daten (*Between-Methods-Triangulation*; vgl. Denzin, 1970, S. 307 u. f.; Flick, 2004, S. 180 u. f.; Flick, 2011, S. 15 u. f.; Kuckartz, 2014, S. 47).

Ergo: Eine Datensorten- und Between-Methods-Triangulation stellt eine gegenstandsangemessene Auswertung der in der Hauptstudie gewonnenen Daten dar. In Phase 3 der Entwicklungsstudie galt es daher abschließend das Design einer solchen Datensorten- und Between-Methods-Triangulation zu entwickeln. Dieses Forschungsdesign wird im Fol-

genden erläutert. Als Grundlage für die Entwicklung dieses Designs dienten die allgemeinen Empfehlungen der Literatur zur Auswahl und Gestaltung eines Mixed-Methods-Triangulationsdesigns (z. B. Mayring, 2001; Creswell & Plano Clark, 2007, S. 58 u. f.; Flick, 2011, S. 80 u. f.; Kuckartz, 2014, S. 57 u. f.). Des Weiteren wurden hier bereits Empfehlungen zur praktischen Auswertung von Daten, die mit Hilfe des lauten Denkens bzw. einer retrospektiven Befragung gewonnen wurden mit berücksichtigt (z. B. Ericsson & Simon, 1985, S. 261 u. f.; van Someren et al., 1994, S. 115 u. f.; Chi, 1997; A. Green, 1998, S. 68 u. f.; C. Green & Gilhooly, 2002, S. 60 u. f.; Heine & Schramm, 2007, S. 195 u. f.). Zu Beachten ist, dass das im Folgenden vorgestellte Forschungsdesign lediglich eine Skizze des geplanten Auswertungsprozesses in der Hauptstudie darstellt, in der noch keine Detailentscheidungen über die konkrete qualitative und/oder quantitative Analyse eines bestimmten Teildatensatzes getroffen wurden. Diese Entscheidungen wurden erst im Rahmen der Hauptstudie der vorliegenden Arbeit getroffen. Grund hierfür ist, dass insbesondere in Bezug auf die geplante qualitative Auswertung vor der eigentlichen Analyse zunächst ein „sich Vertraut machen“ mit dem Datenmaterial unabdingbar ist (vgl. Altheide, 1996, S. 23 u. f.; Mayring, 2015, S. 29 u. f.) und daher endgültige Entscheidungen über die Gegenstandsangemessenheit verschiedener Auswertungsmethoden erst im Lichte der tatsächlich in der Hauptstudie gewonnenen Daten getroffen werden konnten.

Die finale Version des Forschungsdesings zur geplanten Auswertung der in der Hauptstudie erhobenen Daten ist in Abbildung 5.9 schematisch dargestellt. Die durchgezogenen Pfeile symbolisieren die Analyseschritte zur Beantwortung von Forschungsfrage (F1), die gestrichelten Pfeile jene zu Forschungsfrage (F2). Wie Abbildung 5.9 verdeutlicht, umfasst die geplante Auswertung der in der Laborsituation gewonnenen Daten zunächst eine getrennte Auswertung der Laut-Denk-Daten der Teilnehmer_innen und jenen der retrospektiven Befragungen. Dabei sollen in beiden Teilauswertungen bezüglich Forschungsfrage (F1) und (F2) sowohl qualitative als auch quantitative Teilbefunde gewonnen werden. Erst im Anschluss erfolgt eine Integration der Befunde, die bei der Analyse von Teildatensätzen gewonnen wurden, indem diese zusammengeführt, miteinander verglichen und kontrastiert werden.

Vor der qualitativen Analyse erfolgt in beiden Teilauswertungen eine Aufbereitung der erhobenen Verbaldaten. Zunächst werden die Audiographien der Erwartungshorizonterstellung, die der Korrekturarbeit der teilnehmenden Physiklehrkräfte und die der retrospektiven Befragungen regelgeleitet transkribiert. Die erhobenen Teilnehmerprodukte und die Leitungsmitschriften dienen hierbei als Unterstützung, um möglichst detailgetreue Transkripte anfertigen zu können. Im Falle der Laut-Denk-Daten erfolgt zudem eine Segmentierung der angefertigten Transkripte in die einzelnen von den Teilnehmer_innen mitvokalisierten Gedankenschritte. Diese Segmentierung dient in der anschließenden Auswertung dazu, Rückschlüsse auf die Denkprozesse der befragten Physiklehrkräfte während der Materialbearbeitung vornehmen zu können (vgl. Heine & Schramm, 2007, S. 197). Dabei ist zu erwarten, dass die teilnehmenden Lehrkräfte beim lauten Denken in der Regel

keine „wohlgeformten“ Sätze formulieren, sondern ihr lautes Mitvokalisieren meist eher fragmentarischen Charakter aufweist (vgl. van Someren et al., 1994, S. 46).

Die qualitative Analyse in beiden Teilauswertungen ist als ein inhaltsanalytisches Vorgehen geplant. Die Entscheidung hierfür erfolgte aufgrund der folgenden (pragmatischen) Argumente:

1. Bei inhaltsanalytischen Auswertungsverfahren wird oftmals zwischen qualitativer und quantitativer Inhaltsanalyse unterschieden (z. B. Kracauer, 1952; Georg, 1959; Mathes, 1992; Altheide, 1996, S. 14 u. f.; Kuckartz, 2016, S. 13 u. f.). Diese Unterscheidung ist allerdings keine strikte Dichotomie, sondern beschreibt die Pole eines Kontinuums für verschiedenste Realisierungsmöglichkeiten eines inhaltsanalytischen Auswertungsverfahren (vgl. Holsti, 1969, S. 5 u. f.; Boyatzis, 1998, S. 4 u. f.; Krippendorff, 2004, S. 87 u. f.; Schreier, 2014a, S. 172). Im Allgemeinen handelt es sich bei der Inhaltsanalyse also um eine „gemischte“ Auswertungsmethode, weswegen sie besonders passgenau zu den in der Laborsituation gewählten „gemischten“ Erhebungsmethoden ist (vgl. Abschnitt 5.2.3).
2. Im Sinne der geplanten Datensorten-Triangulation ist es sinnvoll, bei den qualitativen Auswertung beider Teildatensätze auf dieselbe Analysemethode zurückzugreifen, um die anschließende Integration der gewonnenen Teilbefunde zu begünstigen (vgl. Denzin, 1970, S. 301 u. f.; Flick, 2011, S. 13).
3. In der methodentheoretischen Literatur herrscht ein weitgehender Konsens darüber, dass für die Analyse von Laut-Denk-Daten ein inhaltsanalytisches Vorgehen zu bevorzugen ist, da...

   ... auf Grundlage eines geeigneten Codierschemas die (mentalen, interaktionalen oder aktionalen) Handlungsabläufe in einem Laut-Denk-Protokoll systematisch identifiziert bzw. aus diesem extrahiert werden können (vgl. van Someren et al., 1994, S. 118 u. f.; Chi, 1997, S. 282 u. f.),

   ... ein inhaltsanalytisches Vorgehen eine systematische Abstraktion und Zusammenfassung umfangreicher Verbaldaten ermöglicht, bei der aber „die wesentlichen Inhalte erhalten bleiben“ (vgl. Mayring, 2015, S. 67) und

   ... die Resultate einer inhaltsanalytischen Auswertung zusätzliche quantitativen Analysen ermöglichen, indem die vorgenommenen Codierungen in Zahlen umgewandelt werden (z. B. die Häufigkeit des Auftretens eines bestimmten Codes oder einer Codekombination) (Kuckartz, 2014, S. 87 u. f.).
4. Rekonstruktive qualitative Auswertungsverfahren wie z. B. die objektive Hermeneutik oder die dokumentarische Methode erfordern im Vergleich zu einem inhaltsanalystischen Vorgehen deutlich langwierigeren Analyseprozesse (vgl. Mathes, 1992, S. 405; Paseka, 2010, S. 159). Rekonstruktive qualitative Auswertungsverfahren erscheinen daher für die Auswertung der in der Hauptstudie erhobenen Daten zu zeitintensiv, da diese nicht ausschließlich qualitativ, sondern methodenplural erfolgen soll (vgl. Abschnitt 5.2.1).

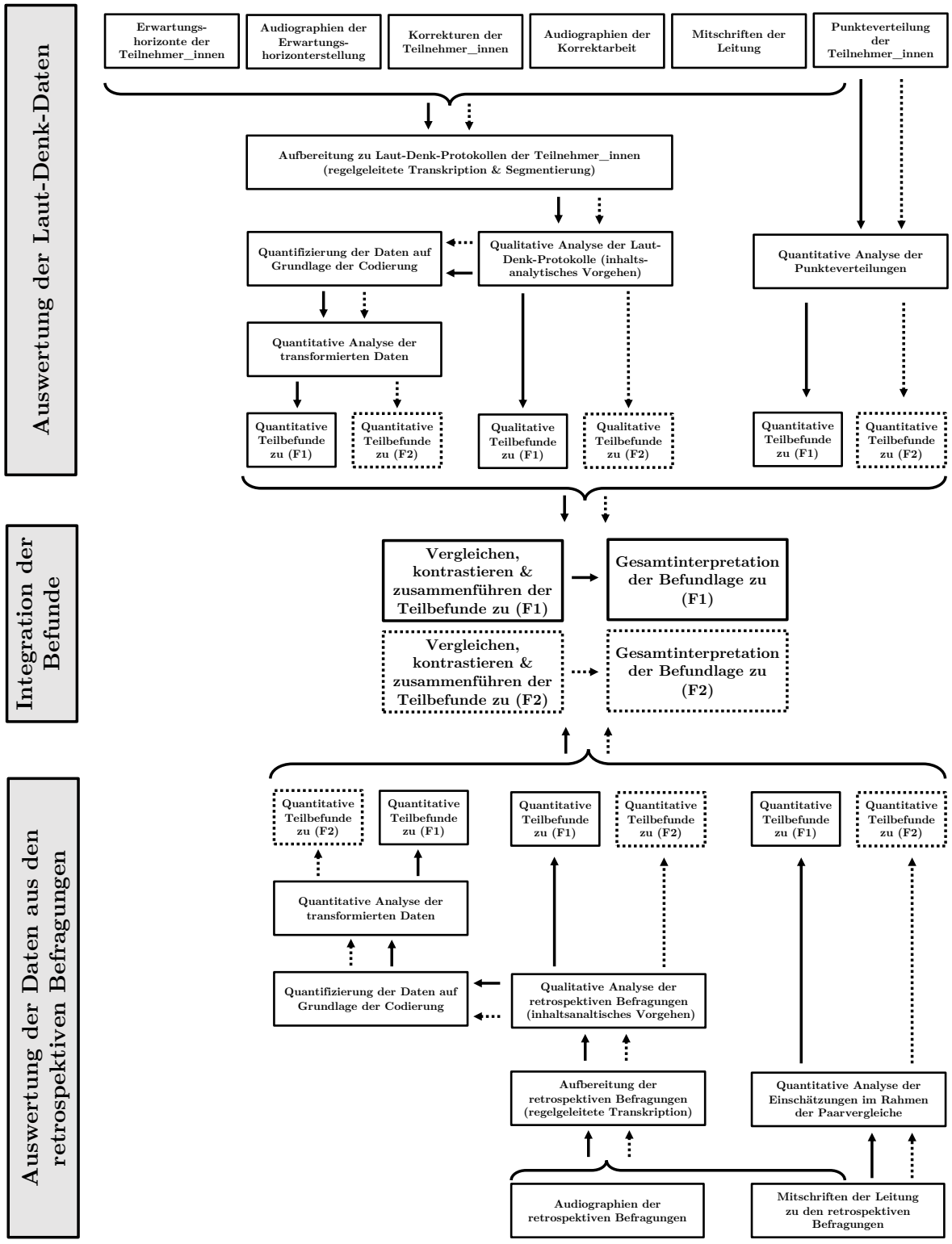


Abbildung 5.9.: Mixed-Methods-Triangulationsdesign zur geplanten Auswertung der in der Hauptstudie erhobenen Daten. Die durchgezogenen Pfeile symbolisieren die Analyseschritte zur Beantwortung von Forschungsfrage (F1), die gestrichelten Pfeile jene zu Forschungsfrage (F2).

Für die quantitative Analyse der erhoben Daten wird auf die von den teilnehmenden Physiklehrkräften vorgenommene Bepunktung der 4 Schülerlösungstexte, sowie auf ihre von der Leitung mitprotokollierten Einschätzungen in den Paarvergleichen im Rahmen der retrospektiven Befragung zurückgegriffen. Die Analyse erfolgt dabei vor allem mit Hilfe deskriptiver statistischer Methoden (Bestimung von geeigneten Lage-, Streuungs- und Regressionsmaßen). Ferner besteht die Möglichkeit, die in Zahlen umgewandelten Codierungen (Quantifizierung) aus den Inhaltsanalysen der Laut-Denk-Protokolle und der retrospektiven Befragungen ebenfalls einer quantitativen Analyse zu unterziehen (siehe oben). Welche quantitativen Analysen dabei allerdings sinnvoll und möglich sind, hängt im wesentlichen von der Ausgestaltung und der Güte der Kategoriensysteme ab, mit denen diese Inhaltsanalysen durchgeführt werden (vgl. Schreier, 2012, S. 231 u. f.). Da die Entwicklung dieser Kategoriensysteme erst an den in der Hauptstudie erhobenen Daten erfolgt, wurden die zusätzlichen quantitativen Analysen der inhaltsanalytischen Codierungen im Design der geplanten Auswertung zunächst nicht weiter spezifiziert.

5.5. Zusammenfassung

Das Erkenntnisinteresse von Forschungsfrage (F1) und (F2) umfasst die Deskription und Exploration der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile von Physiklehrkräften über schriftliche, aus einer Klassenarbeit stammende Schülerleistungen, sowie die einer bei dieser Genese eventuell auftretende Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile. Um dieses Erkenntnisinteresse eruieren zu können, galt es die Leitfrage zu klären, wie eine entsprechende empirische Untersuchung gegenstandsangemessen anzulegen ist. Die Klärung dieser Leitfrage erfolgte in der im vorangegangen Kapitel vorgestellten Entwicklungsstudie. Auf wesentliche Punkte zusammengefasst führte die Klärungsarbeit in der Entwicklungsstudie zu folgendem Ergebnis:

1. Die empirische Hauptstudie besteht aus einer Untersuchung von im Schuldienst aktiven Physiklehrkräften. Die Datenerhebung erfolgt in einer Laborsituation mit kontrolliertem Ablaufplan, die der Alltagspraxis von Physiklehrer_innen bei der Leistungsfeststellung und -beurteilung im Rahmen einer Klassenarbeit nachempfunden ist (vgl. Abschnitt 5.2.2).
2. In der Laborsituation erhalten die teilnehmenden Lehrkräfte die Aufgabe, ihren eigenen Gewohnheiten entsprechend und laut denkend einen Erwartungshorizont zu einer Klassenarbeitsaufgabe zu erstellen und anschließend vier Schülerlösungen zu dieser Aufgabe zu korrigieren (vgl. Abschnitt 5.4.1). Im Anschluss an die Korrekturarbeit unterziehen die Teilnehmer_innen in einer retroskpektiven Befragung die vier von ihnen korrigierten Schülerlösungen Paarvergleichen (vgl. ebd.). In diesen Paarvergleichen nehmen die Teilnehmer_innen Feststellungen und Beurteilungen die vier Schülerlösungstexte bezogen auf ihre fachlich-konzeptuelle Qualität und die Qualität ihrer sprachlichen Realisierung explizit vor (vgl. ebd.).

3. Die Klassenarbeitsaufgabe, mit der die teilnehmenden Lehrkräfte in der Laborsituation konfrontiert werden, wurde aus einem Pool von Aufgaben, die Lehrkräfte tatsächlich in Klassenarbeiten eingesetzt haben, ausgewählt und fordert Schüler_innen dazu auf einen physikalischen Sachverhalt in Form eines Textes zu erklären, der weder zeichnerische noch rechnerische Elemente enthält (vgl. Abschnitt 5.3.1) .
4. Bei den vier Schülerlösungstexten, die die teilnehmenden Lehrkräfte im Rahmen der Laborsituation korrigieren, handelt es sich um Kontrastfälle: Sie unterscheiden sich kriterial bezüglich ihrer fachlich-konzeptuellen Qualität und/oder der Qualität ihrer sprachlichen Realisierung. Diese Komposition aus vier kontrastierenden Schülerlösungstexten wurden in einem mehrschrittigen Codierverfahren aus den Aufgabenbearbeitungen von insgesamt 116 Hamburger Schüler_innen der 9. Jahrgangsstufe identifiziert (vgl. Abschnitt 5.3.2, 5.3.3 und 5.3.4).
5. Im Rahmen der Laborsituation werden von den teilnehmenden Lehrkräften Verbaldaten und Teilnehmerprodukte erhoben (vgl. Abschnitt 5.4.1 und 5.4.2). Hinzu kommt ein Lehrkräftefragebogen, sowie die Mitschriften des_der Forschers_Forscherin, der_die die Laborsituation leitet über besondere Vorkommnisse während der Erhebung (vgl. ebd.).
6. Die geplante Auswertung der in der Hauptstudie gewonnenen Daten ist als Datensorten- und Bewtween-Methods-Triangulation angelegt (vgl. Abschnitt 5.4.2): Zunächst erfolgt eine getrennte, jeweils sowohl qualitative, wie auch quantitative Auswertung der Laut-Denk-Daten der Teilnehmer_innen und jenen, die in der retrospektiven Befragung erhoben wurden. Anschließend erfolgt eine Integration der zuvor gewonnenen Teilbefunde, um so ein kaleidoskopartiges Gesamtbild der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile durch Physiklehrkräfte abzuleiten.

Alles in allem wurde also in der Entwicklungsstudie das methodische Vorgehen in der deskriptiv-explorativen Hauptstudie der vorliegenden Arbeit sowohl präzise, als auch umfassend geklärt und vorbereitet. Im nun folgenden Kapitel wird die durchgeführte Hauptstudie und die hierbei gewonnenen Befunde detailliert beschrieben.

6. Hauptstudie

Das folgende Kapitel dient der Darstellung der empirischen Hauptstudie. Es ist in die folgenden Unterkapitel untergliedert, deren Reihenfolge sich aus dem Mixed-Methods-Triangulationdesign zur geplanten Auswertung der in der Hauptstudie erhobenen Daten ableitet (vgl. Abschnitt 5.4.2):

Unterkapitel 6.1: Beschreibung der Stichprobe aus im Schuldienst aktiven Physiklehrkräften, die an der Hauptstudie der vorliegenden Arbeit teilgenommen haben. Diese Beschreibung basierend auf den Auskünften des Lehrkräftefragebogens.

Unterkapitel 6.2: Erläuterung zum Transkriptions- und Segmentierungssystem, mit dessen Hilfe die von den Teilnehmern_Teilnehmerinnen erhobenen Verbaldaten aufbereitet wurden.

Unterkapitel 6.3: Darstellung des methodischen Vorgehens, der Ergebnisse, der Interpretation und der Limitation der qualitativen und quantitativen Analysen der von den Teilnehmern_Teilnehmerinnen erhobenen Laut-Denk-Daten.

Unterkapitel 6.4: Darstellung des methodischen Vorgehens, der Ergebnisse, der Interpretation und der Limitation der qualitativen und quantitativen Analysen der Daten, die von den Teilnehmern_Teilnehmerinnen in den retrospektiven Befragungen gewonnen wurden.

Unterkapitel 6.5: Vergleichen, kontrastieren und zusammenführen der Teilbefunde zu Forschungsfrage (F1) und (F2), die in Unterkapitel 6.3 und 6.4 dargestellt wurden.

Die Gliederung von Kapitel 6 stellt also ein systematisches Raster dar, dass zwar grob dem zeitlichen Verlauf der Hauptstudie folgt (Datenerhebung, -aufbereitung, Teilanalysen, Integration der Befunde), nicht aber als Versuch einer möglichst wirklichkeitsgetreuen Abbildung der Chronologie des tatsächlichen Forschungsprozesses missverstanden werden darf.

6.1. Stichprobengewinnung und -beschreibung

Die Erhebung der Hauptstudie, gemäß des in Abschnitt 5.4.1 beschriebenen Vorgehens, fand von April bis September 2016 statt. Die Stichprobengewinnung stellte sich aus unterschiedlichen Gründen als recht schwierig dar. Erstens kann vermutet werden, dass die Bereitschaft vieler Lehrkräfte an der Studie teilzunehmen durch den entsprechenden Zeitaufwand (ca. 90 Minuten pro Lehrkraft) und zusätzliche Planungshürden (Erhebung in einem eigens hierfür präparierten Raum; vgl. Abschnitt 5.4.1) nur bedingt vorhanden war. Zweitens kann angenommen werden, dass sich Lehrkräfte zum Teil „gehemmt fühlten" an der Untersuchung teilzunehmen, da sie diese selbst als Leistungssituation wahrgenommen haben[99]. Drittens erschwerte die behördliche Auflage, Teilnehmer_innen nicht über Incentives gewinnen zu dürfen, die Stichprobengewinnung.

Aufgrund der eben benannten Schwierigkeiten wurde im Rahmen der Hauptstudie eine Gelegenheitsstichprobe von 21 Physiklehrkräfte befragt, die zum Erhebungszeitpunkt an 16 verschiedenen Hamburger Gymnasien und Stadtteilschulen im Schuldienst aktiv waren[100] (vgl. Tabelle 6.1). Aufgrund dieser vergleichsweise geringen Stichprobengröße konnte für die quantitative Analyse der erhobenen Daten (lediglich) auf nicht-parametrische Methoden und Verfahren zurückgegriffen werden (vgl. Bortz & Lienert, 2008).

Für die Rekrutierung von im Schuldienst aktiven Physiklehrkräften wurden die Schulleitungen aller öffentlichen und privaten Schulen im Bundesland Hamburg in einem Anschreiben mit Rückmeldebogen gebeten, bei den Physiklehrkräften ihrer Schule für eine Teilnahme an der Untersuchung zu werben. Zusätzlich wurden Physiklehrkräfte, zu denen aufgrund von bereits abgeschlossener Forschungsvorhaben der Arbeitsgruppe Physikdidaktik der Universität Hamburg Kontakt bestand, gezielt angeschrieben und um Teilnahme an der Untersuchung gebeten. Ferner wurden Lehrkräfte, die neben Physik ein sprachliches Fach unterrichten, gezielt als Teilnehmer_innen rekrutiert, da anzunehmen war, dass sich deren sprachliche Leistungsurteilsgenese möglicherweise anders gestaltet, als bei Physiklehrkräften, die kein sprachliches Fach unterrichten.

Da sich die anonymen Codes der Teilnehmer_innen zum Teil stark ähnelten, wurde jeder Physiklehrkraft ein Pseudonym[101] inklusiver einer entsprechenden Abkürzung zugewiesen (vgl. Tabelle 6.1). Diese Pseudonyme dienen im weiteren Verlauf der vorliegenden Arbeit der besseren Lesbarkeit.

[99] Diese Annahme stützt sich auf die Ergebnisse der Untersuchung von Kalthoff (1996), die darauf hinweisen, dass Lehrkräfte bei der Feststellung und Beurteilung von Schülerleistungen „mit dem Effekt des eigenen Unterrichtens konfrontiert werden [...] [und daher hierbei] immer auch ihre eigene Leistung konstruieren" (ebd., S. 109-115).

[100] Dabei wurden maximal 2 Physiklehrkräfte pro Gymnasium oder Stadtteilschule befragt.

[101] In Anlehnung an das Vorgehen von Gogolin (2008, S. 216), wurde für die Umbenennung der Teilnehmer_innen aus dem Onlineportal www.dastelefonbuch.de (Abruf: 18.08.2016) jeweils ein Nachname mit dem Anfangsbuchstaben A bis U ausgewählt (jeweils der 100, 200, 300, usw. Eintrag). Anschließend wurden diese 21 Nachnamen den Teilnehmer_innen bezüglich ihrer fortlaufenden Nummer (vgl. Tabelle 6.1) nacheinander zugewiesen.

Lehrkraft (fortlaufende Nummer)	**Anonymer Code**	**Pseudonym**	**Abkürzung**	**Geburtsjahr**	**Studium**	**Berufs-erfahrung** (Jahre)	**Schulform**	**Schulsozial-index**	**Unterrichts-fächer**
1	PET1853	Herr Abney	A	1987	Physik (vollfach)	3.0	StS	1	Ch, Ma, Ph, Se
2	HAN1744	Herr Balke	B	1977	Physik (vollfach)	3.0	Gym	5	Ma, Ph
3	REI1848	Herr Carboni	C	1953	Lehramt (Gym)	37.0	Gym	5	Ma, Nw, Ph, Sp
4	MAN1796	Herr Dassow	D	1988	Lehramt (Gym)	3.5	Gym	5	Ma, Nw, Ph
5	HAN1936	Herr Einert	E	1979	Physik (vollfach)	8.0	Gym	5	Ma, Ph
6	HAN1885	Herr Feldner	F	1963	Lehramt (Gym)	28.0	Gym	5	In, Ma, Ph, Th
7	ERN1867	Herr Geppert	G	1985	Lehramt (Gym)	2.5	StS	5	En, Ph
8	WOL1817	Herr Hastedt	H	1968	Physik (vollfach)	4.5	Gym	5	In, Ph
9	RAI1868	Herr Iezzi	I	1979	Lehramt (Gym)	10.0	StS	4	De, Nw, Ph, Se
10	FEL1757	Herr Jonuzi	J	1957	Physik (vollfach)	23.0	Gym (privat)	---	Ma, Ph
11	RUD1637	Frau Kirik	K	1966	Lehramt (Sek. I)	18.0	StS	3	Ge, Ph, Sw
12	STE1697	Herr Lemos	L	1957	Lehramt (Sek. I)	33.0	StS (privat)	---	Ch, Ma, Nw
13	REI1817	Herr Mehlert	M	1984	Lehramt (Gym)	2.5	Gym	6	Ch, Ph
14	JÜR1690	Frau Novack	N	1980	Lehramt (Gym)	11.5	Gym	5	Ma, NP, Nw, Ph
15	MAN1748	Herr Onne	O	1969	Lehramt (Gym)	23.0	Gym	3	In, Ma, Ph
16	WER1677	Frau Pinna	P	1967	Lehramt (Gym)	25.0	Gym	6	Ma, Ph
17	HER1877	Herr Quezada	Q	1984	Lehramt (Gym)	6.0	StS	5	Ma, Ph
18	HAR1835	Herr Rittershaus	R	1984	Lehramt (Gym)	3.5	StS	5	Bio, Nw, Ma, Ph
19	UWE1762	Frau Sohm	S	1972	Lehramt (Sek. I)	17.0	Gym & StS	4	Ma, Ph
20	CHR1786	Herr Trummer	T	1954	Lehramt (Gym)	30.0	Gym	6	Ma, NP, Ph
21	HEI1867	Herr Uckermark	U	1963	Lehramt (Gym)	21.0	Gym	6	En, NP, Ph

Tabelle 6.1.: Anonyme Codes, zugewiesene Pseudonyme und ausgewählte soziodemographische Eckdaten der Physiklehrkräfte, die an der Hauptstudie teilgenommen haben (Abkürzungen der Unterrichtsfächer: Bio – Biologie; Ch – Chemie; De – Deutsch; En – Englisch; Ge – Gesellschaft; In – Informatik; Ma – Mathematik; NP – Naturwissenschaftliches Praktikum; Nw – Naturwissenschaften; Ph – Physik; Se – Seminar; Sp – Sport; Sw – Sprachwerkstatt; Th – Theater).

Weitere Merkmale der Gelegenheitsstichprobe, die sich aus den Auskünften der Teilnehmer_innen im Lehrkräftefragebogen ergeben (vgl. Abschnitt 5.4; Anhang C.2), werden in den folgenden Abschnitten beschrieben.

6.1.1. Soziodemographische Eckdaten der Teilnehmer_innen

An der Hauptstudie haben Physiklehrkräfte mit unterschiedlich langer Berufserfahrung teilgenommen. Der Berufserfahrungsmedian der Teilnehmer_innen betrug zum Erhebungszeitpunkt 11.5 Jahre mit einem Interquartilsabstand (IQR) von 21.5 Jahren (Spannweite: 2.5 bis 37 Jahre). Dementsprechend betrug der Altersmedian (bestimmt aus dem Geburtsjahr) 44 Jahre mit einem IQR von 21 Jahren (Spannweite: 29 bis 63 Jahre).

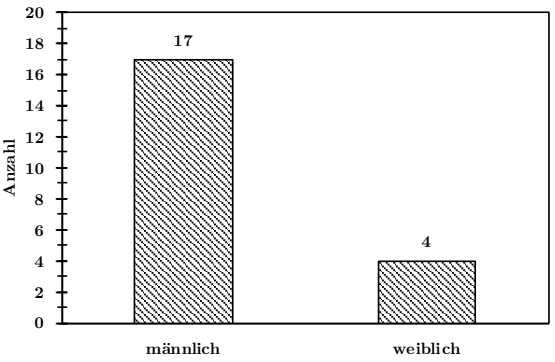


Abbildung 6.1.: Geschlechterverteilung unter den Teilnehmer_innen (absolute Häufigkeiten).

Aus Abbildung 6.1 geht hervor, dass männliche Physiklehrkräfte die Mehrheit unter den Teilnehmer_innen bildeten. Dieses Ungleichgewicht in der Geschlechterverteilung war zu erwarten, da der Frauenanteil an der Physiklehrerschaft in Deutschland (anders als bei der Lehrerschaft in Deutschland im Allgemeinen; vgl. Statistisches Bundesamt , 2018, S. 44 u. f.) deutlich geringer ist als der Männeranteil[102] (vgl. Roisch, 2003, S. 31 u. f.; Richter, Kuhl, Haag, & Pant, 2013, S. 372).

Von den 21 Teilnehmer_innen waren 19 an staatlichen Schulen tätig. Zwei Lehrkräfte (Herr Jonuzi und Herr Lemos) unterrichteten zum Erhebungszeitpunkt an einer Privatschule. 13 Teilnehmer_innen (61.9 %) waren an einem Gymnasium tätig und 7 an einer Stadtteilschule (33.3 %), was in grober Näherung dem Verhältnis der Grundgesamtheit[103] entspricht (vgl. Abbildung 6.2 links). Eine Teilnehmerin (Frau Sohm) gab an, zur Zeit sowohl an einem Gymnasium, als auch an einer Stadtteilschule zu unterrichten. Des Weiteren spiegelt für die 19 Physiklehrkräfte, die an staatlichen Schulen unterrichteten, die Verteilung des Schulsozialindexes[104] ebenfalls grob die Verhältnisse der Grundgesamtheit wieder (vgl. Abbildung 6.2 rechts[105]). Der Interquartilsabstand fiel in der erhobenen Stich-

[102] Aktuelle und der Öffentlichkeit zugängliche Zahlen zum Frauenanteil an der Physiklehrerschaft speziell im Bundesland Hamburg sind m. W. nicht vorhanden.

[103] Die Grundgesamtheit sind alle Physiklehrkräfte, die zum Erhebungszeitpunkt an einem Hamburger Gymnasium oder einer Stadtteilschule im Schuldienst aktiv waren.

[104] Der Schulsozialindex war nicht Teil des Lehrerfragebogens. Alle teilnehmenden Physiklehrkräfte wurden allerdings informell befragt, an welcher Schule sie derzeit unterrichten, woraus sich auf den Sozialindex der entsprechenden Schulen schließen ließ.

[105] Die in den Boxplots der vorliegenden Arbeit angegeben Zahlenwerte sind, sofern nicht explizit anders vermerkt, die zugehörigen Stichprobenmediane.

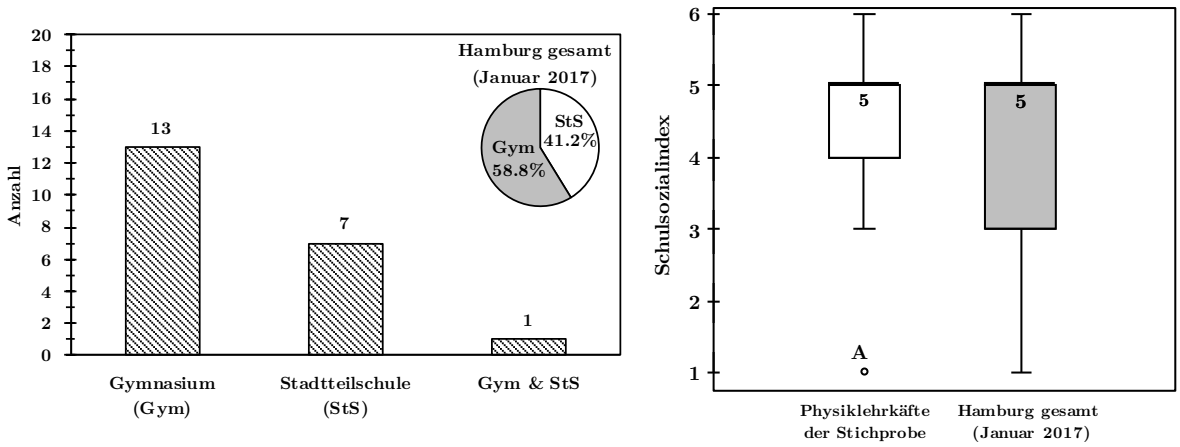


Abbildung 6.2.: Schulform (links) und Schulsozialindex (rechts) der Teilnehmer_innen zum Erhebungszeitpunkt. Die jeweilige Vergleichsgrafik zur gesamten Hamburger Physiklehrerschaft ist aus den von der Stadt Hamburg veröffentlichten Daten zusammengestellt, deren Stand am ehesten mit dem Erhebungszeitpunkt der Hauptstudie übereinstimmt (vgl. Bürgerschaft der Freien und Hansestadt Hamburg , 2013, S. 27 u. f.; Bürgerschaft der Freien und Hansestadt Hamburg , 2017, S. 10 u. f.).

probe allerdings geringer aus als in der Grundgesamtheit. Hieraus lässt sich schließen, dass Physiklehrkräfte, die an Schulen mit geringerer sozialer Belastung unterrichteten, in der Gelegenheitsstichprobe häufiger als in der Grundgesamtheit vertreten sind.

Die meisten Teilnehmer_innen (16) haben als Hochschulausbildung gymnasiales Lehramt (13) oder Lehramt für die Sekundarstufe I (3) studiert. Die übrigen 5 Teilnehmer_innen sind „Quereinsteiger_innen“ in den Lehrerberuf und haben ursprünglich ein Vollfachstudium in Physik absolviert, davon 2 (Herr Balke und Herr Jonuzi) mit Promotion.

Mit Ausnahme von Herrn Lemos unterrichteten alle Teilnehmer_innen zum Erhebungszeitpunkt Physik als eigenständiges Unterrichtsfach (vgl. Abbildung 6.3 links). Mathematik stellte unter den Teilnehmer_innen das am zweithäufigsten genannte Unterrichtsfach dar (15). Besonders häufig wurden zudem andere naturwissenschaftliche Schulfächer (Biologie, Chemie, Naturwissenschaften, naturwissenschaftliches Praktikum) und sprachliche Unterrichtsfächer (Deutsch, Englisch, Sprachwerkstatt) genannt[106]. Ferner erteilten die Physiklehrkräfte im Median 20.0 Unterrichtsstunden pro Woche (IQR = 8.0 $\frac{Stunden}{Woche}$; Spannweite: 10.0 bis 34.0 $\frac{Stunden}{Woche}$). Hiervon sind im Mittel (Median) 42.1 % Physikunterricht oder integrierter Naturwissenschaftsunterricht mit physikalischen Inhaltsanteilen[107] (IQR = 31.3 %; Spannweite: 16.7 bis 75.0 %), wobei die Teilnehmer_innen, von der

[106]Da Lehrkräfte, die neben Physik ein sprachliches Fach unterrichten im Rahmen der Datenerhebung gezielt als Teilnehmer_innen rekrutiert wurden ist es – ohne dass hierzu m. W. aktuelle Zahlen vorliegen – wahrscheinlich, dass der „Sprachlehreranteil“ unter den Teilnehmer_innen höher ist, als in der Grundgesamtheit.

[107]Die Teilnehmer_innen wurden beim Ausfüllen des Lehrkräftefragebogens gebeten, zu ihren erteilten Physikstunden auch integrierten Naturwissenschaftsunterricht zu zählen, wenn sie in diesem gegenwärtig physikalische Sachinhalte thematisieren.

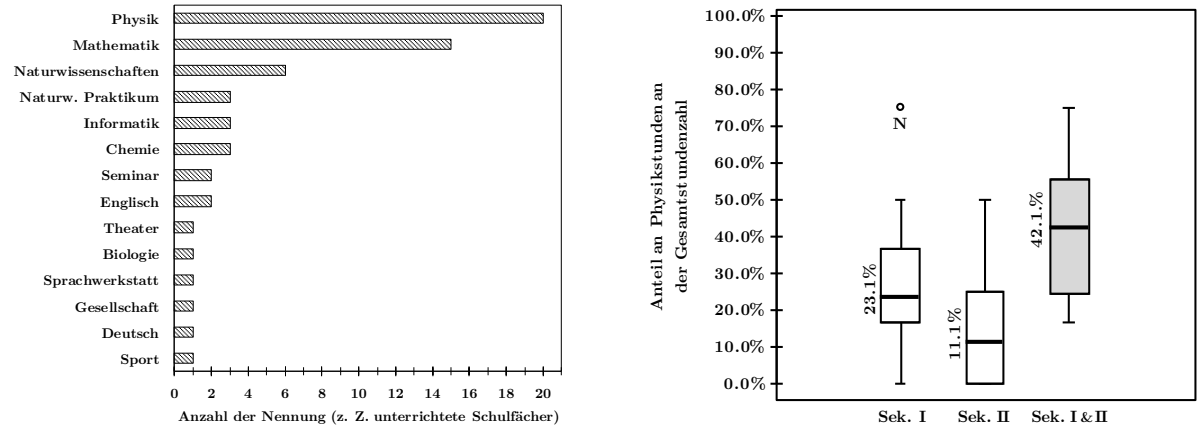


Abbildung 6.3.: Zum Erhebungszeitpunkt unterrichtete Schulfächer der Teilnehmer_innen (links) und Verteilung des relativen Anteils der erteilten Physikstunden an der Gesamtstundenzahl (rechts).

mittleren Tendenz her, mehr Physikunterricht in der Sekundarstufe I erteilten als in der Sekundarstufe II (vgl. Abbildung 6.3 rechts).

6.1.2. Selbstauskünfte der Teilnehmer_innen

Alle teilnehmenden Physiklehrkräfte haben auf dem Lehrkräftefragebogen Fragen zu den Selbstauskunftsskalen „Bewertung nach sozialer Bezugsnorm versus kriterialer Norm“, „Diagnose im Leistungsbereich“ und „Vermittlung der Domänenspezifischen Bildungssprache“ beantwortet, die aus der Dokumentation der Erhebungsinstrumente der *COACTIV*-Studie (vgl. Baumert et al., 2009, S. 169 u. f.) bzw. aus dem Fragebogen der Lehrkräftebefragung zu Sprachbildung im naturwissenschaftlichen Unterricht von Riebling (2013b, S. 108 u. f.) übernommen wurden (vgl. Anhang C). In den entsprechenden Antworten der Teilnehmer_innen lassen sich folgende Tendenzen erkennen[108]:

Die Skala zur „Bewertung nach sozialer Bezugsnorm versus kriterialer Norm“ besteht 4 aus Items. Bei jedem dieser Items wird ein Lehrerhandeln beschrieben, das entweder einer kriterialen oder sozialen Bezugsnormorientierung entspricht. Zu jedem dieser 4 Items wird auf einer vier-stufigen Likertskala erfragt, inwieweit das beschriebene Handeln dem eigenen Lehrerhandeln entspricht. Den Items a) und b), die ein kriterial bezugsnormorientiertes Lehrerhandeln beschreiben[109], stimmte jeweils eine deutliche Mehrheit der Teilnehmer_innen (eher) zu (vgl. Abbildung 6.4 links). Umgekehrt lehnte eine deutliche Mehrheit der Teilnehmer_innen die Aussagen in Item c) und d), in denen ein sozial be-

[108] Aufgrund der geringen Stichprobengröße war im Rahmen der vorliegenden Arbeit eine Neuberechnung der internen Konsistenz dieser drei Skalen (Cronbachs α) nicht möglich. In den entsprechenden Originalarbeiten war diese allerdings für explorative Zwecke jeweils zufriedenstellend ($.73 \leq \alpha \leq .87$; vgl. Baumert et al., 2009, S. 169 u. f.; Riebling, 2013b, S. 116).

[109] Z. B. Item b) „[d]ie Anforderungen einer Notenstufe lege ich vor der Klassenarbeit/Schulaufgabe fest und ändere daran nichts, auch wenn dann relativ viele Arbeiten gut oder schlecht ausfallen“ (Baumert et al., 2009, S. 169).

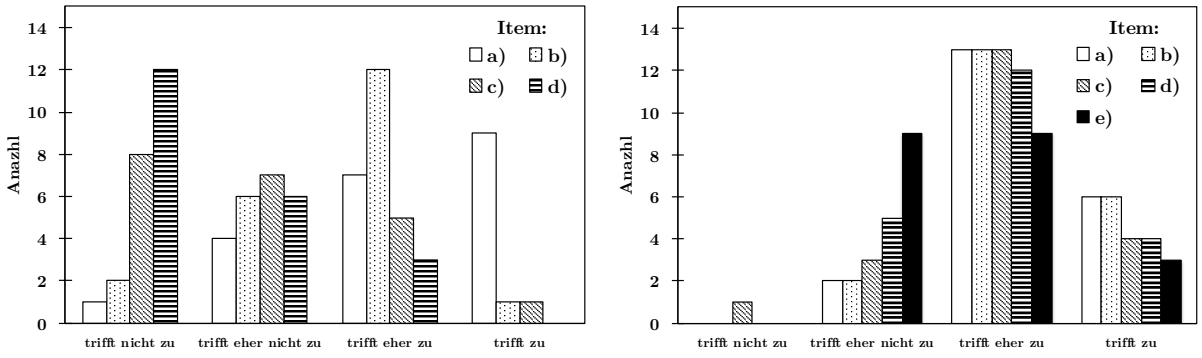


Abbildung 6.4.: Selbstauskünfte der Teilnehmer_innen zur Bewertung nach sozialer versus kriterialer Norm (links) und der Diagnose im Leistungsbereich (rechts).

zugsnormsorientiertes Lehrerhandeln beschrieben wird[110], (eher) ab (vgl. Abbildung 6.4 links). Allerdings zeigte sich auf dieser Skala nur eine schwache Konkordanz der Selbsteinschätzungen der 21 Physiklehrkräfte (Kendalls $W = .349$; $p < .001$; χ^2-Test). Daher lässt sich begründet vermuten, dass sich die Teilnehmer_innen darin unterscheiden, ob sie ihrer Selbsteinschätzung nach eine kriteriale oder eine soziale Bezugsnormorientierung oder beides aufweisen. Jedoch gibt es eine schwache Tendenz in Richtung der kritierialen Bezugsnorm.

Die 5 Items der Skala „Diagnose im Leistungsbereich“ sind Aussagen über die eigene Diagnosesicherheit in verschiedenen Zusammenhängen, zu denen auf einer vier-stufigen Likertskala die persönliche Zustimmung erfragt wird[111]. Eine deutliche Mehrheit der Teilnehmer_innen stimmte jeweils bei allen 5 Aussagen dieser Skala (eher) zu (vgl. Abbildung 6.4 rechts). Allerdings zeigte sich lediglich eine sehr schwache Konkordanz der Selbsteinschätzungen der 21 Physiklehrkräfte (Kendalls $W = .129$; $p = .029$; χ^2-Test). Dies deutet darauf hin, dass es unter den Teilnehmer_innen zwar eine sehr schwache Tendenz in Richtung einer subjektiv eher hoch empfundenen Diagnosesicherheit in verschiedenen Situationen gibt. Jedoch schätzen die Teilnehmer_innen im Vergleich miteinander ihre eigene Diagnosesicherheit deutlich unterschiedlich ein.

Eine ebenfalls schwache Konkordanz zeigte sich bei den Selbsteinschätzungen der 21 Physiklehrkräfte bei den 5 Items der Skala zur „Vermittlung der Domänenspezifischen Bildungssprache“ (Kendalls $W = .399$; $p < .001$; χ^2-Test). Die Items dieser Skala erfragen auf einer 5-stufigen Ordinalskala, wie häufig bestimmte Unterrichtsmethoden, die der Sprachförderung dienen, im eigenen Fachunterricht eingesetzt werden[112]. Die Teilnehmer_innen wurden beim Ausfüllen dieser Skala gebeten, diese 5 Items entsprechend ihres Vorgehens,

[110] Z. B. Item d) „[i]ch orientiere meine Noten am Durchschnitt der Klasse“ (Baumert et al., 2009, S. 169).
[111] Z. B. Item c) „[i]ch weiß, bei welchen Aufgaben die einzelnen Schüler/innen Schwierigkeiten haben“ (Baumert et al., 2009, S. 172).
[112] Z. B. Item c) „[im Fachunterricht...] stelle ich Schülern Aufgaben, die explizit der Einübung des Fachwortschatzes dienen (Skizzen beschriften, Diagramme ergänzen, Lückentexte bearbeiten etc.)“ (Riebling, 2013b, S. 111).

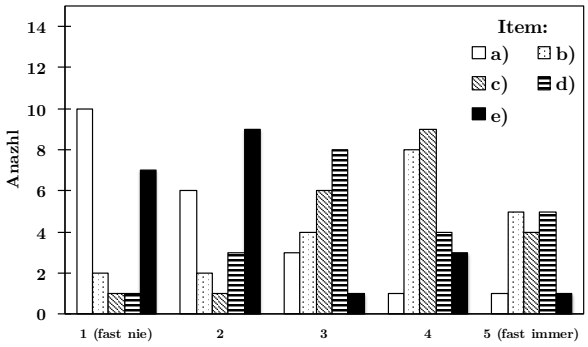


Abbildung 6.5.: Selbstauskünfte der Teilnehmer_innen zur Vermittlung der domänenspezifischen Bildungssprache.

wenn sie Physikunterricht erteilen, zu beantworten. Aus der schwachen Konkordanz der Selbsteinschätzungen lässt sich vermuten, dass sich die Teilnehmer_innen im Vergleich miteinander darin unterscheiden, ob und wenn ja in welchem Umfang sie ihrer Selbsteinschätzung nach im Physikunterricht bestimmte Methoden zur Sprachförderung einsetzen. In Abbildung 6.5 fällt allerdings eine schwache Tendenz auf, nämlich dass eine Mehrheit der Teilnehmer_innen angab, im Physikunterricht selten[113] mit den Schüler_innen Wortschatzlisten anzulegen (Item a) oder Fachsprachengrammatik zu besprechen (Item e) und gleichzeitig eine Mehrheit laut eigener Auskunft neuen (Fach-)Wortschatz häufig[113] ausführlich einführt und einübt (Item b und c).

Zusammengefasst sprechen die eben benannten Tendenzen für eine Heterogenität der im Rahmen der Hauptstudie gewonnenen Gelegenheitsstichprobe. Dies ist für das Erkenntnisinteresse der Hauptstudie bedeutsam, da zum einen in einer (bezüglich der eben benannten Selbsteinschätzungsmerkmalen) heterogenen Stichprobe eher als in einer homogenen erwarten werden kann, unterschiedliche Ressourcen zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen anzutreffen (Forschungsfrage (F1)). Zum anderen kann eher in einer heterogenen Stichprobe erwartet werden, empirische Hinweise zu finden, die für oder gegen eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile von Physiklehrkräften sprechen (Forschungsfrage (F2)).

6.2. Erläuterung zur Aufbereitung der erhobenen Verbaldaten

Am Ende der Erhebungsphase der Hauptstudie lag von allen 21 Physiklehrkräften mindestens eine vollständige Audiographie der Laborsituation vor. Aus diesen Audiographien

[113] Die Begriffe „selten“ und „häufig“ meinen hier, ob die Teilnehmer_innen auf der 5-stufigen Ordinalskala des entsprechenden Items in absoluter Mehrheit einen Wert von 1 oder 2 bzw. 4 oder 5 angegeben haben.

wurden zunächst Basistranskripte[114] der laut-denkenden Aufgabenheftbearbeitungen, sowie der retrospektiven Befragungen angefertigt (vgl. Abschnitt 5.4.1). Die Transkription erfolgte mit Hilfe eines eigens an die Bedürfnisse der Hauptstudie angepassten Transkriptionssystems, für das die Arbeit von Fuß & Karbach (2014), sowie das Transkriptionssystem von Arras (2007, S. 191 u. f.) als Grundlage dienten (vgl. Anhang D.1): Bei der Transkription wurde keine Sprachglättung vorgenommen, das heißt die deutsche Rechtschreibung und Interpunktion wurde zwar berücksichtigt, beibehalten wurden aber Dialekt, umgangssprachliche und fehlerhafte Ausdrucksweisen, Wortabbrüche, ein fehlerhafter Satzbau und Satzabbrüche, sowie Lautäußerungen wie beispielsweise „ähm“ oder „mhm“. Mittranskribiert wurden zudem Pausen im Redefluss, non-verbale Äußerungen der Teilnehmer_innen (z. B. Räuspern) und gegebenenfalls der „auffällige“ Sprachklang einer Äußerung (z. B. eine lachende Betonung der Mitvokalisierung). Ferner wurden Hintergrundgeräusche in den Audiographien (z. B. Telefonklingeln) und hörbare Handlungen der Teilnehmer (z. B. nicht mitvokalisiertes Anfertigen einer handschriftlichen Notiz) in den Basistranskripten als Kommentar[115] vermerkt. Der Auszug aus dem Basistranskript von Herrn Abney, der in der linken Hälfte von Abbildung 6.2 dargestellt ist, illustriert, wie die Audiographien der Teilnehmer_innen gemäß des eben umschriebenen Vorgehens transkribiert wurden.

Die Basistranskripte der retrospektiven Befragungen wurden, außer dass sie in die Fragen der Leitung und Antworten der Teilnehmer_innen segmentiert wurden, nicht weiter aufbereitet. Dies galt jedoch nicht für die Basistranskripte des lauten Denkens der Teilnehmer_innen, wie der Transkriptauszug in der rechte Hälfte von Tabelle 6.2 illustriert. Diese wurden, mit Hilfe eines eigens hierfür entwickelten Segmetierungssystems[116], regelbasiert segmentiert (vgl. Anhang D.2). Einerseits erfolgte dies, um die Lesbarkeit der Transkripte zu erhöhen, indem durch die Segementierung ein Abbild des fragmentarischen Charakters der laut mitvokalisierten Gedankenabfolgen (vgl. Abschnitt 5.4.2) im Transkript erzeugt wird. Andererseits diente die Segmentierung der Vorbereitung der anschließenden inhaltsanalytischen Auswertung, da hier die Transkriptsegmente als Codiereinheiten (vgl. Schreier, 2012, S. 131 u. f.; Kuckartz, 2016, S. 41 u. f.) verwendet wurden (vgl. Abschnitt 6.3.2). Aus selbigen Gründen erfolgte im Rahmen der Segmentierung zudem eine Markierung von „besonderen sprachlichen Ereignissen“ in den Basistranskripten (vgl. Anhang D.2). Beispielsweise wurden Transkriptabschnitte, in denen aus dem Aufgabenheft laut vorgelesen wird, *kursiv* gesetzt.

[114] Die Rohversionen der Basistranskripte von 11 Teilnehmer_innen wurden von einem_einer Zweittranskribierer_in angefertigt. Alle Basistranskripte wurden von dem_der Ersttranskribierer_in mindestens einmal korrekturgelesen.

[115] Um möglichst genaue Kommentare über die Handlungen der Teilnehmer_innen zu vermerken, wurden die ausgefüllten Aufgabenhefte als Unterstützung für die Transkriptanfertigung herangezogen.

[116] Als Grundlage für dieses System dienten die Segmentierungssysteme von Ericsson & Simon (1985, S. 299 u. f.), van Someren et al. (1994, S. 117 u. f.), Chi (1997, S. 284 u. f.), A. Green (1998, S. 75 u. f.), C. Green & Gilhooly (2002, S. 60 u. f.), Hughes & Parkes (2003, S. 129) und Arras (2007, S. 194 u. f.).

Basistranskript	Feintranskript mit Segmentierung
[...] Gut, dann geh- ich jetzt wieder zur Antwort A. Bin jetzt auf Seite 5. Und ähm schau mir des jetzt nochmal etwas genauer an. Ähm, im All ist nichts durch das Ton geht... Ähm, jetzt würde ich sozusagen anmerken, ähm durch das Ton geht ähm, (..) weil wir das ja behandelt haben. Ähm durch das Ton geht sozusagen. Da schrei-... das unterkringel ich jetz- sozusagen. Also unterkringeln heißt bei mir, dass ich so -ne gewellte Linie drunter mache. Das sind für mich immer die Sachen, die an den Ausdruck gehen. Würde am Rand, jetzt hab ich leider keinen Korrekturrand. Des hab- ich jetz- anscheinend schon wieder in -ner (lachend) Klassenarbeit vergessen. (+) (unterringelt „durch das Ton geht") Ähm, ähm durch das Ton geht. Da würd- ich jetzt an der Seite ein A machen und würde Doppelpunkt Medium hinschreiben. Das ist nämlich der Fachbegriff für. A für Ausdruck. (schreibt „A: Medium") [...]	[...] [211] Gut, dann geh- ich jetzt wieder zur Antwort A. [212] Bin jetzt auf Seite 5. [213] Und ähm schau mir des jetzt nochmal etwas genauer an. [214] Ähm, *im All ist nichts durch das Ton geht...* [215] Ähm, jetzt würde ich sozusagen anmerken, [216] ähm *durch das Ton geht* [217] ähm, (..) weil wir das ja behandelt haben. [218] Ähm *durch das Ton geht* sozusagen. [219] Da schrei-... das unterkringel ich jetz- sozusagen. [220] Also unterkringeln heißt bei mir, dass ich so -ne gewellte Linie drunter mache. [221] Das sind für mich immer die Sachen, die an den Ausdruck gehen. [222] Würde am Rand, jetzt hab ich leider keinen Korrekturrand. [223] Des hab- ich jetz- anscheinend schon wieder in -ner (lachend) Klassenarbeit vergessen. (+) [224] (unterringelt „durch das Ton geht") [225] Ähm, ähm *durch das Ton geht.* [226] Da würd- ich jetzt an der Seite ein A machen und würde Doppelpunkt Medium hinschreiben. [227] Das ist nämlich der Fachbegriff für. [228] A für Ausdruck. [229] (schreibt „A: Medium") [...]

Tabelle 6.2.: Auszug aus dem Transkript der laut-denkenden Korrektur von Schülerlösungstext A von Herrn Abney (zirka ab 41 Minuten und 46 Sekunden in der Audiographie).

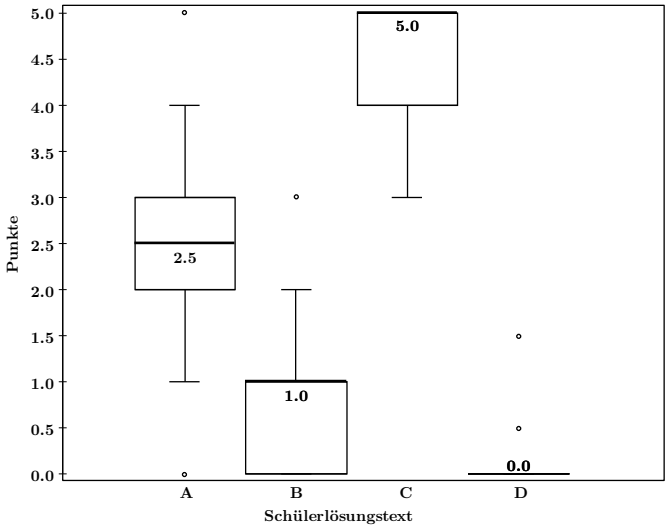
6.3. Analyse der Laut-Denk-Daten

In diesem Unterkapitel werden die Analysen der Laut-Denk-Daten und die hierbei gewonnenen Teilbefunde vorgestellt. Dem Mixed-Methods-Triangulationsdesign zur geplanten Auswertung der in der Hauptstudie erhobenen Daten entsprechend (vgl. Abschnitt 5.4.2), gliedert sich dieses Unterkapitel in die folgenden Abschnitte:

Abschnitt 6.3.1: Eine quantitative Analyse der Punktverteilung, die die 21 Physiklehrkräfte bei der Korrektur der Schülerlösungstexte A bis D vorgenommen haben.

Abschnitt 6.3.2: Eine (qualitative) Analyse der Laut-Denk-Protokolle der 21 Teilnehmer_innen durch ein inhaltsanalytisches Vorgehen.

Jeder dieser Abschnitte beginnt zunächst mit einer methodischen Vorbemerkung, in der das zur Analyse des entsprechenden Teildatensatzes verwendete Auswertungsverfahren beschrieben wird. Dem folgt eine Ergebnisdarstellung mit anschließender Interpretation der gewonnenen Befunde zu Forschungsfrage (F1) und/oder (F2), sowie eine Darstellung der Limitationen der im entsprechenden Abschnitt gewonnenen Teilbefunde.



	A	B	C	D
Herr Abney	3.0	1.0	5.0	0.5
Herr Balke	4.0	0.0	5.0	0.0
Herr Carboni	1.0	0.0	3.0	0.0
Herr Dassow	0.0	0.0	3.5	0.0
Herr Einert	0.0	0.0	5.0	0.0
Herr Feldner	3.0	0.0	4.0	0.0
Herr Geppert	2.0	2.0	5.0	0.0
Herr Hastedt	2.5	1.5	5.0	0.5
Herr Iezzi	2.5	---	3.5	0.0
Herr Jonuzi	2.0	1.0	5.0	0.0
Frau Kirik	4.0	0.0	5.0	0.0
Herr Lemos	5.0	0.0	5.0	0.0
Herr Mehlert	3.0	1.0	5.0	0.0
Frau Novack	2.0	0.0	5.0	0.0
Herr Onne	3.0	1.0	5.0	0.0
Frau Pinna	3.0	3.0	4.0	0 evtl. 1
Herr Quezada	3.0	1.0	5.0	0.0
Herr Rittershaus	2.5	1.0	4.0	1.5
Frau Sohm	2.0	0.0	3.0	0.0
Herr Trummer	3.0	1.0	5.0	0.0
Herr Uckermark	2.5	1.0	5.0	0.5

Abbildung 6.6.: Boxplots (links) und tabellarische Übersicht (rechts) der Punkteverteilung an die Schülerlösungstexte A bis D durch die Teilnehmer_innen.

6.3.1. Quantitative Analyse der Punkteverteilungen

6.3.1.1. Methodische Vorbemerkungen

Abbildung 6.6 liefert eine Übersicht über die Punkte, die die Teilnehmer_innen an die Schülerlösungstexte A bis D vergeben haben. Obwohl jede_r Teilnehmer_in seine_ihre Punktevergabe mit Hilfe eines eigenen Erwartungshorizontes vornahm, weisen diese insgesamt eine moderate intersubjektive Übereinstimmung[117] auf (Kendalls $W = .529$; $p < .001$; χ^2-Test). Ferner zeigen sich in Abbildung 6.6 die folgenden drei Auffälligkeiten: Erstens wurden die vier Schülerlösungstexte von den 21 Physiklehrkräfte im Median erkennbar unterschiedlich bepunktet. Zweitens ist, mit Ausnahme von Schülerlösungstext D, der Interquartilsabstand bei allen Schülerlösungstexten identisch. Erst bei Betrachtung der Spannweite der vergebenen Punkte je Schülerlösungstext zeigen sich Streuungsunterschiede (A: 0 bis 5 Punkte; B 0 bis 3 Punkte; C: 3 bis 5 Punkte; D: 0 bis 1.5 Punkte). Drittens wird aus der tabellarischen Übersicht deutlich, dass Herr Iezzi und Frau Pinna den Schülerlösungstexten B bzw. D jeweils keine (eindeutige) Punktzahl zugewiesen haben[118].

[117] Die unvollständigen Punktevergaben von Herrn Iezzi und Frau Pinna wurden von dieser Berechnung ausgeschlossen.

[118] An den entsprechenden Stellen der Laut-Denk-Protokolle beider Lehrkräfte zeigt sich jeweils eine bemerkenswerte Strategie zur Feststellung und Beurteilung von Schülerleistungen. Möglicherweise speisen sich diese Strategien aus subjektiven Überzeugungen darüber, was faire bzw. milde Zensurenvergabe bedeutet: Herr Iezzi entscheidet, „wenn es [...] während der Klausur [...] keine nachgesteuerten Anweisungen [gab,] [...] die Aufgabe für diesen Schüler aus der Wertung raus[zu]nehmen“ (Seg. 119-121). Er begründet dies damit, dass Schülerlösungstext B „natürlich nicht zu dem, was wir im Unterricht gemacht haben[,] [passt,] [...] [a]ber [...] nicht unlogisch [ist,] [...] auch wenn e[r] [...] nicht ganz stimmt.“ (Seg. 116-118). Frau Pinna ist bei Schülerlösungstext D „nich- ganz klar, was der Schü-

Schülerlösungstext	Punktemedian	mittlerer Rang	Friedman-Test χ^2	df	p
A	2.5	2.9	52.249	3	$< .001$
B	1.0	1.8			
C	5.0	4.0			
D	0.0	1.3			

Tabelle 6.3.: Friedman-Test zur Analyse von Medianunterschieden in der Punkteverteilung für die Schülerlösungstexte A bis D.

Die erste der drei eben benannten Auffälligkeiten wurde mit Hilfe nicht-parametrischer statistischer Methoden genauer untersucht. Bei allen hierbei angewandten Testverfahren wurde das Signifikanzniveau $\alpha = .05$ gewählt. Für die Analyse von Medianunterschieden in der Punkteverteilung wurde zunächst ein Friedman-Test als nicht-parametrisches Äquivalent der einfaktoriellen Varianzanalyse mit Messwiederholung angewendet (vgl. S. Siegel, 1976, S. 159 u. f.; Bortz, Lienert, & Boehnke, 2008, S. 267 u. f.). Um bei einem signifikanten Ausgang des Friedman-Tests darüber Auskunft zu erhalten, welche Mediane sich signifikant voneinander unterscheiden und wie stark diese Unterschiede je Schülerlösungstext-Paar sind, wurden anschließend 2-seitige Wilcoxon-Vorzeichen-Rang-Tests als Post-hoc-Tests durchgeführt (vgl. Corder & Foreman, 2009, S. 87 u. f.), sowie das Effektstärkemaß[119] $ES = \frac{|z|}{\sqrt{n}}$ (vgl. R. Rosenthal, 1991, S. 19) berechnet. Bei diesem Effektstärkemaß gelten – als Faustregel – signifikante Werte ab .10 als „schwacher“, ab .30 als „moderate“ und Werte ab .50 als „starker“ Effekt (vgl. Pallant, 2007, S. 225). Da alle benannten Testverfahren vollständige Datensätze erfordern, wurden die unvollständigen Punktevergaben von Herrn Iezzi und Frau Pinna an entsprechender Stelle von der Datenanalyse ausgeschlossen (vgl. Unterabschnitt 6.3.1.2).

6.3.1.2. Ergebnisse der quantitativen Analyse

Die Ergebnisse der quantitativen Analyse der Punkteverteilungen gemäß dem eben beschriebenen Vorgehen sind in Tabelle 6.3 und Abbildung 6.7 zusammengefasst. Um die Ergebnisse der einzelnen Wilcoxon-Vorzeichen-Rang-Tests besser interpretieren zu können, sind diese in einem Blasendiagramm veranschaulicht (vgl. Abbildung 6.7). In diesem Diagramm sind die Schülerlösungstexte A bis D entsprechend ihrer Vorauswahl in der Entwicklungsstudie (vgl. Unterkapitel 5.3.2) in einem zweidimensionalen Koordinatensystem mit den Achsen „fachlich-konzeptuelle Qualität“ und „Qualität der sprachlichen Realisierung“ verortet. Die Blasengrößen im Diagramm repräsentieren den Median der von den 21 Physiklehrkäften vergebenen Punkte an den jeweiligen Schülerlösungstext (als Zah-

ler unter Frequenz nicht gut genug meint“ (Seg. 389). Sie würde daher, wenn sie „die ganze Arbeit korrigiert ha[t] [und] [...] 1 Punkt jetz- noch für die bessere Note notwengich is-[,] [...] da die Augen zudrücken und ihm den Punkt geben“ (Seg. 410-411).

[119] Beim Effektstärkemaß ES ist z die z-transformierte Teststatistik des Wilcoxon-Vorzeichen-Rang-Tests und n die Gesamtzahl der Beobachtungen ($= 2 \times$ Probandenanzahl N) (vgl. Pallant, 2007, S. 225).

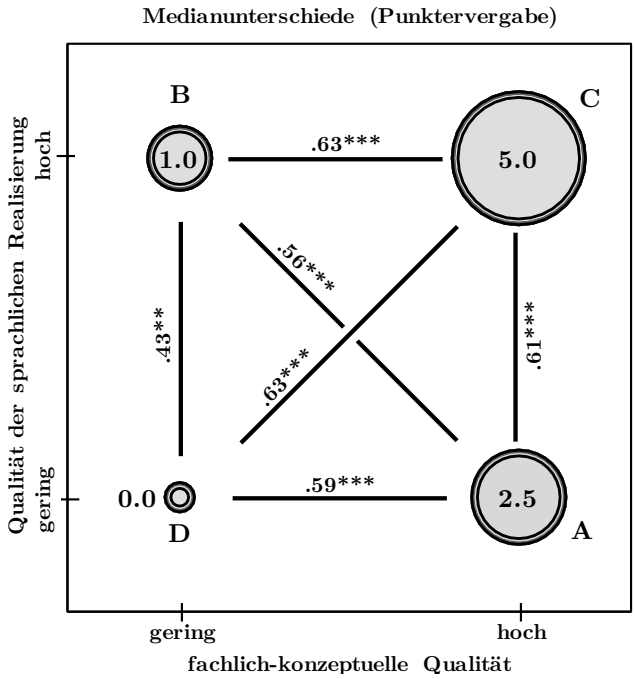


Abbildung 6.7.: Blasendiagramm für die Medianunterschiede (Effektstärke) in der Punkteverteilung für die die Schülerlösungstexte A bis D (***:$p \leq .001$; **:$p \leq .01$; 2-seitiger Wilcoxon-Vorzeichen-Rang-Test).

lenwert mit angegeben). Die Zahlenwerte über den im Blasendiagramm eingezeichneten Linien sind die Effektstärken für die Medianunterschiede in der Punkteverteilung.

6.3.1.2.1. Ergebnis des Friedman-Tests

Wie aus Tabelle 6.3 deutlich wird, hatte der Friedman-Test, bei dem die Punktevergaben von Herrn Iezzi und Frau Pinna von den Berechnungen ausgeschlossen wurden, einen signifikanten Ausgang ($p < .001$). Die mittleren Ränge der von den Teilnehmer_innen vergebenen Punkte unterscheiden sich also bei mindestens zwei der vier Schülerlösungstexte signifikant voneinander.

6.3.1.2.2. Ergebnisse der Wilcoxon-Vorzeichen-Rang-Tests

Die im Anschluss an den Friedman-Test durchgeführten Wilcoxon-Vorzeichen-Rang-Tests je Schülerlösungstext-Paar hatten zum Ergebnis, dass sich alle vier Schülerlösungstexte paarweise bezüglich ihres Punktemedians signifikant voneinander unterscheiden ($p \leq .01$ für Schülerlösungstext B und D; $p \leq .001$ bei allen anderen Schülerlösungstext-Paaren; vgl. Abbildung 6.7). Die Punktevergaben von Herrn Iezzi und Frau Pinna wurden von den Berechnungen ausgeschlossen, in denen die Punkteverteilung von Schülerlösungstext B und/oder D einen Teil des analysierten Datensatzes bildeten. Ferner zeigt sich in den berechneten Effektstärken, dass es sich hierbei nicht um geringfügige, sondern um moderate bis starke Unterschiede handelt ($.43 \leq ES \leq .63$).

6.3.1.3. Interpretation: quantitative Teilbefunde

Die Ergebnisse der quantitativen Analyse der Punkteverteilung lassen sich vor allem im Sinne von Forschungsfrage (F1) interpretieren:

Zunächst lässt sich aus dem signifikanten Ausgang des Friedman-Tests und aller Wilcoxon-Vorzeichen-Rang-Tests schlussfolgern, dass die Physiklehrkräfte die Schülerlösungstexte A bis D unterschiedlich bepunkteten und damit also die Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie global gesehen die Punktevergabe der Physiklehrkräfte beeinflusste.

Um die Ergebnisse im Sinne von Forschungsfrage (F1) genauer interpretieren zu können, gilt es sich ins Gedächtnis zu rufen, dass es sich bei den Schülerlösungstexten A bis D um eine kontrastierende Auswahl handelt (vgl. Unterkapitel 5.3.4): Schülerlösungstext B und D (A und C) weisen eine geringe (hohe) fachlich-konzeptuelle Qualität auf und Schülerlösungstext A und D (B und C) eine geringe (hohe) Qualität der sprachlichen Realisierung. Dementsprechend lässt der signifikante Ausgang des Friedman-Tests und aller Wilcoxon-Vorzeichen-Rang-Tests die Interpretation zu, dass sowohl die fachlich-konzeptuelle Qualität, als auch die Qualität der sprachlichen Realisierung der Schülerlösungstexte die Punktevergabe der Physiklehrkräfte beeinflusste. Ferner zeigen sich in den berechneten Medianen zwei Tendenzen, die die eben aufgeführte Interpretation zusätzlich unterstützen[120]:

Tendenz 1: Bei zwei Schülerlösungstexten mit vergleichbarer Qualität in der sprachlichen Realisierung, ist der Punktemedian desjenigen Schülerlösungstextes größer, dessen fachlich-konzeptuelle Qualität höher ist (vgl. Medianunterschiede zwischen Schülerlösungstext A und D bzw. B und C).

Tendenz 2: Bei zwei Schülerlösungstexten, deren fachlich-konzeptuelle Qualität vergleichbar ist, ist der Punktemedian desjenigen Schülerlösungstextes größer, dessen sprachliche Realisierung eine höhere Qualität aufweist (vgl. Medianunterschiede Schülerlösungstext A und C bzw. B und D).

Des Weiteren zeigt sich in den berechneten Effektstärken eine dritte Tendenz[120]:

Tendenz 3: Der Medianunterschied zwischen den Schülerlösungstexten B und D ist deutlich geringer (moderater Effekt), als zwischen Schülerlösungstext A und C (starker Effekt). Diese Schülerlösungstext-Paare gleichen sich darin, dass die fachlich-konzeptuelle Qualität beider Texte eines Paares jeweils vergleichbar ist und dass sich beide bezüglich der Qualität ihrer sprachlichen Realisierung

[120] Diese Tendenzen sprechen dafür, in einer zukünftigen Studie mit Hilfe einer zweifaktoriellen Varianzanalyse mit Messwiederholung in beiden Faktoren, die Haupteffekte der fachlich-konzeptuellen Qualität und der Qualität der sprachlichen Realisierung eines Schülerlösungstextes auf die Bepunktung durch Physiklehrkräfte zu untersuchen. Zudem sollte hierbei auch der Interaktionseffekt beider Qualitäten untersucht werden, da sich das Vorhanden- bzw. Nichtvorhandenseins eines solchen im Sinne von Forschungsfrage (F2) interpretieren ließe. In der Hautstudie der vorliegenden Arbeit war dies aufgrund der geringen Stichprobengröße nicht möglich und weil es m. W. gegenwärtig kein nicht-parametrisches Äquivalent zur zweifaktoriellen Varianzanalyse mit Messwiederholung in beiden Faktoren gibt.

voneinander unterscheiden. Die Schülerlösungstext-Paare unterscheiden sich jedoch darin, dass die fachlich-konzeptuelle Qualität von Schülerlösungstext B und D geringer ist, als jene von Schülerlösungstext A und C.

Diese dritte Tendenz lässt sich im Sinne von Forschungsfrage (F2) interpretieren: Der Unterschied in den Effektstärken deutet darauf hin, dass die Teilnehmer_innen bei den Schülerlösungstexten mit höherer fachlich-konzeptueller Qualität sprachliche Merkmale bei der Punktevergabe stärker berücksichtigten, als bei den Schülerlösungstexten mit einer geringeren fachlich-konzeptuellen Qualität. In den Punktevergaben der Teilnehmer_innen zeigt sich also ein empirischen Hinweis auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile.

6.3.1.4. Limitationen

Die eben vorgenommene Interpretation spricht im Sinne der Assessment Literacy Konzeption aus Abschnitt 2.2.4 dafür, dass die Physiklehrkräfte im Rahmen der laut-denkenden Korrekturarbeit auf Wissen und Können und/oder berufsbezogene Überzeugungen zur Leistungsfeststellung und -beurteilung zurückgegriffen haben, bei dem bzw. bei denen sowohl fachlich-konzeptuelle, als auch sprachliche Merkmale von Schülerlösungstexten eine Rolle spielen. Allerdings kann die hier durchgeführte Analyse der Medianunterschiede in der Punkteverteilung nicht aufklären, ob diese Ressourcen zur Leistungsurteilsgenese von den Teilnehmer_innen für eine kriteriale Beurteilung herangezogen wurden und/oder ob sie den Teilnehmer_innen dazu dienten, die vier Schülerlösungstexte durch unmittelbaren Vergleich miteinander zu beurteilen (kriteriale vs. soziale Bezugsnormorientierung; vgl. Abschnitt 2.1.2).

Des Weiteren müssen Teile der aufgeführten Interpetation an dieser Stelle der vorliegenden Arbeit als vorläufig gelten. Grund hierfür ist, dass die Punkteverteilung der Teilnehmer_innen lediglich „finale Gesamturteile“ darstellen, aufgrund derer sich nur unter Vorbehalt auf den Prozess der Genese von fachlich-konzeptuellen und sprachlichen Teilleistungsurteilen schließen lässt.

6.3.2. Inhaltsanalytische Auswertung der Laut-Denk-Protokolle

6.3.2.1. Methodische Vorbemerkungen zum gesamten Auswertungsprozess

Die rechte Spalte von Tabelle 6.4 liefert einen schlagwortartigen Überblick über die Selbstauskünfte der Lehrkräfte, inwieweit sie das laute Denken bei der Bearbeitung des Aufgabenhefts als belastend und/oder unterbrechend empfunden haben (diese Selbstauskünfte wurden unmittelbar nach der Bearbeitung des Aufgabenhefts durch die Teilnehmer_innen erfragt; vgl. Abschnitt 5.4.1). Hier zeigt sich, dass die deutliche Mehrheit der Teilnehmer_innen (17) das laute Denken während der Aufgabenheftbearbeitung nicht als belastend

Pseudonym	**Dauer der Laut-Denk-Phase**	**Segmentanzahl der Laut-Denk-Protokolle**	**Belastung/Unterbrechung durch das laute Denken?** (Selbstauskunft)
Herr Abney	26 min 39 s	411	keine
Herr Balke	15 min 49 s	245	keine
Herr Carboni	38 min 58 s	432	keine
Herr Dassow	38 min 42 s	508	erfordert Multitasking
Herr Einert	47 min 39 s	842	keine
Herr Feldner	26 min 14 s	507	keine
Herr Geppert	54 min 26 s	827	keine
Herr Hastedt	29 min 54 s	545	keine
Herr Iezzi	14 min 11 s	198	nein
Herr Jonuzi	12 min 47 s	245	keine
Frau Kirik	29 min 50 s	523	keine
Herr Lemos	23 min 2 s	255	permanentes Sprechen ist fordernd
Herr Mehlert	41 min 27 s	832	keine
Frau Novack	31 min 3 s	497	keine
Herr Onne	15 min 6 s	239	keine
Frau Pinna	27 min 21 s	418	erfordert Multitasking
Herr Quezada	23 min 36 s	313	keine
Herr Rittershaus	36 min 57 s	697	permanentes Sprechen ist fordernd
Frau Sohm	26 min 49 s	472	keine
Herr Trummer	36 min 14 s	599	keine
Herr Uckermark	40 min 13 s	669	keine
Median	29 min 50 s	497	
IQR	15 min 6 s	350	
Summe	10 h 36 min 57 s	10 274	

Tabelle 6.4.: Übersicht über die Laut-Denk-Daten der 21 Teilnehmer_innen.

oder unterbrechend empfand[121]. Eine Minderheit der Teilnehmer_innen (4) gab an, dass sie das Multitasking[122] oder das permanente Sprechen[123] beim lauten Denken als fordernd empfunden haben, jedoch nicht als hochgradig belastend oder unterbrechend. Vor allem bei diesen vier Teilnehmer_innen ist also davon auszugehen, dass das laute Denken einen reaktiven Effekt hatte. Allerdings kann auf Grundlage der Selbstauskünfte dieser reaktive Effekt als vertretbar angesehen werden, weswegen die Laut-Denk-Daten dieser vier Teilnehmer_innen nicht von der weiteren Analyse ausgeschlossen wurden.

In den beiden mittleren Spalten von Tabelle 6.4 ist für jede teilnehmende Lehrkraft die Dauer der laut-denkenden Aufgabenheftbearbeitung, sowie die Anzahl der Segmente im Laut-Denk-Protokoll angegeben. Was hier unmittelbar auffällt, ist der erhebliche Umfang[124] der (Verbal-)Daten, die von den Teilnehmer_innen bei der Bearbeitung des Aufgabenhefts erhoben wurden. Zum einen galt es dieses umfangreiche Datenmaterial im

[121] Z. B. äußerte Herr Trummer: „War nich- schwer, [...] weil das -ne Tätigkeit is-, die mir nich- fremd is- [das Korrigieren einer Klassenarbeit; M. S. F.]. [...] Und der Übergang zum lauten Denken is da so völlig problemlos gewesen“ (Seg. 606).

[122] Z. B. äußerte Herr Dassow: „Es ist anstrengend zu reden und gleichzeitig zu denken, zu schreiben und hin und her blättern. Das ist ja dann schon so'n bisschen Multitasking. [...] Also viel länger hätte d[as] jetzt nicht sein dürfen für mich. Aber es ging soviel“ (Seg. 514).

[123] Z. B. äußerte Herr Rittershaus: „Is- schon -n -ne zusätzliche zusätzliche Belastung für mich. Ähm ja. N- kleine zumindest. Also f-... Ich merk-, dass das für meine Stimme zum Beispiel anstrengend is-“ (Seg. 710).

[124] Zudem fällt auf, dass sich die erhobenen Verbaldaten bzw. die Laut-Denk-Protokolle der Teilnehmer_innen bezüglich ihres Umfangs zum Teil deutlich voneinander unterscheiden (Audiographien: IQR = 15 min 6 s; Spannweite: 12 min 57 s bis 54 min 26 s. Laut-Denkprotokolle: IQR = 350 Segmente; Spannweite: 198 bis 842 Segmente). Diese Unterschiede zwischen den Teilnehmer_innen galt es bei

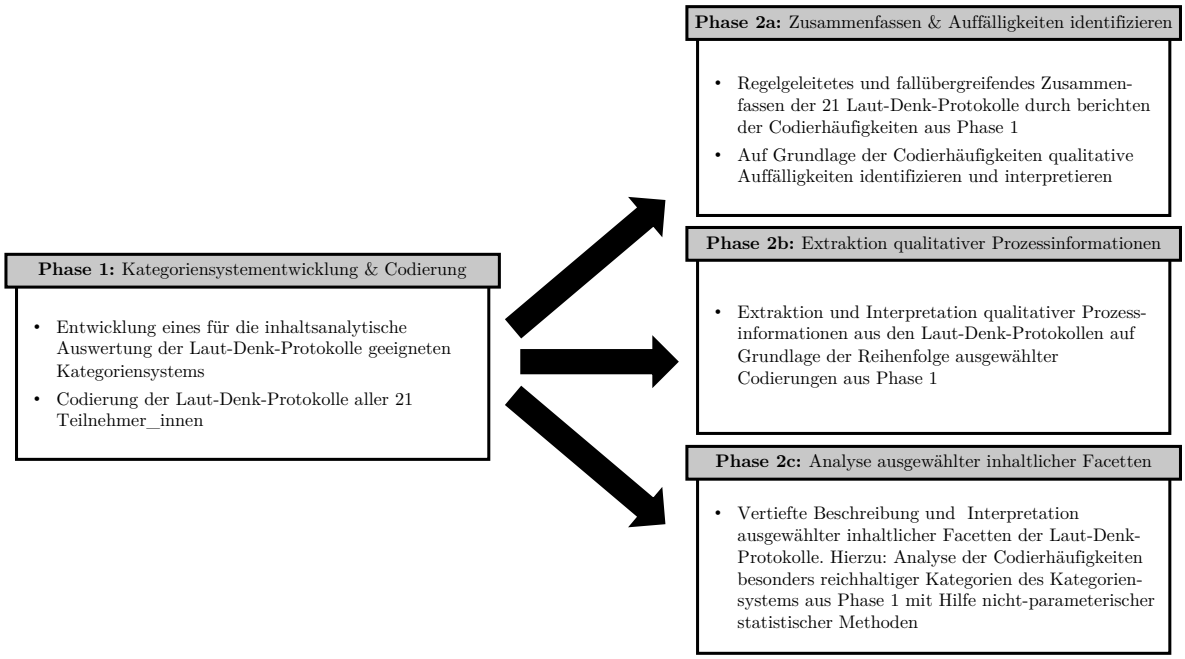


Abbildung 6.8.: Phasen der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle.

Rahmen der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle zusammenzufassen, um so dem deskriptiven Charakter der Hauptstudie gerecht zu werden (vgl. Unterkapitel 4.2). Zum anderen galt es allerdings auch ausgewählte inhaltliche Aspekte und Aspekte des Geneseprozesses von Leistungsurteilen durch die Teilnehmer_innen im Datenmaterial zu identifizieren und systematisch zu beschreiben, um so dem explorativen Charakter der Hauptstudie gerecht zu werden (vgl. Unterkapitel 4.2). Ein Rückgriff auf eine der in der Literatur beschriebenen Standardvarianten der qualitativen Inhaltsanalyse (vgl. Schreier, 2014b) war dabei nicht möglich, da diese entweder auf eine Zusammenfassung des zu untersuchenden Materials oder auf die Identifikation und Interpretation ausgewählter Aspekte (durch Explikation, Extraktion oder Strukturierung) ausgelegt sind (vgl. Lissmann, 2001, S. 58 u. f.; Gläser & Laudel, 2010, S. 199 u. f.; Mayring, 2015, S. 67). Die inhaltsanalytische Auswertung der Laut-Denk-Daten der Teilnehmer_innen erfolgte daher anhand eines eigens an die Bedürfnisse der Hauptstudie angepassten Ablaufschemas, in der unterschiedliche Analysetechniken zum Einsatz kamen und das sich grob in zwei Phasen gliedern lässt (vgl. Abbildung 6.8):

Phase 1: Zunächst wurde in einem deduktiv-induktiven Verfahren ein inhaltsanalytisches Kategoriensystem entwickelt, mit dessen Hilfe für jedes Segment in den Laut-Denk-Protokollen die wesentliche inhaltliche Bedeutung erfasst werden konnte. Anschließend erfolgte die Codierung der Laut-Denk-Protokolle aller 21 Teilnehmer_innen mit dem entwickelten Kategoriensystem.

der inhaltsanalytischen Auswertung der laut-Denk-Protokolle besonders zu berücksichtigen, was im Folgenden an entsprechender Stelle erläutert wird.

Phase 2: Die Codierung der Laut-Denk-Protokolle aller 21 Teilnehmer_innen in Phase 1 bildete die Grundlage für drei anschließende Analysen:

Phase 2a: Erstens wurden die 21 Laut-Denk-Protokolle durch Berichten der Codierhäufigkeiten regelbasiert und fallübergreifend zusammengefasst. Ferner wurden in den berichteten Codierhäufigkeiten (globale) qualitative Auffälligkeiten identifiziert und unter Anderem durch Rückgriff auf die codierten Verbaldaten dem Erkenntnisinteresse der vorliegenden Arbeit entsprechend interpretiert.

Phase 2b: Zweitens erfolgte eine *Extraktion*[125] und Interpretation qualitativer Facetten des Geneseprozesses von Leistungsurteilen durch die Teilnehmer_innen. Dies geschah durch eine Analyse der chronologischen Abfolge, in der Segmente mit bestimmten Codierungen in den Laut-Denk-Protokollen auftraten.

Phase 2c: Drittens wurden die Codierhäufigkeiten besonders reichhaltiger Kategorien des in Phase 1 entwickelten Kategoriensystems mit Hilfe nichtparametrischer statistischer Methoden analysiert. Hierdurch konnten ausgewählte inhaltliche Facetten in den erhoben Daten identifiziert und systematisch im Sinne des Erkenntnisinteresse der vorliegenden Arbeit beschrieben werden.

Entsprechend dieser Phasen gliedern sich die nachfolgenden Unterabschnitte der vorliegenden Arbeit. Als erstes wird in Unterabschnitt 6.3.2.2 die Entwicklung und der Aufbau des Kategoriensystems erläutert, mit dem die Laut-Denk-Protokolle der Teilnehmer_innen inhaltsanalytisch ausgewertet wurden (Phase 1). In den drei darauf folgenden Unterabschnitten erfolgt jeweils die Darstellung des methodischen Vorgehens, der Ergebnisse, der Interpretation und der Limitation der drei Analysen der Laut-Denk-Protokolle auf Grundlage ihrer Codierung mit dem zuvor entwickelten Kategoriensystem (Phase 2a, 2b und 2c).

6.3.2.2. Phase 1: Kategoriensystementwicklung und Codierung der Laut-Denk-Protokolle

Vor der Entwicklung des Kategoriensystems für die inhaltsanalytische Auswertung der Laut-Denk-Protokolle wurden die Transkriptsegmente als *Codiereinheiten* (vgl. Schreier, 2012, S. 131 u. f.; Kuckartz, 2016, S. 41 u. f.) festgelegt, um eine möglichst feinkörnige

[125] Im Rahmen der vorliegenden Arbeit ist „Extraktion“ im Sinne des Begriffsverständnisses von Gläser & Laudel (2010) zu verstehen: „Wir verwenden den Begriff Extraktion, um den Unterschied zum ‚Kodieren‘ von Texten deutlich zu machen: Das Kodieren indiziert den Text, um ihn auswerten zu können. Es macht also Text und Index zum gemeinsamen Gegenstand der Auswertung. Mit der Extraktion entnehmen wir dem Text Informationen und werten diese Informationen aus. [...] Extraktion heißt, den Text zu lesen und zu entscheiden, welche der in ihm enthaltenen Informationen für die Untersuchung relevant sind. Diese Informationen werden den Kategorien des Suchrasters [das Kategoriensystem, mit dem die Inhaltsanalyse vorgenommen wird; M. S. F.] zugeordnet“ (ebd., S. 199-200).

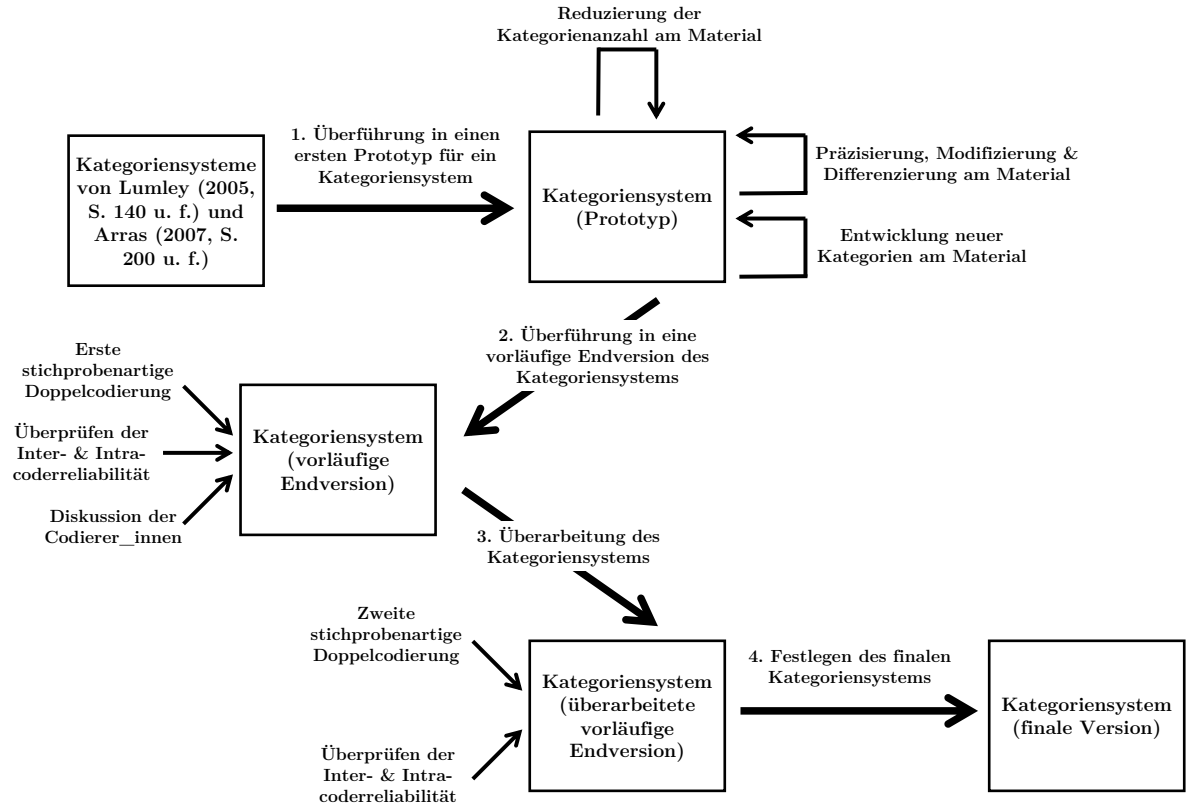


Abbildung 6.9.: Ablaufschema des deduktiv-induktiven Verfahrens zur Entwicklung des Kategoriensystems für die inhaltsanalytische Auswertung der Laut-Denk-Protokolle.

Analyse der von den Teilnehmer_innen mitvokalisierten Gedankenschritte zu ermöglichen. Zusätzlich wurde die Codierregel festgelegt, dass bei der Codierung im Zweifelsfall auch die zum zu codierenden Segment unmittelbar benachbarten Segmente zu beachten sind. Diese zusätzliche Codierregel war notwendig, da aufgrund des zum Teil fragmentarischen Charakters der Segmente oftmals der Kontext ausschlaggebend dafür ist, welche inhaltliche Bedeutung einem Transkriptsegment zuzuordnen ist.

Wie Abbildung 6.9 schematisch veranschaulicht, wurde für die Entwicklung dieses Kategoriensystems – ähnlich wie bei der Kriterienrasterentwicklung im Rahmen der Entwicklungsstudie (vgl. Abschnitt 5.3.2) – eine deduktiv-induktive Verfahrensweise gewählt (vgl. Boyatzis, 1998, S. 37 u. f.; Schreier, 2012, S. 89 u. f.; Kuckartz, 2016, S. 95 u. f.; Saldaña, 2016, S. 74): Im ersten Schritt wurde die bisherige erziehungswissenschaftliche Forschungsliteratur nach Studien durchsucht, in denen Kategoriensysteme zur Analyse von laut-denkend vorgenommenen Feststellungen und Beurteilungen schriftlicher (Schüler-)Leistungen entwickelt wurden. Bei dieser Literaturrecherche stellten sich die Kategoriensysteme, die Lumley (2005, S. 140 u. f.) und Arras (2007, S. 200 u. f.) in ihren Laut-Denk-Studien zur Feststellung und Beurteilung schriftlicher Leistungen in einer Fremdsprache entwickelten, als besonders detailliert und gut dokumentiert heraus.

Aus den Dokumentationen dieser Kategoriensysteme wurde a-priori ein erster Prototyp eines Kategoriensystems für die inhaltsanalytische Auswertung der Laut-Denk-Protokolle der Teilnehmer_innen entwickelt. Dieses A-Priori-Kategoriensystem wurde aus mehreren Kategorien aufgebaut, welche die beim lauten Denken mitvokalisierten (mentalen, interaktionalen oder aktionalen) Handlungen der Teilnehmer_innen grob zusammenfassen (z. B. „lesen/erfassen eines Textes“) (vgl. Boyatzis, 1998, S. 29 u. f.). Ferner bestand jede dieser Kategorie aus mehreren Subkategorien, welche die inhaltliche Bedeutung einer Kategorie ausdifferenzieren und weiter präzisieren (z. B. „lesen eines Schülerlösungstextes“ oder „Paraphrasierung von Anweisungen im Aufgabenheft“).

Im zweiten Schritt wurden die zuvor a-priori entwickelten Kategorien und Subkategorien in einem zirkulären Prozess am gesamten Datenmaterial (den Laut-Denk-Protokolle der Teilnehmer_innen) pilotiert und überarbeitet:

In einem Zyklus dieses Prozesses wurde der Kategoriensystem-Prototyp auf die Laut-Denk-Protokolle von zwei bis vier Teilnehmer_innen angewendet. Dabei wurden Transkriptsegemente, denen bezüglich einer bestimmten Kategorie keine der definierten Subkategorien eindeutig zugeordnet werden konnten, in die zusätzliche Kategorie „sonstige Transkriptsegmente“ eingeordnet. Des Weiteren wurden Transkriptsegemente, die einer bestimmten Subkategorien zugeordnet werden konnten, deren inhaltliche Bedeutung durch die entsprechende Subkategorie aber augenscheinlich noch zu grobkörnig erfasst wurde, mit einem dementsprechenden Memo versehen[126] (vgl. Saldaña, 2016, S. 44 u. f.).

Nachdem auf alle Laut-Denk-Protokolle des gewählten Teildatensatzes der Kategoriensystem-Prototyp angewandt wurde, wurde dieser folgendermaßen überarbeitet:

1. Die Kategorien und ihre Subkategorien wurden so lange präzisiert, modifiziert und differenziert, bis alle Transkriptsegemente, die zuvor der Kategorie „sonstige Transkriptsegmente“ zugeordnet wurden, eindeutig einer Subkategorie zugeordnet werden konnten.
2. Die Subkategorien, denen keine oder nur wenige Transkriptsegemente zugeordnet wurden, wurden mit anderen, inhaltlich ähnlichen Subkategorien zusammengelegt, um durch eine Reduzierung der Kategorienanzahl die finale Version des Kategoriensystems so handhabbar wie möglich zu gestalten. Selbiges geschah mit inhaltlich ähnlichen Subkategorien, die sich bei der Anwendung des Kategoriensystem-Prototyps auf das Material (augenscheinlich) nicht hinreichend präzise voneinander unterscheiden ließen.

[126] Zum Beispiel äußerte Herr Abney bei der Korrektur von Schülerlösungstext A, dass „[d]es is ja v-... per se erstmal nicht falsch, was da steht“ (Seg. 250), was bei der Anwendung des Kategoriensystem-Prototyps der vorläufigen Subkategorie „Kommentar zum Schülerlösungstext“ zugeordnet wurde. Bei der Codierung wurde als Memo allerdings vermerkt, dass es sich bei Herrn Abneys Kommentar um eine wertende Äußerung handelt („[...] [es ist] per se erstmal nicht falsch“). Aus diesem und weiteren Memos in den Laut-Denk-Protokollen anderer Teilnehmer_innen wurde im Verlauf der Kategoriensystementwicklung die Subsubkategorie „3.0.3 Art der Äußerung“ entwickelt (vgl. Kurzfassung der finalen Version des Kategoriensystems in Abbildung 6.6).

3. Aus den Trankriptsegmenten, die bei der Codierung mit einem Memo versehen worden waren, wurden für die entsprechenden Kategorien bzw. Subkategorien zusätzliche (Subsub-)Subkategorien entwickelt, die den zuvor nur grobkörnig erfassten Inhalt der Transkriptsegmente präziser erfassen[126] und diese am gewählten Teildatensatz erprobt.

Anschließend begann der Prozess von vorne, indem der überarbeitete Kategoriensystem-Prototyp auf die Laut-Denk-Protokolle von zwei bis vier anderen Teilnehmer_innen angewandt wurde. Dieser zirkuläre Prozess wurde so lange durchlaufen, bis allen Transkriptsegmenten in den Laut-Denk-Protokollen der Teilnehmer_innen in jeder Kategorie eindeutig eine (Subsub-)Subkategorie zugeordnet werden konnte und keine Codierung mit einem Memo versehen war. Abschließend wurde der Kategoriensystem-Prototyp zu einer vorläufigen Endversion verschriftlicht und alle 21 Laut-Denk-Protokolle mit Hilfe dieses Kategoriensystems codiert.

Im dritten und vierten Schritt der Kategoriensystementwicklung wurden zirka 10 % des gesamten Datenmaterials[127] insgesamt zwei Zweitcodierungen unterzogen, um die Reliabilität der vorläufigen Endversion des Kategoriensystems zu überprüfen, diese zu verbessern und um das Kategoriensystem in eine finale Version zu überführen:

Für die erste Zweitcodierung wurde ein_e Zweitcodierer_in[128] in den Aufbau der vorläufigen Endversion des Kategoriensystems eingewiesen und dessen Anwendung an einem codierten Beispiel-Laut-Denk-Protokoll[129] erläutert. Anschließend wurde aus den verbleibenden 20 Laut-Denk-Protokollen zufällig das von Herrn Trummer ausgewählt, durch Erst- und Zweitcodier_in unabhängig voneinander (erneut) codiert[130] und schließlich die Inter- und Intracoderreliabilität bestimmt. Die Bestimmung dieser beiden Reliabilitätsmaße diente dazu, zusätzlich eine Aussage über das Ausmaß von *Codierer-Effekten* (vgl. Abschnitt 5.3.4) bei Anwendung des Kategoriensystems vornehmen zu können: Intercoderreliabilität bezeichnet das Ausmaß in dem (mindestens zwei) verschiedene Codierer_innen mit Hilfe eines festgelegten Kategoriensystems zum selben Codier-Ergebnis gelangen und Intracoderreliablität jenes, in dem der_die selbe Codierer_in bei Wiederholung des Codiervorgangs zum selben Codier-Ergebnis gelangt (vgl. Mayring, 2015, S. 124). Der Differenzbetrag der Koeffizienten für Inter- und Intracoderreliabilität ist daher ein Kennwert für das Ausmaß in dem Unterschiede im Codier-Ergebnis auf die Verschieden-

[127] Aufgrund des oftmals erheblichen Umfangs des auszuwertenden Datenmaterial im Rahmen einer Laut-Denk-Studie empfehlen C. Green & Gilhooly (2002, S. 60) die Reliabilität eines zu diesem Zwecke entwickelten Kategoriensystem an einem zufällig ausgewählten Teildatensatz im Umfang von zirka 10 % des gesamten Datenmaterials zu überprüfen. Dieser Empfehlung wurde im Rahmen der vorliegenden Arbeit gefolgt.

[128] Bei dem_der Zweitcodierer_in handelte es sich um eine_n Doktorierende_n der Physikdidaktik.

[129] Hierzu diente das Laut-Denk-Protokoll von Frau Novack, da in diesem nahezu alle (Subsub-)Subkategorien der vorläufigen Endversion des Kategoriensystems auftraten und es sich daher zur Illustration der Anwendung des Kategoriensystems besonders eignete.

[130] Die zweifache Codierung des Laut-Denk-Protokolls von Herrn Trummer durch den_die Erstcodierer_in fand in einem zeitlichen Abstand von einer Woche statt, um sicherzustellen, dass die zweite Codierung nicht durch eine noch deutlich vorhandene Erinnerung an die erste Codierung beeinflusst wurde.

	Erste Zweitcodierung	**Interpretation** (vgl. Wirtz & Caspar, 2002, S. 59)	**Zweite Zweitcodierung**	**Interpretation** (vgl. Wirtz & Caspar, 2002, S. 59)
Intercoderreliabilität κ_{inter}	.61	gut	.70	gut
Intracoderreliabilität κ_{intra}	.77	sehr gut	.87	sehr gut
Differenzbetrag $\lvert\Delta\kappa\rvert$	.16	gering	.17	gering

Tabelle 6.5.: Inter- und Intracoderreliabilität des Kategoriensystems zur inhaltsanalytischen Auwertung der Laut-Denk-Protokolle der Teilnehmer_innen (jeweils Brennans und Predigers κ), sowie Differenzbetrag beider Reliabilitätsmaße als Kennwert für das Ausmaß von Codierer-Effekten.

heit von Erst- und Zweitcodierer_in selbst oder deren „›Interaktion‹ mit der konkreten Codier-Situation [zurückführen sind]“ (Degen, 2015, S. 79).

Als Kennwert für die Inter- und Intracoderreliabilität wurde der von Brennan & Prediger (1981) vorgeschlagene κ-Koeffizient bestimmt, für den – als Faustregel – Werte ab .40 als „akzeptabel“, ab .60 als „gut“ und Werte ab .75 als „sehr gut“ angesehen werden (vgl. Wirtz & Caspar, 2002, S. 59). Wie aus der linken Hälfte von Tabelle 6.5 hervorgeht, zeigte sich für die vorläufige Endversion des Kategoriensystems eine gute Inter- bzw. sehr gute Intracoderreliablität. Ferner konnten bereits für die erste Zweitcodierung Codierer-Effekte als gering eingeschätzt werden, da der Differenzbetrag zwischen beiden κ-Koeffizienten mit $|\Delta\kappa|$=.16 niedriger ausfiel, als die Hälfte des Richtwerts für eine gerade noch akzeptable Inter- bzw. Intracoderreliablität ($\kappa = .40$).

Nach der ersten Zweitcodierung wurden in einer Diskussion zwischen Erst- und Zweitcodierer_in Diskrepanzen in der Codierung besprochen. Im Rahmen dieser Diskussion wurden missverständliche Formulierungen der (Subsub-)Subkategorien der vorläufigen Endversion des Kategoriensystems identifiziert. Diese Formulierungen wurden im Anschluss an die Diskussion von Erst- und Zweitcodier_in durch solche ersetzt, die augenscheinlich besser geeignet sind, um die Bedeutung der (Subsub-)Subkategorien beschreiben zu können.

Um zu überprüfen, ob durch diese Überarbeitung der vorläufigen Endversion des Kategoriensystems eine tatsächliche Verbesserung einherging, wurde erneut eine Zweitcodierung durchgeführt. Für diese zweite Zweitcodierung wurde aus dem gesamten Datenmaterial zufällig das Laut-Denk-Protokoll von Herrn Onne ausgewählt und analog zu dem Vorgehen bei der ersten Zweitcodierung durch Erst- und Zweitcodierer_in zwei- bzw. einmal codiert, sowie schließlich die Inter- und Intracoderreliabilität bestimmt. Wie aus der rechten Hälfte von Tabelle 6.5 hervorgeht, zeigte sich für die Überarbeitung der vorläufigen Endversion des Kategoriensystems eine leichte Verbesserung sowohl in der Inter-, als auch der Intracoderreliabilität. Ferner blieben die Codierer-Effekte (nahezu) unverändert gering ($|\Delta\kappa| = .17$). Da zudem aufgrund des hohen Kennwertes für die Intracoderreliablität ($\kappa_{intra} = .87$) kaum zu erwarten war, dass eine erneute Überarbeitung des Kategoriensystems noch zu einer tatsächlichen Verbesserung führen würde, wurde auf eine erneute Diskussion zwischen Erst- und Zweitcodierer_in verzichtet. Stattdessen wurde das Kate-

Kategorie 1: lesen/erfassen eines Textes

Subkategorie	Subsubkategorien			Codierregel
1.0 zusammenhängendes Lesen	1.0.1 eine Anweisung im Aufgabenheft	1.0.2 die Aufgabe Weltraumspaziergang	1.0.3 einen Schülerlösungstext	Einem Segment wird genau eine Subkategorie mit genau einer Subsubkategorie zugewiesen
1.1 fragmentarisches Lesen	1.1.1 eine Anweisung im Aufgabenheft	1.1.2 die Aufgabe Weltraumspaziergang	1.1.3 einen Schülerlösungstext	
1.3 paraphrasieren	1.2.1 eine Anweisung im Aufgabenheft	1.2.2 die Aufgabe Weltraumspaziergang	1.2.3 einen Schülerlösungstext	

Kategorie 2: erstellen des Erwartungshorizonts zur Aufgabe Weltraumspaziergang

Subkategorie	Codierregel
2.0 Beurteilungskriterien werden benannt/kommentiert/abgewogen/festgelegt (Handlung, Kommentar oder selbstreflektierte Äußerung)	Einem Segment dieser Kategorie wird genau eine Subkategorie zugewiesen
2.1 Zuweisung von Punkten im Erwartungshorizont (Handlung, Kommentar oder selbstreflektierte Äußerung)	

Kategorie 3: Korrektur der Schülerlösungstexte

Subkategorie	Subsubkategorien	Subsubsubkategorien					Codierregel
3.0 Feststellung und Beurteilung eines Schülerlösungstextes (z. B. Kommentare, aushandeln einer Entscheidung)	3.0.1 fokussiertes Merkmal	3.0.1.1 fachlich-konzeptueller Eindruck	3.0.1.2 sprachliche Realisierung	3.0.1.3 sonstiges Merkmal (z. B. Handschrift)	3.0.1.4 mehrere/ uneindeutig		Einem Segment dieser Subkategorie wird pro Subsubkategorie genau eine Subsubsubkategorie zugewiesen
	3.0.2 Bezug der Verortung	3.0.2.1 sachliches Kriterium	3.0.2.2 andere Schülerlösungstexte	3.0.2.3 allgemeine Erfahrung	3.0.2.4 mutmaßliches Personenmerkmal (z. B. Geschlecht)	3.0.2.5 mehrere/ uneindeutig	
	3.0.3 Art der Äußerung	3.0.3.1 positiv wertend/akzeptierend		3.0.3.2 negativ wertend/ ablehnend	3.0.3.3 neutral/gemischt/sonstig		
3.1 Beurteilungskriterien ad hoc benennen/abwägen oder aus dem Erwartungshorizont entnehmen	3.1.1 fachlich-konzeptueller Eindruck	3.1.2 sprachliche Realisierung		3.1.3 sonstiges Merkmal (z. B. Handschrift)	3.1.4 mehrere/uneindeutig		Einem Segment dieser Kategorie wird genau eine Subsubkategorie zugewiesen.

Kategorie 4: Äußerungen außerhalb der eigentlichen Korrektur und Erwartungshorizonterstellung

Subkategorie	Codierregel
4.0 Beschreibung allgemeiner Handlungsstrategien beim Feststellen und Beurteilen von Schülerleistungen oder Äußerungen zum allgemeinen Vorgehen/zu allgemeinen Erfahrungen beim Feststellen und Beurteilen von Schülerleistungen (Kommentar, Kritik oder selbstreflektierte Äußerung)	Einem Segment dieser Kategorie wird genau eine Subkategorie zugewiesen
4.1 Kommentare zur, Kritik an, Fragen zur oder Interpretation der Aufgabe Weltraumspaziergang	
4.2 Kommentare zu, Kritik an, Fragen zum oder Interpretation von Anweisungen oder zum Aufbau des Aufgabenheftes	
4.3 Beschreibung allgemeiner Handlungsstrategien zum Erstellen eines Erwartungshorizonts oder Kommentar zum, Kritik oder selbstkritische Äußerungen am eigenen Erwartungshorizont im Allgemeinen	

Kategorie 5: emotionale Äußerungen und nichtsprachliche Ereignisse

Subkategorie	Codierregel
5.0 Lachen, Stöhnen, Ausdruckspartikel (z. B. So!), Planungsäußerungen und Verzögerungslaute (z. B. ähm), usw.	Einem Segment dieser Kategorie wird genau eine Subkategorie zugewiesen
5.1 Klanggesten (z. B. Fingerschnippen)	

Kategorie 6: sonstige Äußerungen/Artefakte des lauten Denkens/sonstige nichtsprachliche Ereignisse

Subkategorie	Codierregel
6.0 „activity descriptions“, sonstige Handlungen (z. B. blättern) oder Äußerungen zu sonstigen eigenen Handlungen/Verhaltensweisen/Gedanken	Einem Segment dieser Kategorie wird genau eine Subkategorie zugewiesen
6.1 Ankündigungen (z. B. „Also:“); weitere Äußerungen/Transkriptsegemente	

Tabelle 6.6.: Kurzfassung des finalen Kategoriensystems für die inhaltsanalytische Auswertung der 21 Laut-Denk-Protokolle.

goriensystem, das aus der Überarbeitung nach der ersten Zweitcodierung hervorging, als finale Version des Kategoriensystem für die inhaltsanalytische Auswertung der Laut-Denk-Protokolle festgelegt, mit dem abschließend die Laut-Denk-Protokolle aller 21 Teilnehmer_innen einer endgültigen Codierung unterzogen wurden[131].

Die finale Version des Kategoriensystems für die inhaltsanalytische Auswertung der Laut-Denk-Protokolle ist in Tabelle 6.6 überblicksartig dargestellt. Das vollständige Kategoriensystem ist in Anhang E zu finden. Um die hierarchische Anordnung der Kategorien und ihrer (Subsub-)Subkategorien nachvollziehbar zu machen und um im weiteren Verlauf der vorliegenden Arbeit einfacher auf einzelne (Subsubsub-)Kategorien verweisen zu können, sind diese mit eindeutigen Codenummern versehen.

Das finale Kategoriensystem besteht aus den folgenden 6 Kategorien, welche – wie bereits die Kategorien des A-priori-Prototyps – die beim lauten Denken mitvokalisierten (mentalen, interaktionalen oder aktionalen) Handlungen der Teilnehmer_innen grob zusammenfassen (Details zu den (Subsub-)Subkategorien dieser Kategorien werden in Unterabschnitt 6.3.2.3 dargestellt):

Kategorie 1: lesen/erfassen eines Textes

Kategorie 2: erstellen des Erwartungshorizonts zur Aufgabe Weltraumspaziergang

Kategorie 3: Korrektur der Schülerlösungstexte

Kategorie 4: Äußerungen außerhalb der eigentlichen Korrektur und Erwartungshorizonterstellung

Kategorie 5: emotionale Äußerungen und nichtsprachliche Ereignisse

Kategorie 6: sonstige Äußerungen/Artefakte des lauten Denkens/sonstige nichtsprachliche Ereignisse

Bei der Anwendung[132] des Kategoriensystems, die in Tabelle 6.7 an einem Beispieltranskript illustriert ist, wird jedem Segment des Laut-Denk-Protokolls eines_einer Teilnehmers_Teilnehmerin genau eine dieser 6 Kategorien zugewiesen, indem es mit genau einer (Sub-)Subkategorien codiert wird (vgl. Kurzfassung der Codierregeln in Tabelle 6.6). Einzige Ausnahme bildet die Subkategorie „Feststellung und Beurteilung eines Schülerlösungstextes“ (Subkategorie 3.0), in der eine Dreifachcodierung erfolgt: Ist ein Segment des Laut-Denk-Protokolls eines_einer Teilnehmers_Teilnehmerin dieser Subkategorie zuzuweisen, so wird ihm aus jeder der Subsubkategorien 3.0.1, 3.0.2 und 3.0.3 jeweils genau eine Subsubsubkategorie zugewiesen.

Aus Tabelle 6.6 geht hervor, dass es in der finalen Version des Kategoriensystems keine (Subsubsub-)Katgeorien gibt, die beschreiben, an welcher Stelle der Laut-Denk-Protokolle sich mit welchem der vier Schülerlösungstexte auseinandergesetzt wird. Auf derartige

[131] Aus forschungsökonomischen Gründen diente die Codierung der Laut-Denk-Protokolle mit der vorläufigen Endversion des Kategoriensystems als Grundlage für die endgültige Codierung. Diese Codierung wurde mit der finalen Version des Katgegoriensystems überprüft und überarbeitet.

[132] Details hierzu sind den Codierregeln für das Kategoriensystem in Anhang E zu entnehmen.

	Feintranskript mit Segmentierung	Codierung (Codenr.)
	[...]	
[3]	Gut, (.) dann les- ich mir jetz- erstmal Aufgabe 1 durch:	6.0
[4]	*Stellen Sie sich folgende Situation vor: Sie unterrichten eine 9. Klasse in Physik und haben eine Klassenarbeit geschrieben. In dieser Klassenarbeit haben Sie die Aufgabe „Weltraumspaziergang" als Grundwissensaufgabe eingesetzt (siehe Seite 3). Sie haben Ihren Schülerinnen und Schülern zusätzlich folgende Anweisung gegeben: „Schreibt eure Antworten in ganzen Sätzen auf." „Skizzen oder Zeichnungen können bei dieser Aufgabe nicht gezählt werden."*	1.0.1
[5]	(zustimmend) Mhm. (+) Gut.	5.0
[6]	Würd- ich so wahrscheinlich nich- machen.	4.2
	[...]	
[45]	Ich bin jetz- schon gedanklich so'n bisschen dabei die Punkte zu verteilen.	6.0
[46]	Ich würde jetzt vielleicht für den ersten... (.) den ersten Punkt dafür vergeben...	2.1
[47]	Erster... 1 Punkt...	2.1
[48]	(schreibt „1 P." in den Erwartungshorizont)	2.1
[49]	dafür... für ähm die Erklärung warum er ihn nicht hört.	2.0
	[...]	
[147]	So, der oder die Schülerin schreibt:	6.1
[148]	*Im All ist nichts durch das Ton geht und er hört seinen Freund nicht. Dann kommt aber der Ton durch die Helme, da Ton über Glas geht.*	1.0.3
[149]	Okay.	5.0
[150]	*Im All ist nichts das durch Ton geht...*	1.1.3
[151]	Das heißt, der oder die Schülerin hat erkannt, dass Schall ein Medium benöticht.	3.0.1.1 3.0.2.1 3.0.3.1
[152]	Wobei das Wort Schall nicht auftaucht.	3.0.1.2 3.0.2.1 3.0.3.2
[153]	Aber es taucht das Wort Ton auf.	3.0.1.2 3.0.2.1 3.0.3.2
[154]	Also hier ist der falsche Begriff.	3.0.1.2 3.0.2.1 3.0.3.2
	[...]	

Tabelle 6.7.: Auszug aus dem Laut-Denk-Protokoll von Frau Sohm zur Illustration der Anwendung des Kategoriensystems für die inhaltsanalytische Auswertung der Laut-Denk-Protokolle.

(Subsubsub-)Katgeorien wurde verzichtet, da sich diese Information im Anschluss an die Codierung des Gesamtmaterials halb-automatisiert codieren ließ und daher auf eine (mutmaßlich fehleranfällige) Codierung „per Hand“ verzichtet werden konnte: Mittels einer softwarebasierten Wortsuche wurden in den Laut-Denk-Protokollen die Kommentare identifiziert, die angeben, dass ein_e Teilnehmer_in auf die Seite im Aufgabenheft, auf der Schülerlösungstext A, B, C oder D abgedruckt war, blätterte bzw. von dieser Seite weiterblätterte[133]. Anschließend wurden alle Segmente in den Laut-Denk-Protokollen, die zwischen zwei derartigen Kommentaren lagen, mit der zusätzlichen Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „Auseinandersetzung mit Schülerlösungstext B“, „Auseinandersetzung mit Schülerlösungstext C“ oder „Auseinandersetzung mit Schülerlösungstext D“ versehen. Um auf diese zusätzlichen Kategorien im weiteren Verlauf der

[133]Derartige Kommentare stellten eigene Segmente in den Laut-Denk-Protokolle dar und hatten stets den gleichen Wortlaut, nämlich „(blätter auf Seite [...])“. Zur Illustration siehe z. B. Segment 265, 272 und 280 im Transkriptauszug von Herrn Trummer in Tabelle 6.14 (vgl. Unterabschnitt 6.3.2.3).

vorliegenden Arbeit einfacher verweisen zu können, wurden diese mit der Codenummer „A“, „B“, „C“, „D“ versehen.

6.3.2.3. Phase 2a: Zusammenfassen der Laut-Denk-Protokolle und Identifizieren qualitativer Auffälligkeiten

6.3.2.3.1. Methodische Vorbemerkungen

Ziel von Phase 2a der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle war, die gesamten erhobenen Laut-Denk-Daten fallübergreifend und regelbasiert zusammenzufassen, sowie dabei qualitative Auffälligkeiten zu identifizieren und diese zu interpretieren. Hierzu wurde wie folgt vorgegangen:

Die vollständige Codierung mit der finalen Version des Kategoriensystems stellte die Grundlage für die Zusammenfassung der Laut-Denk-Protokolle dar. Bei dieser wurde – wie in Unterabschnitt 6.3.2.2 beschrieben – jedes Segment der Laut-Denk-Protokolle mit (mindestens) einer (Subsub-)Subkategorie codiert, die die wesentliche inhaltliche Bedeutung dieses Segments erfasst. Die Häufigkeiten, in denen die (Subsub-)Subkategorien im Laut-Denk-Protokoll eines_einer bestimmten Teilnehmers_Teilnehmerin codiert wurden, stellen damit also eine durch ein sukzessives und regelbasiertes Vorgehen gewonnene Zusammenfassung dieses Laut-Denk-Protokolls dar.

In Tabelle 6.8 sind daher die absoluten Codehäufigkeiten der einzelnen (Subsub-)Subkategorien für die Laut-Denk-Protokolle aller 21 Teilnehmer_innen zusammengetragen. Da sich allerdings die Laut-Denk-Protokolle der Teilnehmer_innen bezüglich ihrer Segmentgesamtanzahl zum Teil deutlich voneinander unterscheiden (Median = 497 Segmente; IQR = 350 Segmente; Spannweite: 198 bis 842 Segmente; vgl. Tabelle 6.4), war auf Grundlage der absoluten Codehäufigkeiten noch keine fallübergreifende Zusammenfassung aller Laut-Denk-Protokolle möglich. Um dies zu ermöglichen wurden die absoluten Codehäufigkeiten für jede_n der Teilnehmer_innen via Division durch die Segmentgesamtanzahl in ihrem Laut-Denk-Protokoll in prozentuelle Codehäufigkeiten umgewandelt (vgl. rechte Hälfte der Spalten in Tabelle 6.8). Anschließend wurden für jede (Subsub-)Subkategorie die prozentuellen Häufigkeiten aller 21 Teilnehmer_innen als Boxplot aufgetragen. Mit Hilfe dieser Boxplots wurden die Mediane der prozentuellen Häufigkeiten für die einzelnen (Subsub-)Subkategorien (unter Berücksichtigung der zugehörigen Streuungsparameter) qualitativ miteinander verglichen, um so globale Auffälligkeiten in der Codierung aller Laut-Denk-Protokolle zu identifizieren. Diese Auffälligkeiten wurden anschließend durch Rückgriff auf das codierte Datenmaterial und durch ergänzende qualitative Analysen interpretiert. Hierbei wurde darauf verzichtet, Auffälligkeiten in den prozentuellen Codehäufigkeiten mittels geeigneter statistischer Methoden zusätzlich auf Signifikanz zu überprüfen. Grund hierfür war zum einen, dass ein derartiges Vorgehen nicht für alle (Subsub-)Subkategorien aufgrund ihres seltenen Auftretens sinnvoll und möglich war (z. B. für Subkategorie 5.1; vgl. Tabelle 6.8) und zum anderen, dass ein derartiges Vorgehen

Lehrkraft (Abkürzung)

Codenr.	A n	A %	B n	B %	C n	C %	D n	D %	E n	E %	F n	F %	G n	G %	H n	H %	I n	I %	J n	J %
1.0.1	18	4.4	10	4.1	10	2.3	12	2.4	23	2.7	25	4.9	16	1.9	12	2.2	7	3.5	13	5.3
1.0.2	2	0.5	1	0.4	1	0.2	2	0.4	4	0.5	3	0.6	1	0.1	3	0.6	0	0.0	0	0.0
1.0.3	11	2.7	7	2.9	7	1.6	7	1.4	4	0.5	6	1.2	5	0.6	5	0.9	5	2.5	6	2.4
1.1.1	6	1.5	8	3.3	9	2.1	2	0.4	6	0.7	4	0.8	5	0.6	8	1.5	2	1.0	1	0.4
1.1.2	1	0.2	4	1.6	1	0.2	3	0.6	6	0.7	1	0.2	15	1.8	3	0.6	2	1.0	3	1.2
1.1.3	16	3.9	15	6.1	20	4.6	25	4.9	11	1.3	20	3.9	27	3.3	18	3.3	6	3.0	7	2.9
1.2.1	3	0.7	5	2.0	4	0.9	1	0.2	11	1.3	6	1.2	9	1.1	7	1.3	2	1.0	2	0.8
1.2.2	0	0.0	1	0.4	1	0.2	3	0.6	2	0.2	0	0.0	2	0.2	1	0.2	0	0.0	1	0.4
1.2.3	0	0.0	0	0.0	2	0.5	1	0.2	2	0.2	2	0.4	0	0.0	0	0.0	0	0.0	0	0.0
2.0	54	13.1	24	9.8	84	19.4	81	15.9	149	17.7	52	10.3	107	12.9	32	5.9	20	10.1	65	26.5
2.1	11	2.7	14	5.7	18	4.2	20	3.9	11	1.3	23	4.5	27	3.3	25	4.6	10	5.1	15	6.1
3.0.1.1	40	9.7	24	9.8	27	6.3	86	16.9	48	5.7	56	11.0	68	8.2	30	5.5	46	23.2	11	4.5
3.0.1.2	20	4.9	4	1.6	28	6.5	39	7.7	6	0.7	19	3.7	41	5.0	22	4.0	12	6.1	24	9.8
3.0.1.3	2	0.5	4	1.6	11	2.5	7	1.4	10	1.2	6	1.2	4	0.5	6	1.1	0	0.0	0	0.0
3.0.1.4	23	5.6	21	8.6	23	5.3	15	3.0	34	4.0	21	4.1	36	4.4	46	8.4	10	5.1	12	4.9
3.0.2.1	73	17.8	40	16.3	69	16.0	126	24.8	60	7.1	84	16.6	114	13.8	68	12.5	57	28.8	33	13.5
3.0.2.2	0	0.0	0	0.0	3	0.7	4	0.8	8	1.0	0	0.0	2	0.2	2	0.4	2	1.0	0	0.0
3.0.3.3	0	0.0	0	0.0	1	0.2	1	0.2	7	0.8	0	0.0	3	0.4	3	0.6	0	0.0	3	1.2
3.0.2.4	0	0.0	0	0.0	3	0.7	2	0.4	0	0.0	0	0.0	4	0.5	1	0.2	0	0.0	0	0.0
3.0.2.5	12	2.9	13	5.3	13	3.0	14	2.8	23	2.7	18	3.6	26	3.1	30	5.5	9	4.5	11	4.5
3.0.3.1	33	8.0	16	6.5	21	4.9	41	8.1	31	3.7	29	5.7	51	6.2	32	5.9	23	11.6	11	4.5
3.0.3.2	25	6.1	29	11.8	50	11.6	82	16.1	44	5.2	41	8.1	50	6.0	27	5.0	22	11.1	24	9.8
3.0.3.3	27	6.6	8	3.3	18	4.2	24	4.7	23	2.7	32	6.3	48	5.8	45	8.3	23	11.6	12	4.9
3.1.1	15	3.6	4	1.6	7	1.6	17	3.3	19	2.3	18	3.6	33	4.0	4	0.7	7	3.5	2	0.8
3.1.2	3	0.7	0	0.0	1	0.2	8	1.6	0	0.0	0	0.0	7	0.8	1	0.2	0	0.0	0	0.0
3.1.3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
3.1.4	0	0.0	0	0.0	0	0.0	0	0.0	1	0.1	0	0.0	5	0.6	3	0.6	0	0.0	1	0.4
4.0	19	4.6	1	0.4	11	2.5	7	1.4	33	3.9	6	1.2	43	5.2	57	10.5	12	6.1	0	0.0
4.1	8	1.9	6	2.4	4	0.9	6	1.2	28	3.3	11	2.2	91	11.0	13	2.4	1	0.5	6	2.4
4.2	30	7.3	25	10.2	35	8.1	20	3.9	79	9.4	28	5.5	52	6.3	27	5.0	8	4.0	34	13.9
4.3	13	3.2	7	2.9	11	2.5	10	2.0	46	5.5	13	2.6	23	2.8	13	2.4	0	0.0	1	0.4
5.0	10	2.4	13	5.3	32	7.4	26	5.1	81	9.6	27	5.3	51	6.2	48	8.8	9	4.5	12	4.9
5.1	0	0.0	0	0.0	3	0.7	0	0.0	18	2.1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
6.0	96	23.4	38	15.5	66	15.3	66	13.0	169	20.1	128	25.2	110	13.3	139	25.5	28	14.1	17	6.9
6.1	10	2.4	9	3.7	16	3.7	44	8.7	41	4.9	32	6.3	49	5.9	22	4.0	11	5.6	12	4.9

Codenr.	K n	K %	L n	L %	M n	M %	N n	N %	O n	O %	P n	P %	Q n	Q %	R n	R %	S n	S %	T n	T %	U n	U %
1.0.1	15	2.9	9	3.5	14	1.7	10	2.0	14	5.9	6	1.4	15	4.8	16	2.3	7	1.5	11	1.8	3	0.4
1.0.2	3	0.6	0	0.0	1	0.1	2	0.4	1	0.4	1	0.2	3	1.0	4	0.6	2	0.4	2	0.3	2	0.3
1.0.3	10	1.9	12	4.7	18	2.2	9	1.8	8	3.3	6	1.4	7	2.2	13	1.9	7	1.5	17	2.8	12	1.8
1.1.1	10	1.9	2	0.8	13	1.6	0	0.0	9	3.8	4	1.0	0	0.0	5	0.7	1	0.2	25	4.2	4	0.6
1.1.2	1	0.2	2	0.8	8	1.0	1	0.2	1	0.4	1	0.2	1	0.3	4	0.6	3	0.6	6	1.0	0	0.0
1.1.3	10	1.9	12	4.7	41	4.9	17	3.4	13	5.4	16	3.8	4	1.3	31	4.4	14	3.0	18	3.0	11	1.6
1.2.1	6	1.1	11	4.3	7	0.8	4	0.8	8	3.3	9	2.2	3	1.0	21	3.0	8	1.7	4	0.7	6	0.9
1.2.2	1	0.2	1	0.4	5	0.6	0	0.0	0	0.0	4	1.0	0	0.0	3	0.4	1	0.2	2	0.3	0	0.0
1.2.3	1	0.2	1	0.4	6	0.7	2	0.4	0	0.0	0	0.0	2	0.6	5	0.7	2	0.4	0	0.0	1	0.1
2.0	53	10.1	21	8.2	98	11.8	73	14.7	32	13.4	38	9.1	31	9.9	60	8.6	51	10.8	40	6.7	121	18.1
2.1	10	1.9	16	6.3	48	5.8	18	3.6	9	3.8	25	6.0	15	4.8	17	2.4	19	4.0	21	3.5	24	3.6
3.0.1.1	27	5.2	19	7.5	86	10.3	60	12.1	27	11.3	55	13.2	20	6.4	122	17.5	75	15.9	47	7.8	71	10.6
3.0.1.2	11	2.1	10	3.9	23	2.8	13	2.6	3	1.3	18	4.3	7	2.2	12	1.7	41	8.7	18	3.0	28	4.2
3.0.1.3	2	0.4	5	2.0	4	0.5	6	1.2	0	0.0	2	0.5	4	1.3	12	1.7	5	1.1	1	0.2	1	0.1
3.0.1.4	29	5.5	20	7.8	63	7.6	23	4.6	11	4.6	23	5.5	21	6.7	44	6.3	57	12.1	54	9.0	44	6.6
3.0.2.1	27	5.2	31	12.2	118	14.2	77	15.5	31	13.0	73	17.5	35	11.2	157	22.5	109	23.1	74	12.4	108	16.1
3.0.2.2	2	0.4	0	0.0	6	0.7	0	0.0	1	0.4	0	0.0	0	0.0	6	0.9	12	2.5	2	0.3	3	0.4
3.0.3.3	4	0.8	1	0.4	1	0.1	3	0.6	0	0.0	2	0.5	0	0.0	0	0.0	3	0.6	0	0.0	0	0.0
3.0.2.4	13	2.5	4	1.6	0	0.0	1	0.2	0	0.0	4	1.0	2	0.6	2	0.3	26	5.5	7	1.2	0	0.0
3.0.2.5	23	4.4	18	7.1	51	6.1	21	4.2	9	3.8	19	4.5	15	4.8	25	3.6	28	5.9	37	6.2	33	4.9
3.0.3.1	22	4.2	21	8.2	59	7.1	31	6.2	23	9.6	21	5.0	18	5.8	43	6.2	44	9.3	30	5.0	56	8.4
3.0.3.2	22	4.2	26	10.2	56	6.7	50	10.1	9	3.8	39	9.3	17	5.4	81	11.6	49	10.4	44	7.3	42	6.3
3.0.3.3	25	4.8	7	2.7	61	7.3	21	4.2	9	3.8	38	9.1	17	5.4	66	9.5	85	18.0	46	7.7	46	6.9
3.1.1	25	4.8	3	1.2	33	4.0	10	2.0	1	0.4	4	1.0	3	1.0	12	1.7	3	0.6	23	3.8	10	1.5
3.1.2	1	0.2	0	0.0	3	0.4	0	0.0	0	0.0	1	0.2	0	0.0	3	0.4	1	0.2	0	0.0	5	0.7
3.1.3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	2	0.3	0	0.0	0	0.0	0	0.0
3.1.4	0	0.0	0	0.0	1	0.1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	4	0.8	0	0.0	0	0.0
4.0	39	7.5	22	8.6	13	1.6	24	4.8	1	0.4	26	6.2	12	3.8	21	3.0	11	2.3	34	5.7	48	7.2
4.1	14	2.7	0	0.0	7	0.8	10	2.0	5	2.1	3	0.7	7	2.2	25	3.6	12	2.5	14	2.3	1	0.1
4.2	58	11.1	20	7.8	41	4.9	22	4.4	13	5.4	21	5.0	28	8.9	30	4.3	10	2.1	50	8.3	10	1.5
4.3	44	8.4	5	2.0	32	3.8	12	2.4	9	3.8	11	2.6	7	2.2	23	3.3	23	4.9	14	2.3	27	4.0
5.0	70	13.4	22	8.6	47	5.6	60	12.1	24	10.0	53	12.7	13	4.2	62	8.9	32	6.8	96	16.0	89	13.3
5.1	0	0.0	1	0.4	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
6.0	55	10.5	33	12.9	195	23.4	85	17.1	30	12.6	75	17.9	94	30.0	122	17.5	71	15.0	68	11.4	112	16.7
6.1	28	5.4	8	3.1	25	3.0	36	7.2	20	8.4	16	3.8	16	5.1	28	4.0	12	2.5	34	5.7	39	5.8

Tabelle 6.8.: Absolute (n) und prozentuelle Häufigkeit (%), in der die Segmente der Laut-Denk-Protokolle der Teilnehmer_innen mit einer bestimmten (Subsub-)Subkategorie (als Codenummer angegeben) codiert wurden.

in Phase 2c der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle für eine Auswahl besonders reichhaltiger (Subsub-)Subkategorien erfolgte (vgl. Unterabschnitt 6.3.2.1).

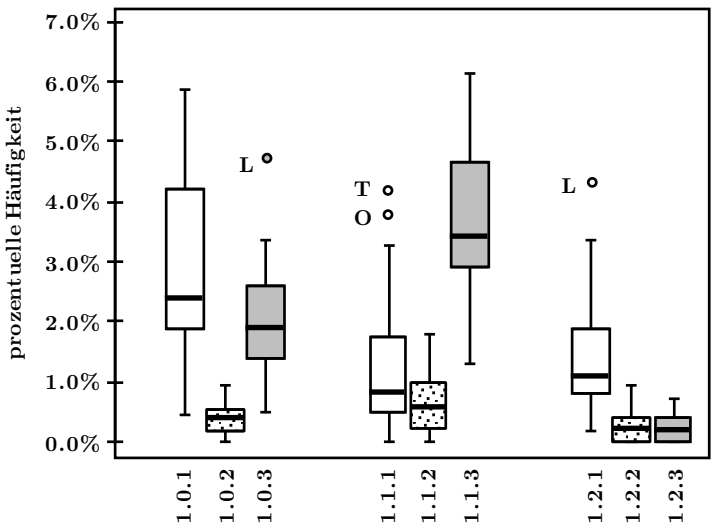
Die Ergebnisse, die durch das eben beschriebene Vorgehen gewonnen wurden, adressieren durchgängig Forschungsfrage (F1). Diese Ergebnisse, sowie ihre Limitationen, werden im Folgenden dargestellt. Zudem erfolgt an dieser Stelle der vorliegenden Arbeit eine genauere Beschreibung der einzelnen (Subsub-)Subkategorien der finalen Version des Kategoriensystems.

6.3.2.3.2. Ergebnisse von Phase 2a der inhaltsanalytischen Auswertung der Laut–Denk-Protokolle

Kategorie 1: lesen/erfassen eines Textes

In der Kategorie „lesen/erfassen eines Textes“ wird in drei Subkategorien unterschieden, ob in einem Segment der Laut-Denk-Protokolle ein Text zusammenhängend oder fragmentarisch (z. B. beim Überfliegen oder Scannen eines Textes zur Suche nach Belegen) gelesen wird oder ob er paraphrasiert wird. Jede dieser drei Subkategorien ist wiederum aus drei Subsubkategorien aufgebaut, die spezifizieren ob es sich bei dem gelesenen/erfassten Text um eine Anweisung im Aufgabenheft (Subsubkategorie 1.0.1, 1.1.1 und 1.2.1), die Aufgabe Weltraumspaziergang (Subsubkategorie 1.0.2, 1.1.2 und 1.2.2) oder einen der vier Schülerlösungstexte handelt (Subsubkategorie 1.0.3, 1.1.3 und 1.2.3).

Beim Vergleich der prozentuellen Häufigkeiten (vgl. Abbildung 6.10) fällt auf, dass Anweisungen im Aufgabenheft von den Teilnehmer_innen im Median am häufigsten zusammenhängend gelesen wurden (Median = 2.4 %), wohingegen die vier Schülerlösungstexte am



Codenr.	Median	IQR
1.0.1	2.4 %	2.3 %
1.0.2	.4 %	.2 %
1.0.3	1.9 %	1.2 %
1.1.1	.8 %	1.2 %
1.1.2	.6 %	.8 %
1.1.3	3.4 %	1.8 %
1.2.1	1.1 %	1.1 %
1.2.2	.2 %	.4 %
1.2.3	.2 %	.4 %

Abbildung 6.10.: Boxplots für die (Sub-)Subkategorien der Kategorie „lesen/erfassen eines Textes“. Die Boxplots Subsubkategorien „eine Anweisung im Textheft“, „die Aufgabe Weltraumspaziergang“ und „einen Schülerlösungstext“ sind weiß, gepunktet und grau hervorgehoben.

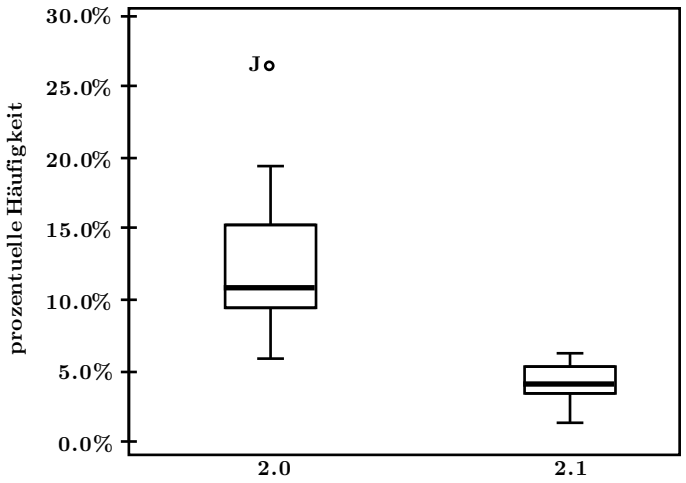
Feintranskript mit Segmentierung		Codierung (Codenr.)
	[...]	
[358]	*Die Helme bestehen aus Glas. Wenn man innerhalb des Helmes spricht (..) kann sich die Stimme ausbreiten, da es Luft gibt.*	1.0.3
[359]	*...die Stimme...*	1.1.3
[360]	Auch wieder sprachlich nicht perfekt.	3.0.1.2 3.0.2.1 3.0.3.2
[361]	Da müsste man sagen...	3.0.1.2 3.0.2.1 3.0.3.2
[362]	Also das unterstreich- ich mal so'n bisschen (.) vorsichtich.	6.0
[363]	(unterstreicht „die Stimme" in Zeile 8 mit einer gestrichelten Linie)	3.0.1.2 3.0.2.1 3.0.3.2
[364]	*...kann sich (..) der Schall...*	1.1.3
[365]	(schreibt) Besser: (.) Der Schall (+)	3.0.1.2 3.0.2.1 3.0.3.2
[366]	das wäre in Zeile 8.	6.0
[367]	*...ausbreiten.*	1.1.3
[368]	Is- für mich jetz- aber nich- wirklich (.) -n Punktabzuch.	3.0.1.2 3.0.2.1 3.0.3.3
[369]	Sondern das geht für mich trotzdem in Ordnung.	3.0.1.2 3.0.2.1 3.0.3.1
[370]	*...da es Luft gibt.*	1.1.3
[371]	Okay.	5.0
	[...]	

Tabelle 6.9.: Auszug aus dem Laut-Denk-Protokoll von Herrn Trummer während der Korrektur von Schülerlösungstext C.

häufigsten fragmentarisch gelesen wurden (Median = 3.4 %). Es lässt sich daher vermuten, dass die Teilnehmer_innen beim lauten Denken unterschiedliche Lesestrategien angewendet haben, je nachdem ob sie im Aufgabenheft eine Anweisung erfassen wollten oder (wie in der Laborsituation gefordert) ihren eigenen Gewohnheiten entsprechend die Schülerlösungstexte A bis D korrigierten. Die codierten Segmente in den Laut-Denk-Protokollen untermauerte diese Vermutung. Anweisungen, die an die Physiklehrkräfte gerichtet waren, wurden z. B. zum Zweck der Sinnerfassung oder eines Sich-Ins-Gedächtnisrufens vorwiegend zusammenhängend gelesen. Demgegenüber wurden die vier Schülerlösungstexte vor allem durch fragmentarisches Lesen nach Belegen oder Informationen durchsucht und dabei beurteilt. Aus letzterem lässt sich folgern, dass fragmentarisches Lesen für die Teilnehmer_innen eine Ressource bei Leistungsfeststellung und Beurteilung im Kontext der Laborsituation der Hauptstudie darstellte (Forschungsfrage (F1)). Besonders deutlich wird dies im Auszug aus dem Laut-Denk-Protokoll von Herrn Trummer, der in Tabelle 6.9 dargestellt ist (ähnliches zeigt sich zudem im Auszug aus dem Laut-Denk-Protokoll von Frau Sohm in Tabelle 6.7 ab Segment 147). In diesem wird ein Teil des Schülerlösungstextes C zunächst zusammenhängend gelesen. Anschließend wird der zuvor zusammenhängend gelesene Textauszug von Herrn Trummer fragmentarisch gelesen, um so einzelne Merkmale feststellen und beurteilen zu können.

Kategorie 2: erstellen des Erwartungshorizonts zur Aufgabe Weltraumspaziergang

Die Kategorie „erstellen des Erwartungshorizonts zur Aufgabe Weltraumspaziergang" betrifft die laut-denkende Bearbeitung der ersten Aufgabe im Aufgabenheft (vgl. Anhang C.1). Sie besteht aus den zwei Subkategorien „Beurteilungskriterien werden benannt/kommentiert/abgewogen/festgelegt" (Subkategorie 2.0) und „Zuweisung von Punkten im Erwartungshorizont" (Subkategorie 2.1).



Codenr.	Median	IQR
2.1	10.8 %	5.9 %
2.2	4.0 %	3.0 %

Abbildung 6.11.: Boxplots für die Subkategorien der Kategorie „erstellen des Erwartungshorizonts zur Aufgabe Weltraumspaziergang“.

Aus Abbildung 6.11 geht hervor, dass die prozentuellen Häufigkeiten beider Subkategorien im Median und in grober Näherung in einem drei zu eins Verhältnis stehen. Dieses Verhältnis deutet darauf hin, dass für die Teilnehmer_innen bei der Erstellung des Erwartungshorizonts die Zuweisung von Punkten im Vergleich zum Benennen, Kommentieren, Abwiegen und Festlegen von Beurteilungskriterien keine nebensächliche Tätigkeit darstellte. Viel mehr scheint das Zuweisen von Punkten in einer längerfristigen gedanklichen (und daher beim lauten Denken umfänglich mitvokalisierten) Auseinandersetzung mit dem Aufgabenheft stattgefunden zu haben. Dies ist ein möglicher Hinweis, dass die Punktezuweisung durch die Teilnehmer_innen bis zu einem bestimmten Grad im Sinne der in Abschnitt 2.2.4 dargestellten Konzeption der Assessment Literacy von Lehrkräften (als bewusstes Kompromissfinden im Rahmen einer Handlungsepisode) stattfand. Aufgrund des grobkörnigen Charakters von Subkategorie 2.1 können an dieser Stelle der vorliegenden Arbeit aber keine Aussagen darüber getroffen werden, inwieweit die Teilnehmer_innen bei der Zuweisung von Punkten sowohl ihre selbst generierten Beurteilungskriterien zur Aufgabe Weltraumspaziergang, als auch die der Kontextualisierung der Laborsituation berücksichtigten (z. B. dass der Erwartungshorizont für eine Klassenarbeit in der neunten Jahrgangsstufe anzufertigen war; vgl. Unterkapitel 5.4.1).

Ferner blieben aufgrund des ebenfalls grobkörnigen Charakters von Subkategorie 2.0 die von den Teilnehmer_innen bei der laut-denkenden Erwartungshorizonterstellung benannten, kommentierten, abgewogenen und festgelegten Beurteilungskriterien zur Aufgabe Weltraumspaziergang unklar. Aus diesem Grund wurden im Rahmen einer Qualifikationsarbeit (vgl. Hackemann, 2017) die Laut-Denk-Protokolle und verschriftlichten Erwartungshorizonte aller Teilnehmer_innen zusätzlich einer typisierend-strukturierenden qualitativen Inhaltsanalyse unterzogen (vgl. Mayring, 2015, S. 103 u. f.). Durch dieses Vorgehen konnten bei den Teilnehmer_innen vier unterschiedliche Arten des Umgangs mit sprachlichen Merkmalen eines Schülerlösungstextes bei der Erstellung des Erwar-

Art des Umgangs mit sprachlichen Merkmalen eines Schülerlösungstextes bei der Erstellung des Erwartungshorizonts zur Aufgabe Weltraumspaziergang	**Lehrkräfte, bei denen sich diese Art des Umgangs identifizieren ließ**	
	Abkürzung	**Anzahl**
Sprachliche Merkmale eines Schülerlösungstextes wurden laut denkend als Beurteilungskriterium mitvokalisiert und schriftlich im Erwartungshorizont fixiert. Im Erwartungshorizont wurde das Auftreten dieser sprachlichen Merkmale mit einer Punktevergabe verknüpft.	G, H, I, J, K	5
Sprachliche Merkmale eines Schülerlösungstextes wurden laut denkend als Beurteilungskriterium mitvokalisiert und/oder schriftlich im Erwartungshorizont fixiert. Im Erwartungshorizont wurde das Nichtauftreten dieser sprachlichen Merkmale mit einem Punkteabzug verknüpft.	A, L, M, O, P, Q, S, U	8
Sprachliche Merkmale eines Schülerlösungstextes wurden nur laut denkend als Beurteilungskriterium mitvokalisiert, im Erwartungshorizont aber nicht schriftlich fixiert. Es bleibt unklar, inwieweit diese sprachlichen Merkmale mit einem Punkteabzug oder einer Punktevergabe verknüpft wurden.	E, F, R, T	4
Sprachliche Merkmale eines Schülerlösungstextes wurden weder laut denkend als Beurteilungskriterium mitvokalisiert noch schriftlich im Erwartungshorizont fixiert. Es bleibt unklar, inwieweit sprachliche Merkmale eines Schülerlösungstextes mit einem Punkteabzug oder einer Punktevergabe verknüpft wurden.	B, C, D, N	4

Tabelle 6.10.: Arten des Umgangs der Teilnehmer_innen mit sprachlichen Merkmalen eines Schülerlösungstextes bei der Erstellung des Erwartungshorizonts (vgl. Hackemann, 2017, S. 58 u. f.).

tungshorizonts zur Aufgabe Weltraumspaziergang identifiziert werden, die als Ressourcen zur Leistungsfeststellung und -beurteilung im Sinne von Forschungsfrage (F1) verstanden werden können. Deren Kurzcharakteristika sind in Tabelle 6.10 dargestellt[134]. Zusätzlich ist in Tabelle 6.10 die Verteilung der 21 Teilnehmer_innen auf diese Arten des Umgangs mit sprachlichen Merkmalen eines Schülerlösungstextes angegeben. Aus dieser Verteilung lässt sich im Sinne der in Abschnitt 2.2.4 dargestellten Konzeption der Assessment Literacy von Lehrkräften vermuten,...

... dass sich mehr als ein Drittel der Teilnehmer_innen (die 8 Lehrkräfte, bei denen sich die dritte und vierte Art des Umgangs zeigte) bei der Erstellung eines Erwartungshorizonts zur Aufgabe Weltraumspaziergang entweder kaum für die Feststellung und Beurteilung sprachlicher Merkmale von Schülerlösungstexten zuständig fühlte und/oder nur wenig über diesbezügliches Wissen und Können verfügte, das beim lauten Denken mitvokalisierbar war. Sie benannten Beurteilungskriterien, die sprachliche Merkmale von Schülerlösungstexten betreffen, nicht oder kaum und es blieb unklar, inwieweit diese Merkmale mit einem Punkteabzug oder einer Punktevergabe verknüpft wurden.

... dass mehr als ein Drittel der Teilnehmer_innen (die 8 Lehrkräfte, bei denen sich die zweite Art des Umgangs zeigte) sprachliche Merkmale von Schülerlösungstexten bei der Erstellung eines Erwartungshorizonts zur Aufgabe Weltraumspaziergang in einer auffällig defizitorientierten Art und Weise berücksichtigte. Sie verknüpften das Nichtauftreten sprachlicher Merkmale in einem Schülerlösungstext mit einem Punkteabzug, nicht aber das Auftreten sprachlicher Merkmale mit einer Punktevergabe.

[134]Eine ausführliche Darstellung findet sich bei Hackemann (2017, S. 58 u. f.).

... dass weniger als ein Drittel der Teilnehmer_innen (die 5 Lehrkräfte, bei denen sich die erste Art des Umgangs zeigte) sprachliche Merkmale von Schülerlösungstexten bei der Erstellung eines Erwartungshorizonts zur Aufgabe Weltraumspaziergang in einer auch fähigkeitsorientierten Art und Weise berücksichtigt, da sie das Auftreten sprachlicher Merkmale in einem Schülerlösungstext mit einer Punktevergabe verknüpften.

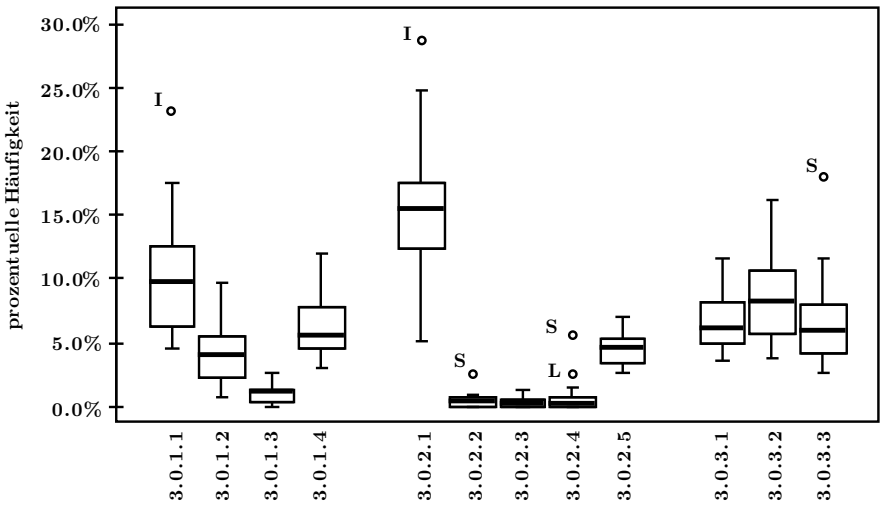
Kategorie 3: Korrektur der Schülerlösungstexte

Die Kategorie „Korrektur der Schülerlösungstexte“ ist in zwei Subkategorien untergliedert: Die Subkategorie „Feststellung und Beurteilung eines Schülerlösungstextes“ (Subkategorie 3.0) und die Subkategorie „Beurteilungskriterien ad hoc benennen/abwägen oder aus dem Erwartungshorizont entnehmen“ (Subkategorie 3.1).

Die Subkategorie 3.0 umfasst das laut-denkende Aushandeln und Kommentieren der Teilnehmer_innen bei der Feststellung und Beurteilung der vier Schülerlösungstexte. Sie ist die einzige Subkategorie im gesamten Kategoriensystem, bei der eine Dreifachcodierung erfolgt (vgl. Unterabschnitt 6.3.2.2). Sie ist aus den folgenden drei Subsubkategorien aufgebaut, die jeweils eine andersartige inhaltliche Bedeutung eines zu codierenden Segments erfassen. Bei der Codierung ist einem Segment aus jeder der drei Subsubkategorien jeweils genau eine Subsubsubkategorie zugewiesen worden:

- Die Subsubkategorie 3.0.1 erfasst, welches Merkmal eines Schülerlösungstextes im zu codierenden Segment fokussiert wird. Dabei wird zwischen den Merkmalsfoki „fachlich-konzeptueller Eindruck“ (Subsubsubkategorie 3.0.1.1), „sprachliche Realisierung“ (Subsubsubkategorie 3.0.1.2), „sonstiges Merkmal“ (Textlänge, Rechtschreibung, Zeichensetzung, Handschrift, Gliederung, usw.) (Subsubsubkategorie 3.0.1.3) und „mehrere/uneindeutig“ unterschieden (Subsubsubkategorie 3.0.1.4).
- Die Subsubkategorie 3.0.2 erfasst den Bezug der Verortung eines Schülerlösungstextes im zu codierenden Segment. Hier wird unterschieden, ob ein Schülerlösungstext in Bezug zu einem sachlichen Kriterium (Subsubsubkategorie 3.0.2.1), einem anderen Schülerlösungstext (Subsubsubkategorie 3.0.2.2), allgemeinen Erfahrungen mit Physiklernenden (Subsubsubkategorie 3.0.2.3) oder mutmaßlichen Personenmerkmalen des_der Schülers_Schülerin (Geschlecht, Herkunft, Alter, intellektuelle Reife, usw.) (Subsubsubkategorie 3.0.2.4) gesetzt wird oder ob der Bezug mehr- bzw. uneindeutig ist (Subsubsubkategorie 3.0.2.5).
- Die Subsubkategorie 3.0.1 erfasst, inwieweit es sich bei dem zu codierenden Segment um eine wertende Äußerung handelt. Dementsprechend wird hier zwischen positiv wertenden/akzeptierenden (Subsubsubkategorie 3.0.3.1), negativ wertenden/ablehnenden (Subsubsubkategorie 3.0.3.2) und neutralen/gemischten/sonstigen Äußerungen (Subsubsubkategorie 3.0.3.3) unterschieden.

Subkategorie 3.1 ist analog zur Subsubkategorie 3.0.1 aufgebaut und erfasst die von den Teilnehmer_innen bei der laut-denkenden Feststellung und Beurteilung eines Schülerlösungstextes mitvokalisierten sachlichen Beurteilungskriterien (ad hoc benannt/abgewogen



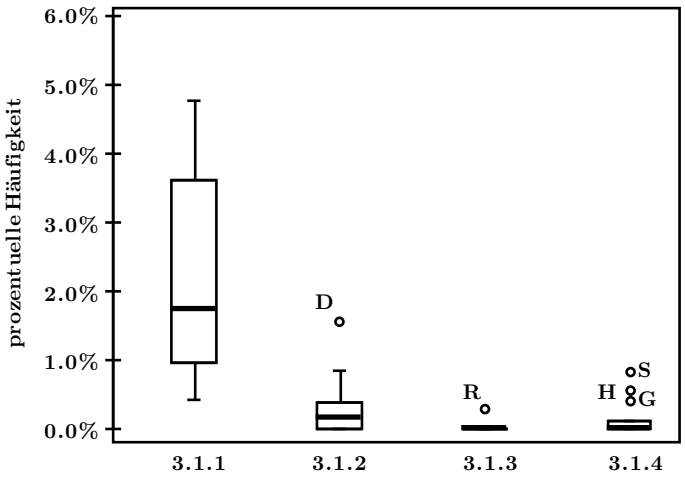
Codenr.	Median	IQR
3.0.1.1	9.8 %	6.3 %
3.0.1.2	3.9 %	3.3 %
3.0.1.3	1.1 %	1.0 %
3.0.1.4	5.5 %	3.1 %
3.0.2.1	15.4 %	5.2 %
3.0.2.2	.4 %	.8 %
3.0.2.3	.2 %	.6 %
3.0.2.4	.2 %	.8 %
3.0.2.5	4.5 %	2.0 %
3.0.3.1	6.2 %	3.2 %
3.0.3.2	8.1 %	5.1 %
3.0.3.3	5.8 %	3.8 %

Abbildung 6.12.: Boxplots für die Subsubsubkategorien der Subkategorie „Feststellung und Beurteilung eines Schülerlösungstextes“.

oder aus dem eigenen Erwartungshorizont entnommen). Diese Beurteilungskriterien werden dabei den Subsubkategorien „fachlich-konzeptueller Eindruck“ (Subsubkategorie 3.1.1), „sprachliche Realisierung“ (Subsubkategorie 3.1.2), „sonstiges Merkmal“ (Textlänge, Rechtschreibung, Zeichensetzung, Handschrift, Gliederung, usw.) (Subsubkategorie 3.1.3) und „mehrere/uneindeutig“ (Subsubsubkategorie 3.1.4) zugewiesen.

Die prozentuellen Häufigkeiten der (Sub-)Subsubkategorien der Kategorie „Korrektur der Schülerlösungstexte“ sind in Abbildung 6.12 und 6.13 dargestellt. Bei diesen zeigen sich die folgenden Auffälligkeiten:

1. Bei der Korrektur der vier Schülerlösungstexte wurde von den Teilnehmer_innen überwiegend deren fachlich-konzeptueller Eindruck fokussiert (Median Subsubsubkategorie 3.0.1.1 = 9.8 %) oder dementsprechende Beurteilungskriterien mitvokalisiert (Median Subsubkategorie 3.1.1 = 1.7 %). Im Vergleich hierzu spielte die sprachliche Realisierung der vier Schülerlösungstexte bei der Korrektur der Teilnehmer_innen eine deutlich geringere Rolle (Median Subsubsubkategorie 3.0.1.2 = 3.9 %; Median Subsubkategorie 3.1.2 = .2 %). Ferner wurden sonstige Merkmale der Schülerlösungstexte von den Teilnehmer_innen bei der Korrektur noch einmal deutlich weniger mit einbezogen (Median Subsubsubkategorie 3.0.1.3 = 1.1 %; Median Subsubkategorie 3.1.3 = .0 %). Hierbei ist allerdings zu beachten, dass die Schülerlösungstexte im Aufgabenheft in einer einheitlichen Handschrift geschrieben und von Rechtschreibfehlern bereinigt waren (vgl. Abschnitt 5.4.1). Da die Teilnehmer_innen hierüber informiert waren, sie aber dennoch sonstige Merkmale wie Handschrift oder Rechtschreibung bei der Korrektur (wenn auch nur wenig) beachteten, lässt sich vermuten, dass derartigen Schülerlösungstextmerkmalen im Berufsalltag der Teilnehmer_innen eher eine bedeutende als eine geringfügige Rolle im Sinne einer Ressource zur Leistungsurteilsgenese zukommt.



Codenr.	Median	IQR
3.1.1	1.7 %	2.6 %
3.1.2	.2 %	.4 %
3.1.3	.0 %	.0 %
3.1.4	.0 %	.1 %

Abbildung 6.13.: Boxplots für die Subsubkategorien der Subkategorie „Beurteilungskriterien ad hoc benennen/abwägen oder aus dem Erwartungshorizont entnehmen“.

2. Die vier Schülerlösungstexte wurden von den Teilnehmer_innen bei der Korrektur überwiegend bezüglich eines sachlichen Kriteriums verortet (Median Subsubsubkategorie 3.0.2.1 = 15.4 %). Diese Auffälligkeit ist vereinbar damit, dass aus den Lehrkräftefragebögen hervorging, dass die Teilnehmer_innen ihrer Selbsteinschätzung nach in schwacher Tendenz eine kriteriale Bezugsnormorientierung aufweisen (vgl. Abschnitt 6.1.2). Andere Schülerlösungstexte zur Aufgabe Weltraumspaziergang, allgemeine Erfahrungen mit Physiklernenden oder Personenmerkmale des_der Schülers_Schülerin (Subsubsubkategorie 3.0.2.2 bis 3.0.2.4) spielten als Bezug für die Verortung der Schülerlösungstexte eine vergleichsweise geringe Rolle. Beim Auftreten von Personenmerkmalen als Bezug für die Verortung handelt es sich aber dennoch um eine besondere Auffälligkeit, da den Teilnehmer_innen im Rahmen der Laborsituation keine Informationen über Personenmerkmale der Schüler_innen, die die Schülerlösungstexte A bis D verfasst haben, zur Verfügung gestellt wurden. Bei den Personenmerkmalen, die die Teilnehmer_innen als Ressource zur Leistungsurteilsgenese herangezogen haben, muss es sich daher um „mutmaßliche Personenmerkmale“ handeln, auf die die Teilnehmer_innen während der Korrektur der Schülerlösungstexte geschlossen haben (woher auch die Bezeichnung dieser Subsubsubkategorie rührt). In den Laut-Denk-Protokollen, in denen die Subsubkategorie 3.0.2.3 codiert wurde, zeigte sich, dass die Teilnehmer_innen hierzu auf Erwartungen zurückgegriffen haben, die eine Verbindung von Merkmalen der Textprodukte von Schüler_innen mit bestimmten Personenmerkmalen betreffen. Tabelle 6.11 veranschaulicht dies exemplarisch. In dieser sind zwei Auszüge aus Laut-Denk-Protokollen dargestellt, in denen ein_e Teilnehmer_in bei der Korrektur eines Schülerlösungstextes seine_ihre eigenen Erwartungen bezogen auf die Unterschiede von Textprodukten, die von einer Schülerin oder einem Schüler im Physikunterricht verfasst wurden, besonders ausführlich mitvokalisiert.

Herr Geppert (Korrektur von Schülerlösungstext B; Segment 592-595)	**Frau Kirik** (Korrektur von Schülerlösungstext A; Segment 443-450)
[...] Ich denk-... fang- jetz- immer an zu sagen Schülerin. Weil ich i-wie das Gefühl hab-, dass is- -n Schülerin. Ähm weil ich äh das Gefühl hab- das... ich hab- das Gefühl das Jungs eher sich an den... an die Standarderklärung halten und sagen was will der Lehrer hörn. Und die... das hab- ich vielleicht so für mich im Gefühl... und Mädchen häufiger dann eher dann noch die kreativere Lösung finden. Oder ähm (...) jetz- ganz ohne werten zu wollen. [...]	[...] (..) Hier würde ich mir immer denken, dass der vielleicht auch ein bisschen schreibfaul ist. Der könnte das noch... denn d-... Eigentlich hat er schon erfasst, aber er beschreibt es einfach nicht genau (.) genug. (.) Das ist ja grade bei Jungs ein (..) eine beliebte Strategie das möglichst äh kurz zu machen und dann stimmt zwar der Kern, aber es fehlt so'n bisschen die Hälfte. Und Mädchen haben eher so die Tendenz sehr ausufernd zu schreiben und teilweise der Inhalt gleich 0. Oder völlig falsch. Oder (.) was ganz anderes. Ähm (.) also wenn ich nett bin geb- ich hier 4 von 5 Punkten. [...]

Tabelle 6.11.: Auszug aus dem Laut-Denk-Protokoll von Herrn Geppert und Frau Kirik, in denen bei der Korrektur eines Schülerlösungstextes (vor allem) Erwartungen an die Merkmale der Textprodukte einer Schülerin oder eines Schülers mitvokalisiert werden (Subsubsubkategorie 3.0.2.3 und 3.0.3.3).

3. In den Laut-Denk-Protokollen der Teilnehmer_innen wurde ein nicht zu vernachlässigender Anteil der Segmente den (Sub-)Subsubkategorien 3.0.1.4, 3.0.2.5 und 3.1.4 zugewiesen, die eine Mehr- und/oder Uneindeutigkeit ausdrücken. Bei einer Durchsicht dieser Segmente offenbarte sich vor allem, dass die Teilnehmer_innen die Leistungen in den Schülerlösungstexten A bis D auch auf Grundlage eines (ersten) eher holistischen Eindrucks und/oder heuristisch feststellten und beurteilten, wie der Auszug aus dem Laut-Denk-Protokoll von Herrn Balke in Tabelle 6.12 illustriert. In diesem wird der Schülerlösungstext C zum ersten Mal gelesen und auf Basis eines ersten positiven Gesamteineindrucks die volle Punktzahl vergeben. Ähnliches zeigt sich auch im Auszug aus dem Laut-Denk-Protokoll von Herrn Trummer, der in Tabelle 6.14 dargestellt ist (siehe unten). Hier werden die Schülerlösungstexte A und B erstmalig gelesen und bereits Beurteilungen aufgrund eines ersten holistischen Eindrucks vorgenommen.

4. Der prozentuelle Anteil von Segmenten, in denen sich die Teilnehmer_innen bei der Korrektur der vier Schülerlösungstexte positiv wertend/akzeptierend oder neutral/gemischt/sonstig äußern, ist im Median in etwa gleich groß (Median Subsubsubkategorie 3.0.3.1 = 6.2 %; Median Subsubsubkategorie 3.0.3.3 = 5.8 %). Die mediane prozentuelle Häufigkeit von Segmenten, in denen sich die Teilnehmer_innen negativ wertend/ablehnend äußern, ist demgegenüber erhöht (Subsubsubkategorie 3.0.3.2 = 8.1 %). Hier zeigt sich also ein Hinweis, dass die Teilnehmer_innen die Leistungen in den vier Schülerlösungstexten in einer tendenziell defizitorientierten Art und Weise feststellten und beurteilten.

Feintranskript mit Segmentierung		Codierung (Codenr.)		
	[...]			
[186]	Dann.	6.1		
[187]	(...) Seite äh 7 jetzt.	6.0		
[188]	*Antwort C*	1.1.2		
[189]	Äh *1. Nicht hörbar...*	1.1.3		
[190]	(.) ähm *noch hörbar,*	1.1.3		
[191]	oder *wieder hörbar*	1.1.3		
[192]	Das is- hier hö-... äh klar unterteil, die beiden Phänomene.	3.0.1.3	3.0.2.1	3.0.3.1
[193]	Das gefällt mir schonmal gut.	3.0.1.3	3.0.2.1	3.0.3.1
[194]	*Im Weltall herrscht ein Vakuum, also können sich die Schwingungen nicht fortbewegen. Sie werden nämlich durch Luft geleitet.*	1.0.3		
[195]	(hustet) Okay.	5.0		
[196]	*Noch hörbar: Die Helme bestehen aus Glas. Wenn man innerhalb des Helmes spricht kann sich die Stimme ausbreiten, da es Luft gibt. Hält man zwei Helme aneinander, werden die Schallwellen durch die Schwingung des Glases weitergegeben.*	1.0.3		
[197]	Ja, is- ja -ne super Antwort.	3.0.1.4	3.0.2.5	3.0.3.1
[198]	Alle beide.	3.0.1.4	3.0.2.5	3.0.3.1
[199]	(schreibt 5/5 unter den Schülerlösungstext)	3.0.1.4	3.0.2.5	3.0.3.1
[200]	So stell- ich mir das vor.	3.0.1.4	3.0.2.5	3.0.3.1
[201]	5 von 5 Punkten druntergeschrieben.	6.0		
[202]	Brauch- ich auch garnich- differenziert noch was ranzuschreiben, wie der Teil is- richtig oder dieser is- richtig.	3.0.1.4	3.0.2.1	3.0.3.3
[203]	(.) Sondern das is- sozusagen -ne (.) -ne Musterlösung.	3.0.1.4	3.0.2.1	3.0.3.1
	[...]			

Tabelle 6.12.: Auszug aus dem Laut-Denk-Protokoll von Herrn Balke bei der Korrektur von Schülerlösungstext C.

Kategorie 4: Äußerungen außerhalb der eigentlichen Korrektur und Erwartungshorizonterstellung

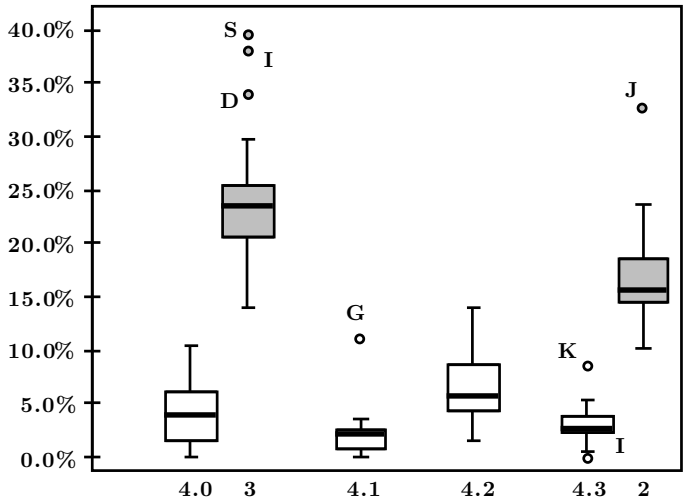
In der Kategorie „Äußerungen außerhalb der eigentlichen Korrektur und Erwartungshorizonterstellung“ werden vier Subkategorien unterschieden. Den Subkategorien 4.1 und 4.2 werden Segmente der Laut-Denk-Protokolle zugewiesen, bei denen es sich um Kommentare zur, Kritik an der, Fragen zur oder Interpretationen der Aufgabe Weltraumspaziergang bzw. von Anweisungen oder zum Aufbau des Aufgabenheftes handelt. Demgegenüber werden den Subkategorien 4.0 und 4.3 Segmente der Laut-Denk-Protokolle zugewiesen, bei denen es sich um Kommentare, Kritik oder selbstreflektierte Äußerungen handelt, in denen die Teilnehmer_innen allgemeine Handlungsstrategien für das Feststellen und Beurteilen von Schülerleistungen bzw. zur Erstellung eines Erwartungshorizonts beschrieben haben oder in denen sie sich zu ihrem allgemeinen Vorgehen/zu ihren allgemeinen Erfahrungen diesbezüglich äußerten. Die prozentuellen Häufigkeiten von Subkategorie 4.0 und 4.3 sind dabei für Forschungsfrage (F1) von besonderer Relevanz, da in Segmenten, die diesen Subkategorien zugewiesen wurden, die Teilnehmer_innen Facetten ihres bis zu einem bestimmten Grad generalisierten Wissens und Könnens, sowie ihrer berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung mitvokalisierten, wie die Auszüge aus den Laut-Denk-Protokollen von Herrn Hastedt und Frau Pinna in Tabelle 6.13 exemplarisch verdeutlichen.

Herr Hastedt (Zu Beginn der Korrektur von Schülerlösungstext B; Segment 412-419)	**Frau Pinna** (Zu Beginn der Erwartungshorizonterstellung; Segment 21-29)
[...] Der Schüler denkt physikalisch. Ich gebe grundsätzlich, wenn physikalisch sinnvolle Antworten da sind, wo jemand auch in eine völlig falsche Richtung gedacht hat. Also was heißt falsch? In eine Richtung die ich nich- erwartet habe. Aber trotzdem sich sinnvolle physikalische Ant- Gedanken herstellt. Gebe ich Punkte. So. [...]	[...] Ich würde für meine (.) Konzeption nur Stichworte nehmen, weil ich den Schülern freie Hand beim Formulieren lasse. Ich lege also nich- Wert auf beschrimt- bestimmte Formulierungen. Allerdings leg- ich Wert auf Fachbegriffe. Ähm, die in bestimmten Zusammenhängen dann auch auftauchen sollten. Und richtich verwendet werden sollten. Zum Beispiel Arbeit wird verrichtet. Also das entsprechende V- Verb dann auch dazu. Gut. [...]

Tabelle 6.13.: Auszug aus dem Laut-Denk-Protokoll von Herrn Hastedt und Frau Pinna, in denen die Teilnehmer_innen (vor allem) allgemeine Handlungsstrategien für das Feststellen und Beurteilen von Schülerleistungen bzw. zur Erstellung eines Erwartungshorizonts beschreiben oder in denen sie sich zu ihrem allgemeinen Vorgehen/zu ihren allgemeinen Erfahrungen diesbezüglich äußern (Subkategorie 4.0 und 4.3).

In Abbildung 6.14 sind die Boxplots für die prozentuellen Häufigkeiten der vier Subkategorien von Kategorie 4 dargestellt. Zudem sind in dieser Abbildung die Boxplots für die prozentuelle Häufigkeit der Segmente, die mit (mindestens einer) (Subsub-)Subkategorie der Kategorien „Erstellung des Erwartungshorizonts zur Aufgabe Weltraumspaziergang“ (Kategorie 2) und „Korrektur der Schülerlösungstexte“ (Kategorie 3) codiert wurden, grau hervorgehoben dargestellt. Diese dienen als Vergleichsgrafiken für die Boxplots der Subkategorien 4.0 und 4.3, da sich zeigte, dass Segmente, die mit diesen beiden Subkategorien zu codieren waren, in den meisten Fällen sich Segmenten anschlossen bzw. von Segmenten gefolgt wurden, die der Kategorie 3 bzw. 2 zugewiesen waren.

Es fällt auf, dass die Mediane der prozentuellen Häufigkeiten von Subkategorie 4.0 und Kategorie 3 („Korrektur der Schülerlösungstexte“), sowie Subkategorie 4.1 und Kategorie 2 („Erstellung des Erwartungshorizonts zur Aufgabe Weltraumspaziergang“) in grober Näherung in einem sechs zu eins bzw. fünf zu eins Verhältnis stehen. Diese Verhältnisse deuten darauf hin, dass die Teilnehmer_innen für die Korrektur der vier Schülerlösungstexte bzw. die Erstellung des Erwartungshorizonts zu einem nicht vernachlässigenden Anteil auf ihrem Bewusstsein zugängliche (und daher beim lauten Denken mitvokalisierte) und generalisierte Wissens- und Könnensaspekte und/oder berufsbezogene Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung zurückgegriffen haben. Ähnlich wie bei der Subkategorie „Zuweisung von Punkten im Erwartungshorizont“ (Subkategorie 2.1; siehe oben) zeigt sich hier also ein möglicher Hinweis, dass die Korrektur der vier Schülerlösungstexte und die Erstellung des Erwartungshorizonts zur Aufgabe Weltraumspaziergang durch die Teilnehmer_innen bis zu einem bestimmten Grad im Sinne der in Abschnitt 2.2.4 dargestellten Konzeption der Assessment Literacy von Lehrkräften (als bewusstes Kompromissfinden unter anderem zwischen Wissen und Können und berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung)



Codenr.	Median	IQR
4.0	3.9 %	4.6 %
4.1	2.2 %	1.6 %
4.2	5.5 %	4.3 %
4.3	2.6 %	1.5 %

Codenr.	Median	IQR
2	15.5 %	4.1 %
3	23.5 %	4.8 %

Abbildung 6.14.: Boxplots für die Subkategorien der Kategorie „Äußerungen außerhalb der eigentlichen Korrektur und Erwartungshorizonterstellung“. Zudem sind die Boxplots für die prozentuelle Häufigkeit der Transkriptsegmente, die mit (Subsub-)Subkategorien der Kategorien 2 und 3 codiert wurden, grau hervorgehoben dargestellt.

stattfand. Aufgrund des grobkörnigen Charakters von Subkategorie 4.0 und 4.3 können an dieser Stelle der vorliegenden Arbeit aber keine Detailaussagen über die von den Teilnehmer_innen mitvokalisierten Wissens- und Könnensaspekte und/oder berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung getroffen werden.

Ferner blieb unklar, inwieweit die Teilnehmer_innen für die Korrektur der vier Schülerlösungstexte bzw. die Erstellung des Erwartungshorizonts auch auf Wissens- und Könnensaspekte und/oder berufsbezogene Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung zurückgegriffen haben, die ihrem Bewusstsein nicht zugänglich waren. Dies begründet sich daraus, dass die vier Subkategorien der Kategorie „Äußerungen außerhalb der eigentlichen Korrektur und Erwartungshorizonterstellung“ lediglich erfassen, welche Inhalte in den Segmenten der Laut-Denk-Protokolle mitvokalisiert wurden, nicht aber die impliziten Regelhaftigkeiten der erhobenen Verbaldaten (d. h. wie bestimmte Inhalte beim Mitvokalisieren dargestellt werden). Aus diesem Grund wurden im Rahmen einer Qualifikationsarbeit (vgl. Kroll, 2017) die vollständig codierten Laut-Denk-Protokolle, sowie die Eingangs- und Abschlusssequenzen der retrospektiven Befragung von 4 Teilnehmer_innen, die aufgrund ihrer Auskünfte im Lehrkräftefragebogen zueinander kontrastierend ausgewählt wurden (Herr Jonuzi, Frau Kirik, Herr Rittershaus und Frau Sohm), einer zusätzlichen rekonstruktiven Analyse mit Hilfe der Dokumentarischen Methode unterzogen[135] (vgl. Nohl, 2017; Sander, 2017, S. 109 u. f.). Aus dieser Analyse gingen zusammengefasst die folgenden explorativen Ergebnisse hervor:

[135]Eine ausführliche Darstellung des methodischen Vorgehens bei dieser rekonstruktiven Analyse findet sich bei Kroll (2017, S. 7 u. f.).

> „[A]us den untersuchten Daten [...] [ließ sich] ein den Lehrkräften gemeinsamer negativer Horizont[136] rekonstruieren [ein negatives Ideal, von dem sich in den impliziten Regelhaftigkeiten der Verbaldaten abgewendet wird; M. S. F; vgl. Przyborski & Wohlrab-Sahr, 2014, S. 296]. Er besteht im Verfehlen der [...] Erwartung [eine professionell handelnde Lehrkraft zu sein,] [...] [dadurch dass] ein Großteil der gegeben[en] Schüler_innenantworten [in einer Leistungsfeststellungs- und -beurteilungssituation ‚schlecht' ausfällt]. Aus den damit einhergehenden Vermeidungshandlungen der Lehrkräfte lässt sich ein Bewusstsein für die eigene Rolle bei einer Beurteilung schriftlicher Schülerleistungen erkennen[,] [...] [sowie insbesondere] eine Tendenz zur Wahrung des Anscheins [eigener Professionalität.] [...] Ressourcen, auf die im Zusammenhang mit dem benannten negativen Horizont zurückgegriffen w[urde], bestehen [...] [beispielsweise in der Reduzierung] der mit einer Aufgabe verbundenen Erwartung an [die] Antworten von als ‚gut' eingeschätzten Schüler_innen[,] [...] [und der] Nutzung von [in der Literatur] empfohlenen Methoden[,] wie [...] [der Arbeit] mit Operatoren[,] [...] mit der eine mögliche Steigerung der Erwartbarkeit von Schüler_innenantworten verbunden wird." (Kroll, 2017, S. 61-62)

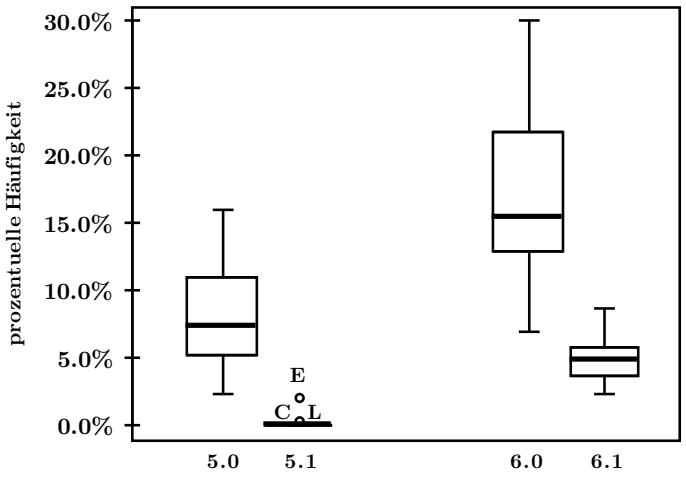
Hier zeigte sich also vor allem ein möglicher Hinweis auf eine wie in Abschnitt 2.2.4 auf theoretischer Ebene dargestellte Verknüpfung zwischen der Assessment Literacy von Lehrkräften und ihrer berufsbezogenen Teilidentität als Assessor of Learning (Assessment Literacy als Entfaltung der berufsbezogenen Teilidentität als Assessor of Learning), da Letztere unter anderem davon gerahmt wird, wie eine Lehrkraft ihre eigene Rolle im Kontext von schulischer Leistungsfeststellung und -beurteilung sieht und inwiefern sich eine Lehrkraft bezogen auf schulische Leistungfeststellung und -beurteilung für souverän hält (vgl. Abschnitt 2.2.4).

Kategorie 5 und 6: emotionale Äußerungen und nichtsprachliche Ereignisse; sonstige Äußerungen/Artefakte des lauten Denkens/sonstige nichtsprachliche Ereignisse

Die Kategorien 5 und 6 erfassen „emotionale Äußerungen und nichtsprachliche Ereignisse" (Kategorie 5), sowie „sonstige Äußerungen/Artefakte des lauten Denkens/sonstige nichtsprachliche Ereignisse" (Kategorie 6). In Kategorie 5 werden die beiden Subkategorien „Lachen, Stöhnen, Ausdruckspartikel, Planungsäußerungen und Verzögerungslaute, usw." (Subkategorie 5.0) und „Klanggesten" (z. B. Fingerschnippen) (Subkategorie 5.1) unterschieden. Kategorie 6 besteht aus den beiden Subkategorien „‚activity descriptions', sonstige Handlungen oder Äußerungen zu sonstigen eigenen Handlungen/Verhalten/Gedanken" (Subkategorie 6.0) und „Ankündigungen; weitere Äußerungen/Transkriptsegemente" (Subkategorie 6.1).

In Abbildung 6.15 sind die Boxplots für die prozentuellen Häufigkeiten der Subkategorien von Kategorie 5 und 6 dargestellt. Diese werden an dieser Stelle der vorliegenden Arbeit

[136] Besonders deutlich wird dieser negative Horizont bei Herrn Rittershaus, der sich zum Abschluss der retrospektiven Befragung unter anderem wie folgt äußert: „Also manchma- mach- ich dann sogar so, dass ich erst mir die Antworten durchgucke und (.) ähm aus dem, was... meim Eindruck den Erwartungshorizont schreibe. [...] Weil (.) ich finde es kann auch nich- sein, dass irgendwie -ne ganze Klasse völlig dran vorbei schreibt. Dann scheint ja auch was im Unterricht schief gelaufen zu sein. [...] Und (.) ähm (.) weil es einfach unfair ist, wenn auch -n guter Schüler kann dran vorbei schreiben, wenn der Lehrer die Aufgabe nicht gut stellt. [...] Und deswegen greif ich gerne auch auf Aufgaben zurück, die schon von anderen (.) guten Leuten gestellt wurden" (Herr Rittershaus, Seg. 767).



Codenr.	Median	IQR
5.0	7.4 %	5.9 %
5.1	.0 %	.0 %
6.0	15.5 %	8.7 %
6.1	4.9 %	2.5 %

Abbildung 6.15.: Boxplots für die Subkategorien der Kategorie „emotionale Äußerungen und nichtsprachliche Ereignisse“, sowie der Katgeorie „sonstige Äußerungen/Artefakte des lauten Denkens/sonstige nichtsprachliche Ereignisse“.

allerdings lediglich der Vollständigkeit halber berichtet und keiner weiterführenden Analyse unterzogen. Grund hierfür ist, dass die Kategorien 5 und 6 für die inhaltsanalytische Auswertung der Laut-Denk-Protokolle lediglich die folgenden zwei dienenden Funktionen erfüllten:

1. Bei der Codierung der Laut-Denk-Protokolle erfüllten die mit Kategorien 5 und 6 codierten Segmente die Funktion, (eindeutiger) auf die Bedeutung noch nicht codierter Segmente schließen zu können. In Tabelle 6.14 ist Segment 271 ein Beispiel für diese Funktion: Aufgrund der zustimmenden Lautäußerung „Mh“ in Segment 270 (codiert mit der Subkategorie 5.0) konnte Segment 271 eindeutiger der Subsubsubkategorie 3.0.3.1 zugewiesen werden.

2. Die Kategorien 5 und 6 dienten für die Analyse der codierten Laut-Denk-Protokolle als „Restekategorien“. Segmente, die den Subkategorien dieser beiden Kategorien zugewiesen wurden, wurden von dieser Analyse ausgeschlossen, da sich ihre durch die Codierung erfasste inhaltliche Bedeutung in vielen Fällen nicht oder kaum im Sinne des Erkenntnisinteresses, das den Forschungsfrage (F1) und (F2) zugrunde liegt, interpretieren ließ (in Tabelle 6.14 z. B. Segment 266) und/oder relevant war (in Tabelle 6.14 z. B. das Blättern im Aufgabenheft in Segment 265, 272 und 280 oder die Ankündigungen in Segment 268, 273 und 281).

6.3.2.3.3. Limitation und Zwischenfazit

Wie im vorangegangenen Unterabschnitt dargestellt, konnten im Rahmen von Phase 2a der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle unterschiedliche Ressourcen der Teilnehmer_innen zur Leistungsfeststellung und -beurteilung identifiziert werden. Als Zwischenfazit hervorzuheben ist, dass die Teilnehmer_innen...

Feintranskript mit Segmentierung		Codierung (Codenr.)
	[...]	
[264]	Okay.	5.0
[265]	(blättert auf Seite 5)	6.0
[266]	Gut!	5.0
[267]	Dann schau- ich mir das mal an, was die lieben Kleinen geschrieben haben.	6.0
[268]	(..) Bei A heißt es:	6.1
[269]	*Im All ist nichts durch das Ton geht und er hört seinen Freund nicht. Dann kommt aber Ton durch die Helme, da Ton über Glas geht.*	1.0.3
[270]	(zustimmend) Mh. (+)	5.0
[271]	Schonma- nich- schlecht.	3.0.1.4 3.0.2.5 3.0.3.1
[272]	(blättert auf Seite 6)	6.0
[273]	B schreibt:	6.1
[274]	Äh *die beiden An-... Astronauten können sich wieder hören, weil der geringe Abstand zwischen den beiden Funkgeräten eine bessere Funkverbindung herstellt. Deswegen kann der jüngere den älteren wieder leise hören.*	1.0.3
[275]	Ach, das- ja drollich!	3.0.1.4 3.0.2.5 3.0.3.2
[276]	(lacht)	5.0
[277]	Ja.	5.0
[278]	(.) (prustet)	5.0
[279]	(.) Ob das noch -n Punkt gibt (.) is- sehr fraglich.	3.0.1.4 3.0.2.5 3.0.3.2
[280]	(blättert auf Seite 7)	6.0
[281]	C (.) sagt:	6.1
	[...]	

Tabelle 6.14.: Auszug aus dem Laut-Denk-Protokoll von Herrn Trummer beim erstmaligen Lesen der Schülerlösungstexte A und B.

... im Rahmen der Laborsituation fragmentarisches Lesen nutzten, um hierdurch einen Schülerlösungstext nach Belegen oder Informationen durchsuchen zu können.

... bei der Erstellung eines Erwartungshorizonts sprachliche Merkmale eines Schülerlösungstextes in einem unterschiedlichen Ausmaß (z. B. explizites Fixieren im Erwartungshorizont) und einer unterschiedlichen Art und Weise (z. B. defizit- oder fähigkeitsorientiert) berücksichtigten.

... bei der Korrektur unterschiedliche Merkmale der Schülerlösungstexte fokussierten. Es deutet sich allerdings an, dass der fachlich-konzeptuelle Eindruck über einen Schülerlösungstext hierbei eine größere Rolle spielt, als die sprachliche Realisierung eines Schülerlösungstextes.

... überwiegend sachliche Kriterien als Bezug für die Verortung eines Schülerlösungstextes nutzten. Bemerkenswert ist allerdings, dass die Teilnehmer_innen hierfür – wenn auch nur vereinzelt – auch auf mutmaßliche Personenmerkmale von Schüler_innen (z. B. ihr Geschlecht) zurückgegriffen haben.

... in Teilen die Leistungen in den vier Schülerlösungstexten auch auf Grundlage eines (ersten) eher holistischen Eindrucks und/oder heuristisch feststellten und beurteilten.

... die Leistungen in den vier Schülerlösungstexten in einer tendenziell defizitorientierten Art und Weise feststellten und beurteilten.

... bei der Korrektur der vier Schülerlösungstexte und der Erstellung des Erwartungshorizonts auch auf bis zu einem bestimmten Grad generalisierte (möglicherweise aber nur zum Teil ihrem Bewusstsein zugängliche) Wissens- und Könnensaspekte und/oder berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung zurückgegriffen haben.

Die identifizierten Auffälligkeiten haben dabei keinen Anspruch auf Vollständigkeit. Viel mehr ist es plausibel anzunehmen, dass es weitere Ressourcen zur Leistungsfeststellung und Beurteilung gibt, auf die allerdings lediglich nur von einzelnen Lehrkräften im Rahmen der Laborsituation zurückgriffen wurde. Derartige Ressourcen konnten im Rahmen der vorgestellten Zusammenfassung der Laut-Denk-Protokolle nicht abgebildet werden, da das hierbei gewählte Vorgehen auf die fallübergreifende Reduktion und Verdichtung des untersuchten Datenmaterials ausgelegt war.

6.3.2.4. Phase 2b: Extrahieren qualitativer Prozessinformationen aus den Laut-Denk-Protokollen

6.3.2.4.1. Methodische Vorbemerkungen

Ziel von Phase 2b der inhaltsanalystischen Auswertung der Laut-Denk-Protokolle war, qualitative Prozessaspekte aus den erhobenen Daten zu extrahieren und dem Erkenntnisinteresse der vorliegenden Arbeit entsprechend zu interpretieren. Da Forschungsfrage (F1) und (F2) die Genese fachlich-konzeptueller und sprachlicher Leistungsurteile durch Physiklehrkräfte im Rahmen einer Klassenarbeit adressieren, wurden für Phase 2b nicht die vollständig codierten Laut-Denk-Protokolle herangezogen. Stattdessen dienten die Codierungen mit folgenden (Subsubsub-)Kategorien als Grundlage für die Analyse:

- Subsubsubkategorie 3.0.1.1, mit der die Segmente in den Laut-Denk-Protokollen codiert wurden, in denen der fachlich-konzeptuelle Eindruck eines Schülerlösungstextes festgestellt und beurteilt wurde.
- Subsubsubkategorie 3.0.1.2, mit der die Segmente in den Laut-Denk-Protokollen codiert wurden, in denen die sprachliche Realisierung eines Schülerlösungstextes festgestellt und beurteilt wurde.
- Subsubsubkategorie 3.0.3.1, mit der die Segmente in den Laut-Denk-Protokollen codiert wurden, in denen sich bei der Feststellung und Beurteilung eines Schülerlösungstextes positiv wertend/akzeptierend geäußert wurde.
- Subsubsubkategorie 3.0.3.2, mit der die Segmente in den Laut-Denk-Protokollen codiert wurden, in denen sich bei der Feststellung und Beurteilung eines Schülerlösungstextes negativ wertend/ablehnend geäußert wurde.
- Subsubsubkategorie 3.0.3.3, mit der die Segmente in den Laut-Denk-Protokollen codiert wurden, in denen sich bei der Feststellung und Beurteilung eines Schülerlösungstextes neutral/gemischt/sonstig geäußert wurde.

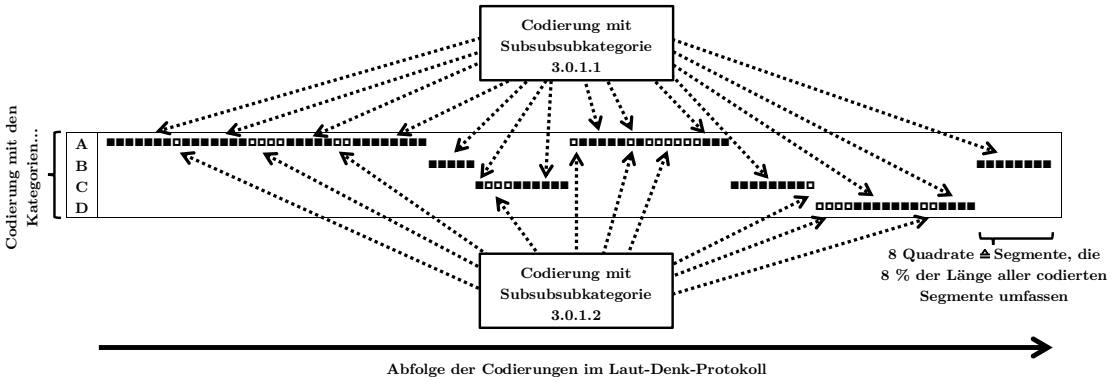


Abbildung 6.16.: Veranschaulichung des Aufbaus der Dokumenten-Portraits an einem Beispielportrait (Dokumenten-Portrait von Herrn Mehlert).

- Kategorie „A“, „B“, „C“ und „D“, mit der die Segmente in den Laut-Denk-Protokollen codiert wurden, in denen eine Auseinandersetzung mit Schülerlösungstext A, B, C oder D stattfand.

Für die Analyse selbst wurde zunächst die Abfolge, in der Codierungen mit den Subsubsubkategorien 3.0.1.1 und 3.0.1.2 in Kombination mit Codierungen mit den Kategorien „A“, „B“, „C“ und „D“ in den 21 Laut-Denk-Protokollen auftraten als sogenannte *Dokumenten-Portraits* visualisiert[137]. In diesen sind nur die Codierungen mit den eben benannten (Subsubsub-)Kategorien in der Abfolge ihres Auftretens im entsprechenden Laut-Denk-Protokoll grafisch dargestellt (zur Veranschaulichung vgl. Abbildung 6.16). Die Codierungen mit den Subsubsubkategorien 3.0.1.1 und 3.0.1.2 sind dabei als schwarze bzw. weißen Quadrate symbolisiert. Jedes Dokumenten-Portrait besteht dabei aus insgesamt 100 schwarzen bzw. weißen Quadraten und die Anzahl an schwarzen bzw. weißen Quadraten entspricht der Länge der entsprechend codierten Segmente im jeweiligen Laut-Denk-Protokoll (vgl. VERBI Software. Consult. Sozialforschung. GmbH , 2018, S. 251). Die schwarzen und weißen Quadrate sind zudem auf vier Zeilen verteilt, je nachdem ob die mit Subsubsubkategorie 3.0.1.1 und 3.0.1.2 codierten Segmente zusätzlich mit der Kategorie „A“, „B“, „C“ oder „D“ codiert waren. Dabei ist zu beachten, dass sich aus den Zeilen den Dokumentenportraits nicht auf die Reihenfolge schließen lässt, in der die Schülerlösungstexte A bis D von den Teilnehmer_innen korrigiert wurden. Grund hierfür ist, dass in den Dokumentenportraits nicht alle Segmente, die mit den Kategorien „A“, „B“, „C“ und „D“ codiert wurden, symbolisch dargestellt sind[138]. Kurz charakterisiert stellen die Dokumentenportraits also ein verjüngtes Abbild des chronologischen Verlaufs ausge-

[137] Diese Dokumentenportraits wurden mit Hilfe eines entsprechenden Visualisierungstools des Softwarepakets MAXQDA Plus (Version 12.3.2) erstellt (vgl. VERBI Software. Consult. Sozialforschung. GmbH , 2018, S. 250 u. f.) und manuell in die Form, in der sie in der vorliegenden Arbeit dargestellt sind, nachbearbeitet.

[138] Bei einer Durchsicht der Laut-Denk-Protokolle aller Teilnehmer_innen zeigte sich allerdings, dass die deutliche Mehrheit der Lehrkräfte die Schülerlösungstexte (zum Teil zyklisch) in der Reihenfolge korrigierten, wie sie im Aufgabenheft abgedruckt waren (alphabetische Reihenfolge). Bei dieser Korrekturreihenfolge handelt es sich allerdings möglicherweise um ein Artefakt der Laborsituation. Beispielsweise waren die Seiten der Aufgabenhefte nummeriert und durch Heftklammern fest miteinander

wählter Gedankenabfolgen dar, die die Teilnehmer_innen im Rahmen der Laborsituation laut-denkend mitvokalisiert haben.

Die Dokumenten-Portraits aller Teilnehmer_innen sind in Abbildung 6.17 und 6.18 dargestellt. Wie unmittelbar erkennbar ist, sind sich einige Dokumenten-Portraits sehr ähnlich (z. B. die Dokumenten-Portraits von Herrn Einert und Herrn Onne), wohingegen andere sich deutlich voneinander unterscheiden (z. B. die Dokumenten-Portraits von Herrn Trummer und Herrn Uckermark). Durch paarweises Vergleichen der 21 Dokumentportraits miteinander und durch Rückgriff auf die entsprechend codierten Segmente in den Laut-Denk-Protokollen zeigte sich, dass sich „ähnliche“ Dokumentenportraits von „unähnlichen“ anhand der folgenden zwei Vergleichsdimensionen voneinander unterscheiden lassen:

1. Bei welchen Schülerlösungstexten findet eine Feststellung und Beurteilung der sprachlichen Realisierung statt (Auftreten von Subsubsubkategorie 3.0.1.2)?
2. Wie oft findet bei der Korrektur der Schülerlösungstexte ein Wechsel zwischen dem Feststellen und Beurteilen des fachlich-konzeptuellen Eindrucks und dem Feststellen und Beurteilen der sprachlichen Realisierung statt (Wechsel von Subsubsubkategorie 3.0.1.1 zu 3.0.1.2 oder umgekehrt)?

Bezüglich dieser beiden Vergleichsdimensionen konnten verschiedene Muster in den Dokumentenportraits und letztendlich in den Laut-Denk-Daten der Teilnehmer_innen identifiziert werden. Diese lassen sich daher als unterschiedliche Bewertungslogiken der Teilnehmer_innen im Sinne von Arten des Umgangs mit den Schülerlösungstexten A bis D bei der laut-denkenden Korrektur interpretieren (Ressourcen zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen; Forschungsfrage (F1)). Die Charakteristika und Interpretation dieser Muster werden im folgenden vorgestellt.

verbunden, was bei den Lehrkräften den Eindruck erweckte haben könnte, das Aufgabenheft müsse im Rahmen der Laborsituation von vorne nach hinten bearbeitet werden.

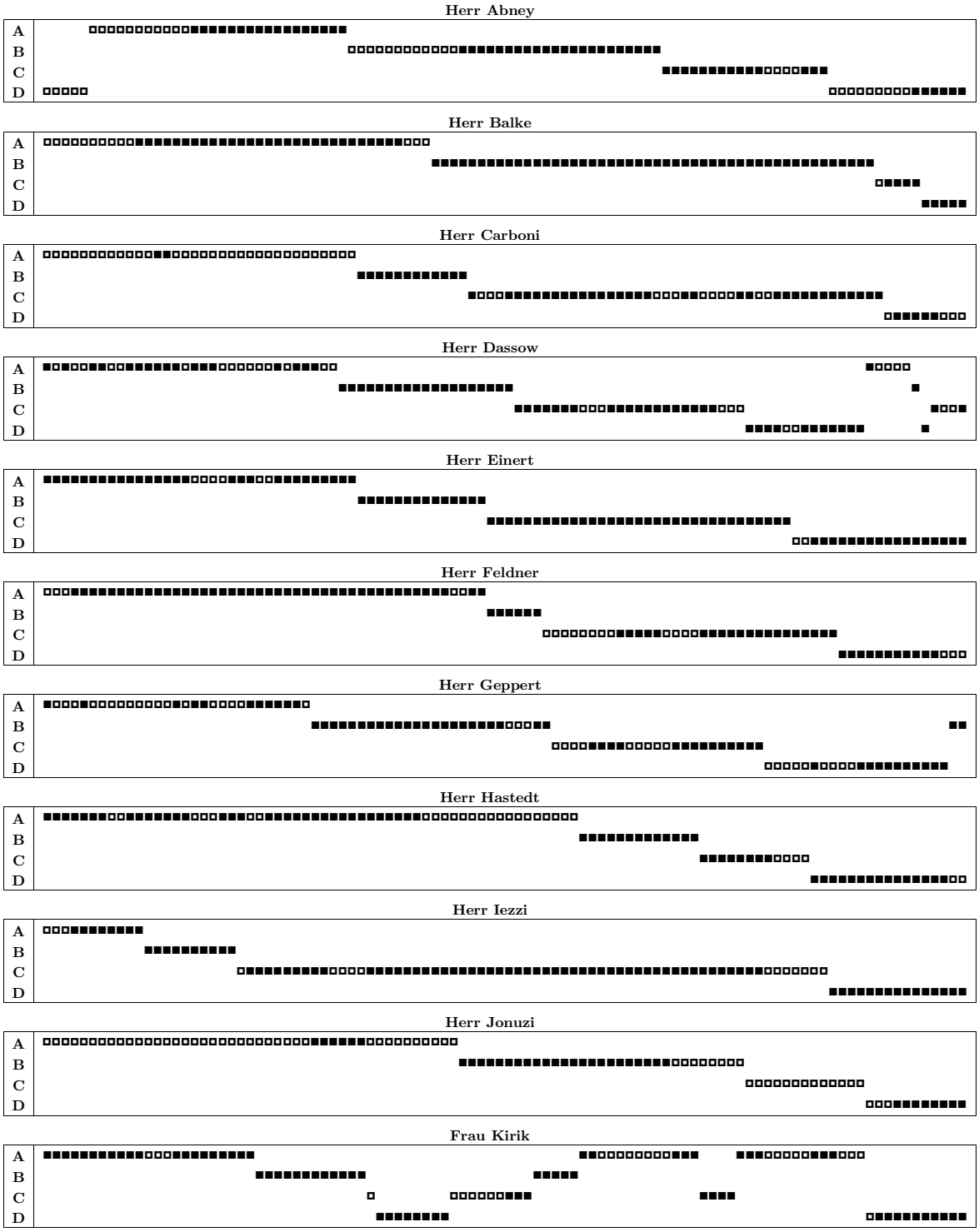


Abbildung 6.17.: Dokumentenportraits der teilnehmenden Physiklehrkräfte mit den Abkürzungen A bis K. Die weißen (schwarzen) Quadrate repräsentieren die chronologische Abfolge der Segmente, in denen die sprachliche Realisierung (der fachlich-konzeptuelle Eindruck) eines Schülerlösungstextes festgestellt und beurteilt wird. Die Zeilen der Dokumentenportraits differenzieren, welcher der vier Schülerlösungstexte A bis D an welchen Stellen der Laut-Denk-Protokolle korrigiert wird.

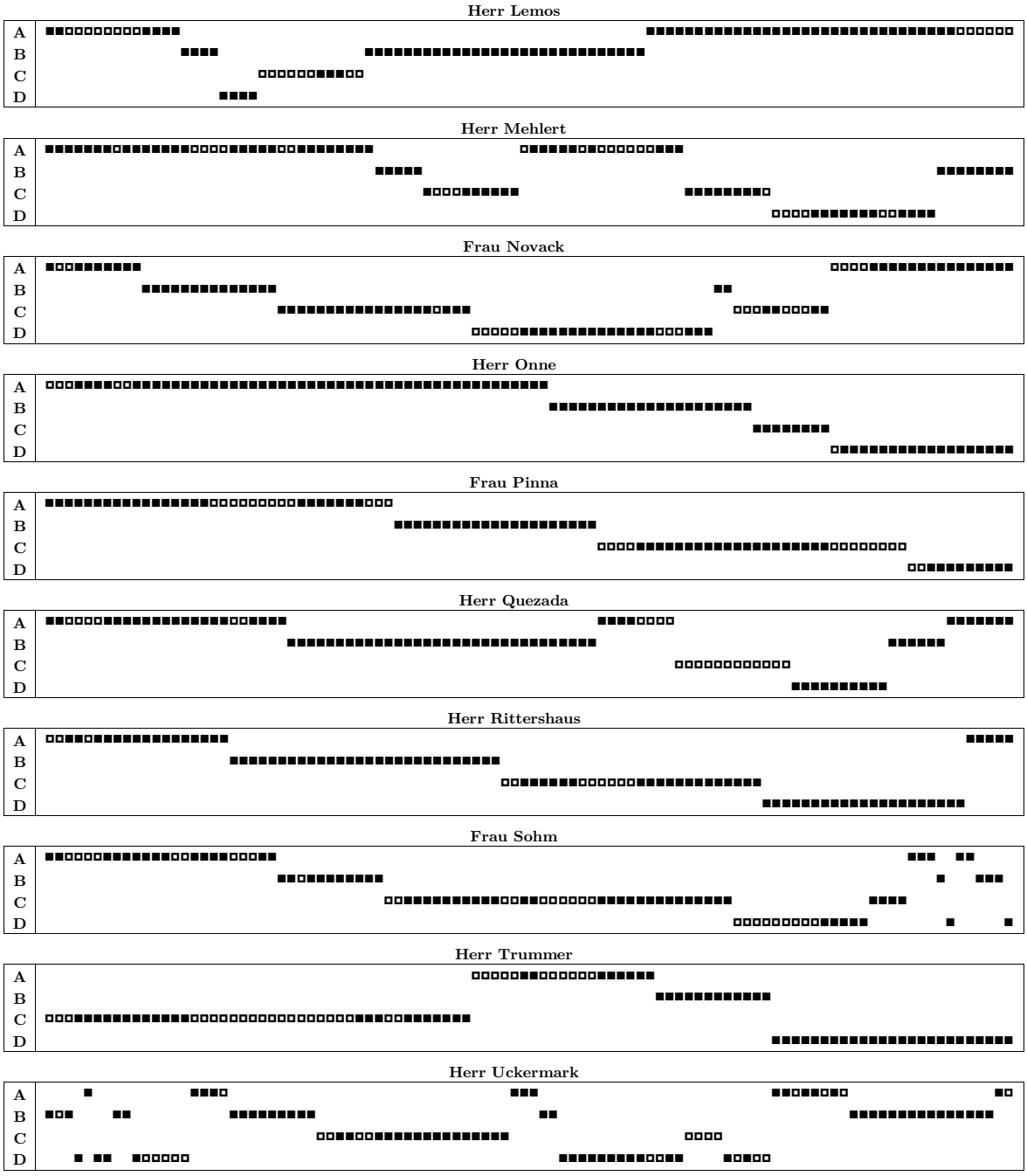


Abbildung 6.18.: Dokumentenportraits der teilnehmenden Physiklehrkräfte mit den Abkürzungen L bis U. Die weißen (schwarzen) Quadrate repräsentieren die chronologische Abfolge der Segmente, in denen die sprachliche Realisierung (der fachlich-konzeptuelle Eindruck) eines Schülerlösungstextes festgestellt und beurteilt wird. Die Zeilen der Dokumentenportraits differenzieren, welcher der vier Schülerlösungstexte A bis D an welchen Stellen der Laut-Denk-Protokolle korrigiert wird.

6.3.2.4.2. Muster im Feststellen und Beurteilen der sprachlichen Realisierung

Durch einen Vergleich der Dokumentenportraits danach, bei welchen der vier Schülerlösungstexte ein Feststellen und Beurteilen der sprachlichen Realisierung stattfand (auftreten von weißen Quadraten in den Zeilen der Dokumentenportraits), konnten insgesamt vier verschiedene Muster identifiziert werden:

Muster (1a): Festellen und Beurteilen der sprachlichen Realisierung nur bei Schülerlösungstexten mit geringer Qualität in der sprachlichen Realisierung
(Lehrkraft: E, O)

Bei diesem Muster wurde eine Feststellung und Beurteilung der sprachlichen Realisierung nur bei den beiden Schülerlösungstexten mit einer diesbezüglich geringen Qualität vorgenommen (Schülerlösungstext A und D). Bei Schülerlösungstext C und B, die gemäß der Vorauswahl in der Entwicklungsstudie eine hohe Qualität in ihrer sprachlichen Realisierung aufweisen, fand dies nicht statt. Muster (1a) lässt sich daher als Hinweis auf die Tendenz einer Defizitorientierung bei der Feststellung und Beurteilung der sprachlichen Realisierung schriftlicher Schülerleistungen interpretieren: es werden vor allem sprachliche Mängel von Schülerlösungstexten festgestellt und beurteilt, weswegen eine Feststellung und Beurteilung der sprachlichen Realisierung nur bei sprachlich defizitären Schülerlösungstexten stattfand. Gestützt wird diese Interpretation von entsprechend codierten Segmenten in den Laut-Denk-Protokollen der beiden Lehrkräfte, bei denen sich dieses Muster zeigte. Herr Einert äußerte sich durchgehend negativ wertend/ablehnend bezüglich der sprachlichen Realisierung von Schülerlösungstext A und D (z. B. „Also sprachlich is- das alles ir-... auch jetz- nich- so -ne (.) Meisterleistung“ (Seg. 598) oder „Also sprachlich- Mängel. Im Satzaufbau“ (Seg. 757-758)). Ähnliches gilt auch für Herrn Onne. In zwei Segmenten seines Laut-Denk-Protokolls äußerte er allerdings bezüglich Schülerlösungstext A, dass dieser „prinzipiell [...] zu verstehen [ist]“ (Seg. 146) bzw. „man [das] im Grund genommen sagen [kann]“ (Seg. 150). Hierbei handelt es sich zwar um akzeptierende Äußerungen bezüglich der sprachlichen Realisierung von Schülerlösungstext A, gleichzeitig weisen diese Äußerungen durch das Adjektiv „prinzipiell“ bzw. die Redewendung „im Grunde genommen“ eine negative Konnotation auf. Daher sind auch die positiv wertenden/akzeptierenden Äußerungen von Herrn Onne mit der Interpretation einer Tendenz zur Defizitorientierung bei der Feststellung und Beurteilung der sprachlichen Realisierung schriftlicher Schülerlösungen vereinbar.

Muster (1b): Festellen und Beurteilen der sprachlichen Realisierung nur bei Schülerlösungstexten mit hoher fachlich-konzeptueller Qualität
(Lehrkraft: B, I, L, Q, R, T)

Lehrkräfte, bei denen sich dieses Muster zeigte, nahmen eine Feststellung und Beurteilung der sprachlichen Realisierung nur bei den Schülerlösungstexten A und C vor, also denjenigen mit einer hohen fachlich-konzeptuellen Qualität. Bei Schülerlösungstext B und D wurde hingegen nur deren fachlich-konzeptuelle Qualität festgestellt und beurteilt. Dieses

Lehrkraft	Feintranskript mit Segmentierung (alle Segmente beziehen sich auf Schülerlösungstext B)	Feintranskript mit Segmentierung (alle Segmente beziehen sich auf Schülerlösungstext D)
Herr Balke	[181] Also, beide äh Aufgabenteile nich- richtig beantwortet.	[229] Ähm und mit der Frequenz hat das ja alles schon mal gar nichts zu tun.
Herr Iezzi	[116] (...) Also es passt natürlich nicht zu dem, was wir im Unterricht gemacht haben.	[185] Ähm (...) das stimmt einfach nicht.
Herr Lemos	[129] Öh der hat gar nichts verstanden, worum es geht.	[148] D- mit Frequenz hat das auch nichts zu tun.
Herr Quezada	[196] (...) Okay, also das is- eine An-... Au-... k-... eine Antwort die nicht unbedingt zu dem Kontext (.) passt, so wie die Aufgabenstellung gemeint is-.	[282] Und das ist schlicht und einfach falsch.
Herr Rittershaus	[385] Und irgendwie (.) auch ganz anders als ich das erwarte... (.) erwartet habe.	[581] Also ich finde... (.) Ich finde das (..) mhm es is- nicht die richtige Erklärung.
Herr Trummer	[529] Äh nicht Themen äh... nicht themengerecht beantwortet.	[555] Und äh die Frequenz kann nicht der Grund sein.

Tabelle 6.15.: Transkriptauszüge aus den Laut-Denk-Protokollen von Herrn Balke, Iezzi, Lemos, Quezada, Ritterhaus und Trummer während der Korrektur von Schülerlösungstext B und D. In allen Transkriptauszügen wird expliziert, dass Schülerlösungstext B bzw. D nicht den eigenen (fachlich-konzeptuellen) Erwartungen genügt.

Muster lässt sich auf zweierlei Arten interpretieren: Zum einen ist denkbar, dass die entsprechenden Lehrkräfte eine Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes nur dann vornehmen, wenn dieser bis zu einem bestimmten Grad auch ihren fachlich-konzeptuellen Erwartungen entspricht. Eine andere Interpretation ist die einer Fokussierung auf die Feststellung und Beurteilung fachlich-konzeptueller Merkmale eines Schülerlösungstextes, wenn dieser die eigenen diesbezüglichen Erwartungen verfehlt. Dementsprechend treten Merkmale der sprachlichen Realisierung eines Schülerlösungstextes bei der Leistungsfeststellung und -beurteilung in den Hintergrund. Die codierten Segmente in den Laut-Denk-Protokollen sprechen vor allem für die letzte der beiden eben genannten Interpretationen. In den codierten Segmente zeigte sich, dass sich – außer Herr Iezzi[139] – alle 6 Lehrkräfte, die Muster (1b) zugeordnet wurden, vor allem negativ wertend/ablehnend zum fachlich-konzeptuellen Eindruck von Schülerlösungstext B und D äußerten (mehrheitliche Codierung der für diesen Teil der Auswertung ausgewählten Segmente mit Subkategorie 3.0.3.2). Ferner hoben alle 6 Lehrkräfte besonders hervor, dass beide Schülerlösungstexte ihre fachlich-konzeptuellen Erwartungen (deutlich) verfehlen. Um Letzteres zu veranschaulichen ist in Tabelle 6.15 für jede der 6 Lehrkräfte ein dementsprechend prägnantes Transkriptsegment jeweils für Schülerlösungstext B und D dargestellt.

[139] Herr Iezzi äußert sich vor allem neutral zum fachlich-konzeptuellen Eindruck von Schülerlösungstext B (mehrheitliche Codierung der für diesen Teil der Auswertung ausgewählten Segmente mit Subkategorie 3.0.3.3), was damit zusammenhängt, dass er entschließt „die Aufgabe für diesen Schüler aus der Wertung raus[zu]nehmen“ (Seg. 121; vgl. auch Abschnitt 6.3.1). Gerade hieran ist aber deutlich zu erkennen, dass Schülerlösungstext B nicht Herrn Iezzis fachlich-konzeptuellen Erwartungen entspricht.

Muster (1c): Festellen und Beurteilen der sprachlichen Realisierung nur bei Schülerlösungstexten mit geringer Qualität in der sprachlichen Realisierung <u>oder</u> bei Schülerlösungstexten mit hoher fachlich-konzeptueller Qualität
(Lehrkraft: C, D, F, H, K, M, N, P)

Bei diesem Muster fand bei fast allen Schülerlösungstexten eine Feststellung und Beurteilung der sprachlichen Realisierung statt. Die einzige Ausnahme stellte Schülerlösungstext B dar, der gemäß der Vorauswahl in der Entwicklungsstudie eine geringe fachlich-konzeptuelle Qualität aufweist, jedoch eine hohe Qualität in seiner sprachlichen Realisierung. Dieses Muster lässt sich als eine Mischform von Muster (1a) und (1b) interpretieren: Eine Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes wurde nur dann vorgenommen, wenn...

... der Schülerlösungstext zu einem bestimmten Grad bezüglich sprachlicher Merkmale defizitär ist oder

... der Schülerlösungstext die eigenen fachlich-konzeptuellen Erwartungen nicht verfehlt.

Wie Tabelle 6.16 und 6.17 verdeutlichen, lässt sich diese Interpretation ebenfalls durch entsprechend codierte Segmente in den Laut-Denk-Protokolle der 8 Lehrkräfte, bei denen sich Muster (1c) zeigte, stützen. In Tabelle 6.16 sind analog zu Tabelle 6.15 für jede der 8 Lehrkräfte Transkriptsegmente dargestellt, in denen besonders prägnant hervorgehoben wird, dass Schülerlösungstext B nicht den eigenen (fachlich-konzeptuellen) Erwartungen entspricht. In den codierten Segmente zeigte sich zudem, dass sich – außer Herr Hastedt[140] und Herr Mehlert[141] – alle 8 Lehrkräfte, die Muster (1c) zugeordnet wurden, vor allem negativ wertend/ablehnend zum fachlich-konzeptuellen Eindruck von Schülerlösungstext B äußerten (mehrheitliche Codierung der für diese Teil der Auswertung ausgewählten Segmente mit Subkategorie 3.0.3.2).

In Tabelle 6.17 hingegen ist für Schülerlösungstexte A, C und D, sowie für jede der 8 Lehrkräfte die Anzahl an Segmenten angegeben, in denen sich positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig zu deren sprachliche Realisierung geäußert wurde (Anzahl der Segmente, die mit Subsubsubkategorie 3.0.1.2 und zusätzlich mit der Subsubsubkategorie 3.0.3.1, 3.0.3.2 oder 3.0.3.3 codiert wurden). Aus den

[140] Herr Hastedt äußerte sich vor allem positiv zum fachlich-konzeptuellen Eindruck von Schülerlösungstext B (mehrheitliche Codierung der für diesen Teil der Auswertung ausgewählten Segmente mit Subkategorie 3.0.3.1). Dies lässt sich auf sein Honorieren, dass „[d]er Schüler [...] physikalisch [denkt]“ (Seg. 412; vgl. auch Tabelle 6.13) zurückführen. Bei der Korrektur machte Herr Hastedt aber dennoch deutlich, dass Schülerlösungstext B nicht seinen fachlich-konzeptuellen Erwartungen entspricht (vgl. Tabelle 6.16).

[141] Herr Mehlert äußerte sich vor allem neutral zum fachlich-konzeptuellen Eindruck von Schülerlösungstext B (mehrheitliche Codierung der für diesen Teil der Auswertung ausgewählten Segmente mit Subkategorie 3.0.3.3). Dies lässt sich darauf zurückführen, dass er vor der eigentlichen Korrektur von Schülerlösungstext B ausführlich und neutral beschreibt, dass seiner Vermutung nach „der Schüler [...] von -nem anderen Problem ausgegangen [ist,] [...] und dachte die hören sich praktisch über die Funkverbindung [...] [u]nd nich- über die Ausbreitung von Schall“ (Seg. 485-487). Anschließend machte Herr Mehlert jedoch deutlich, dass Schülerlösungstext B nicht seinen fachlich-konzeptuellen Erwartungen entspricht (vgl. Tabelle 6.16).

Lehrkraft	Feintranskript mit Segementierung (alle Segmente beziehen sich auf Schülerlösungstext B)	
Herr Carboni	[322]	Im Deutschen hieße das Thema verfehlt!
Herr Dassow	[320]	(...) Das ähm is- macht eigentlich so auf den ersten Blick schon mal klar, dass... oder oder wenn auf den ersten Blick wird schon mal deutlich, dass das Problem an sich eben nicht erkannt wurde.
Herr Feldner	[383]	Das ist ähm... od-... das halt ich für eine nicht zutreffende Erklärung.
Herr Hastedt	[406]	So, das is- komplett (...) falsch!
	[407]	(.) Im Sinne der erwarteten Aufgabenstellung.
Frau Kirik	[326]	Das ist, würd- ich sagen, aber (.) Quark, ne?
Herr Mehlert	[516]	Also nach mein- Bewertungshorizont sind das 0 Punkte.
Frau Novack	[288]	(.) Ähm trotzdem is- des nach (.) der Aufgabenstellung (.) und nach meim Erwartungshorizont erstma- 0 Punkte.
Frau Pinna	[243]	(.) Ähm (...) is- nich- ganz falsch.
	[244]	Ähm es geht aber (..) äh aus dem Text, aus meiner Sicht, nicht wirklich so hervor.

Tabelle 6.16.: Transkriptauszüge aus den Laut-Denk-Protokollen von Herrn Carboni, Herrn Dassow, Herrn Feldner, Herrn Hastedt, Frau Kirik, Herrn Mehler, Frau Novack und Frau Pinna während der Korrektur von Schülerlösungstext B. In allen Transkriptauszügen wird expliziert, dass der Schülerlösungstext B nicht den eigenen (fachlich-konzeptuellen) Erwartungen entspricht.

Spalten dieser Tabelle wird deutlich, dass sich – außer Herr Feldner[142] – alle 8 Lehrkräfte bei Schülerlösungstext A und D am häufigsten negativ wertend/ablehnend bezüglich der sprachlichen Realisierung äußerten, was sich als Hinweis auf eine diesbezüglich tendenziell defizitorientierte Feststellung und Beurteilung beider Schülerlösungstexte deuten lässt. Ferner äußerten sich vier der acht Lehrkräfte (Herr Carboni, Herr Dassow, Herr Feldner und Frau Pinna) auch bei Schülerlösungstext C, der gemäß der Vorauswahl in der Entwicklungsstudie eine hohe Qualität in seiner sprachlichen Realisierung aufweist, am häufigsten negativ wertend/ablehnend bezogen auf seine sprachlichen Realisierung, was bei diesen vier Lehrkräften ebenfalls für die eben benannte Tendenz zur Defizitorientierung spricht. Bei den übrigen vier Lehrkräften (Herr Hastedt, Frau Kirik, Herr Mehlert und Frau Novack) überwiegt hingegen die Anzahl positiver wertender/akzeptierender Äußerungen bezüglich der sprachlichen Realisierung von Schülerlösungstext C. Bei diesen vier Lehrkräften zeigt sich also ein vorsichtiger Hinweis einer fähigkeitsorientierten Feststellung und Beurteilung der sprachlichen Realisierung bei Schülerlösungstexten, die sowohl eine hohe fachlich-konzeptuelle Qualität aufweisen, als auch eine hohe Qualität bezüglich ihrer sprachlichen Realisierung.

[142] In den drei Transkriptsegmenten, in denen sich Herr Feldner positiv wertend/akzeptierend zur sprachlichen Realisierung äußert zeigte sich – ähnlich wie bei Herrn Onne (siehe oben) – eine negative Konnotation. Er äußerte sich wie folgt: „(..) Ka- man das so grammatikalisch (.) gelten lassen? Joar, könnte man noch. Akzeptiert. Also an der Grammatik jetz- nichts zu ändert“ (Seg. 273-276).

Lehrkraft	Anzahl der Segment, in denen sich zur sprachliche Realisierung positiv wertend/akzeptierend (+), negativ wertend/ablehnend (-) oder neutral/gemischt/sonstig (n) geäußert wurde								
	Schülerlösungstext A			Schülerlösungstext C			Schülerlösungstext D		
	+	-	n	+	-	n	+	-	n
Herr Carboni	0	18	1	3	4	0	0	2	0
Herr Dassow	0	26	1	0	6	1	0	4	1
Herr Feldner	3	1	2	0	6	4	0	3	0
Herr Hastedt	1	10	8	2	0	0	0	1	0
Frau Kirik	2	4	0	3	0	1	0	1	0
Herr Mehlert	2	13	2	3	0	0	0	3	0
Frau Novack	0	4	0	2	1	0	0	5	1
Frau Pinna	0	7	3	2	3	2	0	1	0

Tabelle 6.17.: Anzahl an Segmenten in den Laut-Denk-Protokollen von Herrn Carboni, Herrn Dassow, Herrn Feldner, Herrn Hastedt, Frau Kirik, Herrn Mehler, Frau Novack und Frau Pinna, in denen sich positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig zur sprachlichen Realisierung von Schülerlösungstext A, C und D geäußert wurde (Anzahl der Segmente, die pro Schülerlösungstext mit Subsubsubkategorie 3.0.1.2 und zusätzlich mit der Subsubsubkategorie 3.0.3.1, 3.0.3.2 oder 3.0.3.3 codiert wurden).

Muster (1d): Festellen und Beurteilen der sprachlichen Realisierung bei allen vier Schülerlösungstexten
(Lehrkraft: A, G, J, S, U)

Bei diesem Muster wurde bei allen vier Schülerlösungstexten eine Feststellung und Beurteilung der sprachlichen Realisierung vorgenommen. Es lässt sich daher vermuten, dass die Lehrkräfte, die Muster (1d) zugeordnet wurden, stets auch die sprachlichen Leistungen eines Schülerlösungstextes feststellen und beurteilen. Allein aufgrund der Dokumenten-Portraits ließen sich allerdings keine Hinweise identifizieren, die für eine Tendenz zu einer Defizit- und/oder Fähigkeitsorientierung bei der Feststellung und Beurteilung der sprachlichen Realisierung der Schülerlösungstexte sprechen. Aus Tabelle 6.18 geht aber hervor, dass sich eher eine Defizitorientierung annehmen lässt. Fast alle Lehrkräfte (Herr Abney, Herr Jonuzi, Frau Sohm und Herr Uckermark) äußerten sich bei drei der vier Schülerlösungstexte vor allem negativ wertend/ablehnend bezogen auf deren sprachliche Realisierung. Lediglich Herr Geppert äußerte sich bei Schülerlösungstexten mit einer geringen Qualität in der sprachlichen Realisierung diesbezüglich vor allem negativ wertend/ablehnend (Schülerlösungstext A und D), wohingegen er sich bei Schülerlösungstexten mit einer hohen Qualität in der sprachlichen Realisierung vor allem positiv wertend/akzeptierend äußerte (Schülerlösungstext B und C). Bei Herrn Geppert lässt sich daher weder die Tendenz zu einer Defizit-, noch die einer Fähigkeitsorientierung der Feststellung und Beurteilung der sprachlichen Realisierung schriftlicher Schülerleistungen vermuten.

Lehrkraft	Anzahl der Segmente, in denen sich zur sprachliche Realisierung positiv wertend/akzeptierend (+), negativ wertend/ablehnend (-) oder neutral/gemischt/sonstig (n) geäußert wurde											
	Schülerlösungstext A			Schülerlösungstext B			Schülerlösungstext C			Schülerlösungstext D		
	+	-	n	+	-	n	+	-	n	+	-	n
Herr Abney	0	7	0	0	4	0	1	0	0	0	8	0
Herr Geppert	0	14	6	2	0	0	7	0	4	0	7	1
Herr Jonuzi	1	10	2	3	0	0	2	3	2	0	1	0
Frau Sohm	0	8	4	0	1	0	2	5	9	1	8	3
Herr Uckermark	0	7	0	0	1	0	9	0	1	2	6	2

Tabelle 6.18.: Anzahl an Segmenten in den Laut-Denk-Protokollen von Herrn Abney, Herrn Geppert, Herrn Jounzi, Frau Sohm und Herrn Uckermark, in denen sich positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig zur sprachlichen Realisierung von Schülerlösungstext A, B, C und D geäußert wurde (Anzahl der Segmente, die pro Schülerlösungstext mit Subsubsubkategorie 3.0.1.2 und zusätzlich mit der Subsubsubkategorie 3.0.3.1, 3.0.3.2 oder 3.0.3.3 codiert wurden).

6.3.2.4.3. Muster in der Häufigkeit des Wechsels zwischen fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung

Beim Vergleichen der Dokumentenportraits bezüglich der Anzahl, in der zwischen dem Feststellen und Beurteilen des fachlich-konzeptuellen Eindrucks und dem Feststellen und Beurteilen der sprachlichen Realisierung gewechselt wurde (Anzahl der Wechsel von schwarzen auf weiße Quadraten oder umgekehrt), ließ sich zunächst eine Polarität mit zahlreichen Abstufungen identifizieren. Wie Tabelle 6.19 veranschaulicht, zeigt sich in den Dokumenten-Portraits eine Verteilung ausgehend von Teilnehmer_innen, die beim Korrigieren der vier Schülerlösungstexte nur wenig zwischen fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung wechselten (z. B. Herr Balke und Herr Jonuzi) hin zu Teilnehmer_innen, bei denen ein solcher Wechsel häufig stattfand (z. B. Herr Dassow und Herr Uckermark).

Aufgrund der Auffälligkeit im von Tajmel (2017b) untersuchten Fallbeispiel, dass positive (aber negativ konnotierte) Äußerungen zu fachlich-konzeptuellen Aspekten eines Schülerlösungstextes zum Teil gepaart mit negativen Anmerkungen bezüglich ihrer sprachlichen Realisierung auftreten (vgl. Abschnitt 3.2.2), ließ sich vermuten, dass sich bei den Teilnehmer_innen, die häufig zwischen fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung wechseln und bei denen diese Wechsel in ihrem Laut-Denk-Protokoll unmittelbar aufeinander folgend sind, auch Muster zu finden wären, die für eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile sprechen (Forschungsfrage (F2)). Da es sich bei den Dokumenten-Portraits allerdings um verjüngte Abbilder des chronologischen Verlaufs der Laut-Denk-Protokolle handelt, ließ sich keine Aussage darüber treffen, bei welchen Wechseln zwischen fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung in den Dokumenten-Portraits es sich auch um unmittelbar aufeinander folgende, mitvokalisierte Gedankenschritte der Teilnehmer_innen handelte. Aufgrund dessen wurden die codierten Laut-Denk-Protokolle aller

	Anzahl der Wechsel im Dokumenten-Portrait zwischen dem Feststellen und Beurteilen des fachlich-konzeptuellen Eindrucks und dem Feststellen und Beurteilen der sprachlichen Realisierung				
	1 bis 5	**6 bis 10**	**11 bis 15**	**15 bis 20**	**21 bis 25**
Lehrkraft	B, J, O	A, E, F, I, L, P, Q, R, T	C, H, K, N, S	G, M	D, U

Tabelle 6.19.: Anzahl der Wechsel in den Dokumenten-Portraits zwischen dem Feststellen und Beurteilen des fachlich-konzeptuellen Eindrucks der Schülerlösungstexte und dem Feststellen und Beurteilen der sprachlichen Realisierung.

Teilnehmer_innen gezielt nach unmittelbar aufeinander folgenden Segmenten durchsucht, in denen fachliche-konzeptionelle (Codierung mit Subsubkategorie 3.0.1.1) und sprachbezogene Eindrücke (Codierung mit Subsubkategorie 3.0.1.2) in der einen oder anderen Reihenfolge unmittelbar aufeinander geäußert wurden. Wie aus Tabelle 6.20 hervorgeht, fanden sich insgesamt 69 Fundstellen für derartig codierte, unmittelbar aufeinanderfolgende Segmente. Bei der Mehrheit dieser Fundstellen (26+10=36) handelte es sich...

... entweder um eine positiv wertende/akzeptierende Äußerung zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes (Codierung des ersten Segments mit Subsubsubkategorie 3.0.1.1 und 3.0.3.1) gefolgt von einer negativ wertenden/ablehnenden Äußerung bezüglich seiner sprachlichen Realisierung (Codierung des zweiten Segments mit Subsubsubkategorie 3.0.1.2 und 3.0.3.2)...

... oder um eine negativ wertende/ablehnende Äußerung zur sprachlichen Realisierung eines Schülerlösungstextes (Codierung des ersten Segments mit Subsubsubkategorie 3.0.1.2 und 3.0.3.2) gefolgt von einer positiv wertenden/akzeptierenden Äußerung bezüglich seines fachlich-konzeptuellen Eindrucks (Codierung des zweiten Segments mit Subsubsubkategorie 3.0.1.1 und 3.0.3.1).

Bei diesen 36 Fundstellen zeigte sich, dass sich diese entweder auf Schülerlösungstext A oder C bezogen, also auf diejenigen Schülerlösungstexte, die gemäß der Vorauswahl in der Entwicklungsstudie eine hohe fachlich-konzeptuelle Qualität aufweisen (zur Illustration vgl. Tabelle 6.21 und 6.22; für die Teilnehmer_innen mit mehreren Fundstellen ist in Tabelle 6.21 bzw. 6.22 jeweils die für alle Fundstellen am ehesten repräsentative dargestellt). Wie aus Tabelle 6.20 ferner hervorgeht, handelt es sich bei den benannten 36 Fundstellen um die einzigen, die in den Laut-Denk-Protokollen mehrerer Lehrkräfte mehrfach auffindbar waren. Diese 36 Fundstellen ließen sich daher fallübergreifend miteinander vergleichen. Hierdurch konnten die folgenden drei Muster identifiziert werden, wobei insbesondere die mit den Nummern (2a) und (2c) versehenen Muster einen Hinweis auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile darstellen:

Äußerung zum/zur... (Art der Äußerung)		Äußerung zum/zur... (Art der Äußerung)	Anzahl aller Fundstellen	Lehrkräfte mit mindestens einer Fundstelle	Lehrkräfte mit mehreren Fundstelle
fachlich-konzeptuellen Eindruck (negativ wertend/ablehnend)	**...gefolgt von...**	sprachlichen Realisierung (negativ wertend/ablehnend)	8	C, E, G, K, M, P, S, U	---
sprachlichen Realisierung (negativ wertend/ablehnend)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (negativ wertend/ablehnend)	6	D, E, J, K, S	S
fachlich-konzeptuellen Eindruck (positiv wertend/akzeptierend)	**...gefolgt von...**	sprachlichen Realisierung (negativ wertend/ablehnend)	26	B, C, D, G, H, J, K, N, Q, R, S, T, U	C, D, N, Q, S, T, U
sprachlichen Realisierung (negativ wertend/ablehnend)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (positiv wertend/akzeptierend)	10	D, G, K, N, T	D, G, K
fachlich-konzeptuellen Eindruck (negativ wertend/ablehnend)	**...gefolgt von...**	sprachlichen Realisierung (positiv wertend/akzeptierend)	4	I, J, U	U
sprachlichen Realisierung (positiv wertend/akzeptierend)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (negativ wertend/ablehnend)	2	I, U	---
fachlich-konzeptuellen Eindruck (positiv wertend/akzeptierend)	**...gefolgt von...**	sprachlichen Realisierung (positiv wertend/akzeptierend)	0	---	---
sprachlichen Realisierung (positiv wertend/akzeptierend)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (positiv wertend/akzeptierend)	6	C, G, I, K, N, O	---
fachlich-konzeptuellen Eindruck (neutral/gemischt/sonstig)	**...gefolgt von...**	sprachlichen Realisierung (positiv wertend/akzeptierend)	0	---	---
sprachlichen Realisierung (positiv wertend/akzeptierend)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (neutral/gemischt/sonstig)	0	---	---
fachlich-konzeptuellen Eindruck (neutral/gemischt/sonstig)	**...gefolgt von...**	sprachlichen Realisierung (negativ wertend/ablehnend)	1	D	---
sprachlichen Realisierung (negativ wertend/ablehnend)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (neutral/gemischt/sonstig)	0	---	---
fachlich-konzeptuellen Eindruck (positiv wertend/akzeptierend)	**...gefolgt von...**	sprachlichen Realisierung (neutral/gemischt/sonstig)	0	---	---
sprachlichen Realisierung (neutral/gemischt/sonstig)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (positiv wertend/akzeptierend)	4	N, D, H	H
fachlich-konzeptuellen Eindruck (negativ wertend/ablehnend)	**...gefolgt von...**	sprachlichen Realisierung (neutral/gemischt/sonstig)	0	---	---
sprachlichen Realisierung (neutral/gemischt/sonstig)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (negativ wertend/ablehnend)	1	T	---
fachlich-konzeptuellen Eindruck (neutral/gemischt/sonstig)	**...gefolgt von...**	sprachlichen Realisierung (neutral/gemischt/sonstig)	0	---	---
sprachlichen Realisierung (neutral/gemischt/sonstig)	**...gefolgt von...**	fachlich-konzeptuellen Eindruck (neutral/gemischt/sonstig)	1	M	---

Tabelle 6.20.: Anzahl der unmittelbaren Wechsel in den Laut-Denk-Protokollen der Teilnehmer_innen zwischen dem Feststellen und Beurteilen des fachlich-konzeptuellen Eindrucks eines Schülerlösungstextes und dem Feststellen und Beurteilen seiner sprachlichen Realisierung, differenziert nach Art der Äußerung.

Muster in den Fundstellen, in denen einer positiv wertenden/akzeptierenden Äußerung zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes eine negativ wertende/ablehnende Äußerung zur sprachlichen Realisierung folgt (vgl. Tabelle 6.21)

(2a) Es gibt Fundstellen, in denen dem_der Verfasser_in von Schülerlösungstext A oder C zunächst zugestanden wird, fachlich-konzeptuell die Aufgabe Weltraumspaziergang bzw. Teile dieser gelöst zu haben (in Tabelle 6.21: Herr Carboni, Seg. 239-240,

391-392; Herr Dassow, Seg. 262-263, Seg. 390-391; Herr Geppert[143], Seg. 452-454; Herr Jonuzi, Seg. 185-186; Frau Kirik, Seg. 240-241; Herr Quezada, Seg. 176-177; Frau Sohm, Seg. 151-152). Zum Teil ist dieses Zugeständnis eher widerwillig, da es zugleich negativ konnotiert ist und/oder sprachlich relativiert wird (z. B. „[m]ir ist zwar (.) klar, was der Schüler meint“ (Herr Carboni, Seg. 236) oder „[i]m Grunde ist es ja richtig“ (Frau Kirik, Seg. 424)). Diesem Zugeständnis wird anschließend entweder eine global formulierte, negativ wertende/ablehnende Äußerung bezüglich der sprachlichen Realisierung des Schülerlösungstextes (z. B. „aber natürlich (.) ist das von der Formulierung her nicht wirklich das, mhm was man von äh ihm oder ihr erwartet“ (Herr Carboni, Seg. 240)) oder das Fehlen bestimmter Fachbegriffe als Gegensatz gegenübergestellt (z. B. „wobei der Fachbegriff [Medium; M. S. F.] auch da nich- genannt wurde“ (Herr Dassow, Seg. 263)). Hier deutet sich also die Bewertungslogik an, fachlich-konzeptuell richtige oder anschlussfähige Denkfiguren in einem Schülerlösungstext zu relativieren, wenn in diesem Schülerlösungstext die Sprachgebrauchserwartung einer Lehrkraft nicht hinreichend erfüllt wurde.

(2b) Es gibt weitere Fundstellen, in denen dem_der Verfasser_in von Schülerlösungstext A oder C ebenfalls zunächst (eher widerwillig) zugestanden wird, fachlich-konzeptuell die Aufgabe Weltraumspaziergang bzw. Teile dieser gelöst zu haben (in Tabelle 6.21: Herr Balke, Seg. 149-150; Herr Hastedt, Seg. 282-283; Frau Novack, Seg. 240-241, Seg. 492-493; Herr Rittershaus, Seg. 276-277; Frau Sohm, Seg. 272-273; Herr Trummer, Seg. 435-436, 376-377; Herr Uckermark, Seg. 339-340). Die sich anschließende negativ wertende/ablehnende Äußerung zur sprachlichen Realisierung des Schülerlösungstextes wird allerdings von diesem Zugeständnis abgegrenzt, sie wird einräumend untergeordnet, sowie zum Teil auch relativiert (z. B. „also nur’n bisschen komisch ausgedückt“ (Herr Balke, Seg. 150) oder „[a]uch wenn die Fachsprache [...] ganz schlecht ist“ (Herr Uckermark, Seg. 340)). Hier deutet sich daher die Bewertungslogik an, fachlich-konzeptuelle Merkmale von Merkmalen der sprachlichen Realisierung zu trennen und sprachliche Merkmale eines Schülerlösungstextes zu relativieren, wenn die Feststellung und Beurteilung fachlich-konzeptueller Merkmale momentan im Vordergrund steht.

Muster in den Fundstellen, in denen einer negativ wertenden/ablehnenden Äußerung zur sprachlichen Realisierung eines Schülerlösungstextes eine positiv wertende/akzeptierende Äußerung zum fachlich-konzeptuellen Eindruck folgt (vgl. Tabelle 6.22)

(2c) In den Fundstellen, die mit einer negativ wertenden/ablehnenden Äußerung zur sprachlichen Realisierung von Schülerlösungstext A oder C beginnen, wird zunächst hervorgehoben, dass bestimmte Aspekte ihrer sprachlichen Realisierung nicht den

[143]Im Laut-Denk-Protokoll von Herrn Geppert trat die Besonderheit auf, dass eine negativ wertende/ablehnende Äußerung zur sprachlichen Realisierung von Schülerlösungstext A eine positiv wertende Äußerung bezüglich seines fachlich-konzeptuellen Eindruck folgte, die wiederum von einer negativ wertenden/ablehnenden Äußerung zur seiner sprachlichen Realisierung gefolgt wurde. In Ihrer Gesamtheit sind diese drei aufeinanderfolgenden Transkriptsegmente am besten dem Muster (2a) zuordenbar.

eigenen Erwartungen genügen (z. B. „es fehlt eben die Fachsprache“ (Frau Kirik, Seg. 430)). Zum Teil werden diese Erwartungen relativiert („[z]war nich- mein[e] äh speziellen Fachwörter“ (Frau Novack, Seg. 493) oder „vom Aufbau her -n bisschen merkwürdich“ (Herr Trummer, Seg. 286)). Insbesondere wird der negativ wertenden/ablehnenden Äußerung zur sprachlichen Realisierung – außer in der Fundstelle im Laut-Denk-Protokoll von Herrn Geppert[143] –, eine (global formulierte) positiv wertende/akzeptierende Äußerung zum fachlich-konzeptuellen Eindruck als Gegensatz gegenübergestellt (z. B. „[a]ber im All ist nichts das ist ja sowas wie Vakuum.“ (Frau Kirik, Seg. 431) oder „obwohl sie ja schon äh das Ganze schon ganz gut erklärt hat“ (Herr Dassow, Seg. 396)). Hier deutet sich daher die Bewertungslogik an, sprachliche Mängel eines Schülerlösungstextes zu relativieren, wenn in diesem die fachlich-konzeptuellen Erwartungen der Lehrkraft (in Teilen) erfüllt wurden.

6.3.2.4.4. Limitation und Zwischenfazit

In Phase 2b konnten verschiedene Muster in den Laut-Denk-Protokollen identifiziert werden. Bei diesen Mustern handelt es sich um qualitative Prozessaspekte, die sich als unterschiedliche Bewertungslogiken der Teilnehmer_innen bei ihrer laut-denkenden Korrektur der Schülerlösungstexte interpretieren lassen. Zentrale Befunde sind...

... die Tendenz zur Defizitorientierung bei der Feststellung und Beurteilung der sprachlichen Realisierung der Schülerlösungstexte (Muster (1a) bis (1d)).

... die Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile (Bewertungslogik einer Relativierung des fachlich-konzeptuellen Eindrucks, aufgrund der sprachlichen Realisierung bzw. umgekehrt; Muster (2a) und (2c)).

... die Bewertungslogik einer Trennung fachlich-konzeptueller und sprachlicher Leistungsurteile (Muster (2b)).

Anzumerken ist, dass sich an dieser Stelle der vorliegenden Arbeite keine Aussagen darüber treffen lassen, ob die in den Laut-Denk-Protokollen identifizierten Muster im gesamten Leistungsurteilsgeneseprozess der Teilnehmer_innen auch eine tragende Rolle eingenommen haben. Grund hierfür ist, dass die Dokumenten-Portraits, die als Grundlage für die in Phase 2b vorgenommene Analyse dienten, lediglich ein verjüngtes (und damit nicht vollständiges) Abbild des chronologischen Verlaufs ausgewählter Gedankenfolgen darstellen, die die Teilnehmer_innen im Rahmen der Laborsituation laut-denkend mitvokalisiert haben.

Lehrkraft	Feintranskript mit Segementierung (positive wertende/akzeptierende Äußerung zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes gefolgt von einer negativ wertenden/ablehenden Äußerung zur seiner sprachlichen Realisierung)	
	Schülerlösungstext A	**Schülerlösungstext C**
Herr Balke	[149] Im Grunde (.) is- das Ph- hm ähm also hm is- (.) Ph- Phänomen 1 is alles ähm (.) alles gesagt worden. [150] Also nur'n bisschen komisch ausgedrückt.	---
Herr Carboni	[239] (...) Mir ist zwar (.) klar, was der Schüler meint, [240] aber natürlich (.) ist das von der Formulierung her nicht wirklich das, mhm was man von äh ihm oder ihr erwartet.	[391] Äh innerhal- inhaltlich ist das ja auch nicht falsch. [392] Man kann's vielleicht besser formulieren.
Herr Dassow	[262] Es wird zwar schon erkannt, dass die, ähm (.) dass äh im All kein Medium is-, [263] wobei der Fachbegriff auch da nich- genannt wurde.	[390] Ähm das is- richtig. [391] Ähm auf der anderen Seite w- würde hier jetz- auch wieder der Fachbegriff d- des Mediums, also dass das Medium fehlt, in dem es sich ausbreiten kann...
Herr Geppert	[453] Also generell (.) würd- ich hier sagen, der Schüler hat das Phänomen... also e- er kennt die Erklärung, [454] aber sagt ähm (.) aber benutzt nich- die... die physikalischen Fachbegriffe und die... und äh wie man das schreiben würde.	---
Herr Hastedt	[282] is- im Prinzip... (..) mhm (..) den Punkt kann ich geben. [283] Es fällt zwar der Begriff Vakuum nich-.	---
Herr Jonuzi	[185] weil ich finde, dass er ähm erkannt hat, was das eigentliche Problem ist. [186] Hat's schlecht formuliert.	---
Frau Kirik	[424] (..) Im Grunde ist es ja richtig. [425] Also was hier nicht verwendet wurde is- Fachsprache.	---
Frau Novack	[240] Das wäre ja so (.) mein (.) Medium. [241] Auch wenn dann jetz- hier als äh Fachbegriff nich- kommt.	[429] Und ähm ja alle wesentlichen Punkte enthalten sind. [430] (.) Zwar nich- mit mein äh speziellen Fachworten, die ich in der (.) im Erwartungshorizont benutzt habe.
Herr Quezada	[176] Okay, die Situation wird verstanden. [177] Aber die Verbalisierung des Ganzen ist einfach noch nich- da.	---
Herr Rittershaus	[276] Also hier wurden auf jeden Fall [„Phänomen und Erklärung genannt" (Seg. 279); M. S. F.], [277] wenn auch (.) sehr fachsprachlich knapp und f-...	---
Frau Sohm	[151] Das heißt, der oder die Schülerin hat erkannt, dass Schall ein Medium benöticht. [152] Wobei das Wort Schall nicht auftaucht.	[272] Warum der nich- gehört wird, dafür gibt es 1 Punkt. [273] Geleitet is- nich- ganz richtich, aber das füg- ich gleich beim Tipp dazu.
Herr Trummer	[435] Also derjenige hat schon verstanden worum es geht. [436] Is- aber jetz- rein von der Ausdrucksweise zu bemängeln.	[376] Insofern weiß derjenige wovon er redet. [377] Auch wenn er die Stimme schreibt.
Herr Uckermark	[339] Weil ähm ich erkenne... also er oder sie hat doch gemerkt, dass es hier um die (.) Schallausbreitung geht. [340] Auch wenn die Fachsprache ähm ganz äh ganz schlecht ist.	---

Tabelle 6.21.: Ausgewählte Fundstellen aus den Laut-Denk-Protokollen für eine positiv wertende/akzeptierende Äußerung zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes gefolgt von einer negativ wertenden/ablehnenden Äußerung zur seiner sprachlichen Realisierung.

Lehrkraft	**Feintranskript mit Segementierung** (negativ wertende/ablehnende Äußerung zur sprachlichen Realisierung eines Schülerlösungstextes gefolgt von einer positiv wertenden/akzeptierenden Äußerung zur seinem fachlich-konzeptuellen Eindruck)	
	Schülerlösungstext A	**Schülerlösungstext C**
Herr Dassow	[303] Ne, des is- ja (.) die sprachliche Unklarheiten. [304] Er meint allerdings das richtige.	[395] Ähm würde da jetz- f- mhm noch fehlen [Begriff Medium, M. S. F], [396] obwohl sie ja schon äh das Ganze schon ganz gut erklärt hat.
Herr Geppert	[452] Is- auch wieder dieses Ton. [453] Also generell (.) würd- ich hier sagen, der Schüler hat das Phänomen... also e-... er kennt die Erklärung.	---
Frau Kirik	[430] (..) Na ja, also diese Begründung ist... es fehlt eben die Fachsprache. [431] Aber im All ist nichts das ist ja sowas wie Vakuum.	---
Frau Novack	---	[430] (.) Zwar nich- mit mein äh speziellen Fachworten, die ich in der (.) im Erwartungshorizont benutzt habe. [431] Aber inhaltlich damit (.) ähm meines Erachtens voll erfasst.
Herr Trummer	---	[286] Na (.) gut, so jetz- von -ner... vom Aufbau her -n bisschen merkwürdich. [287] Aber man weiß was gemeint is-.

Tabelle 6.22.: Ausgewählte Fundstellen aus den Laut-Denk-Protokollen für eine negativ wertende/ablehnende Äußerung zur sprachlichen Realisierung eines Schülerlösungstextes gefolgt von einer positiv wertenden/akzeptierenden Äußerung zur seinem fachlich-konzeptuellen Eindruck.

6.3.2.5. Phase 2c: Quantitative Analyse ausgewählter inhaltlicher Facetten der Laut-Denk-Protokolle[144]

6.3.2.5.1. Methodische Vorbemerkungen

Ziel von Phase 2c der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle war eine vertiefte Beschreibung und Interpretation ausgewählter inhaltlicher Facetten der Laut-Denk-Protokolle aller Teilnehmer_innen. Dies geschah mittels einer quantitativen Analyse der Codierhäufigkeiten besonders reichhaltiger (Subsubsub-)Kategorien des Kategoriensystems aus Phase 1. Grundlage für diese Analyse bildeten die Segmente in den Laut-Denk-Protokollen, die der Subkategorie „Feststellung und Beurteilung eines Schülerlösungstextes“ (Subkategorie 3.0) zugewiesen wurden. Grund hierfür ist, dass diese Segmente insgesamt mit vier verschiedenen (Subsubsub-)Kategorien codiert wurden[145] und somit – wie in Phase 2b bereits auf qualitativer Ebene geschehen – die Suche nach auffälligen Muster bei verschiedensten Codingkombinationen ermöglichten. Die zentralen Schritte, durch welche die Codierung dieser Segmente aufbereitet und mit Hilfe nichtparametrischer statistischer Methoden analysiert wurden, lassen sich wie folgt zusammenfassen:

- Erstens wurden für jedes der 21 Laut-Denk-Protokolle die absoluten Häufigkeiten der Segmente bestimmt,...
  - (α) ... die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) codiert wurden.
  - (β) ... die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden.
  - (γ) ... die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden.
  - (δ) ... die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden.
  - (ϵ) ... die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden.

[144] Teile dieses Abschnitts stellen eine überarbeitete und erweiterte Fassung von Feser & Höttecke (2019), sowie Feser & Höttecke (im Druck) dar.

[145] Jedem dieser Segment wurde aus jeder der Subsubkategorien 3.0.1, 3.0.2 und 3.0.3 jeweils genau eine Subsubsubkategorie zugewiesen, sowie eine der zusätzlichen Kategorien „Auseinandersetzung mit Schülerlösungstext A“, „Auseinandersetzung mit Schülerlösungstext B“, „Auseinandersetzung mit Schülerlösungstext C“ oder „Auseinandersetzung mit Schülerlösungstext D“ (vgl. Unterabschnitt 6.3.2.2).

(ζ) ... die mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden.

- Zweitens wurde jede dieser absoluten Häufigkeiten von Codingkombinationen pro Laut-Denk-Protokoll in geeignete prozentuelle Häufigkeiten umgewandelt (Details siehe Ergebnisdarstellung). Diese Umwandlungen waren notwendig, da sich die Laut-Denk-Protokolle der Teilnehmer_innen bezüglich ihrer Segmentgesamtanzahl zum Teil deutlich voneinander unterscheiden (vgl. Tabelle 6.4) und daher eine fallübergreifende Analyse der absoluten Häufigkeiten von Codingkombinationen nicht möglich war.
- Drittens wurden der Stichprobenmedian und der Interquartilsabstand als Lage- bzw. Streuungsmaß für die prozentuellen Häufigkeiten jeder dieser Codingkombinationen bestimmt.
- Viertens wurden 2-seitige Wilcoxon-Vorzeichen-Rang-Test durchgeführt, sowie das Effektstärkemaß ES berechnet (vgl. auch Abschnitt 6.3.1). Dies geschah, um darüber Auskunft zu erhalten, bei welchen Codingkombinationen, die inhaltlich sinnvoll miteinander vergleichbar sind (Details siehe Ergebnisdarstellung), sich deren Mediane signifikant voneinander unterscheiden und wie stark diese Unterschiede sind. Bei allen durchgeführten Tests wurde das Signifikanzniveau $\alpha = .05$ gewählt.
- Fünftens wurden die in den vorherigen Schritten gewonnenen statistischen Kennwerte für die einzelnen Codingkombination-Häufigkeiten nach Auffälligkeiten durchsucht, sowie diese im Sinne von Forschungsfrage (F1) und (F2) interpretiert.

Die Ergebnisse, die durch die eben beschriebene Verfahrensweise gewonnenen wurden, werden im Folgenden dargestellt.

6.3.2.5.2. Quantitative Analyse der medianen prozentuellen Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) codiert wurden

In Tabelle 6.23 sind für die Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) codiert wurden, die Ergebnisse der in Phase 2c vorgenommenen Aufbereitung und Analyse zusammengefasst. Tabelle 6.23 ist dabei wie folgt zu lesen:

1. **In die Tabellendiagonale** sind die Stichprobenmediane (die Interquartilsabstände) für die prozentuellen Häufigkeiten der einzelnen Codingkombinationen eingetragen. Zum Beispiel bedeutet der Eintrag am linken Ende der Tabellendiagonalen, dass von den Segmenten, die je Laut-Denk-Protokoll mit Kategorie „A“ codiert wurden, im Median 37.2 % zusätzlich mit Subsubsubkategorie 3.0.1.1 (Fokussierung des fachlich-konzeptuellen Eindrucks) codiert wurden, sowie dass der Interquartils-

abstand der pronzentuellen Häufigkeit dieser Codingkombination 23.1 % beträgt. Für die übrigen Einträge in der Tabellendiagonalen gilt Analoges.

2. **Unterhalb der Tabellendiagonalen** ist für die prozentuellen Häufigkeiten der Codingkombinationen, die inhaltlich sinnvoll miteinander vergleichbar sind, das Effektstärkemaß ES als Kennwert für das Ausmaß des Medianunterschieds der jeweiligen Codingkombinationen angegeben, sowie ob es sich hierbei um einen signifikanten Unterschied handelt (Ergebnis des entsprechenden 2-seitigen Wilcoxon-Vorzeichen-Rang-Tests). Zum Beispiel bedeutet der in der ersten Spalte von Tabelle 6.23 von oben gezählt...

   ... zweite Eintrag, dass sich zwischen den prozentuellen Häufigkeiten der Codingkombination aus Kategorie „A“ und Subsubsubkategorie 3.0.1.1 (Fokussierung des fachlich-konzeptuellen Eindrucks), sowie der Codingkombination aus Kategorie „A“ und Subsubsubkategorie 3.0.1.2 (Fokussierung der sprachlichen Realisierung) ein schwacher, statistisch nicht signifikanter Medianunterschied zeigt ($ES = .29$; $p > .05$).

   ... fünfte Eintrag, dass sich zwischen den prozentuellen Häufigkeiten der Codingkombination aus Kategorie „A“ und Subsubsubkategorie 3.0.1.1 (Fokussierung des fachlich-konzeptuellen Eindrucks), sowie der Codingkombination aus Kategorie „B“ und Subsubsubkategorie 3.0.1.1 (Fokussierung des fachlich-konzeptuellen Eindrucks) ein starker, statistisch signifikanter Medianunterschied zeigt ($ES = .58$; $p \leq .001$).

   Für die übrigen Einträge unterhalb der Tabellendiagonalen gilt Analoges.

Auffälligkeiten in den deskriptiven Kennwerten und Ergebnissen der Wilcoxon-Vorzeichen-Rang-Tests

In den in Tabelle 6.23 aufgeführten Kennwerten zeigen sich die folgenden Auffälligkeiten:

(α1) Bei Schülerlösungstext A wurden von den Teilnehmer_innen der fachlich-konzeptuelle Eindruck, die sprachliche Realisierung, sowie mehrere/uneindeutige Merkmale im Median in vergleichbarer Häufigkeit herangezogen. Zwischen der prozentuellen Häufigkeit dieser Merkmalsfoki zeigt sich jeweils kein statistisch signifikanter Medianunterschied. Sonstige Merkmale spielten hingegen mit einer medianen prozentuellen Häufigkeit von 0.0 % de facto keine Rolle. Ferner zeigt sich zwischen sonstigen Merkmalen und den übrigen codierten Merkmalsfoki jeweils ein starker, statistisch signifikanter Medianunterschied ($.60 \leq ES \leq .62$).

(α2) Bei Schülerlösungstext B wurde von den Teilnehmer_innen im Median am häufigsten der fachlich-konzeptuelle Eindruck herangezogen (Median = 66.7 %). Zwischen dem fachlich-konzeptuellen Eindruck und den übrigen codierten Merkmalsfoki zeigt

		A				B				C				D			
		Fach (3.0.1.1)	**Sprache** (3.0.1.2)	**sonstig** (3.0.1.3)	**mehrere** (3.0.1.4)	**Fach** (3.0.1.1)	**Sprache** (3.0.1.2)	**sonstig** (3.0.1.3)	**mehrere** (3.0.1.4)	**Fach** (3.0.1.1)	**Sprache** (3.0.1.2)	**sonstig** (3.0.1.3)	**mehrere** (3.0.1.4)	**Fach** (3.0.1.1)	**Sprache** (3.0.1.2)	**sonstig** (3.0.1.3)	**mehrere** (3.0.1.4)
A	**Fach** (3.0.1.1)	**37.2% (23.1%)**															
	Sprache (3.0.1.2)	.29	**19.1% (25.4%)**														
	sonstig (3.0.1.3)	.61***	.60***	**0.0% (5.3%)**													
	mehrere (3.0.1.4)	.25	.03	.62***	**25.0% (23.5%)**												
B	**Fach** (3.0.1.1)	.58***				**66.7% (18.8%)**											
	Sprache (3.0.1.2)		.62***			.62***	**0.0% (0.0%)**										
	sonstig (3.0.1.3)			.32*		.62***	.11	**0.0% (0.0%)**									
	mehrere (3.0.1.4)				.12	.60***	.55***	.60***	**21.4% (20.0%)**								
C	**Fach** (3.0.1.1)	.13				.62***				**35.9% (25.3%)**							
	Sprache (3.0.1.2)		.31*				.55***			.37*	**18.9% (15.3%)**						
	sonstig (3.0.1.3)			.39**				.53***		.56***	.44**	**8.1% (9.2%)**					
	mehrere (3.0.1.4)				.18				.38*	.07	.47**	.61***	**37.1% (21.8%)**				
D	**Fach** (3.0.1.1)	.34*				.59***				.42**				**52.4% (19.4%)**			
	Sprache (3.0.1.2)		.39**				.50***				.15			.62***	**11.1% (22.6%)**		
	sonstig (3.0.1.3)			.03				.31				.35*		.61***	.39*	**0.0% (5.3%)**	
	mehrere (3.0.1.4)				.04				.17				.28	.41**	.36*	.59***	**28.1% (18.4%)**

Tabelle 6.23.: Zusammenfassung der Lageparameter-Analyse für die prozentuellen Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) codiert wurden. Kennwerte in der Tabellendiagonalen: Median (Interquartilsabstand) der jeweiligen Codingkombination. Kennwerte unterhalb der Tabellendiagonalen: Effektstärkemaß *ES* für den Medianunterschied der jeweiligen Codingkombinationen (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$; 2-seitiger Wilcoxon-Vorzeichen-Rang-Test).

sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.60 \leq ES \leq .62$). Dem fachlich-konzeptuellen Eindruck folgen mehrere/uneindeutige Merkmale mit einer medianen prozentuellen Häufigkeit von 21.4 %. Die sprachliche Realisierung, sowie sonstige Merkmale spielten hingegen mit einer medianen prozentuellen Häufigkeit von jeweils 0.0 % de facto keine Rolle.

(α3) Bei Schülerlösungstext C wurde von den Teilnehmer_innen vor allem entweder der fachlich-konzeptuelle Eindruck (Median = 35.9 %), oder mehrere/uneindeutige Merkmale herangezogen (Median = 37.1 %). Zwischen der prozentuellen Häufigkeit dieser Merkmalsfoki zeigt sich kein statistisch signifikanter Medianunterschied; zu den übrigen codierten Merkmalsfoki zeigt sich jeweils ein moderat bis starker, statistisch signifikanter Medianunterschied ($.37 \leq ES \leq .61$). Diesen beiden Merkmalsfoki folgen die sprachliche Realisierung mit einer medianen prozentuellen Häufigkeit von 18.9 %, sowie sonstige Merkmale mit einer medianen prozentuellen Häufigkeit von 8.1 %. Zwischen der prozentuellen Häufigkeit der beiden zuletzt genannten Merkmalsfoki zeigt sich zudem ein moderater, statistisch signifikanter Medianunterschied ($ES = .44$).

(α4) Bei Schülerlösungstext D zeigt sich zwischen den prozentuellen Häufigkeiten aller codierten Merkmalsfoki ein moderater bis starker, statistisch signifikanter Medianunterschied ($.36 \leq ES \leq .62$). Der fachlich-konzeptuelle Eindruck wurde dabei mit einer medianen prozentuellen Häufigkeit von 52.4 % am häufigsten für die Leistungsurteilsgenese herangezogen. Dem folgen in absteigender Reihenfolge mehrere/uneindeutige Merkmale (Median = 28.1 %), die sprachliche Realisierung (Median = 11.1 %) und sonstige Merkmale (Median = 0.0 %).

(α5) Der fachlich-konzeptuelle Eindruck wurde von den Teilnehmer_innen bei Schülerlösungstext B im Median am häufigsten fokussiert (Median = 66.7 %). Zu der prozentuellen Häufigkeit, in der der fachlich-konzeptuelle Eindruck bei den Schülerlösungstexten A, C und D fokussiert wurde, zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.58 \leq ES \leq .62$). Am seltensten wurde der fachlich-konzeptuelle Eindruck bei Schülerlösungstext A und C fokussiert (Median Schülerlösungstext A = 37.2 %; Median Schülerlösungstext C = 35.9 %; kein statistisch signifikanter Medianunterschied). Diesbezüglich im Mittelfeld liegt Schülerlösungstext D (Median 52.4 %), mit einem moderaten bis starken, statistisch signifikanten Medianunterschied zu den Schülerlösungstexten A, B und C ($.34 \leq ES \leq .59$).

(α6) Die sprachliche Realisierung wurde von den Teilnehmer_innen bei Schülerlösungstext A im Median am häufigsten fokussiert (Median = 19.1 %). Zu der prozentuellen Häufigkeit, mit der die sprachliche Realisierung bei den Schülerlösungstexten B, C und D fokussiert wurde, zeigt sich jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.31 \leq ES \leq .62$). Am seltensten wurde die sprachliche Realisierung bei Schülerlösungstext B fokussiert (Median = 0.0 %; jeweils ein starker, statistisch signifikanter Medianunterschied zu den Schülerlösungstexten A, C und D). Diesbezüglich im Mittelfeld liegen Schülerlösungstext C und D (Median

Schülerlösungstext C = 18.9 %; Median Schülerlösungstext D = 11.1 %; kein statistisch signifikanter Medianunterschied).

(α7) Sonstige Merkmale wurden von den Teilnehmer_innen im Median lediglich bei Schülerlösungstext C fokussiert (Median = 8.1 %). Zu der prozentuellen Häufigkeit, in der sonstige Merkmale bei den Schülerlösungstexten A, B und D fokussiert wurden, zeigt sich jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.32 \leq ES \leq .53$). Bei den übrigen drei Schülerlösungstexten spielten sonstige Merkmale mit einer medianen prozentuellen Häufigkeit 0.0 % de facto keine Rolle. Der signifikante Ausgang des Wilcoxon-Vorzeichen-Rang-Tests zwischen den prozentuellen Häufigkeiten, in denen sonstige Merkmale bei Schülerlösungstext A und B fokussiert wurden, lässt sich auf den entsprechenden Streuungsunterschied zwischen diesen beiden Schülerlösungstexten zurückführen (IQR Schülerlösungstext A = 5.3 %; IQR Schülerlösungstext B = 0.0 %). Analoges gilt für Schülerlösungstext B und D, bei denen der entsprechende Wilcoxon-Vorzeichen-Rang-Tests allerdings einen knapp nicht signifikanten Ausgang hatte ($p = .063$).

(α8) Mehrere/Uneindeutige Merkmale wurden von den Teilnehmer_innen im Median bei allen vier Schülerlösungstexte in vergleichbarer Häufigkeit fokussiert (Median Schülerlösungstext A = 25.0 %; Median Schülerlösungstext B = 21.4 %; Median Schülerlösungstext C = 37.1 %; Median Schülerlösungstext D = 28.1 %). Lediglich zwischen Schülerlösungstext B und C zeigt sich diesbezüglich ein moderater, statistisch signifikanter Medianunterschied ($ES = .38$).

Interpretation: quantitative Teilbefunde

Die eben aufgeführten Auffälligkeiten (α1) bis (α8) lassen alles in allem die folgenden Interpretationen zu. Diese decken sich zum Teil mit jenen Interpretationen, die im Rahmen von Phase 2a und 2b der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle vorgenommen wurden (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.4):

1. Aus den Auffälligkeiten (α1) bis (α4) lässt sich schlussfolgern, dass die Teilnehmer_innen bei der Korrektur der vier Schülerlösungstext den fachlich-konzeptuellen Eindruck, die sprachliche Realisierung, sonstige Merkmale und mehrere/uneindeutige Merkmale jeweils unterschiedlich umfangreich fokussierten. Da die Schülerlösungstexte A bis D im Rahmen der Entwicklungsstudie als eine Komposition aus vier konstrastierenden Schülerlösungstexten ausgewählt wurden, ist es plausibel anzunehmen, dass die fachlich-konzeptuelle Qualität, sowie die Qualität der sprachlichen Realisierung der vier Schülerlösungstexte beeinflusst, in welchem Umfang die Teilnehmer_innen welche Merkmale eines Schülerlösungstextes im Rahmen ihrer Leistungsurteilsgenese berücksichtigten, bzw. wie umfangreich sie bei welchem Schülerlösungstext auf Lehrerwissen und -können zur Feststellung und Beurteilung bestimmter Merkmale einer Schülerlösung zurückgegriffen haben.

2. Aus Auffälligkeit (α5) geht hervor, dass die Teilnehmer_innen den fachlich-konzeptuellen Eindruck signifikant häufiger bei den Schülerlösungstexten fokussierten, die eine geringe fachlich-konzeptuelle Qualität aufweisen (Median Schülerlösungstext B = 66.7 %; Median Schülerlösungstext D = 52.4 %), als bei den Schülerlösungstexten, die sich durch eine hohe fachlich-konzeptuelle Qualität auszeichnen (Median Schülerlösungstext A = 37.2 %; Median Schülerlösungstext C = 35.9 %). Hier zeigt sich also ein Hinweis darauf, dass die Teilnehmer_innen die fachlich-konzeptuellen Leistungen in den vier Schülerlösungstexten in einer zum Teil defizitorientierten Art und Weise feststellten und beurteilten.

3. Aus Auffälligkeit (α6) geht zweierlei hervor: Zum einen zeigt sich die Tendenz, dass die Teilnehmer_innen – außer beim direkten Vergleich von Schülerlösungstext C und D – die sprachliche Realisierung signifikant häufiger bei den Schülerlösungstexten fokussierten, die eine hohe fachlich-konzeptuelle Qualität aufweisen (Median Schülerlösungstext A = 19.1 %; Median Schülerlösungstext C = 18.1 %), als bei den Schülerlösungstexten mit einer geringen fachlich-konzeptuellen Qualität (Median Schülerlösungstext B = 0.0 %; Median Schülerlösungstext D = 11.1 %). Hierin zeigt sich ein Hinweis, dass die Teilnehmer_innen eine Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes vor allem dann vornehmen, wenn dieser bis zu einem bestimmten Grad auch ihren fachlich-konzeptuellen Erwartungen entspricht. Allgemein ausgedrückt lässt sich hier also ein empirischer Hinweis auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung ausfindig machen (Forschungsfrage F2). Zum anderen deutet sich bei den Teilnehmer_innen auch eine zum Teil defizitorientierte Feststellung und Beurteilung der sprachlichen Realisierung schriftlicher Schülerleistungen an. Dies zeigt sich darin, dass die Teilnehmer_innen die sprachliche Realisierung am häufigsten bei Schülerlösungstext A beachteten und dass sich, entgegen der eben zuerst genannten Tendenz, zwischen den prozentuellen Häufigkeiten, in denen die sprachliche Realisierung bei Schülerlösungstext C und D fokussiert wurde, kein signifikanter Medianunterschied zeigt.

4. In Auffälligkeit (α7) zeigt sich, dass von den Teilnehmer_innen sonstige Merkmale (z. B. Handschrift oder Textlänge) lediglich bei Schülerlösungstext C in nennenswerter Häufigkeit fokussiert wurden (Median 8.1 %), bzw. dass sonstige Merkmale bei der Korrektur der übrigen drei Schülerlösungstexte de facto keine Rolle spielten (mediane prozentuelle Häufigkeit jeweils 0.0 %). Da Schülerlösungstext C gemäß der Vorauswahl in der Entwicklungsstudie sowohl eine hohe fachlich-konzeptuelle Qualität aufweist, als auch eine hohe Qualität bezogen auf seine sprachliche Realisierung, lässt sich aufgrund von Auffälligkeit (α7) vorsichtig vermuten, dass die Teilnehmer_innen eine Feststellung und Beurteilung sonstiger Merkmale eines Schülerlösungstextes vor allem dann vornehmen, wenn dieser bis zu einem bestimmten Grad sowohl ihren fachlich-konzeptuellen, als auch ihren sprachlichen Erwartungen entspricht.

5. In Auffälligkeit ($\alpha 8$) zeigt sich, dass die Teilnehmer_innen mehrere/uneindeutige Merkmale bei allen vier Schülerlösungstext in vergleichbarer Häufigkeit fokussierten. Dies lässt sich als möglicher Hinweis darauf deuten, dass die Teilnehmer_innen die Leistungen in allen vier Schülerlösungstexten zu einem bestimmten Anteil auch auf Grundlage eines eher holistischen Eindrucks und/oder heuristisch feststellten und beurteilten.

6.3.2.5.3. Quantitative Analyse der medianen prozentuellen Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden

In Tabelle 6.24 sind für die Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden, die Ergebnisse der in Phase 2c vorgenommenen Aufbereitung und Analyse zusammengefasst. Tabelle 6.24 ist analog zu Tabelle 6.23 aufgebaut, und daher in identischer Art und Weise zu lesen.

Auffälligkeiten in den deskriptiven Kennwerten und Ergebnissen der Wilcoxon-Vorzeichen-Rang-Tests

In den in Tabelle 6.24 aufgeführten Kennwerten zeigen sich die folgenden Auffälligkeiten:

($\beta 1$) Schülerlösungstext A wurde von den Teilnehmer_innen im Median vor allem entweder bezüglich sachlicher Kriterien verortet (Median = 50.0 %), oder der Bezug der Verortung war mehr- und/oder uneindeutig (Median = 37.5 %). Zwischen der prozentuellen Häufigkeit dieser Arten der Verortung zeigt sich kein statistisch signifikanter Medianunterschied. Andere Schülerlösungstexte, allgemeine Erfahrungen mit Physiklernenden, sowie mutmaßliche Personenmerkmale des_der Schülers_Schülerin spielten mit einer medianen prozentuellen Häufigkeit von jeweils 0.0 % de facto keine Rolle.

($\beta 2$) Bei den Schülerlösungstexten B, C und D zeigt sich ein nahezu identisches Bild wie bei Schülerlösungstext A. Der einzige wesentliche Unterschied besteht darin, dass bei diesen drei Schülerlösungstexten sachliche Kriterien signifikant häufiger als Bezug der Verortung herangezogen wurden, als dass der Bezug der Verortung mehr- und/oder uneindeutig war ($.53 \leq ES \leq .61$). Der signifikante Ausgang der Wilcoxon-Vorzeichen-Rang-Tests zwischen den prozentuellen Häufigkeiten, in denen Schülerlösungstext C bezüglich anderer Schülerlösungstexte bzw. mutmaßlicher Personenmerkmale des_der Schülers_Schülerin verortet wurde, lässt sich auf den entsprechenden Streuungsunterschied zwischen diesen beiden Arten des Bezugs der Verortung zurückführen (IQR andere Schülerlösungstexte = 5.6 %; IQR mutmaßliche Personenmerkmale = 0.0 %).

		A					**B**					**C**					**D**				
		Kriterium (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)	**Kriterium** (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)	**Kriterium** (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)	**Kriterium** (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)
A	**Kriterium** (3.0.2.1)	**50.0% (24.0%)**																			
	Texte (3.0.2.2)	.61***	**0.0% (0.0%)**																		
	Erfahrung (3.0.2.3)	.58***	.12	**0.0% (4.3%)**																	
	Person (3.0.2.4)	.59***	.28	.23	**0.0% (8.3%)**																
	mehrere (3.0.2.5)	.13	.62***	.62***	.61***	**37.5% (20.5%)**															
B	**Kriterium** (3.0.2.1)	.50***					**68.6% (18.8%)**														
	Texte (3.0.2.2)		.05				.62***	**0.0% (0.0%)**													
	Erfahrung (3.0.2.3)			.10			.62***	.03	**0.0% (0.0%)**												
	Person (3.0.2.4)				.06		.62***	.19	.14	**0.0% (5.7%)**											
	mehrere (3.0.2.5)					.52***	.61***	.53***	.54***	.42**	**18.8% (18.0%)**										
C	**Kriterium** (3.0.2.1)	.48***					.09					**74.4% (13.5%)**									
	Texte (3.0.2.2)		.20					.18				.62***	**0.0% (5.6%)**								
	Erfahrung (3.0.2.3)			.34*					.29			.62***	.37*	**0.0% (0.0%)**							
	Person (3.0.2.4)				.36*					.30		.62***	.30	.28	**0.0% (0.0%)**						
	mehrere (3.0.2.5)					.45**					.10	.53***	.59***	.62***	.59***	**21.9% (11.0%)**					
D	**Kriterium** (3.0.2.1)	.39**					.04					.15					**66.7% (31.3%)**				
	Texte (3.0.2.2)		.22					.16					.04				.62***	**0.0% (5.9%)**			
	Erfahrung (3.0.2.3)			.15					.05					.23			.62***	.22	**0.0% (0.0%)**		
	Person (3.0.2.4)				.03					.06					.31*		.62***	.05	.21	**0.0% (5.0%)**	
	mehrere (3.0.2.5)					.40**					.07					.05	.55***	.61***	.59***	.60***	**22.2% (20.8%)**

Tabelle 6.24.: Zusammenfassung der Lageparameter-Analyse für die prozentuellen Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden. Kennwerte in der Tabellendiagonalen: Median (Interquartilsabstand) der jeweiligen Codingkombination. Kennwerte unterhalb der Tabellendiagonalen: Effektstärkemaß *ES* für den Medianunterschied der jeweiligen Codingkombinationen (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$; 2-seitiger Wilcoxon-Vorzeichen-Rang-Test).

(β3) Am seltensten wurden sachliche Kriterien als Bezug der Verortung bei Schülerlösungstext A herangezogen (Median = 50.0 %). Zu den prozentuellen Häufigkeiten, in denen sachliche Kriterien als Bezug der Verortung bei Schülerlösungstext B, C und D herangezogen wurden, zeigt sich jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.39 \leq ES \leq .50$). Zwischen Schülerlösungstext B, C und D zeigt sich diesbezüglich hingegen kein statistisch signifikanter Medianunterschied.

(β4) Am häufigsten wurden mehr- und/oder uneindeutige Bezüge der Verortung bei Schülerlösungstext A herangezogen (Median = 37.5 %). Zu den prozentuellen Häufigkeiten, in denen mehr- und/oder uneindeutige Bezüge der Verortung bei Schülerlösungstext B, C und D herangezogen wurden, zeigt sich jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.40 \leq ES \leq .52$). Zwischen Schülerlösungstext B, C und D zeigt sich diesbezüglich hingegen kein statistisch signifikanter Medianunterschied.

(β5) Andere Schülerlösungstexte, allgemeine Erfahrungen mit Physiklernenden, sowie mutmaßliche Personenmerkmale des_der Schülers_Schülerin wurden von den Teilnehmer_innen bei keinem Schülerlösungstext in nennenswerter Häufigkeit als Bezug der Verortung herangezogen (mediane prozentuelle Häufigkeit jeweils 0.0 %). Der signifikante Ausgang einzelner Wilcoxon-Vorzeichen-Rang-Tests zwischen entsprechenden prozentuellen Häufigkeiten (vgl. Tabelle 6.24) lassen sich auf die jeweiligen Streuungsunterschiede zwischen den betrachten Arten des Bezugs der Verortung zurückführen.

Interpretation: quantitative Teilbefunde

Die eben aufgeführten Auffälligkeiten (β1) bis (β5) lassen alles in allem die folgenden Interpretationen zu. Diese decken sich zum Teil mit jenen Interpretationen, die im Rahmen von Phase 2a und 2b der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle vorgenommen wurden (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.4):

1. In den Auffälligkeiten (β1) und (β2) zeigt sich, dass die Teilnehmer_innen bei der Korrektur aller vier Schülerlösungstexte überwiegend sachliche Kriterien als Bezug der Verortung heranzogen (Median Schülerlösungstext A = 50.0 %; Median Schülerlösungstext B = 68.6 %; Median Schülerlösungstext C = 74.4 %; Median Schülerlösungstext D = 66.7 %). Hieraus lässt sich vermuten, dass die Teilnehmer_innen Leistungen in allen vier Schülerlösungstexten in einer tendenziell deutlich an der kriterialen Bezugsnorm orientierten Art und Weise feststellten und beurteilten. Zusätzlich gestützt wird diese Interpretation zum einen dadurch, dass andere Schülerlösungstexte, allgemeine Erfahrungen mit Physiklernenden, sowie mutmaßliche Personenmerkmale des_der Schülers_Schülerin bei der Leistungsfeststellung und -beurteilung der einzelnen Schülerlösungstexte jeweils de facto keine Rolle spielten (mediane prozentuelle Häufigkeit jeweils 0.0 %; vgl. auch Auffälligkeit (β5)) und zum anderen, dass die Teilnehmer_innen im Median bei der Korrektur aller vier

Schülerlösungstexte zwar zu einem bedeutenden, jeweils aber nicht mehrheitlichen Anteil auch mehr- und/oder uneindeutige Bezüge der Verortung heranzogen (Median Schülerlösungstext A = 37.5 %; Median Schülerlösungstext B = 18.8 %; Median Schülerlösungstext C = 21.9 %; Median Schülerlösungstext D = 22.2 %). Insbesondere dass Personenmerkmale als Bezug für die Verortung der vier Schülerlösungstexte de facto nicht herangezogen wurden, könnte allerdings ein Artefakt der Erhebung im Rahmen der Laborsituation darstellen. Grund hierfür ist, dass den Teilnehmer_innen im Rahmen der Laborsituation keine Informationen über Personenmerkmale der Schüler_innen, die die Schülerlösungstexte A bis D verfasst haben, zur Verfügung gestellt wurden.

2. In den Auffälligkeiten (β3) und (β4) zeigt sich, dass die eben aufgeführte Tendenz Leistungen in den Schülerlösungstexten in einer Art und Weise festzustellen und zu beurteilen, die an der kriterialen Bezugsnorm orientiert ist, bei Schülerlösungstext A signifikant weniger stark ausgeprägt ist als bei den Schülerlösungstexten B, C, und D. Zum einen ist es naheliegend anzunehmen, dass sich dieser Umstand auf die Merkmale von Schülerlösungstext A zurückführen lässt, anhand derer er im Rahmen der Entwicklungsstudie ausgewählt wurde. Eine plausible Interpretation der Auffälligkeiten (β3) und (β4) ist daher, dass die Teilnehmer_innen über vergleichsweise wenig Wissen und Können zur kriterialen Feststellung und Beurteilung von Schülerleistungen mit einer hohen fachlich-konzeptuellen Qualität und gleichzeitig geringer Qualität in der sprachlichen Realisierung verfügten. Sie verorteten Schülerlösungstext A daher vergleichsweise wenig bezüglich sachlicher Kriterien bzw. haben bei Schülerlösungstext A daher vergleichsweise häufig mehr- und/oder uneindeutige Bezüge der Verortung heranzogen. Zum anderen ist aber nicht auszuschließen, dass es sich bei den Auffälligkeiten (β3) und (β4) um ein Artefakt der Erhebung im Rahmen der Laborsituation handelt: Die Schülerlösungstexte A bis D waren im Aufgabenheft, dass die Lehrkräfte bearbeiteten, in alphabetischer Reihenfolge abgedruckt (vgl. Anhang C.1). Schülerlösungstext A stellte daher bei jedem_jeder Teilnehmer_in den ersten dar, mit dem diese im Rahmen der Laborsituation konfrontiert wurden. Dementsprechend ist eine weitere mögliche Erklärung für die Auffälligkeiten (β3) und (β4), dass sich die Teilnehmer_innen zu Beginn der laut-denkenden Korrekturarbeit – also vor allem bei der Korrektur von Schülerlösungstext A – zunächst häufiger mehr- und/oder uneindeutige Bezüge der Verortung herangezogen haben, im weiteren Verlauf dann aber in zunehmender Häufigkeit die vier Schülerlösungstexte anhand sachlicher Kriterien verorteten.

6.3.2.5.4. Quantitative Analyse der medianen prozentuellen Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden

In Tabelle 6.25 sind für die Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden, die Ergebnisse der in Phase 2c vorgenommenen Aufbereitung und Analyse zusammengefasst. Tabelle 6.25 ist analog zu Tabelle 6.23 aufgebaut, und daher in identischer Art und Weise zu lesen.

Auffälligkeiten in den deskriptiven Kennwerten und Ergebnissen der Wilcoxon-Vorzeichen-Rang-Tests

In den in Tabelle 6.25 aufgeführten Kennwerten zeigen sich die folgenden Auffälligkeiten:

(γ1) Bei Schülerlösungstext A halten sich die medianen prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig äußerten, nahezu die Waage. Lediglich zwischen der prozentuellen Häufigkeit, in der sich die Teilnehmer_innen positiv wertend/akzeptierend zu Schülerlösungstext A äußerten, und jener, in denen sie sich neutral/gemischt/sonstig äußerten, zeigt sich ein moderater, statistisch signifikanter Medianunterschied ($ES = .37$).

(γ2) Zu Schülerlösungstext B wurde sich von den Teilnehmer_innen am wenigsten positiv wertend/akzeptierend geäußert (Median = 10.5 %). Zwischen der prozentuellen Häufigkeit positiv wertender/akzeptierender Äußerungen und negativ wertender/ablehnender, bzw. neutraler/gemischter/sonstiger Äußerungen zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.52 \leq ES \leq .59$). Kein statistisch signifikanter Medianunterschied zeigt sich hingegen zwischen der prozentuellen Häufigkeit, in der sich die Teilnehmer_innen negativ wertend/ablehend zu Schülerlösungstext B äußerten (Median = 45.7 %), und jener, in der sie sich neutral/gemischt/sonstig äußerten (Median = 34.3 %).

(γ3) Zu Schülerlösungstext C wurde sich von den Teilnehmer_innen vor allem positiv wertend/akzeptierend geäußert (Median = 66.7 %). Zwischen der prozentuellen Häufigkeit positiv wertender/akzeptierender Äußerungen und negativ wertender/ablehnender bzw. neutraler/gemischter/sonstiger Äußerungen zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.57 \leq ES \leq .58$). Kein statistisch signifikanter Medianunterschied zeigt sich hingegen zwischen der prozentuellen Häufigkeit, in der sich die Teilnehmer_innen negativ wertend/ablehend zu Schülerlösungstext C äußerten (Median = 14.3 %), und jener, in der sie sich neutral/gemischt/sonstig äußerten (Median = 15.2 %).

		A			**B**			**C**			**D**		
		+ (3.0.3.1)	**-** (3.0.3.2)	**n** (3.0.3.3)	**+** (3.0.3.1)	**-** (3.0.3.2)	**n** (3.0.3.3)	**+** (3.0.3.1)	**-** (3.0.3.2)	**n** (3.0.3.3)	**+** (3.0.3.1)	**-** (3.0.3.2)	**n** (3.0.3.3)
A	**+** (3.0.3.1)	**27.8% (12.6%)**											
	- (3.0.3.2)	.12	**29.5% (20.2%)**										
	n (3.0.3.3)	.37*	.13	**37.0% (16.9%)**									
B	**+** (3.0.3.1)	.52***			**10.5% (16.7%)**								
	- (3.0.3.2)		.45**		.59***	**45.7% (44.4%)**							
	n (3.0.3.3)			.13	.52***	.26	**34.3% (31.8%)**						
C	**+** (3.0.3.1)	.58***			.61***			**66.7% (28.2%)**					
	- (3.0.3.2)		.42**			.56***		.58***	**14.3% (25.9%)**				
	n (3.0.3.3)			.54***			.39**	.57***	.06	**15.2% (10.6%)**			
D	**+** (3.0.3.1)	.61***			.34*			.62***			**0.0% (6.7%)**		
	- (3.0.3.2)		.60***			.37*			.62***		.62***	**72.7% (24.1%)**	
	n (3.0.3.3)			.41**			.30*			.19	.51***	.56***	**20.0% (22.2%)**

Tabelle 6.25.: Zusammenfassung der Lageparameter-Analyse für die prozentuellen Häufigkeiten der Segmente, die mit der Kategorie „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ oder „D“ und zusätzlich mit den Subsubsubkategorien der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden. Kennwerte in der Tabellendiagonalen: Median (Interquartilsabstand) der jeweiligen Codingkombination. Kennwerte unterhalb der Tabellendiagonalen: Effektstärkemaß *ES* für den Medianunterschied der jeweiligen Codingkombinationen (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$; 2-seitiger Wilcoxon-Vorzeichen-Rang-Test).

(γ4) Bei Schülerlösungstext D zeigt sich zwischen den prozentuellen Häufigkeiten aller codierten Arten der Äußerungen jeweils ein starker, statistisch signifikanter Medianunterschied ($.51 \leq ES \leq .62$). Negativ wertende/Ablehnende Äußerungen traten dabei im Median am häufigsten auf (Median = 72.7 %), gefolgt von neutralen/gemischten/sonstigen Äußerungen (Median = 20.0 %). Positiv wertende/Akzeptierende Äußerungen spielten bei Schülerlösungstext D mit einer medianen prozentuellen Häufigkeit von 0.0 % de facto keine Rolle.

(γ5) Zwischen den prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen positiv wertend/akzeptierend zu Schülerlösungstext A, B, C oder D äußerten, zeigt sich jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.34 \leq ES \leq .62$). Im Median am häufigsten wurde sich zu Schülerlösungstext C positiv wertend/akzeptierend geäußert (Median = 66.7 %), den zweiten diesbezüglichen Rangplatz nimmt Schülerlösungstext A ein (Median = 27.8 %), den dritten Schülerlösungstext B (Median = 10.5 %) und Schülerlösungstext D bildet mit einer medianen prozentuellen Häufigkeit von 0.0 % das Schlusslicht.

(γ6) Zwischen den prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen negativ wertend/ablehnend zu Schülerlösungstext A, B, C oder D äußerten, zeigt sich jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.37 \leq ES \leq .62$). Im Median am häufigsten wurde sich zu Schülerlösungstext D negativ wertend/ablehnend geäußert (Median = 72.7 %), den zweiten diesbezüglichen Rangplatz nimmt Schülerlösungstext B ein (Median = 45.7 %), den dritten Schülerlösungstext A (Median = 29.5 %) und Schülerlösungstext C bildet mit einer medianen prozentuellen Häufigkeit von 14.3 % das Schlusslicht.

(γ7) Am häufigsten neutral/gemischt/sonstig äußerten sich die Teilnehmer_innen bei Schülerlösungstext A und B (Median Schülerlösungstext A = 37.0 % ; Median Schülerlösungstext B = 34.3 %). Zwischen den prozentuellen Häufigkeiten, in denen sich bei diesen beiden Schülerlösungstexten neutral/gemischt/sonstig geäußert wurde, zeigt sich kein statistisch signifikanter Medianunterschied. Gleiches gilt für die prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen zu Schülerlösungstext C und D neutral/gemischt/sonstig äußerten (Median Schülerlösungstext C = 15.2 % ; Median Schülerlösungstext D = 20.0 %). Ein moderater bis starker, statistisch signifikanter Medianunterschied diesbezüglich zeigt sich hingegen jeweils zwischen Schülerlösungstext A bzw. B und Schülerlösungstext C bzw. D ($.30 \leq ES \leq .54$).

Interpretation: quantitative Teilbefunde

Die eben aufgeführten Auffälligkeiten (γ1) bis (γ7) lassen alles in allem die folgenden Interpretationen zu. Diese decken sich zum Teil mit jenen Interpretationen, die im Rahmen der quantitativen Analyse der Punkte, die die Teilnehmer_innen an die Schülerlösungstexte A bis D vergeben haben, vorgenommen wurden (vgl. Unterabschnitt 6.3.1):

1. Aus den Auffälligkeiten (γ1) bis (γ4) lässt sich schlussfolgern, dass sich die Teilnehmer_innen bei der Korrektur der vier Schülerlösungstexte in unterschiedlichem Umfang positiv wertend/akzeptierend, negativ wertend/ablehnend und neutral/gemischt/sonstig äußerten. Da die Schülerlösungstexte A bis D im Rahmen der Entwicklungsstudie als eine Komposition aus vier konstrastierenden Schülerlösungstexten ausgewählt wurden, ist es pausibel anzunehmen, dass die fachlich-konzeptuelle Qualität, sowie die Qualität der sprachlichen Realisierung der vier Schülerlösungstexte beeinflusst, in welchem Umfang sich die Teilnehmer_innen in welcher Art und Weise zu einem Schülerlösungstext äußerten. In den Auffälligkeiten (γ5) und (γ6) zeigen sich vier Tendenzen, die diese Interpretation zusätzlich unterstützen:

   Tendenz 1: Bei zwei Schülerlösungstexten, deren fachlich-konzeptuelle Qualität vergleichbar ist (Schülerlösungstext A und C, bzw. B und D), wurde sich bei dem Schülerlösungstext, dessen sprachliche Realisierung eine höhere Qualität aufweist, im Median häufiger positiv wertend/akzeptierend geäußert (starker, statistisch signifikanter Medianunterschied zwischen Schülerlösungstext A und C, bzw. moderater, statistisch signifikanter Medianunterschied zwischen Schülerlösungstext B und D).

   Tendenz 2: Bei zwei Schülerlösungstexten, deren fachlich-konzeptuelle Qualität vergleichbar ist (Schülerlösungstext A und C, bzw. B und D), wurde sich bei dem Schülerlösungstext, dessen sprachliche Realisierung eine geringere Qualität aufweist, im Median häufiger negativ wertend/ablehnend geäußert (moderater, statistisch signifikanter Medianunterschied zwischen Schülerlösungstext A und C, bzw. B und D).

   Tendenz 3: Bei zwei Schülerlösungstexten mit vergleichbarer Qualität in der sprachlichen Realisierung (Schülerlösungstext A und D, bzw. B und C), wurde sich bei dem Schülerlösungstext, dessen fachlich-konzeptuelle Qualität höher ist, im Median häufiger positiv wertend/akzeptierend geäußert (starker, statistisch signifikanter Medianunterschied zwischen Schülerlösungstext A und D, bzw. B und C).

   Tendenz 4: Bei zwei Schülerlösungstexten mit vergleichbarer Qualität in der sprachlichen Realisierung (Schülerlösungstext A und D, bzw. B und C), wurde sich bei dem Schülerlösungstext, dessen fachlich-konzeptuelle Qualität geringer ist, im Median häufiger negativ wertend/ablehnend geäußert (starker, statistisch signifikanter Medianunterschied zwischen Schülerlösungstext A und D, bzw. B und C).

2. Ferner zeigen sich in Auffälligkeit (γ5) und (γ6) zwei weitere Tendenzen, wenn man die Ergebnisse für Schülerlösungstexte B und D mit jenen für Schülerlösungstext A und C vergleicht:

   Tendenz 5: Der Medianunterschied zwischen den prozentuellen Häufigkeiten positiv wertender/akzeptierender Äußerungen ist bei den Schülerlösungstexten

B und D deutlich geringer (moderater Effekt; $ES = .34$), als zwischen Schülerlösungstext A und C (starker Effekt; $ES = .58$).

Tendenz 6: Der Medianunterschied zwischen den prozentuellen Häufigkeiten negativ wertender/ablehnender Äußerungen ist bei den Schülerlösungstexten B und D geringer (moderater Effekt; $ES = .37$), als zwischen Schülerlösungstext A und C (moderater Effekt; $ES = .42$).

Diese Schülerlösungstext-Paare gleichen sich darin, dass die fachlich-konzeptuelle Qualität beider Texte jeweils vergleichbar ist und dass sich beide Texte bezüglich der Qualität ihrer sprachlichen Realisierung voneinander unterscheiden. Die Schülerlösungstext-Paare unterscheiden sich jedoch darin, dass die fachlich-konzeptuelle Qualität von Schülerlösungstext B und D geringer ist, als jene von Schülerlösungstext A und C. Die eben aufgeführte fünfte und sechste Tendenz lässt sich daher im Sinne von Forschungsfrage (F2) interpretieren: Die Unterschiede in den Effektstärken deuten darauf hin, dass die Teilnehmer_innen während der laut-denkenden Korrekturarbeit bei den Schülerlösungstexten, die eine hohe fachlich-konzeptuelle Qualität aufweisen,...

... sprachliche Qualitäten stärker positiv wertend/akzeptierend berücksichtigten (Tendenz 5),

... sprachliche Mängel stärker negativ wertend/ablehnend berücksichtigten (Tendenz 6),

... als bei den Schülerlösungstexten mit einer geringen fachlich-konzeptuellen Qualität. In den medianen prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen bei der laut-denkenden Korrektur der vier Schülerlösungstexte positiv wertend/akzeptierend, bzw. negativ wertend/ablehnend äußerten, zeigt sich damit also ein empirischen Hinweis auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteilsgenese.

6.3.2.5.5. Quantitative Analyse der medianen prozentuellen Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden

In Tabelle 6.26 sind für die Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden, die Ergebnisse der in Phase 2c vorgenommenen Aufbereitung und Analyse zusammengefasst. Tabelle 6.26 ist analog zu Tabelle 6.23 aufgebaut, und daher in identischer Art und Weise zu lesen.

		Fach (3.0.1.1)					**Sprache** (3.0.1.2)					**sonstig** (3.0.1.3)					**mehrere** (3.0.1.4)				
		Kriterium (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)	**Kriterium** (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)	**Kriterium** (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)	**Kriterium** (3.0.2.1)	**Texte** (3.0.2.2)	**Erfahrung** (3.0.2.3)	**Person** (3.0.2.4)	**mehrere** (3.0.2.5)
Fach (3.0.1.1)	**Kriterium** (3.0.2.1)	**94.3%** **(8.8%)**																			
	Texte (3.0.2.2)	.62***	**0.0%** **(4.1%)**																		
	Erfahrung (3.0.2.3)	.62***	.19	**0.0%** **(0.0%)**																	
	Person (3.0.2.4)	.62***	.18	.27	**1.6%** **(4.3%)**																
	mehrere (3.0.2.5)	.62***	.29	.10	.37*	**0.0%** **(0.0%)**															
Sprache (3.0.1.2)	**Kriterium** (3.0.2.1)	.22					**100.0%** **(4.3%)**														
	Texte (3.0.2.2)		.36*				.64***	**0.0%** **(0.0%)**													
	Erfahrung (3.0.2.3)			.12			.64***	.29	**0.0%** **(2.6%)**												
	Person (3.0.2.4)				.23		.64***	.23	.02	**0.0%** **(0.0%)**											
	mehrere (3.0.2.5)					.25	.64***	.15	.34*	.25	**0.0%** **(0.0%)**										
sonstig (3.0.1.3)	**Kriterium** (3.0.2.1)	.16					.30					**100.0%** **(40.0%)**									
	Texte (3.0.2.2)		.07					.28				.57***	**0.0%** **(0.0%)**								
	Erfahrung (3.0.2.3)			.28					.34*			.60***	.28	**0.0%** **(0.0%)**							
	Person (3.0.2.4)				.22					.06		.59***	.15	.21	**0.0%** **(0.0%)**						
	mehrere (3.0.2.5)					.06					.15	.60***	.22	.15	.16	**0.0%** **(0.0%)**					
mehrere (3.0.1.4)	**Kriterium** (3.0.2.1)	.62***					.62***					.57***					**14.3%** **(19.1%)**				
	Texte (3.0.2.2)		.06					.34*					.05				.53***	**0.0%** **(3.2%)**			
	Erfahrung (3.0.2.3)			.08					.01					.34*			.53***	.06	**0.0%** **(1.6%)**		
	Person (3.0.2.4)				.05					.02					.06		.49***	.03	.03	**0.0%** **(0.0%)**	
	mehrere (3.0.2.5)					.62***					.62***					.62***	.62***	.62***	.62***	.62***	**73.9%** **(23.2%)**

Tabelle 6.26.: Zusammenfassung der Lageparameter-Analyse für die prozentuellen Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.2 (Bezug der Verortung) codiert wurden. Kennwerte in der Tabellendiagonalen: Median (Interquartilsabstand) der jeweiligen Codingkombination. Kennwerte unterhalb der Tabellendiagonalen: Effektstärkemaß *ES* für den Medianunterschied der jeweiligen Codingkombinationen (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$; 2-seitiger Wilcoxon-Vorzeichen-Rang-Test).

Auffälligkeiten in den deskriptiven Kennwerten und Ergebnissen der Wilcoxon-Vorzeichen-Rang-Tests

In den in Tabelle 6.26 aufgeführten Kennwerten zeigen sich die folgenden Auffälligkeiten:

(δ1) Sowohl der fachlich-konzeptuelle Eindruck, als auch die sprachliche Realisierung, sowie ferner sonstige Merkmale eines Schülerlösungstextes wurden von den Teilnehmer_innen nahezu vollständig bezüglich sachlicher Kriterien verortet (Median fachlich-konzeptueller Eindruck = 94.3 %; Median sprachliche Realisierung = 100.0 %; Median sonstige Merkmale = 100.0 %). Bei allen eben benannten Merkmalsfoki zeigt sich zwischen der prozentuellen Häufigkeit, in der sachliche Kriterien als Bezug der Verortung herangezogen wurden und der prozentuellen Häufigkeit aller übrigen codierten Arten der Verortung, ein starker, statistisch signifikanter Medianunterschied ($.59 \leq ES \leq .64$). Ferner zeigt sich beim fachlich-konzeptuellen Eindruck eines Schülerlösungstextes zwischen der prozentuellen Häufigkeit, in denen dieser bezüglich mutmaßlicher Personenmerkmale verortet wurde, und jener, in der dieser mehr- und/oder uneindeutig verortet wurde, ebenfalls ein moderater, statistisch signifikanter Medianunterschied ($ES = .37$). Des Weiteren hatten die durchgeführten Wilcoxon-Vorzeichen-Rang-Tests für die prozentuellen Häufigkeiten, in denen die sprachliche Realisierung bezüglich allgemeiner Erfahrungen mit Physiklernenden bzw. mehr- und/oder uneindeutigen Bezügen verortet wurde, einen signifikanten Ausgang. Dies lässt sich allerdings auf die Streuungsunterschiede zwischen diesen beiden Arten des Bezugs der Verortung zurückführen (IQR allgemeine Erfahrungen mit Physiklernenden = 2.6 %; IQR mehr- und/oder uneindeutige Bezüge der Verortung = 0.0 %).

(δ2) Bei mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes wurden von den Teilnehmer_innen am häufigsten mehr- und/oder uneindeutige Bezüge der Verortung herangezogen (Median = 73.9 %). Zwischen der prozentuellen Häufigkeit dieser und den übrigen codierten Arten der Verortung zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied (jeweils $ES = .62$). Den zweiten diesbezüglichen Rangplatz nehmen sachliche Kriterien als Bezug der Verortung ein (Median = 14.3 %). Andere Schülerlösungstexte, allgemeine Erfahrungen mit Physiklernenden, sowie mutmaßliche Personenmerkmale des_der Schülers_Schülerin spielten bei mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes mit einer medianen prozentuellen Häufigkeit von 0.0 % de facto keine Rolle.

(δ3) Sachliche Kriterien wurden von den Teilnehmer_innen bei fast allen codierten Merkmalsfoki in vergleichbarer Häufigkeit als Bezug der Verortung herangezogen. Lediglich zwischen der prozentuellen Häufigkeit, in der mehrere/uneindeutige Merkmale eines Schülerlösungstextes bezüglich sachlicher Kriterien verortet wurden, und jener, in der dies bei den übrigen codierten Merkmalsfoki geschah, zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.57 \leq ES \leq .62$).

(δ4) Andere Schülerlösungstexte, sowie allgemeine Erfahrungen mit Physilernenden wurden von den Teilnehmer_innen bei keinem der codierten Merkmalsfoki in nennens-

werter Häufigkeit als Bezug der Verortung herangezogen (mediane prozentuelle Häufigkeit jeweils 0.0 %). Der signifikante Ausgang einzelner Wilcoxon-Vorzeichen-Rang-Tests zwischen entsprechenden prozentuellen Häufigkeiten (vgl. Tabelle 6.26) lässt sich auf die jeweiligen Streuungsunterschiede zwischen den betrachteten codierten Merkmalsfoki zurückführen.

(δ5) Mutmaßliche Personenmerkmale des_der Schülers_Schülerin wurden von den Teilnehmer_innen bei nahezu keinem der codierten Merkmalsfoki in nennenswerter Häufigkeit als Bezug der Verortung herangezogen (Median fachlich-konzeptueller Eindruck = 1.6 %; alle übrigen medianen prozentuellen Häufigkeiten jeweils 0.0 %). Alle durchgeführten Wilcoxon-Vorzeichen-Rang-Tests zwischen den entsprechenden prozentuellen Häufigkeiten hatten einen nicht signifikanten Ausgang (vgl. Tabelle 6.26).

(δ6) Mehr- und/oder uneindeutige Bezüge der Verortung wurden von den Teilnehmer_innen am häufigsten bei der Feststellung und Beurteilung mehrerer/uneindeutiger Merkmale eines Schülerlösungstextes herangezogen (Median = 73.9 %). Zu den prozentuellen Häufigkeiten, in denen mehr- und/oder uneindeutige Bezüge der Verortung bei den übrigen codierten Merkmalsfoki herangezogen wurden, zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied (jeweils $ES = .62$). Zwischen den übrigen codierten Merkmalsfoki zeigt sich diesbezüglich kein statistisch signifikanter Medianunterschied.

Interpretation: quantitative Teilbefunde

Die eben aufgeführten Auffälligkeiten (δ1) bis (δ6) lassen alles in allem die folgende Interpretation zu: In den Auffälligkeiten (δ1), sowie (δ3) bis (δ5) zeigt sich, dass die Teilnehmer_innen de facto ausschließlich sachliche Kriterien als Bezug der Verortung heranzogen, wenn sie ein bestimmtes Schülerlösungstextmerkmal bei ihrer Leistungsurteilsgenese in den Fokus rückten (Median fachlich-konzeptueller Eindruck = 94.3 %; Median sprachliche Realisierung = 100.0 %; Median sonstige Merkmale = 100.0 %). Lediglich bei mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes war der Bezug der Verortung signifikant häufiger mehr- und/oder uneindeutig (vgl. Auffälligkeiten (δ2) und (δ6)). Ferner spielten bei keinem der codierten Merkmalsfoki andere Schülerlösungstexte, allgemeine Erfahrungen mit Physiklernenden, sowie mutmaßliche Personenmerkmale des_der Schülers_Schülerin de facto eine Rolle (Median mutmaßliche Personenmerkmale als Bezug der Verortung des fachlich-konzeptuellen Eindrucks = 1.6 %; alle übrigen medianen prozentuellen Häufigkeiten jeweils 0.0 %). Insgesamt lässt sich hieraus schlussfolgern, dass die Teilnehmer_innen bestimmte Schülerlösungstextmerkmale (den fachlich-konzeptuellen Eindruck, die sprachliche Realisierung und sonstige Merkmale) in einer tendenziell deutlich an der kriterialen Bezugsnorm orientierten Art und Weise feststellten und beurteilten. Allein bei mehr- und/oder uneindeutigen Merkmalen eines Schülerlösungstextes lässt sich keine Tendenz zu einer bestimmten Bezugsnorm ausfindig machen. Ferner wird aus den Auffälligkeiten (δ1) bis (δ6) deutlich, dass sich bei den Teilnehmer_innen kei-

ne Hinweise auf eine Orientierung an der sozialen und/oder der individuellen Bezugsnorm im Rahmen der Laborsituation zeigten (verschwinden geringe Häufigkeit, in der andere Schülerlösungstexte, sowie mutmaßliche Personenmerkmale des_der Schülers_Schülerin als Bezug der Verortung bei allen codierten Merkmalsfoki). Das Fehlen von Hinweisen für eine Orientierung an der individuellen Bezugsnorm könnte es allerdings ein Erhebungsartefakt darstellen, da den Teilnehmer_innen keine Informationen über Personenmerkmale der Schüler_innen, die die Schülerlösungstexte A bis D verfasst haben, zur Verfügung gestellt wurden.

6.3.2.5.6. Quantitative Analyse der medianen prozentuellen Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden

In Tabelle 6.27 sind für die Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden, die Ergebnisse der in Phase 2c vorgenommenen Aufbereitung und Analyse zusammengefasst. Tabelle 6.27 ist analog zu Tabelle 6.23 aufgebaut, und daher in identischer Art und Weise zu lesen.

Auffälligkeiten in den deskriptiven Kennwerten und Ergebnissen der Wilcoxon-Vorzeichen-Rang-Tests

In den in Tabelle 6.27 aufgeführten Kennwerten zeigen sich die folgenden Auffälligkeiten:

(ϵ1) Zwischen den prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig äußerten, zeigt sich jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.31 \leq ES \leq .59$). Im Median am häufigsten wurde sich dabei positiv wertend/akzeptierend zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes geäußert (Median = 51.4 %), am zweithäufigsten traten negativ wertende/ablehnende Äußerungen auf (Median = 27.3%) und am seltensten kam es zu neutralen/gemischten/sonstigen Äußerungen (Median = 19.2 %).

(ϵ2) Zur sprachlichen Realisierung eines Schülerlösungstextes wurde sich von den Teilnehmer_innen vor allem negativ wertend/ablehnend geäußert (Median = 69.6 %). Zwischen der prozentuellen Häufigkeit negativ wertender/ablehnender Äußerungen und allen übrigen codierten Arten der Äußerung zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.56 \leq ES \leq .59$). Kein statistisch signifikanter Medianunterschied zeigt sich hingegen zwischen der prozentuellen Häufigkeit, in

		Fach (3.0.1.1)			**Sprache** (3.0.1.2)			**sonstig** (3.0.1.3)			**mehrere** (3.0.1.4)		
		+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)	+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)	+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)	+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)
Fach (3.0.1.1)	+ (3.0.3.1)	**51.4% (18.8%)**											
	- (3.0.3.2)	.53***	**27.3% (17.3%)**										
	n (3.0.3.3)	.59***	.31*	**19.2% (15.1%)**									
Sprache (3.0.1.2)	+ (3.0.3.1)	.61***			**13.6% (17.0%)**								
	- (3.0.3.2)		.61***		.56***	**69.6% (34.5%)**							
	n (3.0.3.3)			.22	.02	.59***	**10.7% (19.1%)**						
sonstig (3.0.1.3)	+ (3.0.3.1)	.34*			.28			**25.0% (50.0%)**					
	- (3.0.3.2)		.10			.54***		.10	**14.3% (50.0%)**				
	n (3.0.3.3)			.08			.17	.11	.01	**20.0% (50.0%)**			
mehrere (3.0.1.4)	+ (3.0.3.1)	.62***			.19			.09			**25.0% (16.1%)**		
	- (3.0.3.2)		.47***			.61***			.14		.27	**17.5% (16.3%)**	
	n (3.0.3.3)			.63***			.60***			.50***	.51***	.59***	**54.5% (17.9%)**

Tabelle 6.27.: Zusammenfassung der Lageparameter-Analyse für die prozentuellen Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.1 (fokussiertes Merkmal) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden. Kennwerte in der Tabellendiagonalen: Median (Interquartilsabstand) der jeweiligen Codingkombination. Kennwerte unterhalb der Tabellendiagonalen: Effektstärkemaß ES für den Medianunterschied der jeweiligen Codingkombinationen (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$; 2-seitiger Wilcoxon-Vorzeichen-Rang-Test).

der sich die Teilnehmer_innen positiv wertend/akzeptierend zur sprachlichen Realisierung eines Schülerlösungstextes äußerten (Median = 13.6 %), und jener, in der sie sich diesbezüglich neutral/gemischt/sonstig äußerten (Median = 10.7 %).

(ϵ3) Zwischen den prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen zu sonstigen Merkmalen eines Schülerlösungstextes positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig äußerten, zeigt sich jeweils kein statistisch signifikanter Medianunterschied (Median positiv wertende/akzeptierende Äußerungen = 25.0 %; Median negativ wertende/ablehnende Äußerungen = 14.3 %; Median neutrale/gemischte/sonstige Äußerungen = 20.0 %).

(ϵ4) Zu mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes wurde sich von den Teilnehmer_innen vor allem neutral/gemischt/sonstig geäußert (Median = 54.5 %). Zwischen der prozentuellen Häufigkeit neutraler/gemischter/sonstiger Äußerungen und allen übrigen codierten Arten der Äußerung zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.51 \leq ES \leq .59$). Kein statistisch signifikanter Medianunterschied zeigt sich hingegen zwischen der prozentuellen Häufigkeit, in der sich die Teilnehmer_innen positiv wertend/akzeptierend zu mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes äußerten (Median = 25.0 %), und jener, in der sie sich diesbezüglich negativ wertend/ablehnend äußerten (Median = 17.5 %).

(ϵ5) Am häufigsten positiv wertend/akzeptierend äußerten sich die Teilnehmer_innen zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes (Median = 51.4 %). Zwischen der prozentuellen Häufigkeit, in der sich positiv wertend/akzeptierend zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes geäußert wurde, und jener, in der sich die Teilnehmer_innen zu den übrigen codierten Merkmalsfoki positiv wertend/akzeptierend äußerten, zeigt sich jeweils ein moderater bist starker, statistisch signifikanter Medianunterschied ($.34 \leq ES \leq .62$). Kein statistisch signifikanter Medianunterschied zeigt sich hingegen jeweils zwischen der prozentuellen Häufigkeit, in der sich die Teilnehmer_innen positiv wertend/akzeptierend zur sprachlichen Realisierung, zu sonstigen Merkmalen oder mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes äußerten.

(ϵ6) Am häufigsten negativ wertend/ablehnend äußerten sich die Teilnehmer_innen zur sprachlichen Realisierung eines Schülerlösungstextes (Median = 69.6 %). Zwischen der prozentuellen Häufigkeit, in der sich negativ wertend/ablehnend zur sprachlichen Realisierung eines Schülerlösungstextes geäußert wurde, und jener, in der sich die Teilnehmer_innen zu den übrigen codierten Merkmalsfoki negativ wertend/ablehnend äußerten, zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.54 \leq ES \leq .61$). Des Weiteren zeigt sich zwischen der prozentuellen Häufigkeit, in der sich negativ wertend/ablehnend zum fachlich-konzeptuellen Eindruck eines Schülerlösungstextes geäußert wurden, und jener, in der sich negativ wertend/ablehnend bezüglich mehrerer/uneindeutiger Merkmale eines Schüler-

lösungstextes geäußert wurde, ein moderater, statistisch signifikanter Medianunterschied ($ES = .47$).

(ϵ7) Am häufigsten neutral/gemischt/sonstig äußerten sich die Teilnehmer_innen zu mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes (Median = 51.4 %). Zwischen der prozentuellen Häufigkeit, in der sich neutral/gemischt/sonstig zu mehreren/uneindeutigen Merkmalen eines Schülerlösungstextes geäußert wurde, und jener, in der sich die Teilnehmer_innen zu den übrigen codierten Merkmalsfoki neutral/gemischt/sonstig äußerten, zeigt sich jeweils ein starker, statistisch signifikanter Medianunterschied ($.50 \leq ES \leq .63$). Kein statistisch signifikanter Medianunterschied zeigt sich hingegen jeweils zwischen der prozentuellen Häufigkeit, in der sich die Teilnehmer_innen neutral/gemischt/sonstig zum fachlich-konzeptuellen Eindruck, zur sprachlichen Realisierung oder zu sonstigen Merkmalen eines Schülerlösungstextes äußerten.

Interpretation: quantitative Teilbefunde

Die eben aufgeführten Auffälligkeiten (ϵ1) bis (ϵ7) lassen alles in allem die folgenden Interpretationen zu. Diese decken sich zum Teil mit jenen Interpretationen, die im Rahmen von Phase 2b der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle vorgenommen wurden (vgl. Unterabschnitt 6.3.2.4):

1. Aus den Auffälligkeiten (ϵ1) bis (ϵ4) wird deutlich, dass sich die Teilnehmer_innen jeweils unterschiedlich umfangreich positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig äußerten, je nachdem, ob sie den fachlich-konzeptuellen Eindruck, die sprachliche Realisierung, sonstige Merkmale und mehrere/uneindeutige Merkmale der vier Schülerlösungstexte fokussierten. Hier zeigt sich also ein Hinweis darauf, dass die Leistungsurteilsgenese der Teilnehmer_innen, bei verschiedenen Schüerlösungstextmerkmalen auch unterschiedliche Tendenzen hinsichtlich einer Fähigkeiten- oder Defizitorientierung aufweist.
2. Aus den Auffälligkeiten (ϵ1) und (ϵ5) geht hervor, dass sich die Teilnehmer_innen zum fachlich-konzeptuellen Eindruck der vier Schülerlösungstexte überwiegend positiv wertend/akzeptierend äußerten (Median = 51.4 %) und dies signifikant häufiger taten, als bei den übrigen codierten Merkmalsfoki ($.34 \leq ES \leq .62$). Hier zeigt sich also ein Hinweis auf eine deutliche Tendenz zu einer fähigkeitsorientierten Feststellung und Beurteilung fachlich-konzeptueller Schülerleistung durch die Teilnehmer_innen.
3. Aus den Auffälligkeiten (ϵ2) und (ϵ6) geht hervor, dass sich die Teilnehmer_innen zur sprachlichen Realisierung der vier Schülerlösungstexte überwiegend negativ wertend/ablehnend äußerten (Median = 69.6 %) und dies signifikant häufiger taten, als bei den übrigen codierten Merkmalsfoki ($.54 \leq ES \leq .61$). Hier zeigt sich also ein Hinweis auf eine deutliche Tendenz zu einer defizitorientierten Feststellung und Beurteilung sprachlicher Schülerleistung durch die Teilnehmer_innen.

4. Aus den Auffälligkeiten ($\epsilon 3$) und ($\epsilon 5$) bis ($\epsilon 7$) geht hervor, dass sich die medianen prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig zu sonstigen Merkmalen der vier Schülerlösungstexte äußerten, die Waage halten (Median positiv wertende/akzeptierende Äußerungen = 25.0 %; Median negativ wertende/ablehnende Äußerungen = 14.3 %; Median neutrale/gemischte/sonstige Äußerungen = 20.0 %; jeweils kein statistisch signifikanter Medianunterschied). Dies lässt sich als Hinweise darauf deuten, dass die Feststellung und Beurteilung sonstiger Schülerlösungstextmerkmale durch die Teilnehmer_innen weder in einer tendenziell defizitorientierten, noch einer fähigkeitsorientierten Art und Weise stattfand.
5. Aus den Auffälligkeiten ($\epsilon 4$) bis ($\epsilon 7$) geht hervor, dass sich die medianen prozentuellen Häufigkeiten, in denen sich die Teilnehmer_innen positiv wertend/akzeptierend oder negativ wertend/ablehnend zu mehreren/uneindeutigen Merkmalen der vier Schülerlösungstexte äußerten, die Waage halten (Median positiv wertende/akzeptierende Äußerungen = 25.0 %; Median negativ wertende/ablehnende Äußerungen = 17.5 %; kein statistisch signifikanter Medianunterschied), neutrale/gemischte/sonstige Äußerungen zu mehreren/uneindeutigen Schülerlösungstextmerkmalen aber signifikant häufiger auftraten (Median = 54.5 %; $.50 \leq ES \leq .63$). Der zuerst genannte Befund lässt sich als Hinweise darauf deuten, dass die Feststellung und Beurteilung mehrerer/uneindeutiger Schülerlösungstextmerkmale durch die Teilnehmer_innen weder in einer tendenziell defizitorientierten, noch einer fähigkeitsorientierten Art und Weise stattfand. Im zuletzt genannten Befund zeigt sich ferner die Tendenz, dass die Teilnehmer_innen, wenn sie bei der Korrektur der vier Schülerlösungstexte mehrere/uneindeutige Schülerlösungstextmerkmale fokussierten, vergleichsweise häufig nicht und/oder uneindeutig normative Gedankenschritte mitvokalisierten.

6.3.2.5.7. Quantitative Analyse der medianen prozentuellen Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden

In Tabelle 6.28 sind für die Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden, die Ergebnisse der in Phase 2c vorgenommenen Aufbereitung und Analyse zusammengefasst. Tabelle 6.28 ist analog zu Tabelle 6.23 aufgebaut, und daher in identischer Art und Weise zu lesen.

		Kriterium (3.0.2.1)			**Texte** (3.0.2.2)			**Erfahrung** (3.0.2.3)			**Person** (3.0.2.4)			**mehrere** (3.0.2.5)		
		+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)	+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)	+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)	+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)	+ (3.0.3.1)	- (3.0.3.2)	n (3.0.3.3)
Kriterium (3.0.2.1)	+ (3.0.3.1)	**34.5% (10.0%)**														
	- (3.0.3.2)	.29	**44.0% (21.7%)**													
	n (3.0.3.3)	.57***	.61***	**18.6% (11.7%)**												
Texte (3.0.2.2)	+ (3.0.3.1)	.34*			**0.0% (40.0%)**											
	- (3.0.3.2)		.40**		.15	**0.0% (0.0%)**										
	n (3.0.3.3)			.11	.13	.19	**0.0% (60.0%)**									
Erfahrung (3.0.2.3)	+ (3.0.3.1)	.58***			.28			**0.0% (0.0%)**								
	- (3.0.3.2)		.43**			.03		.19	**0.0% (0.0%)**							
	n (3.0.3.3)			.16			.11	.41**	.24	**0.0% (100.0%)**						
Person (3.0.2.4)	+ (3.0.3.1)	.25			.03			.30			**0.0% (50.0%)**					
	- (3.0.3.2)		.56***			.04			.05		.17	**0.0% (0.0%)**				
	n (3.0.3.3)			.05			.11			.21	.03	.19	**0.0% (33.3%)**			
mehrere (3.0.2.5)	+ (3.0.3.1)	.48***			.01			.40**			.01			**19.0% (20.8%)**		
	- (3.0.3.2)		.61***			.36*			.21			.22		.02	**21.7% (23.1%)**	
	n (3.0.3.3)			.62***			.43**			.30			.45**	.55***	.56***	**64.9% (13.7%)**

Tabelle 6.28.: Zusammenfassung der Lageparameter-Analyse für die prozentuellen Häufigkeiten der Segmente, die mit den Subsubsubkategorien der Subsubkategorie 3.0.2 (Bezug der Verortung) und zusätzlich mit denen der Subsubkategorie 3.0.3 (Art der Äußerung) codiert wurden. Kennwerte in der Tabellendiagonalen: Median (Interquartilsabstand) der jeweiligen Codingkombination. Kennwerte unterhalb der Tabellendiagonalen: Effektstärkemaß ES für den Medianunterschied der jeweiligen Codingkombinationen (***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$; 2-seitiger Wilcoxon-Vorzeichen-Rang-Test).

Auffälligkeiten in den deskriptiven Kennwerten und Ergebnissen der Wilcoxon-Vorzeichen-Rang-Tests

In den in Tabelle 6.28 aufgeführten Kennwerten zeigen sich die folgenden Auffälligkeiten:

(ζ1) Wurde ein Schülerlösungstext bezüglich eines sachlichen Kriteriums verortet, wurde sich von den Teilnehmer_innen vor allem entweder positiv wertend/akzeptierend (Median = 34.5 %) oder negativ wertend/ablehnend geäußert (Median = 44.0 %). Es zeigt sich kein signifikanter Medianunterschied zwischen der prozentuellen Häufigkeit, in der ein Schülerlösungstext bezüglich eines sachlichen Kritierums verortet wurde und sich dabei positiv wertend/akzeptierend geäußert wurde, und jener, in der ein Schülerlösungstext bezüglich eines sachlichen Kriteriums verortet wurde und sich dabei negativ wertend/ablehend geäußert wurde. Ein starker, statistisch signifikanter Medianunterschied zeigt sich hingegen zwischen der prozentuellen Häufigkeit, in der ein Schülerlösungstext bezüglich eines sachlichen Kriteriums verortet wurde und sich dabei neutral/gemischt/sonstig geäußert wurde, und jener, in der ein Schülerlösungstext bezüglich eines sachlichen Kriteriums verortet wurde und sich dabei negativ wertend/ablehnend geäußert wurde (ES = .61). Gleiches gilt für den Medianunterschied zwischen der prozentuellen Häufigkeit, in der ein Schülerlösungstext bezüglich eines sachlichen Kriteriums verortet wurde und sich dabei neutral/gemischt/sonstig geäußert wurde, und jener, in der ein Schülerlösungstext bezüglich eines sachlichen Kriteriums verortet wurde und sich dabei positiv wertend/akzeptierend geäußert wurde (ES = .57).

(ζ2) Wurde ein Schülerlösungstext bezüglich eines anderen Schülerlösungstextes, allgemeinen Erfahrungen mit Physiklernenden oder mutmaßlichen Personenmerkmale der_des Schülers_Schülerin verortet, wurde sich weder vor allem positiv wertend/akzeptierend, noch vor allem negativ wertend/ablehnend oder vor allem neutral/gemischt/sonstig geäußert (mediane prozentuelle Häufigkeit jeweils 0.0 %). Lediglich zwischen der prozentuellen Häufigkeit, in der die Teilnehmer_innen einen Schülerlösungstext bezüglich allgemeiner Erfahrungen mit Physiklernenden verorteten und sich dabei positiv wertend/akzeptierend äußerten, und jener, in der sie sich dabei neutral/gemischt/sonstig äußerten, hatte der durchgeführte Wilcoxon-Vorzeichen-Rangtest einen signifikanten Ausgang. Dies lässt sich allerdings auf die Streuungsunterschiede zwischen diesen beiden Arten der Äußerung zurückführen (IQR positiv wertende/akzeptierende Äußerungen = 0.0 %; IQR negativ wertende/ablehnende Äußerungen = 100.0 %).

(ζ3) Wurde ein Schülerlösungstext bezüglich mehr- und/oder uneindeutiger Bezüge verortet, wurde sich dabei am häufigsten neutral/gemischt/sonstig geäußert (Median = 64.9%). Zwischen der prozentuellen Häufigkeit, in der ein Schülerlösungstext bezüglich mehr- und/oder uneindeutiger Bezüge verortet wurde und sich dabei neutral/gemischt/sonstig geäußert wurde, und jener, in der ein Schülerlösungstext bezüglich mehr- und/oder uneindeutiger Bezüge verortet wurde und sich dabei negativ wertend/ablehnend geäußert wurde, zeigt sich ein starker, statistisch signifikanter

Medianunterschied ($ES = .56$). Gleiches gilt für die prozentuelle Häufigkeit, in der ein Schülerlösungstext bezüglich mehr- und/oder uneindeutiger Bezüge verortet wurde und sich dabei neutral/gemischt/sonstig geäußert wurde, und jener, in der ein Schülerlösungstext bezüglich mehr- und/oder uneindeutiger Bezüge verortet wurde und sich dabei positiv wertend/akzeptierend geäußert wurde ($ES = .55$). Kein statistisch signifikanter Medianunterschied zeigt sich hingegen zwischen der prozentuellen Häufigkeit, in der ein Schülerlösungstext bezüglich mehr- und/oder uneindeutiger Bezüge verortet wurde und sich dabei positiv wertend/akzeptierend geäußert wurde, und jener, in der ein Schülerlösungstext bezüglich mehr- und/oder uneindeutiger Bezüge verortet wurde und sich dabei negativ wertend/ablehnend geäußert wurde.

($\zeta 4$) Am häufigsten positiv wertend/akzeptierend äußerten sich die Teilnehmer_innen, wenn sie einen Schülerlösungstext bezüglich eines sachlichen Kriteriums verorteten (Median = 34.5 %). Zwischen der prozentuellen Häufigkeit, in der sich positiv wertend/akzeptierend geäußert wurde und dabei ein Schülerlösungstexte bezüglich eines sachlichen Kriteriums verortet wurde, sowie jener, in der sich positiv wertend/akzeptierend geäußert wurde und dabei ein Schülerlösungstext auf eine der anderen codierten Arten verortet wurde, zeigt sich – mit Ausnahme eines Bezugs zu mutmaßlichen Personenmerkmalen des_der Schüler_in – jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.34 \leq ES \leq .58$). Des Weiteren zeigt sich ein moderater, statistisch signifikanter Medianunterschied zwischen der prozentuellen Häufigkeit, in der sich positiv wertend/akzeptierend geäußert wurde und dabei ein Schülerlösungstextes bezüglich allgemeiner Erfahrungen mit Physiklernenden verortet wurde, sowie jener, in der sich positiv wertend/akzeptierend geäußert wurde und dabei die Verortung eines Schülerlösungstext mehr- und/oder uneindeutig war ($ES = .40$).

($\zeta 5$) Am häufigsten negativ wertend/ablehnend äußerten sich die Teilnehmer_innen, wenn sie einen Schülerlösungstext bezüglich eines sachlichen Kriteriums verorteten (Median = 44.0 %). Zwischen der prozentuellen Häufigkeit, in der sich negativ wertend/ablehnend geäußert wurde und dabei ein Schülerlösungstext bezüglich eines sachlichen Kriteriums verortet wurde, sowie jener, in der sich negativ wertend/ablehnend geäußert wurde und dabei ein Schülerlösungstext auf eine der anderen codierten Arten verortet wurde, zeigt sich jeweils ein moderater, statistisch signifikanter Medianunterschied ($.40 \leq ES \leq .61$). Des Weiteren zeigt sich ein moderater, statistisch signifikanter Medianunterschied zwischen der prozentuellen Häufigkeit, in der sich negativ wertend/ablehnend geäußert wurde und dabei ein Schülerlösungstext bezüglich eines anderen Schülerlösungstextes verortet wurde, sowie jener, in der sich negativ wertend/ablehnend geäußert wurde und dabei die Verortung eines Schülerlösungstext mehr- und/oder uneindeutig war ($ES = .36$).

($\zeta 6$) Am häufigsten neutral/gemischt/sonstig äußerten sich die Teilnehmer_innen, wenn sie bei einen Schülerlösungstext mehr- und/oder uneindeutige Bezüge der Verortung heranzogen (Median = 64.9 %). Zwischen der prozentuellen Häufigkeit, in

der sich neutral/gemischt/sonstig geäußert wurde und dabei mehr- und/oder uneindeutige Bezüge der Verortung heranzogen wurden, sowie jener, in der sich neutral/gemischt/sonstig geäußert wurde und dabei ein Schülerlösungstext auf eine der anderen codierten Arten verortet wurde, zeigt sich – mit Ausnahme eines Bezugs zu allgemeinen Erfahrungen mit Physiklernenden – jeweils ein moderater bis starker, statistisch signifikanter Medianunterschied ($.43 \leq ES \leq .62$).

Interpretation: quantitative Teilbefunde

Die eben aufgeführten Auffälligkeiten ($\zeta 1$) bis ($\zeta 6$) lassen alles in allem die folgenden Interpretationen zu:

1. In Auffälligkeit ($\zeta 1$) zeigt sich, dass sich die Häufigkeit positiv wertender/akzeptierender und negativ wertender/ablehnender Äußerungen in etwa die Waage hält, unabhängig davon, ob der Bezug der Verortung sachliche Kriterien sind, oder ob ein mehr- und/oder uneindeutiger Bezug der Verortung herangezogen wurde. Dies lässt sich als Hinweis darauf deuten, dass die Leistungsfeststellungen und -beurteilungen der Teilnehmer_innen bezüglich dieser beiden Arten der Verortung weder in einer tendenziell defizitorientierten, noch einer fähigkeitsorientierten Art und Weise stattfand.
2. Aus den Auffälligkeiten ($\zeta 4$) und ($\zeta 5$) wird deutlich, dass positiv wertende/akzeptierende bzw. negativ wertende/ablehnende Äußerungen bei mehr- und/oder uneindeutigen Bezügen der Verortung signifikant weniger häufig auftraten (Median positiv wertende/akzeptierende Äußerungen = 19.0 %; Median negativ wertende/ablehnende Äußerungen = 21.7 %), als dies bei einem Bezug zu sachlichen Kriterien der Fall war (Median positiv wertende/akzeptierende Äußerungen = 34.5 %; Median negativ wertende/ablehnende Äußerungen = 44.0 %). Hier zeigt sich also die Tendenz, dass die Teilnehmer_innen, wenn sie sachliche Kriterien als Bezug der Verortung heranzogen, zu diesen auch vergleichsweise häufig eindeutig normative Gedankenschritte mitvokalisierten.
3. Aus Auffälligkeit ($\zeta 3$) und ($\zeta 6$) wird deutlich, dass neutrale/gemischte/sonstige Äußerungen bei mehr- und/oder uneindeutigen Bezügen der Verortung signifikant häufiger auftraten (Median = 64.9 %), als dies bei einem Bezug zu sachlichen Kriterien der Fall war (Median = 18.6 %). Hier zeigt sich also die Tendenz, dass die Teilnehmer_innen, wenn sie mehr- und/oder uneindeutige Bezügen der Verortung heranzogen, zu diesen auch vergleichsweise häufig nicht und/oder uneindeutig normative Gedankenschritte mitvokalisierten.
4. In Auffälligkeit ($\zeta 2$) zeigt sich, dass bezüglich anderer Schülerlösungstexte, allgemeinen Erfahrungen mit Physiklernenden und mutmaßlichen Personenmerkmalen des_der Schülers_Schülerin als Bezüge der Verortung keine Aussagen getroffen werden können. Grund hierfür ist, dass die prozentuelle Häufigkeit, in der sich die Teilnehmer_innen im Median bei einer der eben benannten Arten der Verortung positiv

wertend/akezptierend, negativ wertend/ablehenend oder neutral/gemischt/sonstig geäußert wurde jeweils 0.0 % beträgt.

6.3.2.5.8. Limitation und Zwischenfazit

Wie im vorangegangenen Unterabschnitt dargestellt, konnten mittels einer quantitativen Analyse der Häufigkeit besonders reichhaltiger Codierungen ausgewählte inhaltliche Facetten der Laut-Denk-Protokolle aller Teilnehmer_innen systematisch beschrieben und im Sinne der Forschungsfragen (F1) und (F2) interpretiert werden. Als besonders reichhaltig wurden die Codingkombinationen aus den (Sub-)Subsubkategorien der Subkategorie „Feststellung und Beurteilung eines Schülerlösungstextes“, sowie den zusätzlichen Kategorien „Auseinandersetzung mit Schülerlösungstext A“, „B“, „C“ und „D“ angesehen. Im Rahmen von Phase 2c der inhaltsanalytischen Auswertung der Laut-Denk-Protokolle, konnten die im Folgenden zusammengefassten Befunde gewonnen werden:

- Erstens lieferte die Analyse empirische Hinweise auf eine Beeinflussung des Prozesses der Leistungsurteilsgenese durch die fachlich-konzeptuelle Qualität und die sprachliche Realisierung der Schülerlösungstexte.
    - Zum einen beeinflusste die fachlich-konzeptuelle Qualität und die sprachliche Realisierung, in welchem Umfang die Teilnehmer_innen welche Merkmale eines Schülerlösungstextes (fachlich-konzeptueller Eindruck, sprachliche Realisierung, sonstige Merkmale) im Rahmen ihrer Leistungsurteilsgenese berücksichtigen. Hierbei weist die im Median häufigere Fokussierung des fachlich-konzeptuellen Eindrucks bei den Schülerlösungstexten mit geringer fachlich-konzeptueller Qualität auf eine zum Teil defizitorientierte Feststellung und Beurteilung fachlich-konzeptueller Schülerleistungen hin. Des Weiteren deutet sich in der im Median häufigeren Fokussierung der sprachlichen Realisierung bei Schülerlösungstexten mit hoher fachlich-konzeptueller Qualität an, dass die Teilnehmer_innen die Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes vor allem dann vornehmen, wenn dieser bis zu einem bestimmten Grad auch ihren fachlich-konzeptuellen Erwartungen entspricht. Dies lässt sich als empirischer Hinweis auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung interpretieren (Forschungsfrage (F2)).
    - Zum anderen hatten die fachliche-konzeptuelle Qualität und die sprachliche Realisierung einen Einfluss darauf, inwieweit sich die Teilnehmer_innen bei der Korrektur eines positiv wertend/akzeptierend, negativ wertend/ablehnend oder neutral/gemischt/sonstig äußerten. Insbesondere zeigt sich in den Medianunterschieden der prozentuellen Häufigkeit positiv wertender/akzeptierender bzw. negativ wertender/ablehender Äußerungen je Schülerlösungstext analog zur Punkteverteilung der Teilnehmer_innen (vgl. Unterabschnitt 6.3.1) ein empirischer Hinweis auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteilsgenese.

- Zweitens zeigten sich bei den Teilnehmer_innen deutliche empirische Hinweise für eine Orientierung an der kriterialen Bezugsnorm bei der Leistungsurteilsgenese über einen bestimmten Schülerlösungstext, bzw. wenn sie bestimmtes Schülerlösungstextmerkmal (fachlich-konzeptueller Eindruck, sprachliche Realisierung, sonstige Merkmale) feststellen und beurteilen. Nicht auszuschließen ist allerdings, dass es sich hierbei aber in Teilen um Artefakte der Laborsituation handelt.
- Drittens ließen sich empirische Hinweise auf unterschiedliche Tendenzen hinsichtlich einer Fähigkeiten- oder Defizitorientierung identifizieren, je nachdem welche Schülerlösungstextmerkmale die Teilnehmer_innen bei der Leistungsfeststellung und -beurteilung fokussieren. Für die Feststellung und Beurteilung fachlich-konzeptueller Schülerleistungen zeigt sich dabei eine deutliche Tendenz zur einer Fähigkeitsorientierung (im Median waren 51.4 % der Äußerungen zum fachlich-konzeptuellen Eindruck der vier Schülerlösungstexte positiv wertend/akzeptierend). Für sprachliche Schülerleistungen konnte hingegen eine deutliche Tendenz zur Defizitorientierung ausfindig gemacht werden (im Median waren 69.6 % der Äußerungen zur sprachlichen Realisierung der vier Schülerlösungstexte negativ wertend/ablehnend).

Bei den im vorangegangenen Unterabschnitt dargestellten Befunden handelt es sich um mittlere Tendenzen in der Logik des Handelns aller 21 Teilnehmer_innen im Rahmen der Laborsituation der Hauptstudie. Aus diesen Befunden lassen sich daher keine Aussagen über die Leistungsurteilsgenese einzelner Teilnehmer_innen treffen. Des Weiteren ist anzumerken, dass sich keinen Aussagen darüber treffen lassen, inwieweit die dargestellten Tendenzen für den gesamten Leistungsurteilsgenese der Teilnehmer_innen im Rahmen der Laborsituation von tragender Bedeutung waren. Grund hierfür ist, dass für die in Phase 2c vorgenommenen Analyse lediglich ausgewählte Codierungen der Laut-Denk-Protokolle der Teilnehmer_innen herangezogen wurden.

6.4. Analyse der Daten aus den retrospektiven Befragungen

In diesem Unterkapitel werden die Analysen der Daten aus den retrospektiven Befragungen und die hierbei gewonnenen Teilbefunde vorgestellt. Dem Mixed-Methods-Triangulationsdesign zur geplanten Auswertung der in der Hauptstudie erhobenen Daten entsprechend (vgl. Abschnitt 5.4.2) gliedert sich dieses Unterkapitel in die folgenden Abschnitte:

Abschnitt 6.4.1: Eine qualitative Analyse der Verbaldaten der 21 Teilnehmer_innen zu den Paarvergleichen mittels eines inhaltsanalytischen Vorgehens.

Abschnitt 6.4.2: Eine quantitative Analyse der Entscheidungen der 21 Teilnehmer_innen bezüglich der fachlich-konzeptuellen Qualität bzw. der Qualität der sprachlichen Realisierung der Schülerlösungstexte in den Paarvergleichen.

Analog zum Vorgehen in Unterkapitel 6.3 beginnt jeder dieser Abschnitte zunächst mit methodischen Vorbemerkungen, in denen das zur Analyse des entsprechenden Teildatensatzes verwendete Auswertungsverfahren beschrieben wird. Anschließend erfolgt eine Darstellung der Ergebnisse, sowie eine Interpretation der gewonnenen Befunde zu Forschungsfrage (F1) und/oder (F2). Jeder Abschnitt endet mit einer Darstellung der Limitationen der gewonnenen Teilbefunde.

6.4.1. Qualitative Analyse der Verbaldaten aus den retrospektiven Befragungen[146]

6.4.1.1. Methodische Vorbemerkungen

Wie in Abschnitt 5.4.1 beschrieben, wurden den Teilnehmer_innen in der retrospektiven Befragung die von ihnen bereits korrigierten vier Schülerlösungstexte noch einmal paarweise vorgelegt. Dabei erhielten sie (erstmalig explizit) die Instruktion einzuschätzen und zu begründen, ob einer der beiden ihnen vorliegenden Schülerlösungstexte „fachlich [bzw. sprachlich] besser ist, oder ob sie fachlich [bzw. sprachlich] gleich gut sind“ (Anhang C.3 Durchführungsmanual, S. 9). Zudem sollten die Teilnehmer_innen hierbei außer Acht lassen, „[o]b evtl. eine der beiden Antworten sprachlich [bzw. fachlich] besser ist“ (ebd.; zur Illustration siehe Transkriptauszug in Tabelle 6.29). Dieser Teil der retrospektiven Befragung wird im Folgenden als *fachlich-konzeptuelle* und *sprachliche Paarvergleiche* bezeichnet.

Ziel der qualitativen Analyse der Daten aus den retrospektiven Befragungen war es, die Einschätzungen der Teilnehmer_innen im Rahmen der fachlich-konzeptuellen und sprachlichen Paarvergleiche, sowie die Beurteilsungskriterien, welche die Teilnehmer_innen für die Begründung ihrer Einschätzung herangezogen haben, aus den erhobenen Verbaldaten herauszuarbeiten. Aufgrund dessen erfolgte die Analyse durch eine *inhaltlich strukturierende qualitative Inhaltsanalyse* nach Mayring (2015), da diese Variante der qualitativen Inhaltsanalyse speziell dafür ausgelegt ist...

> „[...] bestimmte Themen, Inhalte, Aspekte aus dem Material herauszufiltern und zusammenzufassen. Welche Inhalte aus dem Material extrahiert werden sollen wird durch theoriegeleitet entwickelte Kategorien und (sofern notwendig) Unterkategorien bezeichnet. Nach der Bearbeitung des Textes mittels des Kategoriensystems [...] wird das in Form von Paraphrasen extrahierte Material zunächst pro Unterkategorie, dann pro Hauptkategorie zusammengefasst.“ (ebd., S. 103)

In den folgenden Unterabschnitten wird zunächst das genaue Vorgehen bei der inhaltlich strukturierenden qualitativen Inhaltsanalyse der erhobenen Verbaldaten beschrieben. Anschließend erfolgt die Darstellung der durch dieses Vorgehen gewonnenen Erkenntnisse.

[146] Teile dieses Abschnitts stellen eine überarbeitete und erweiterte Fassung von Feser & Höttecke (2017a) dar.

Transkript der retrospektiven Befragung mit Segmentierung		
		[...]
[473]	**Leitung:**	**Beurteilen Sie, ob eine der beiden Antworten sprachlich besser ist, oder ob sie sprachlich gleich gut sind. Ob eventuell eine der beiden Antworten fachlich besser ist, soll hierbei komplett unberücksichtigt bleiben. Bitte begründen Sie Ihre Entscheidung.**
[474]	**Leitung:**	**Also es geht jetzt nicht mehr darum...**
[475]	Herr Carboni:	Es geht um die Sprache, nicht um die Inhalte.
[476]	**Leitung:**	**Genau. Und wir fangen wieder an mit A und B.**
[477]	Herr Carboni:	Sprachlich ist auf jeden Fall B besser. A ist sehr legere Alltagssprache. Formal... Also Satzbau beziehungsweise Ausdruck ist da wesentlich schlechter. Zum Teil... Also sprachlich ist auf jeden Fall Antwort B besser.
[478]	**Leitung:**	**Okay. A und C jetzt mal im Vergleich.**
[479]	Herr Carboni:	Auch hier ist für mich eindeutig Antwort C, die sprachlich qualitativ bessere Version. Vollständige Sätze, kurze Sätze, Syntax, Grammatik, alles in Ordnung.
		[...]

Tabelle 6.29.: Auszug aus dem segmentierten Transkript der retrospektiven Befragung von Herrn Carboni.

6.4.1.2. Erläuterungen zum Vorgehen bei der inhaltlich strukturierenden qualitativen Inhaltsanalyse

Wie Abbildung 6.19 schematisch veranschaulicht, lässt sich die inhaltlich strukturierende qualitative Inhaltsanalyse der Verbaldaten aus der retrospektiven Befragung in insgesamt 5 Schritte unterteilen:

Im ersten Schritt wurde der Leitfaden der retrospektiven Befragung (vgl. Anhang C.3 Durchführungsmanual, S. 8 u. f.) in ein deduktives Kategoriensystem überführt. Dieses Kategoriensystem ist in Tabelle 6.30 dargestellt. Dessen hierarchische Struktur entspricht den beiden Instruktionen, die die Teilnehmer_innen für die Paarvergleiche der Schülerlösungstexte erhalten haben (vgl. Anhang C.3 Durchführungsmanual, S. 9), den 6 Paarvergleichen, die die Teilnehmer_innen pro Instruktion vorgenommen haben, sowie den drei möglichen Einschätzungen, die ein_e Teilnehmer_in je Paarvergleich vornehmen konnte (z. B. Schülerlösungstext A ist sprachlich besser als B; Schülerlösungstext A ist sprachlich schlechter als B; Schülerlösungstext A und B sind sprachlich gleich gut). Das Kategoriensystem besteht daher aus insgesamt 2 Kategorien, mit jeweils 6 Subkategorien, die wiederum aus jeweils 3 Subsubkategorien aufgebaut sind (vgl. Tabelle 6.30).

Im zweiten Schritt wurden die Transkripte der retrospektiven Befragung aller Teilnehmer_innen mit Hilfe des zuvor entwickelten Kategoriensystems codiert. Wie in Unterkapitel 6.2 beschrieben und im Transkriptauszug in Tabelle 6.29 illustriert, wurden die Basistranskripte der retrospektiven Befragung in die Fragen der Leitung und Antworten der Teilnehmer_innen segmentiert. Für die Codierung der Transkripte wurden die Segmente, in denen ein_e Teilnehmer_in einen Paarvergleich vornimmt (also die vollständigen Antworten der Teilnehmer_innen zu einer Frage der Leitung) als Codiereinheiten definiert. Zudem wurde die Codierregel festgelegt, dass jeder Codiereinheit genau eine

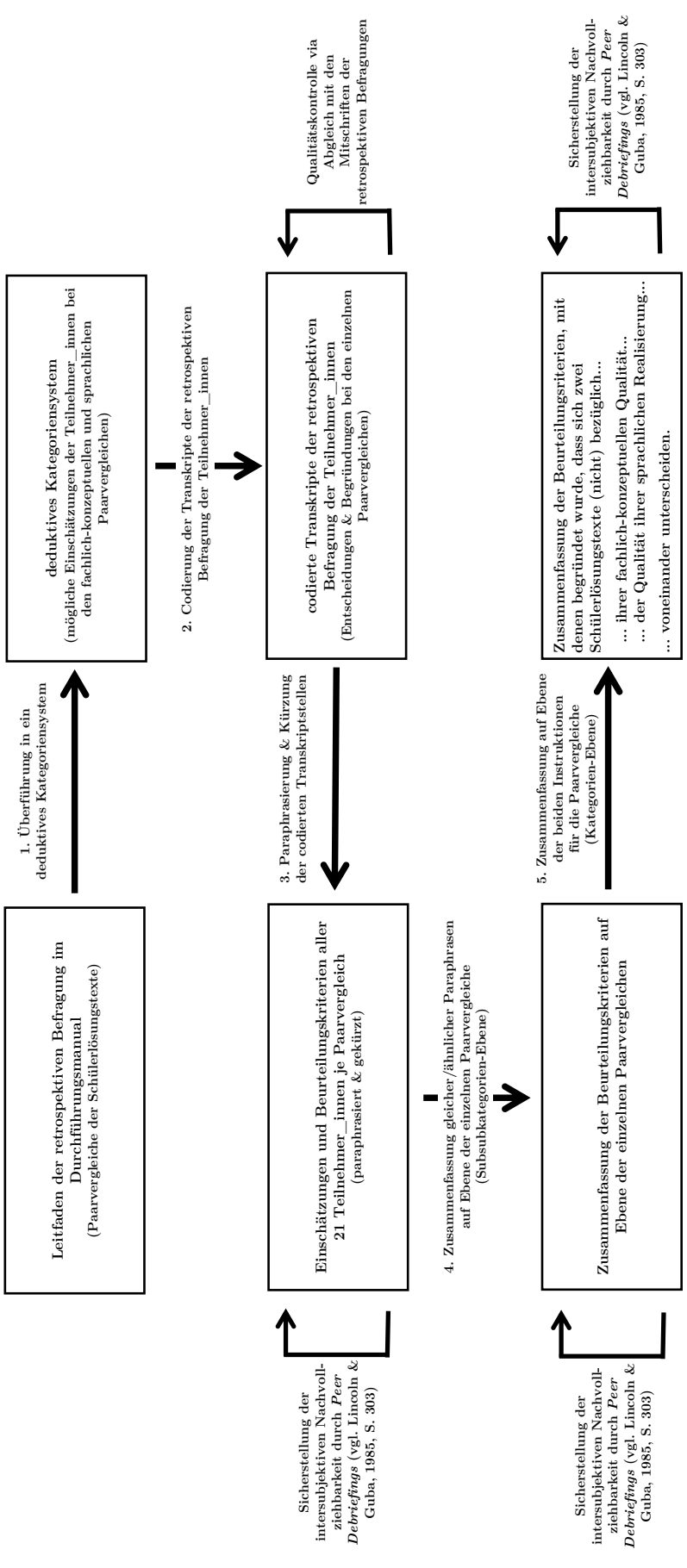


Abbildung 6.19.: Ablaufschema des Vorgehens bei der inhaltlichen strukturierenden qualitativen Inhaltsanalyse der Verbaldaten der retrospektiven Befragung.

Kategorien	Subkategorien	Subsubkatgeorien
Paarvergleiche der Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität	Schülerlösungstext A und B	A ist besser als B (A>B)
		A ist schlechter als B (A<B)
		A & B sind gleich gut (A=B)
	Schülerlösungstext A und C	A ist besser als C (A>C)
		A ist schlechter als C (A<C)
		A & C sind gleich gut (A=C)
	Schülerlösungstext A und D	A ist besser als D (A>D)
		A ist schlechter als D (A<D)
		A & D sind gleich gut (A=D)
	Schülerlösungstext B und C	B ist besser als C (B>C)
		B ist schlechter als C (B<C)
		B & C sind gleich gut (B=C)
	Schülerlösungstext B und D	B ist besser als D (B>D)
		B ist schlechter als D (B<D)
		B & D sind gleich gut (B=D)
	Schülerlösungstext C und D	C ist besser als D (C>D)
		C ist schlechter als D (C<D)
		C & D sind gleich gut (C=D)
Paarvergleiche der Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung	Schülerlösungstext A und B	A ist besser als B (A>B)
		A ist schlechter als B (A<B)
		A & B sind gleich gut (A=B)
	Schülerlösungstext A und C	A ist besser als C (A>C)
		A ist schlechter als C (A<C)
		A & C sind gleich gut (A=C)
	Schülerlösungstext A und D	A ist besser als D (A>D)
		A ist schlechter als D (A<D)
		A & D sind gleich gut (A=D)
	Schülerlösungstext B und C	B ist besser als C (B>C)
		B ist schlechter als C (B<C)
		B & C sind gleich gut (B=C)
	Schülerlösungstext B und D	B ist besser als D (B>D)
		B ist schlechter als D (B<D)
		B & D sind gleich gut (B=D)
	Schülerlösungstext C und D	C ist besser als D (C>D)
		C ist schlechter als D (C<D)
		C & D sind gleich gut (C=D)

Tabelle 6.30.: Deduktives Kategoriensystem zur Analyse der Verbaldaten der retrospektiven Befragung.

Subsubkategorie des Kategoriensystems zuzuweisen ist. Beispielsweise wurden die Segmente 477 und 479 im in Tabelle 6.29 dargestellt Transkriptauszug mit den Subsubkategorien „A ist schlechter als B“ bzw. „A ist schlechter als C“ der Kategorie „Paarvergleiche der Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung“ codiert.

Anschließend wurde die Codierung der Paarvergleiche in den Transkripten einer Qualitätskontrolle unterzogen. Dies geschah, indem überprüft wurde, ob die Codierungen mit den dementsprechenden Mitschriften der Leitung während der Paarvergleiche (vgl. Abschnitt 5.4.1; Anhang C.3 Durchführungsmanual, S. 9) übereinstimmen[147]. Als Maß für diese Überstimmung wurde sowohl für die fachlich-konzeptuellen, als auch für die sprachlichen Paarvergleiche – also für jede Kategorie des Kategoriensystems – jeweils der von Brennan & Prediger (1981) vorgeschlagene κ-Koeffizient bestimmt (vgl. auch Unterabschnitt 6.3.2.2). Für die fachlich-konzeptuellen Paarvergleiche ergab sich eine perfekte Übereinstimmung zwischen der Codierung der Transkripte und den entsprechenden Leitungsmitschriften während der retrospektiven Befragung ($\kappa = 1.00$). Bei den sprachlichen Paarvergleichen zeigten sich hingegen minimale Abweichungen ($\kappa = .97$). Diese Abweichungen konnten auf einen Fehler im Transkript von Herrn Balke und eine einzelne fehlerhafte Mitschrift bei den sprachlichen Paarvergleichen von Herrn Jonuzi zurückgeführt werden. Nach Bereinigung dieser Fehler ergab sich auch für die sprachlichen Paarvergleiche eine perfekte Übereinstimmung zwischen der Codierung der Transkripte und den entsprechenden Leitungsmitschriften.

Entsprechend dem von Mayring (2015) vorgeschlagenen Vorgehen erfolgte im dritten bis fünften Schritt der inhaltlich strukturierenden qualitativen Inhaltsanalyse das Herausfiltern und Zusammenfassen der von den Teilnehmer_innen bei den Paarvergleichen verwendeten Beurteilungskriterien aus den zuvor codierten Transkriptstellen:

Im dritten Schritt wurde zunächst eine regelgeleitete Paraphrasierung vorgenommen. Wie in Abbildung 6.20 exemplarisch veranschaulicht wurden hierbei in jeder zuvor codierten Transkriptstelle „nicht (oder wenig) inhaltstragende[148] Textbestandteile [ausgelassen] [...] [und nur] die inhaltstragenden Textstellen [wurden] auf eine einheitliche Sprachform[,] [...] [in] grammatikalische[r] Kurzform [transformiert]“ (Mayring, 2015, S. 72). Hierdurch wurde für jeden Paarvergleich aller Teilnehmer_innen eine paraphrasierte und gekürzte Version ihrer vorgenommenen Einschätzung und der von ihnen hierbei angeführten Beurteilungskriterien gewonnen.

Anschließend wurden alle Paraphrasen, die aus Transkriptstellen hervorgingen und mit einer bestimmten Subsubkategorie des Kategoriensystems codiert wurden, fallübergreifend zusammengefasst (vierter Schritt). Dabei wurden bedeutungsgleiche aber unterschiedlich formulierte Paraphrasen sprachlich vereinheitlicht, sowie bedeutungsähnliche Paraphrasen gebündelt und in eine generalisierte Paraphrase überführt (vgl. Mayring, 2015, S. 72). Hierdurch wurde für jede Einschätzung, die die Teilnehmer_innen bei den

[147] Das Ergebnis der Codierung nach der durchgeführten Qualitätskontrolle ist in Anhang F für die einzelnen Teilnehemer_innen tabellarisch zusammengefasst.

[148] Als nicht (oder wenig) inhaltstragende Textteile wurden solche festgelegt, in denen der_die Teilnehmer_in keine Beurteilungskriterien für seine_ihre vorgenommene Einschätzung expliziert.

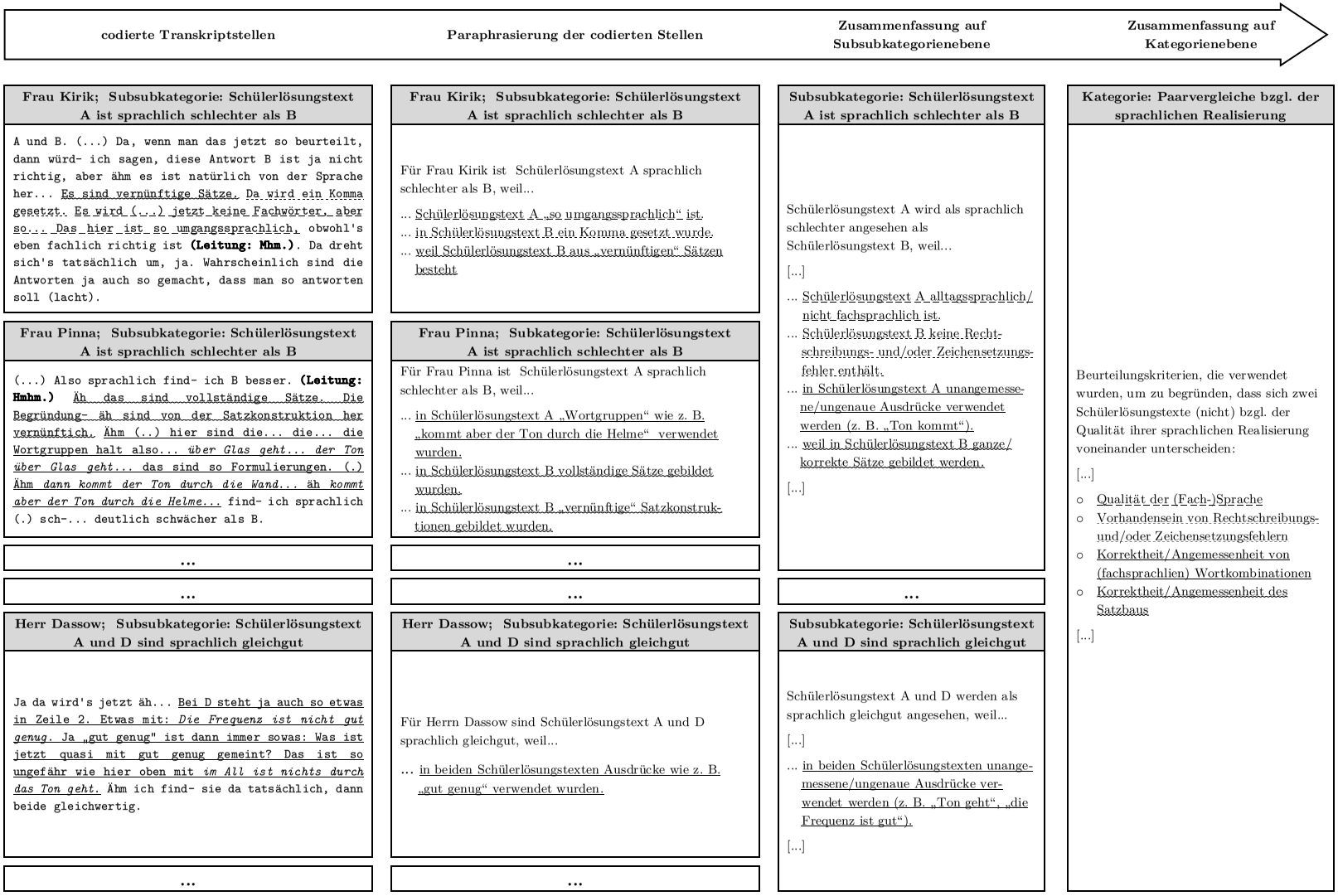


Abbildung 6.20.: Exemplarische Veranschaulichung des Vorgehens, mit dessen Hilfe aus den codierten Transkriptstellen die bei den Paarvergleichen verwendeten Beurteilungskriterien gewonnen wurden. Die verschiedenen Unterstreichungen verdeutlichen, welche Textteile im Analyseprozess paraphrasiert bzw. zusammengefasst wurden.

Paarvergleichen vorgenommen haben, eine erste fallübergreifende Zusammenfassung der herangezogenen Beurteilungskriterien gewonnen (exemplarisch veranschaulicht in Abbildung 6.20).

Im fünften und letzten Schritt wurden die zuvor gewonnenen ersten fallübergreifenden Zusammenfassungen weiter zusammengefasst. Diese zweite fallübergreifende Zusammenfassung erfolgte auf Ebene der beiden Kategorien des Kategoriensystems, mit dem die Transkripte codiert wurden. Dabei wurde analog zu dem im vierten Schritt der inhaltlich strukturierenden qualitativen Inhaltsanalyse angewandten Vorgehen vorgegangen (sprachliche Vereinheitlichungen bedeutungsgleicher Paraphrasen; Bündelung bedeutungsähnlicher Paraphrasen). Hierdurch wurde eine fallübergreifende Zusammenfassung der Beurteilungskriterien gewonnen, mit denen die Teilnehmer_innen begründeten, dass sich zwei Schülerlösungstexte (nicht) bezüglich...

... ihrer fachlich-konzeptuellen Qualität...

... der Qualität ihrer sprachlichen Realisierung...

... voneinander unterscheiden (exemplarisch veranschaulicht in Abbildung 6.20; in Unterabschnitt 6.4.1.3 erfolgt eine detaillierte Darstellung).

Um im dritten, vierten und fünften Schritt der inhaltlich strukturierenden qualitativen Inhaltsanalyse die intersubjektive Nachvollziehbarkeit der Analyse sicherzustellen, wurden die Zwischenergebnisse, die in diesen Schritten gewonnen wurden, jeweils einem sogenannten *Peer Debriefing* unterzogen. Hierbei handelt es sich um eine diskursive Form der Herstellung von intersubjektiver Nachvollziehbarkeit (vgl. Steinke, 1999, S. 214), die sich nach Lincoln & Guba (1985) allgemein beschreiben lässt als...

> „[...] the process of exposing oneself to a disinterested peer in a manner paralleling an analytic session and for the purpose of exploring aspects of the inquiry that might otherwise remain only implicit within the inquirer's mind." (ebd., S. 308)

Die im Anschluss an den dritten, vierten und fünften Schritt der inhaltlich strukturierenden qualitativen Inhaltsanalyse durchgeführten Peer Debriefings erfolgten in Zusammenarbeit mit Mitgliedern der Arbeitsgruppe Physikdidaktik der Universität Hamburg, sowie Mitgliedern der Arbeitsgemeinschaft qualitative Inhaltsanalyse der Graduiertenschule der Fakultät für Erziehungswissenschaft der Universität Hamburg. Den Teilnehmer_innen der Peer Debriefings wurde zunächst die Vorgehensweise, mit dem die zu diskutierenden Zwischenergebnisse gewonnen wurden, erläutert. Anschließend wurde gemeinsam diskutiert, inwieweit die vorliegenden Zwischenergebnisse hinreichend präzise, sowie umfänglich formuliert sind, sodass sie eine „(kritische) Verständigung [...] zwischen Forscher (der eine Studie durchführt) und Leser (der Studie) [...] ermöglichen" (Steinke, 1999, S. 207), sowie dementsprechend gemeinschaftlich umformuliert, ergänzt oder gekürzt.

6.4.1.3. Befunde der inhaltlich strukturierenden qualitativen Inhaltsanalyse

Die Ergebnisse der inhaltlich strukturierenden qualitativen Inhaltsanalyse sind in den Tabellen 6.31 bis 6.34 dargestellt.

In Tabelle 6.31 und 6.32 sind die Beurteilungskriterien aufgelistet, die die Teilnehmer_innen im Rahmen der fachlich-konzeptuellen und sprachlichen Paarvergleiche verwendeten, um ihre Entscheidungen zu begründen. Ferner ist sowohl in Tabelle 6.31, als auch in Tabelle 6.32 für jedes Beurteilungskriterium jeweils ein exemplarischer fachlich-konzeptueller bzw. sprachlicher Paarvergleich eines_einer Teilnehmers_Teilnehmerin angegeben, in dem dieses Beurteilungskriterium für die Begründung einer Einschätzung herangezogen wurde.

In den Tabellen 6.33 und 6.34 sind erstens die absoluten Codehäufigkeiten der Subsubkategorien des Kategoriensystems, mit dem die Paarvergleiche codiert wurden, aufgeführt. Diese entsprechen der Anzahl an Teilnehmer_innen, die im Rahmen der fachlich-konzeptuellen bzw. sprachlichen Paarvergleiche zu einer bestimmten Einschätzung gelangt sind. Zweitens ist in Tabellen 6.33 und 6.34 aufgeführt, wie viele der 21 Teilnehmer_innen ein bestimmtes Beurteilungskriterien für welche Einschätzungen im Rahmen der Paarvergleiche herangezogen haben. Diese Information wurde aus den Zwischenergebnissen der inhaltlich strukturierenden qualitativen Inhaltsanalyse gewonnen, indem zurückverfolgt wurde, welche Paraphrasen aus dem dritten Analyseschritt, im fünften Schritt zu welchem Beurteilungskriterium zusammengefasst wurden. Drittens ist in beiden Tabellen hervorgehoben, welche Einschätzungen bei den fachlich-konzeptuellen bzw. den sprachlichen Paarvergleichen, mit der kontrastierenden Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie übereinstimmen (vgl. Abschnitt 5.3.4).

Alles in allem zeigen sich in den Tabellen 6.31 bis 6.34 die folgenden Befunde:

1. Bei den fachlich-konzeptuellen Paarvergleichen haben die Teilnehmer_innen insgesamt 13 verschiedene Beurteilungskriterien für die Begründung ihrer Einschätzung herangezogen (vgl. Tabelle 6.31), wohingegen sie für die sprachlichen Paarvergleiche insgesamt 20 verschiedene Beurteilungskriterien verwendeten (vgl. Tabelle 6.32). Insbesondere haben die Teilnehmer_innen die folgenden 7 Beurteilungskriterien sowohl bei den fachlich-konzeptuellen, als auch bei den sprachlichen Paarvergleichen eingesetzt:

   (i) Die Differenziertheit/Komplexität des Textes

   (ii) Die Entsprechung der persönlichen Erwartungen

   (iii) Die Qualität der (Fach-)Sprache

   (iv) Die Quantität an Fachwörtern

   (v) Die Strukturiertheit/Gliederung des Textes

   (vi) Der/Die Verdichtungsgrad/Präzision des Textes

   (vii) Das Vorhandensein von Redundanz

Es lässt sich daher vermuten, dass die Teilnehmer_innen auch in ihrem Berufsalltag für die Feststellung und Beurteilung von fachlich-konzeptuellen und sprachlichen Schülerleistungen zum Teil auf dieselben Beurteilungskriterien zurückgreifen (Forschungsfrage (F1)), womit insbesondere das Potenzial für eine Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen einhergeht (Forschungsfrage (F2)).

2. Wie durch einen Vergleich von Tabelle 6.31 und 6.32 deutlich wird, sind einige der von den Teilnehmer_innen verwendeten Beurteilungskriterien eher als holistisch zu bezeichnen. Dies betrifft insbesondere diejenigen Beurteilungskriterien, die sowohl bei den fachlich-konzeptuellen, als auch bei den sprachlichen Paarvergleichen eingesetzt wurden (z. B. das Kriterium „Entsprechung der persönlichen Erwartungen“). Die meisten der von den Teilnehmer_innen verwendeten Beurteilungskriterien sind jedoch eher als analytisch zu bezeichnen (z. B. das Kriterium „Quantität an Fachwörtern“). *Holistisch*[149] bedeutet in diesem Zusammenhang, dass ein Beurteilungskriterium dazu dient, einen (eher) globalen/generellen Eindruck über einen Schülerlösungstext zum Ausdruck zu bringen. Demgegenüber dienen *analytische*[149] Beurteilungskriterien dazu, bestimmte Teilleistungen differenziert erfassen zu können (bei einem Textprodukt z. B. den Wortschatz, den Satzbau, die Rechtschreibung, usw.). Besonders hervorzuheben sind allerdings die Beurteilungskriterien „Qualität der (Fach-)Sprache“ und „Quantität an Fachwörtern“. Bei beiden handelt es sich eindeutig um Beurteilungskriterien, die die sprachliche Realisierung eines Schülerlösungstextes betreffen, was umso bemerkenswerter ist, als dass diese Kriterien von den befragten Lehrkräften (trotz explizit anderer Aufforderung!) auch dazu verwendet wurden, Unterschiede zwischen zwei Schülerlösungstexten bezüglich ihrer fachlich-konzeptuellen Qualität zu begründen. Anders ausgedrückt zeigt sich hier also ein sehr deutlicher empirischer Hinweis auf eine Konfundierung von fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung durch die Teilnehmer_innen, wie in Forschungsfrage (F2) vermutet.

3. Wie sich in Tabelle 6.33 zeigt, weisen die Einschätzungen der Teilnehmer_innen bei den fachlich-konzeptuellen Paarvergleichen eine sehr hohe intersubjektive Übereinstimmung auf. Lediglich beim Paarvergleich von Schülerlösungstext B und D, also den beiden Schülerlösungstexten, die gemäß der Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie eine geringe fachlich-konzeptuelle Qualität aufweisen, zeigt sich eine nennenswerte Nichtübereinstimmung der Einschätzungen (für 14 Teilnehmer_innen ist Schülerlösungstext B „besser“ als Schülerlösungstext D; für 7 Teilnehmer_in sind beide Texte „gleich gut“). Als Kennwert für diese sehr hohe intersubjektive Übereinstimmung lässt sich der von von Eye (2006) vorgeschlagene κ-Koeffizient[150] angeben. Für diesen ergibt sich ein Wert von $\kappa_{von\,Eye} = .91$.

[149] Für eine detaillierte Begriffsbestimmung, sowie eine kritische Auseinandersetzung mit der Dichotomie „holistische versus analytische Beurteilungskriterien“ siehe Arras (2007, S. 80 u. f.).

[150] Von-Eyes-κ ist eine Verallgemeinerung des von Brennan & Prediger (1981) vorgeschlagenen κ-Koeffizient für mehr als 2 Beurteiler_innen (hier: die 21 Teilnehmer_innen der Hauptstudie). Für

Aus den Spalten von Tabelle 6.33 geht allerdings auch hervor, dass jeweils nur von einem Bruchteil der Teilnehmer_innen, die bei einem der fachlich-konzeptuellen Paarvergleiche zu derselben Entscheidung gekommen sind, auch dieselben Beurteilungskriterien zur Entscheidungsbegründung herangezogen wurden. Alles in allem weisen also (lediglich) die Einschätzungen der Teilnehmer_innen im Rahmen der fachlich-konzeptuellen Paarvergleiche eine sehr hohe intersubjektive Übereinstimmung auf, was jedoch nicht für die Begründungen dieser Einschätzungen gilt.

4. Des Weiteren zeigt sich in Tabelle 6.33, dass beim fachlich-konzeptuellen Paarvergleich von Schülerlösungstext A und C die Einschätzungen aller Teilnehmer_innen von der kontrastierenden Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie abweicht: Alle Teilnehmer_innen sind zu der Einschätzung gelangt, dass Schülerlösungstext C „besser“ als Schülerlösungstext A ist. Bemerkenswert ist dabei vor allem, dass eine Mehrheit der Teilnehmer_innen (12) die „Qualität der (Fach-)Sprache“ – also ein nicht die fachlich-konzeptuelle Qualität eines Schülerlösungstextes betreffendes Merkmal – als Beurteilungskriterium bei der Begründung dieser Einschätzung genannt haben. Hier zeigt sich also ein deutlicher empirischer Hinweis auf eine Konfundierung von fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung durch die Teilnehmer_innen (Forschungsfrage (F2)). Insbesondere die Äußerungen von Frau Novack beim fachlich-konzeptuellen Paarvergleich von Schülerlösungstext A und C stützen diese Interpretation: „[Man kann] bei Antwort A ähm schon dies- Konzept ähm finden[,] [...] aber (.) das gehört für mich schon dazu, dass er das auch dann ähm formuliert.“ (Seg. 530). Sie relativiert also fachlich-konzeptuelle Merkmale des Schülerlösungstextes A und begründet dies damit, dass der Text zu viele Mängel bezüglich einer sprachlichen Norm aufweist, deren Beherrschung von ihr vorausgesetzt wird.

5. Wie sich in Tabelle 6.34 zeigt, weisen die Einschätzungen der Teilnehmer_innen bei den sprachlichen Paarvergleichen ebenfalls eine sehr hohe intersubjektive Übereinstimmung auf. Die Nichtübereinstimmung ist hier allerdings ausgeprägter als bei den fachlich-konzeptuellen Paarvergleichen. Bei beiden Paarvergleichen, in denen zwei Schülerlösungstexte mit vergleichbarer Qualität in ihrer sprachlichen Realisierung miteinander verglichen werden sollten (Paarvergleich von Schülerlösungstext A und D, sowie Schülerlösungstext B und C), zeigt sich jeweils eine nennenswerte Nichtübereinstimmung der Einschätzungen. Aus diesem Grund ist auch der Kennwert für die intersubjektive Übereinstimmung zwischen den Teilnehmer_innen bei den sprachlichen Paarvergleichen geringer als bei den fachlich-konzeptuellen Paarvergleichen. Für die sprachlichen Paarvergleiche ergibt sich ein Wert von $\kappa_{von\,Eye} = .79$. Wie zudem aus den Spalten von Tabelle 6.34 hervorgeht, hat jeweils nur ein Bruchteil der Teilnehmer_innen, die bei einem der sprachlichen Paarvergleiche zu derselben Einschätzung gekommen sind, auch dieselben Beurteilungskriterien zur Entscheidungsbegründung herangezogen. Zusammengefasst weisen also auch bei den

von-Eyes-κ gelten – als Faustregel – Werte ab .40 als „akzeptabel“, ab .60 als „gut“ und Werte ab .75 als „sehr gut“ (vgl. Wirtz & Caspar, 2002, S. 59).

sprachlichen Paarvergleichen (lediglich) die Einschätzungen der Teilnehmer_innen eine sehr hohe intersubjektive Übereinstimmung auf, nicht jedoch die Begründungen dieser Einschätzungen.

6. Ferner zeigt sich in Tabelle 6.34, dass bei sprachlichen Paarvergleichen, in denen zwei Schülerlösungstexte miteinander verglichen werden sollen, die beide eine geringe bzw. hohe Qualität in ihrer sprachlichen Realisierung aufweisen (sprachlicher Paarvergleich von Schülerlösungstext A und D, bzw. von Schülerlösungstext B und C), jeweils eine Mehrheit der Teilnehmer_innen eine Einschätzung vorgenommen hat (15 bzw. 13 Teilnehmer_innen), die von der kontrastierenden Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie abweicht. Bei beiden Paarvergleichen ist dabei bemerkenswert, dass das – eher holistische – Beurteilungskriterium „Qualität der (Fach-)Sprache“ von auffällig vielen Teilnehmer_innen verwendet wurde, auch wenn sie zu unterschiedlichen Einschätzungen gekommen sind. Dies lässt vermuten, dass die einzelnen Teilnehmer_innen ein unterschiedliches Verständnis davon haben, wodurch sich die Qualität der sprachlichen Realisierung eines Schülerlösungstextes auszeichnet. Hier zeigt sich also ein Hinweis darauf, dass die Teilnehmern_innen bei der Feststellung und Beurteilung sprachlicher Schülerleistungen zum Teil auf unterschiedliche Ressourcen zurückgreifen (Forschungsfrage (F1)).

6.4.1.4. Limitation und Zwischenfazit

Die Vielzahl der im Rahmen der inhaltlich strukturierenden qualitativen Inhaltsanalyse identifizierten Beurteilungskriterien sprechen dafür, dass die Teilnehmer_innen in der retrospektiven Befragung insgesamt auf ein facettenreiches Wissen und Können zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen zurückgegriffen haben (Forschungfrage (F1)). Die sehr hohe intersubjektive Übereinstimmung der Einschätzungen im Rahmen der Paarvergleiche, dass aber gleichartige Einschätzungen zum Teil sehr unterschiedlich begründet wurden, spricht zudem dafür, dass die Teilnehmer_innen zu vergleichbaren Gesamteinschätzungen der fachlich-konzeptuellen und/oder sprachlichen Leistungen von Schüler_innen gelangten, obwohl sie sich in ihrem Vorgehen bei der Leistungsfeststellung und Beurteilung teilweise deutlich voneinander unterscheiden (Heranziehen verschiedener Beurteilungskriterien). Vor allem aber konnten im Rahmen der fachlich-konzeptuellen und sprachlichen Paarvergleiche die folgenden Befunde identifiziert werden:

- Insgesamt 7 Beurteilungskriterien wurden von den Teilnehmer_innen sowohl bei den fachlich-konzeptuellen Paarvergleichen, als auch bei den sprachlichen Paarvergleichen für die Begründung ihrer Einschätzungen eingesetzt.
- Insbesondere wurden die Beurteilungskriterien „Quantität an Fachwörtern“ und „Qualität der (Fach-)Sprache“ von den Teilnehmer_innen unter anderem auch da-

zu verwendet, um zwei Schülerlösungstexte hinsichtlich ihrer fachlich-konzeptuellen Qualität miteinander zu vergleichen.

- Alle 21 Teilnehmer_innen sind im Rahmen der fachlich-konzeptuellen Paarvergleiche – entgegen der Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie – zu der Einschätzung gelangt, dass sich die Schülerlösungstexte A und C in ihrer fachlich-konzeptuellen Qualität voneinander unterscheiden. Bemerkenswert ist dabei, dass 12 Teilnehmer_innen bei der Begründung dieser Einschätzung (unter anderem) das Beurteilungskriterium „Qualität der (Fach-)Sprache“ genannt haben.

Diese drei Befunde lassen sich als sehr deutliche empirische Hinweise auf eine Konfundierung von fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung durch die Teilnehmer_innen interpretieren (Forschungsfrage (F2)). Darüber, wie ausgeprägt diese Konfundierung ist, lässt sich jedoch auf Grundlage dieser rein qualitativen Befunde keine Aussage treffen. Aus diesem Grund wurden die in der retrospektiven Befragung gewonnenen Daten einer zusätzlichen quantitativen Analyse unterzogen. Diese Analyse wird im folgenden Abschnitt dargestellt.

Kriterium	exemplarischer Transkriptauszug
Angemessenheit des Fachwissens in Bezug zur Aufgabenstellung	C und D...[...] C ist besser, ähm weil D inadäquater Weise mit der *Frequenz* argumentiert wird und Medium nicht benannt... also das Schall -n Medium braucht [...] und das wird bei C geleistet. (Herr Mehlert, Paarvergleich Schülerlösungstext C und D, Seg. 872)
Anwendung von Fachmethoden	Ähm B und D [...] Ob sie... ob sie fachlich besser is-. Dann zähl- ich jetz- das fachmethodische Arbeiten zu dem Fachlichem mit zu [...] und würde dann sagen, Antwort B ist besser als Antwort D, weil da die Fachmethode, das ähm... (.) die Probleme zu trennen geleistet wurde und bei D nich-. (Herr Mehlert, Paarvergleich Schülerlösungstext B und D, Seg. 866-870)
Fachliche Korrektheit des Inhalts	Also ist A auf alle Fälle eben (..) besser. B hat eben äh gar nicht verstanden, worum es geht. Dass eben die Funkverbindung eben abgerissen ist [...] und eben auch nicht dadurch wieder zustande kommt, dass sich eben die Helme berühren. (Herr Lemos, Paarvergleich Schülerlösungstext A und B, Seg. 267)
Inhaltsreichtum/ Quantität von Fachwissen	Ja, also C war ja für mich das Highlight. [...] Insofern is- also ganz eindeutich klar, C is- fachlich besser. Öh weil es fachlich auch ausführlicher is-. Also der Grund... die Grundidee von äh... is- in A auch zu finden, die fachliche Grundidee, aber es is- fachlich ausführlicher und deswegen fachlich besser. Also es is- C. (Herr Trummer, Paarvergleich Schülerlösungstext A und C, Seg. 616)
Qualität des Bezugs zur Aufgabenstellung	Also, ich finde Antwort A besser, weil es ähm zum Thema gehört. Offenbar ist ja Thema „Schall". Und da ist es ja... ähm (.) ist es treffender. [...] Hier ist die Aufgabe falsch verstanden worden. Und auch da gibt es ein mhm mhm Prinzip, aber das ist nur sehr knapp ausgeführt und ähm passt nicht zu... also „Thema verfehlt" würde man knapp sagen. (Herr Uckermark, Paarvergleich Schülerlösungstext A und B, Seg. 694)
Vollständigkeit in Bezug zur Aufgabenstellung	Ich habe die Antwort A besser bewertet, zumindest mit 1 Punkt, als die Antwort B und äh die Begründung ist, dass in Antwort A beide Phänomene, also beide Fälle berücksichtigt worden sind während in äh Antwort B nur die zweite Situation kommentiert wurde. (Herr Carboni, Paarvergleich Schülerlösungstext A und B, Seg. 452)
Differenziertheit/ Komplexität des Textes	So, A, C. Dann ist für mich Antwort C die fachlich bessere. [...] Ähm, weil es noch differenzierter (.) beantwortet wurde. Insbesondere der 2. Teil, wann es denn sozusagen hörbar ist. Das ähm der Schall sich ähm ausbreitet. [...] Also da wird tatsächlich noch -ne Begründung versucht (.) zu geben, warum sich Schall dann über das Glas ausbreiten kann. (Frau Sohm, Paarvergleich Schülerlösungstext A und C, Seg. 490)
Entsprechung der persönlichen Erwartungen	A und C. Okay, da fand ich natürlich C besser, das war ja quasi gelöst wie meine Musterlösung. Von daher würd- ich sagen, da ist es gerechtfertigt, dass des besser ist. (Herr Abney, Paarvergleich Schülerlösungstext A und C, Seg. 427)
Qualität der (Fach-)Sprache	Ja. Ähm bei A und C ähm sehe ich auch -nen Unterschied. Da ist C äh fachlich besser. [...] Ähm, also es sind vor allem... ist vor allem die Fachsprache. (Herr Uckermark, Paarvergleich Schülerlösungstext A und C, Seg. 696)
Quantität an Fachwörtern	B und C. [...] Ja da is- ganz klar C besser. [...] Ähm und da is- ja ganz klar in Aufgabe C viel mehr Fachwörter, die aus dem... aus dem Bereich der Akustik kommen. Und deswegen würd- ich das hier besser bewerten fachlich. (Herr Rittershaus, Paarvergleich Schülerlösungstext B und C, Seg. 734-736)
Strukturiertheit/ Gliederung des Textes	B und C? [...] Ähm da fand ich auch C deutlich besser ähm, weil da des einfach strukturiert aufgeschrieben... [ist.] (Herr Abney, Paarvergleich Schülerlösungstext B und C, Seg. 431)
Verdichtungsgrad/ Präzision des Textes	C und D? Dann wieder C. C ist einfach sehr viel präziser, sehr viel genauer und D hat ja das Problem auch nicht erkannt. (Herr Dassow, Paarvergleich Schülerlösungstext C und D, Seg. 533)
Vorhandensein von Redundanz	C und D. [...] Äh. (...) Gut, ein Manko, was ich jetzt hier ähm gar nicht erwähnt hatte, bisher bei... bei D is- auch die Wiederholung, nich-? Also der erste und der letzte Teil sind ja praktisch äh praktisch identisch. (Herr Uckermark, Paarvergleich Schülerlösungstext C und D, Seg. 706-708)

Tabelle 6.31.: Beurteilungskriterien, die die Teilnehmer_innen bei den fachlich-konzeptuellen Paarvergleichen der Schülerlösungstexte verwendeten, um ihre Einschätzungen zu begründen.

Kriterium	exemplarischer Transkriptauszug
Angemessenheit der Zeitform	A und [...] B. (..) Ja also A sprachlich schlechter als B. [...] Also hier wern die richtigen äh Zeiten und Formen verwendet. Nein also Aufgabe B äh (.) besser als A. (Herr Onne, Paarvergleich Schülerlösungstext A und B, Seg. 287)
Attraktivität des Textes	Äh C und D. C. [...] Also des is- einfach -n schöner Text. So. Also der hat auch seine Mängel, aber wir sind ja auch in der Schule und nich-... (Herr Einert, Paarvergleich Schülerlösungstext C und D, Seg. 890-891)
Eindeutigkeit von Bezügen	B is- trotzdem sprachlich besser, [...] weil auch zum Beispiel sowas wie *sie haben sich nicht gehört.* Ähm is- es schon schöner auch mit *die beiden Astronauten* anzufangen damit klar ist, was mit *sie* gemeint ist. Auch das is- -ne Aufgabe. (Herr Geppert, Paarvergleich Schülerlösungstext B und D, Seg. 916)
Korrektheit/Angemessenheit des Satzbaus	(...) Also sprachlich find- ich B besser. [...] Äh das sind vollständige Sätze. Die Begründung- äh sind von der Satzkonstruktion her vernünftich. (Frau Pinna, Paarvergleich Schülerlösungstext A und B, Seg. 448)
Korrektheit/Angemessenheit von (Fach-)Wörtern	A und C. Ähm (...) auf jeden Fall C, weil das äh ... weil das auch die Eindeutichkeit der Sprache is-. Also da werden die eindeutigen Begriffe formuliert und auch richtich verwendet. (Herr Geppert, Paarvergleich Schülerlösungstext A und C, Seg. 896)
Korrektheit/Angemessenheit von (fachsprachlichen) Wortkombinationen	Also sin- beide nicht perfekt. [...] Aber -n leichter (..) leichter Vorteil gegenüber... bei B gegenüber A, was die (..) äh sprachliche Ausdrucksfähigkeit (.) betrifft. Also (.) das sieht man auch an solchen (.) Wörtern wie *der Ton geht...* (..) äh... *durch das Ton geht.* [...] -So sprachlich -n sehr, sehr einfaches Deutsch. Und das andere: *stellt eine Verbindung her.* Is- dann schon etwas höher. Also B is- besser. (Herr Feldner, Paarvergleich Schülerlösungstext A und B, Seg. 555)
Länge des Textes	B und C. [...] Also ähm ja ich würde jetz- einfach sprachlich doch ma- C besser bewerten, weil ich finde, dass es einfach mehr Text is- und er sich (.) auf so -nem hohen Niveau hält. (Herr Rittershaus, Paarvergleich Schülerlösungstext B und C, Seg. 759)
Lesbarkeit des Textes	Okay A und B. [...] Ganz klar Antwort B. Hier sind komplexere Sätze und irgendwie schön... sch-... sch-... schöner zu lesen. Also fällt mir leichter das zu lesen. (Herr Rittershaus, Paarvergleich Schülerlösungstext A und B, Seg. 747)
(mutmaßliche) Gewissenhaftigkeit des_der Schülers_Schülerin während des Schreibens	(...) Das is- irgendwie so -ne völlig unsinnige Wiederholung. (.) Äh da hat derjenige... (..) würd- ich einfach mal so reininterpretieren, der könnte viel besser, der könnte es auch weglassen, wenn er's nochmal durchgelesen hätte. War einfach nur schlampich. [...] Hat's hingeschrieben, zack weg damit. [...] So, dass is- das... das vermute ich dahinter. (Herr Trummer, Paarvergleich Schülerlösungstext A und D, Seg. 654)
Raffinesse des Satzbaus	A und C. [...] Ähm bei C ist äh (.) positiv auch hier die Konjunktion *also, nämlich.* Also das ist äh... gut. Hier die Voraussetzungen... Jeweils die Voraussetzungen am Anfang genannt. Das ist äh sehr sprachlogisch aufgebaut. [...] Ja, sogar hier so diese, diesen sogenannten Verb-Erstsatz. Meine Frau ist Linguistin, deshalb haben wir das mal untersucht. [...] Das ist ein ganz hohes sprachliches Niveau. (Herr Uckermark, Paarvergleich Schülerlösungstext A und C, Seg. 712)
Verwendung von angemessenen Bindewörter	B und D. Äh da find- ich B deutlich besser. Eben, wie gesagt, vollständige Sätze, äh begründende, also ver-... verbindende Konjunktionen. (Herr Uckermark, Paarvergleich Schülerlösungstext B und D, Seg. 724)
Verwendung der richtigen Artikel	A und D. [...] Also sprachlich äh D über A. Weil [...] die richtige äh äh äh grammatikalische Form benutzt wird, ne? Also hier: *das Ton...* fängt ja... der Artikel fängts an. (Herr Onne, Paarvergleich Schülerlösungstext A und D, Seg. 291)
Vorhandensein von Rechtschreibungs- und/oder Zeichensetzungsfehlern	Ähm (.) also bei A hatte ich auch angestrichen [...] Komma fehlt. (Herr Uckermark, Paarvergleich Schülerlösungstext A und B, Seg. 724)

Tabelle 6.32.: Beurteilungskriterien, die die Teilnehmer_innen bei den sprachlichen Paarvergleichen der Schülerlösungstexte verwendeten, um ihre Einschätzungen zu begründen.

Kriterium	exemplarischer Transkriptauszug
Differenziertheit/ Komplexität des Textes	A und C. Äh (..) C... die Antwort C is- sehr viel differenzierter dargestellt als die Antwort A. Darum find- ich die Antwort C sprachlich... und d- die is- auch sprachlich und vor allem fachsprachlich besser. (Herr Quezada, Paarvergleich Schülerlösungstext A und C, Seg. 349)
Entsprechung der persönlichen Erwartungen	Also die A hat [...] für mich nicht irgendwie das, was ich in der Physik gerne hätte. [...] Ist B knapp besser. (Herr Jonuzi, Paarvergleich Schülerlösungstext A und B, Seg. 280)
Qualität der (Fach-)Sprache	B und C. [...] Also ähm ja ich würde jetz- einfach sprachlich doch ma- C besser bewerten, weil ich finde, dass es einfach mehr Text is- und er sich (.) auf so -nem hohen Niveau hält. Ähm... (liest leise) (...) Ja also da find- ich C leicht auf jeden Fall besser, sprachlich als als B. Warum: Ähm (..) fällt mir schwer das zu begründen muss ich sagen. [...] Und da Antwort C brilliert einfach durch die ho-... das hohe sprachliche Niveau. So! Aus meiner Sicht. (Herr Rittershaus, Paarvergleich Schülerlösungstext B und C, Seg. 759)
Quantität an Fachwörtern	B und C. (...) Schon mhm... C ist sprachlich (..) anspruchsvoller. (..) Ähm (...) [...] Mehr Fachbegriffe benutzt. (..) *Schwingungen*, *Vakuum*, (...) *Schallwellen*, (...) hier wird nur von *besserer Funkverbindung* (..) gesprochen. Also mehr Fachbegriffe [...] als Antwort B. (Herr Mehlert, Paarvergleich Schülerlösungstext B und C, Seg. 880)
Strukturiertheit/Gliederung des Textes	A und C. Ja A hat die Schwächen, die ich gerade schon erklärt habe. Und C is- eigentlich sehr stark. Das is- sehr präzise ähm, sogar noch gegliedert. [...] Und deswegen is- C deutlich besser als A. (Herr Hastedt, Paarvergleich Schülerlösungstext A und C, Seg. 589)
Verdichtungsgrad/Präzision des Textes	Also sprachlich find- ich B besser. Ähm also es sind ja nur 2 Sätze. [...] Insofern is-... Und das alles praktisch so sprachlich relativ kurz zusammengefasst. Insofern find- ich das sprach-... Antwort B sprachlich besser als A. (Herr Mehlert, Paarvergleich Schülerlösungstext A und B, Seg. 876)
Vorhandensein von Redundanz	B und D. B ist besser (...), weil D diesen... diese S- Satzwiederholung [...] eingebaut hat, die zur Verwirrung führt. (Herr Mehlert, Paarvergleich Schülerlösungstext B und D, Seg. 882)

Tabelle 6.32.: Beurteilungskriterien, die die Teilnehmer_innen bei den sprachlichen Paarvergleichen der Schülerlösungstexte verwendeten, um ihre Einschätzungen zu begründen (Fortsetzung).

fachlich-konzeptuelle Paarvergleiche	Schülerlösungstext A und B			Schülerlösungstext A und C			Schülerlösungstext A und D			Schülerlösungstext B und C			Schülerlösungstext B und D			Schülerlösungstext C und D		
Mögliche Einschätzungen	A>B	A<B	A=B	A>C	A<C	A=C	A>D	A<D	A=D	B>C	B<C	B=C	B>D	B<D	B=D	C>D	C<D	C=D
Anzahl an Einschätzungen	**20**	---	**1**	---	**21**	---	**21**	---	---	---	**21**	---	**14**	---	**7**	**21**	---	---
	Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?		
Angemessenheit des Fachwissens in Bezug zur Aufgabenstellung	13	---	1	---	2	---	9	---	---	---	7	---	5	---	3	3	---	---
Anwendung von Fachmethoden	---	---	---	---	---	---	---	---	---	---	1	---	1	---	1	---	---	---
Fachliche Korrektheit des Inhalts	9	---	---	---	3	---	7	---	---	---	7	---	12	---	4	7	---	---
Inhaltsreichtum/ Quantität von Fachwissen	1	---	---	---	8	---	8	---	---	---	3	---	4	---	---	---	---	---
Qualität des Bezugs zur Aufgabenstellung	13	---	1	---	2	---	9	---	---	---	7	---	5	---	3	3	---	---
Vollständigkeit in Bezug zur Aufgabenstellung	2	---	---	---	3	---	2	---	---	---	3	---	---	---	1	1	---	---
Differenziertheit/ Komplexität des Textes	---	---	---	---	5	---	---	---	---	---	1	---	---	---	---	2	---	---
Entsprechung der persönlichen Erwartungen	---	---	---	---	2	---	2	---	---	---	2	---	---	---	1	1	---	---
Qualität der (Fach-)Sprache	5	---	---	---	12	---	4	---	---	---	1	---	2	---	1	1	---	---
Quantität an Fachwörtern	---	---	---	---	1	---	---	---	---	---	1	---	---	---	---	1	---	---
Strukturiertheit/ Gliederung des Textes	---	---	---	---	4	---	---	---	---	---	2	---	---	---	---	---	---	---
Verdichtungsgrad/ Präzision des Textes	---	---	---	---	1	---	---	---	---	---	---	---	---	---	---	1	---	---
Vorhandensein von Redundanz	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	1	---	---

Tabelle 6.33.: Anzahl der Teilnehmer_innen die ein bestimmtes Beurteilungskriterium bei einem bestimmten fachlich-konzeptuellen Paarvergleich verwendeten. Die grau hervorgehobenen Spalten repräsentieren die Einschätzungen bei den fachlich-konzeptuellen Paarvergleichen, die mit der Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie übereinstimmen.

sprachliche Paarvergleiche	Schülerlösungstext A und B			Schülerlösungstext A und C			Schülerlösungstext A und D			Schülerlösungstext B und C			Schülerlösungstext B und D			Schülerlösungstext C und D		
Mögliche Einschätzungen	A>B	A<B	A=B	A>C	A<C	A=C	A>D	A<D	A=D	B>C	B<C	B=C	B>D	B<D	B=D	C>D	C<D	C=D
Anzahl an Einschätzungen	---	20	1	---	21	---	13	2	6	2	11	8	21	---	---	21	---	---
	Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?			Kriterium verwendet?		
Angemessenheit der Zeitform	---	1	---	---	---	---	---	1	---	---	---	---	---	---	---	---	---	---
Attraktivität des Textes	---	1	---	---	2	---	2	---	---	2	---	---	2	---	---	2	---	---
Eindeutigkeit von Bezügen	---	1	---	---	---	---	---	---	---	---	---	---	1	---	---	---	---	---
Korrektheit/Angemessenheit des Satzbaus	---	---	---	---	2	---	---	---	---	---	2	---	2	---	---	1	---	---
Korrektheit/Angemessenheit von (Fach-)Wörtern	---	6	---	---	7	---	3	---	1	1	2	1	4	---	---	3	---	---
Korrektheit/Angemessenheit von (fachsprachlichen) Wortkombinationen	---	11	---	---	1	---	1	1	3	---	---	---	1	---	---	1	---	---
Länge des Textes	---	---	---	---	2	---	---	---	---	---	1	---	3	---	---	1	---	---
Lesbarkeit des Textes	---	1	1	---	3	---	1	---	---	---	1	3	1	---	---	---	---	---
(mutmaßliche) Gewissenhaftigkeit des_der Schülers_Schülerin während des Schreibens	---	---	---	---	---	---	4	---	1	---	---	---	1	---	---	---	---	---
Raffinesse des Satzbaus	---	---	---	---	1	---	---	---	---	---	1	---	---	---	---	---	---	---
Verwendung von angemessenen Bindewörter	---	2	---	---	3	---	---	---	---	---	---	2	1	---	---	1	---	---
Verwendung der richtigen Artikel	---	---	---	---	---	---	---	1	---	---	---	---	---	---	---	---	---	---
Vorhandensein von Rechtschreibungs- und/oder Zeichensetzungsfehlern	---	3	---	---	1	---	---	---	---	---	---	---	---	---	---	---	---	---
Differenziertheit/ Komplexität des Textes	---	---	---	---	2	---	---	---	---	---	3	---	---	---	---	3	---	---
Entsprechung der persönlichen Erwartungen	---	3	---	---	2	---	1	---	---	---	1	---	3	---	---	1	---	---
Qualität der (Fach-)Sprache	---	10	---	---	8	---	4	---	1	2	7	2	1	---	---	1	---	---
Quantität an Fachwörtern	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	1	---	---
Strukturiertheit/ Gliederung des Textes	---	---	---	---	7	---	---	---	---	---	2	1	1	---	---	4	---	---
Verdichtungsgrad/ Präzision des Textes	---	2	---	---	1	---	---	---	1	---	---	---	---	---	---	1	---	---
Vorhandensein von Redundanz	---	---	---	---	---	---	9	---	2	---	---	---	9	---	---	2	---	---

Tabelle 6.34.: Anzahl der Teilnehmer_innen die ein bestimmtes Beurteilungskriterium bei einem bestimmten sprachlichen Paarvergleich verwendeten. Die grau hervorgehobenen Spalten repräsentieren die Einschätzungen bei den sprachlichen Paarvergleichen, die mit der Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie übereinstimmen.

6.4.2. Quantitative Analyse der Einschätzungen der Teilnehmer_innen im Rahmen der Paarvergleiche[151]

6.4.2.1. Methodische Vorbemerkungen

Die Einschätzungen, die die Teilnehmer_innen im Rahmen der fachlich-konzeptuellen und sprachlichen Paarvergleiche vorgenommen haben, wurden nicht nur einer qualitativen, sondern auch einer quantitativen Analyse unterzogen. Die im folgenden dargestellte Analyse entspricht dabei im in Abschnitt 5.4.2 beschriebenen Mixed-Methods-Triangulationdesign sowohl der quantitativen Analyse der Leitungsmitschriften, als auch der quantitativen Analyse der inhaltsanalytischen Codierungen der retrospektiven Befragungen. Grund hierfür ist, dass die Codierungen, die bei der Inhaltsanalyse der Verbaldaten aus den retrospektiven Befragungen vorgenommenen wurden, mit den Leitungsmitschriften übereinstimmen (vgl. Unterabschnitt 6.4.1.2).

Um das methodische Vorgehen, das für die quantitative Analyse der Daten aus den fachlich-konzeptuellen und sprachlichen Paarvergleiche herangezogen wurde, nachvollziehen zu können, gilt es sich ins Gedächtnis zu rufen, dass die vier Schülerlösungstexte A, B, C und D in einem mehrschrittigen Codierfahren als Kontrastfälle gewählt wurden (vgl. Unterkapitel 5.3). Für die Rangfolge der vier Schülerlösungstexte bezüglich...

... ihrer fachlich-konzeptuellen Qualität gilt: B = D < A = C.

... der Qualität ihrer sprachlichen Realisierung gilt: A = D < B = C.

In der linken Hälfte von Abbildung 6.21 sind die vier Schülerlösungstexte entsprechend dieser Rangfolgen in einem zweidimensionalen Koordinatensystem mit den Achsen „fachlich-konzeptuelle Qualität“ und „Qualität der sprachlichen Realisierung“ verortet. Die vier Punkte, die in diesem Koordinatensystem die Schülerlösungstexte A, B, C und D repräsentieren, bilden ein Rechteck, weswegen sich zwischen ihnen kein korrelativer Zusammenhang ergibt (symbolisiert durch die gestrichelte Linie in Abbildung 6.21 links).

Wie in Anhang F vollständig aufgeführt, lassen sich die Einschätzungen, die ein_e Teilnehmer_in im Rahmen der Paarvergleiche vorgenommen hat, ebenfalls in jeweils zwei Rangfolgen der vier Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung zusammenfassen. So ist...

... D < B < A < C die Rangfolge der 4 Schülerlösungstexte, die aus den fachlich-konzeptuellen Paarvergleichen von Herrn Abney hervorgeht,

... D < A < B < C die Rangfolge der 4 Schülerlösungstexte, die aus den sprachlichen Paarvergleichen von Herrn Abney hervorgeht,

... D = B < A < C die Rangfolge der 4 Schülerlösungstexte, die aus den fachlich-konzeptuellen Paarvergleichen von Herrn Balke hervorgeht,

[151] Teile dieses Abschnitts stellen eine überarbeitete und erweiterte Fassung von Feser & Höttecke (2018) dar.

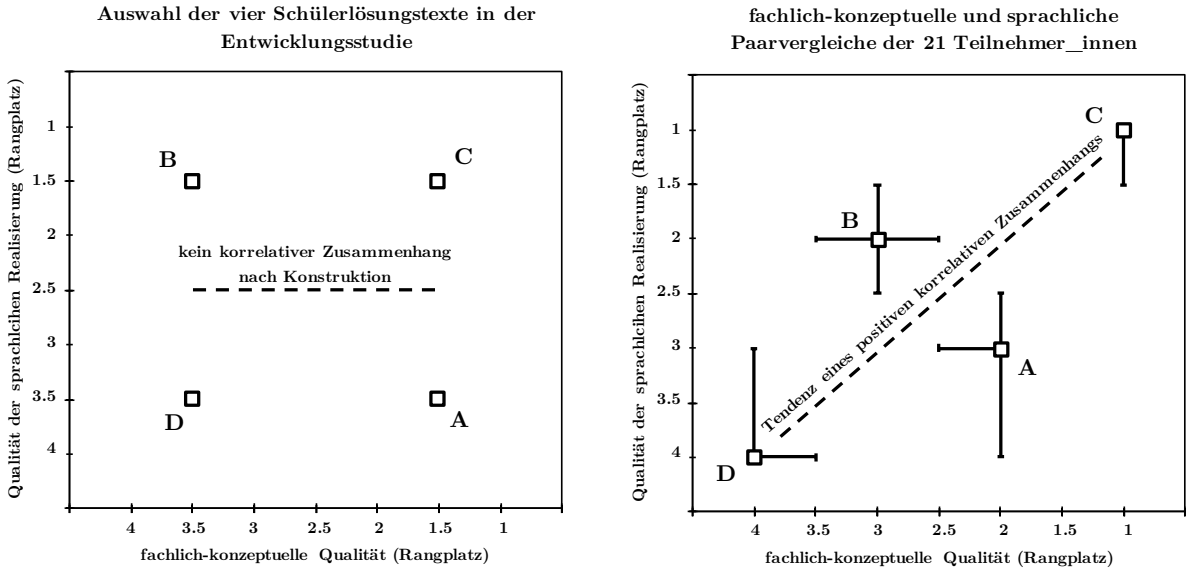


Abbildung 6.21.: Graphische Darstellung der Rangplätze von Schülerlösungstext A bis D bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung. Links sind die vier Schülerlösungstexte entsprechend der kontrastierenden Auswahl in der Entwicklungsstudie angeordnet. Rechts sind die vier Schülerlösungstexte entsprechend der medianen Rangplätze angeordnet, die aus den fachlich-konzeptuellen und sprachlichen Paarvergleichen der Teilnehmer_innen hervorgehen. Die Balken im rechten Diagramm repräsentieren die Rangplatz-Spannweiten der Schülerlösungstexte.

... D < A < C < B die Rangfolge der 4 Schülerlösungstexte, die aus den sprachlichen Paarvergleichen von Herrn Balke hervorgeht,

... usw.

In der rechten Hälfte von Abbildung 6.21 sind die Schülerlösungstexte A, B, C und D entsprechend der medianen Rangplätze angeordnet, die aus den fachlich-konzeptuellen und sprachlichen Paarvergleichen aller 21 Teilnehmer_innen hervorgehen (die Balken repräsentieren die Rangplatz-Spannweiten der Schülerlösungstexte). Wie sich in dieser Teilabbildung andeutet, weichen die Rangfolgen, die aus den Paarvergleichen der Teilnehmer_innen hervorgehen, zum Teil erheblich von der Auswahl der vier Schülerlösungstexte im Rahmen der Entwicklungsstudie ab. Augenscheinlich zeigt sich hier die Tendenz eines positiven korrelativen Zusammenhangs (symbolisiert durch die gestrichelte Linie in Abbildung 6.21 rechts).

Im Rahmen der quantitativen Analyse der Daten aus den fachlich-konzeptuellen und sprachlichen Paarvergleichen wurde überprüft, ob sich diese augenscheinliche Tendenz eines positiven korrelativen Zusammenhangs bestätigen lässt. Hierzu wurde der von Torgerson (1956) und Ludwig (1962) entwickelte Rangkorrelationskoeffizient τ^* bestimmt, da

dieser ermöglicht die folgenden, nicht zu vernachlässigenden Merkmale des zu analysierenden Datensatzes mit zu berücksichtigen:

1. Die insgesamt 252 Einschätzungen, die bei den fachlich-konzeptuellen bzw. bei den sprachlichen Paarvergleichen erhoben wurden, können nicht in eine gemeinsame Rangordnung zusammengefasst werden (vgl. Bortz et al., 2008, S. 439). Grund hierfür ist, dass diese von 21 verschiedenen Lehrer_innen stammen und deshalb die 252 Einschätzungen in 21 Teildatensätze aufzuteilen sind.
2. Insgesamt wurden in der Hauptstudie lediglich 21 Physiklehrer_innen befragt. Für die quantitative Analyse bedarf es daher eines statistischen Verfahrens, dass auch für kleine Probandenzahlen geeignet ist.
3. Für jede Lehrkraft ergibt sich aus den 6 fachlich-konzeptuellen und 6 sprachlichen Paarvergleichen für jeden der vier Schülerlösungstexte ein Rangplatz-Wertepaar. Für die Analyse liegen also pro Lehrkraft lediglich vier Datenpunkte vor.
4. Bei jedem Paarvergleich sollte eingeschätzt werden, ob einer der beiden Schülerlösungstexte besser ist, oder ob beide Texte gleich gut sind. Die Einschätzungen, die bei den fachlich-konzeptuellen und sprachlichen Paarvergleichen erhoben wurden, weisen also lediglich ein ordinales Skalenniveau auf.

τ^* berechnet sich als (mit den Mächtigkeiten der Teildatensätze gewogener) Mittelwert der 21 Rangkorrelationskoeffizienten[152] zwischen den Rangfolgen, die aus den fachlich-konzeptuellen und sprachlichen Paarvergleichen je Teilnehmer_in hervorgehen. Es gilt (vgl. Ludwig, 1962, S. 45):

$$\tau^* = \begin{cases} \frac{1}{k} \cdot \sum\limits_{i=1}^{k} \tau_i & \text{, wenn} \quad n_1 = n_2 = ... = n_k \\ \sum\limits_{i=1}^{k} \frac{n_i \cdot (n_i - 1)}{\sum\limits_{j=1}^{k} n_j \cdot (n_j - 1)} \cdot \tau_i & \text{, im allgemeinen Fall} \end{cases}$$

mit

$k =$ Anzahl der Teildatensätze (hier: Anzahl der Teilnehmer_innen)

$n_i =$ Anzahl der Wertepaare im i-ten Teildatensatz (hier: Anzahl der Schülerlösungstexte)

$\tau_i =$ Kendalls-τ für die Wertepaare im i-ten Teildatensatz[152] (hier: Kendalls-τ zwischen den Rangfolgen, die aus den fachlich-konzeptuellen und sprachlichen Paarvergleichen je Teilnehmer_in hervorgehen).

[152] Hierbei handelt es sich um den Rangkorrelationskoeffizienten nach Kendall für Wertepaare, die keine Rangbindung aufweisen (vgl. Bortz et al., 2008, S. 422 u. f.). Soll τ^* für Wertepaare, die Rangbindungen aufweisen, bestimmt werden, empfiehlt Torgerson (1956, S. 151) allen ranggleichen Wertepaaren zufällig einen der auf sie fallenden Ränge zuzuordnen. Dieser Empfehlung wurde im Rahmen der vorliegenden Arbeit gefolgt.

Ferner gilt, dass für[153] $\sum_{i=1}^{k} n_i > 20$ die Größe

$$z = \frac{\left|\tau^* \cdot \frac{1}{2} \cdot \sum_{i=1}^{k} n_i \cdot (n_i - 1)\right| - 1}{\sqrt{\frac{1}{18} \sum_{i=1}^{k} n_i \cdot (n_i - 1) \cdot (2n_i + 5)}}$$

annähernd standardnormalverteilt ist (vgl. Bortz et al., 2008, S. 439; Ludwig, 1962, S. 46). Bei einem zweiseitigen Test und für das Signifikanzniveau α wird also der Nichtablehnungsbereich der Nullhypothese $\tau^* = 0$ durch das $\frac{\alpha}{2}$- bzw. das $(1 - \frac{\alpha}{2})$-Quantil der Standardnormalverteilung bestimmt.

6.4.2.2. Ergebnis, Interpretation und Limitationen der quantitativen Analyse

Für die Einschätzungen der 21 Teilnehmer_innen im Rahmen der fachlich-konzeptuellen und sprachlichen Paarvergleiche ergibt sich ein Wert von $\tau^* = .44$, der zum Signifikanzniveau $\alpha = .01$ verschieden von 0 ist ($z = 4.08$).

Es ist plausibel, die Stärke dieses korrelativen Zusammenhangs als Indikator für das Ausmaß zu interpretieren[154], in dem die Teilnehmer_innen fachlich-konzeptuelle und sprachliche Teilleistungsurteile miteinander konfundieren. Grund hierfür ist, dass die Schülerlösungstexte A bis D in der Entwicklungsstudie als Kontrastfälle so ausgewählt wurden, dass zwischen ihrer fachlich-konzeptuellen Qualität und der Qualität ihrer sprachlichen Realisierung kein korrelativer Zusammenhang besteht (vgl. Abschnitt 5.2.2), sowie ferner die in Abschnitt 6.4.1 bereits dargestellten qualitativen Befunde, die für eine Konfundierung von fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung durch die Teilnehmer_innen sprechen (z. B. die Verwendung zum Teil gleicher Beurteilungskriterien bei den fachlich-konzeptuellen und sprachlichen Paarvergleichen). Der Wert von $\tau^* = .44$ weist damit also auf eine moderat[155] ausgeprägte Konfundierung der fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung der Teilenehmer_innen hin.

Dieser Befund gilt allerdings nur für die fachlich-konzeptuellen und sprachlichen Paarvergleiche im Rahmen der retrospektiven Befragung, da keine Daten aus der laut-denkenden Korrektur der vier Schülerlösungstexte für die hier vorgenommene Analyse herangezogen wurden. An dieser Stelle der vorliegenden Arbeit kann daher nicht aufgeklärt werden, ob und falls ja wie ausgeprägt die Teilnehmer_innen bei der laut-denkenden Korrektur der

[153] Diese Bedingung ist für die quantitative Analyse der fachlich-konzeptuellen und sprachlichen Paarvergleiche erfüllt, denn es gilt hier: $\sum_{i=1}^{k} n_i = 21 \cdot 4 = 84 > 20$.

[154] Für eine methodentheoretische Diskussion notwendiger Bedingungen einer kausalen Interpretationen korrelativer Zusammenhänge in der erziehungswissenschaftlichen Forschung siehe Renkl (1993).

[155] Da τ^* als (gewogenes) Mittel der Rangkorrelationskoeffizienten nach Kendall der zu analysierenden Teildatensätze berechnet wird (vgl. Unterabschnitt 6.4.2.1), können die gängigen Faustregeln für Kendalls-τ auf τ^* übertragen werden. Für Kendalls-τ gelten Werte ab .10 als „schwacher“, ab .30 als „moderater“ und Werte ab .5 als „starker“ korrelativer Zusammenhang (vgl. J. Cohen, 1988, S. 79 u. f.; Lomax & Hahs-Vaughn, 2012, S. 273 u. f.).

vier Schülerlösungstexte fachlich-konzeptuelle und sprachliche Schülerleistungen miteinander konfundierten.

6.5. Integration der Befunde

In Unterkapitel 6.3 und 6.4 wurde die Analyse der Laut-Denk-Daten, sowie der Daten, die von den Teilnehmern_Teilnehmerinnen in den retrospektiven Befragungen gewonnen wurden, beschrieben. Wie aus einem Vergleich der Ergebnisdarstellungen dieser beiden Unterkapitel hervorgeht, konnten aus den in der Laborsituation erhobenen Daten konvergierende, sowie komplementäre Teilbefunde zu den Forschungsfragen (F1) und (F2) gewonnen werden. Ferner zeigten sich zu keiner der beiden Forschungsfragen sich widersprechend (divergierende) Teilbefunde.

Ziel dieses Unterkapitels ist die Teilbefunde zu Forschungsfrage (F1) und (F2) in ein kaleidoskopartiges Gesamtbild zusammenzuführen. Die Darstellung beschränkt sich dabei auf eine Kurzdarstellung von aus den vorangegangenen Unterkapiteln aggregierten Befunden, die im Sinne der beiden Forschungsfragen von besonderer Relevanz sind.

6.5.1. Integration der Befunde zu Forschungsfrage (F1)

Bezüglich der Frage auf welche Ressourcen die teilnehmenden Physiklehrkräfte bei der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile zurückgreifen (Forschungsfrage (F1)) zeigten sich in den erhobenen Daten sowohl konvergierende, als auch komplementäre Teilbefunde. Diese lassen sich auf 6 Ressourcen zur Leistungsfeststellung und -beurteilung verdichten, die in den folgenden Unterabschnitten genauer beschrieben werden:

1. Im Handeln der Teilnehmer_innen zeigte sich eine Nutzung von Wissen und Können zu schulischer Leistungsfeststellung und -beurteilung, das sich speziell auf fachlich-konzeptuelle und/oder sprachliche Schülerleistungen bezieht.
2. In Teilen erfolgte bei der Leistungsurteilsgenese auch eine holistische Feststellung und Beurteilung der Schülerleistungen.
3. Das Handeln der Teilnehmer_innen bei der Leistungsurteilsgenese erfolgte auch auf Basis von generalisiertem Wissen und Können, sowie berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung.
4. Das Handeln der Teilnehmer_innen bei der Leistungsurteilsgenese ist deutlich an der kriterialen Bezugsnorm orientiert.
5. Bei der Leistungsurteilsgenese wurden von den Teilnehmer_innen vor allem die fachlich-konzeptuellen Merkmale der Schülerlösungstexte beachtet. Sprachliche Merkmale spielten hingegen eine deutlich geringere Rolle. Ferner zeigte sich, dass die

Teilnehmer_innen in unterschiedlichem Ausmaß sprachliche Merkmale bei der Feststellung und Beurteilung der Schülerleistungen berücksichtigten.

6. Die Feststellung und Beurteilung sprachlicher Schülerleistungen erfolgte durch die Teilnehmer_innen tendenziell defizitorientiert, die fachlich-konzeptueller Schülerleistungen hingegen zum Teil defizitorientiert, in Teilen aber auch fähigkeitsorientiert.

Zusammengefasst nutzten die Teilnehmer_innen facettenreiches und überwiegend angemessenes Wissen und Können zu (fachspezifischer) Leistungsfeststellung und -beurteilung. Ferner handelten sie auch auf Basis ihrer berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung, sowie ihrer Bezugsnormorientierungen. Das Handeln der Teilnehmer_innen bei der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile in der Laborsituation ist also vereinbar mit dem in Abschnitt 2.2.4 dargestellten Rahmenkonzept einer Assessment Literacy. Sie haben auf Ressourcen zugriffen, die sich in diesem Konzept als zwei der drei maßgebende Faktoren identifizieren lassen (Wissen und Können, sowie berufsbezogene Überzeugungen zu (fachspezifischer) Leistungsfeststellung und -beurteilung), zwischen denen es für eine Lehrkraft im Rahmen einer schulische Leistungsfeststellung und -beurteilung betreffende Handlungsepisode gilt, einen Kompromiss zu realisieren, da diese Faktoren bis zu einem bestimmten Grad vereinbar sind und/oder einander konfligieren. Ergo: Die Teilnehmer_innen lassen sich bis zu einem bestimmten Grad als bezüglich fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung grundgebildete Physiklehrkräfte charakterisieren.

6.5.1.1. Nutzung von Wissen und Können zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen

Die Teilnehmer_innen nutzten in der Laborsituation Wissen und Können zu schulischer Leistungsfeststellung und -beurteilung, das sich speziell auf fachlich-konzeptuelle und/oder sprachliche Schülerleistungen bezieht. Zusammen mit der moderaten bzw. sehr hohen intersubjektiven Übereinstimmung der Punkteverteilungen und den Einschätzungen während der Paarvergleiche (vgl. Abschnitt 6.3.1 und 6.4.1) spricht dieser Befund für die Validität der Leistungsurteilsgense der Teilnehmer_innen: Die Teilnehmer_innen nutzten in der Laborsituation (unter anderem) derartiges Wissen und Können, um zu einer, bis zu einem bestimmten Grad übereinstimmenden Feststellungen und Beurteilungen schriftlicher Schülerleistungen zu gelangen, die sich kriterial vor allem bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung voneinander unterscheiden lassen.

Die Nutzung von Wissen und Können zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen zeigte sich im Handeln der Teilnehmer_innen in der Laborsituation sowohl direkt, also auch indirekt. Direkt zeigte sich diese Nutzung darin, dass die Teilnehmer_innen...

... bei ihrer laut denkenden Korrekturarbeit fachlich-konzeptuelle bzw. sprachliche Merkmale der Schülerlösungstexte explizit fokussierten (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.5).

... bei ihrer laut denkenden Korrekturarbeit Beurteilungskriterien mitvokalisierten, die explizit fachlich-konzeptuelle bzw. sprachliche Merkmale von Schülerlösungstexten betreffen (vgl. Unterabschnitt 6.3.2.3).

... bei der Erstellung ihrer Erwartungshorizonte in unterschiedlichem Ausmaß neben fachlich-konzeptuellen auch sprachliche Merkmale eines Schülerlösungstextes berücksichtigten (vgl. Unterabschnitt 6.3.2.3; Hackemann, 2017).

... im Rahmen der retrospektiven Befragung eine Vielzahl von Beurteilungskriterien heranzogen, um einzuschätzen und zu begründen, inwieweit sich zwei Schülerlösungstexte hinsichtlich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung voneinander unterscheiden (vgl. Abschnitt 6.4.1).

Indirekt zeigte sich diese Nutzung darin, dass sich Teilnehmer_innen zu den vier Schülerlösungstexten entsprechend ihrer fachlich-konzeptuellen oder sprachlichen Qualitätsunterschiede in unterschiedlichem Umfang positiv wertend/akzeptierend bzw. negativ wertend/ablehnend äußerten (vgl. Unterabschnitt 6.3.2.5) und sie die vier Schülerlösungstexte entsprechend ihrer kontrastierenden Auswahl auch unterschiedlich bepunkteten (vgl. Abschnitt 6.3.1).

6.5.1.2. In Teilen holistische Feststellung und Beurteilung der Schülerleistungen

Im Rahmen der Laborsituation erfolgte auch eine holistische Feststellung und Beurteilung der Schülerleistungen. Sowohl während ihrer laut-denkenden Korrekturarbeit, als auch der retrospektiven Befragung äußerten die Teilnehmer_innen unter anderem Beurteilungskriterien, die dazu dienen, einen (eher) globalen/generellen Eindruck über eine Schülerleistung zum Ausdruck zu bringen (vgl. Unterabschnitt 6.3.2.3 und Abschnitt 6.4.1). Ferner zeigte sich in den Laut-Denk-Protokollen, dass die Teilnehmer_innen die Leistungen in den Schülerlösungstexten auch auf Grundlage eines (ersten) eher holistischen Eindrucks und/oder heuristisch feststellten und -beurteilten (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.5).

Hervorzuheben ist allerdings, dass die Teilnehmer_innen bei ihrer Leistungsurteilsgenese in der Laborsituation vor allem analytisch vorgegangen sind. Sowohl beim lauten Denken, als auch in der retrospektiven Befragung äußerten die Teilnehmer_innen vor allem Beurteilungskriterien, die eine differenzierte Erfassung bestimmter Teilleistungen ermöglichen (vgl. Unterabschnitt 6.3.2.3 und Abschnitt 6.4.1). Ferner fokussierten sie bei ihrer laut-denkenden Korrekturarbeit vornehmlich fachlich-konzeptuelle, sprachliche oder sonstige Merkmale der vier Schülerlösungstexte und verorteten diese in Bezug zu einem sachlichen Kriterium (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.5).

6.5.1.3. Generalisiertes Wissen und berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung

Die Teilnehmer_innen handelten in der Laborsituation unter anderem auch auf Basis von generalisiertem (ihrem Bewusstsein möglicherweise aber nicht zugänglichem) Wissen und Können zu schulischer Leistungsfeststellung und -beurteilung, sowie diesbezüglichen berufsbezogenen Überzeugungen. Während des lauten Denkens nutzten sie bis zu einem bestimmten Grad verallgemeinerte Erfahrungen und allgemeine Handlungsstrategien zum Erstellen eines Erwartungshorizonts und zum Festellen und Beurteilen von Schülerleistungen (vgl. Unterabschnitt 6.3.2.3). Ferner zeigten sich in einer rekonstruktiven Analyse Hinweise, dass das Handeln der Teilnehmer_innen beim lauten Denken, sowie in der retrospektiven Befragung (bewusst oder unbewusst) auch davon beeinflusst war, welche Rolle sich ein_e Teilnehmer_in im Kontext von schulischer Leistungsfeststellung und -beurteilung selbst zuschreibt und inwiefern sich ein_e Teilnehmer_in bezogen auf schulische Leistungsbeurteilung für souverän hält (vgl. Unterabschnitt 6.3.2.3; Kroll, 2017).

6.5.1.4. Orientierung der Leistungsfeststellung und -beurteilung an der kriterialen Bezugsnorm

Bei der laut-denkenden Korrektur der vier Schülerlösungstexte zeigte sich vor allem eine Orientierung der Leistungsfeststellung und -beurteilung an der kriterialen Bezugsnorm (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.5). Die Teilnehmer_innen zogen nahezu ausschließlich sachliche Kriterien als Bezug der Verortung heran, wenn sie fachlich-konzeptuelle, sprachliche oder sonstige Merkmale der Schülerlösungstexte bei ihrer Leistungsurteilsgenese in den Fokus rückten (vgl. Unterabschnitt 6.3.2.5). Ferner zeigte sich bei den Teilnehmer_innen die Tendenz sich zu einer Schülerleistung normativ zu äußern, wenn sie diese in Bezug zu einem sachlichen Kriterium verorteten (vgl. Unterabschnitt 6.3.2.5). Hinweise, die für eine Orientierung an der sozialen und/oder der individuellen Bezugsnorm sprechen (z. B. das Heranziehen anderer Schülerlösungstexte oder mutmaßliche Personenmerkmale der Schüler_innen als Bezug der Verortung), zeigten sich hingegen lediglich vereinzelt (vgl. Unterabschnitt 6.3.2.3).

Einerseits ist bei diesem Befund nicht auszuschließen, dass es sich hierbei in Teilen um ein Artefakt der Laborsituation handelt (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.5). Andererseits ist dieser Befund mit den Auskünften der Teilnehmer_innen aus dem Lehrkräftefragebogen bis zu einem bestimmten Grad vereinbar. Aus diesen ging hervor, dass die Teilnehmer_innen ihrer Selbsteinschätzung nach in schwacher Tendenz eine kriteriale Bezugsnormorientierung aufweisen (vgl. Abschnitt 6.1.2).

6.5.1.5. Beachtung fachlich-konzeptueller und sprachlicher Merkmale bei der Leistungsurteilsgenese

Die Teilnehmer_innen beachteten bei der laut-denkenden Korrektur der vier Schülerlösungstexte überwiegend fachlich-konzeptuelle Merkmale und äußerten dementsprechende Beurteilungskriterien (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.5). Sprachliche Merkmale der vier Schülerlösungstexte spielten im Vergleich hierzu eine mit Abstand geringere Rolle (vgl. Unterabschnitt 6.3.2.3 und 6.3.2.5). Hinzu kommt ein unterschiedlicher Umfang, in dem sprachliche Merkmale von den Teilnehmer_innen in der Laborsituation berücksichtigt wurden. So zeigte sich, dass die Teilnehmer_innen bei der Erstellung ihres Erwartungshorizonts sprachliche Merkmale der Schülerlösungstexte in unterschiedlichem Ausmaß berücksichtigten. Das Spektrum reicht hier von Teilnehmer_innen, die bei der Erwartungshorizonterstellung keine Beurteilungskriterien explizierten, die sprachliche Merkmale der Schülerlösungstexte betreffen, bis hin zu Teilnehmer_innen, die derartige Beurteilungskriterien explizit in ihrem Erwartungshorizont fixierten (vgl. Unterabschnitt 6.3.2.3; Hackemann, 2017, S. 58 u. f.). Des Weiteren hat die Mehrheit der Teilnehmer_innen eine Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes vor allem dann vorgenommen, wenn dieser zu einem bestimmten Grad bezüglich sprachlicher Merkmale defizitär ist und/oder dieser die eigenen fachlich-konzeptuellen Erwartungen nicht verfehlt (vgl. Anzahl der Teilnehmer_innen, die in Unterabschnitt 6.3.2.4 Muster (1a) und (1c) zugeordnet wurden). Lediglich von einer Minderheit der Teilnehmer_innen wurde bei allen vier Schülerlösungstexten auch eine Feststellung und Beurteilung der sprachlichen Realisierung vorgenommen (vgl. Anzahl der Teilnehmer_innen, die in Unterabschnitt 6.3.2.4 Muster (1d) zugeordnet wurden).

6.5.1.6. Defizit- bzw. fähigkeitsorientierte Feststellung und Beurteilung sprachlicher und fachlich-konzeptueller Schülerleistungen

Betrachtet man die Häufigkeit, in der sich die Teilnehmer_innen beim lauten Denken positiv wertend/akzeptierend, negativ wertend/ablehend oder neutral/gemischt/sonstig zu den vier Schülerlösungstexten äußerten, deutet sich eine Tendenz zur defizitorientierten Feststellung und Beurteilung der Schülerleistungen an (vgl. Unterabschnitt 6.3.2.3). Bezogen auf die Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen ist diese Tendenz allerdings differenziert zu betrachten:

Die Feststellung und Beurteilung fachlich-konzeptueller Schülerleistungen erfolgte zum Teil defizitorientiert, in Teilen aber auch in einer fähigkeitsorientierten Art und Weise. Eine Tendenz zur Fähigkeitsorientierung zeigte sich darin, dass sich die Teilnehmer_innen, wenn sie sich beim lauten Denken zum fachlich-konzeptuellen Eindruck der vier Schülerlösungstexte äußerten, sie dies signifikant am häufigsten positiv wertend/akzeptierend taten (im Median 51.4 % dementsprechender Äußerungen; vgl. Unterabschnitt 6.3.2.5). Eine Defizitorientierung der Feststellung und Beurteilung fachlich-konzeptueller Schülerleistungen zeigte sich hingegen darin, dass sich die Teilnehmer_innen bei Schülerlösungs-

texten mit geringer fachlich-konzeptueller Qualität, im Median signifikant häufiger zum fachlich-konzeptuellen Eindruck äußerten, als sie dies bei Schülerlösungstexten mit hoher fachlich-konzeptueller Qualität taten (vgl. Unterabschnitt 6.3.2.5). Zudem zeigten sich bei einer Mehrheit der Teilnehmer_innen Hinweise darauf, dass sie sich bei der Korrektur eines Schülerlösungstextes auf die Feststellung und Beurteilung fachlich-konzeptueller Merkmale fokussierten, wenn der Schülerlösungstext die eigenen diesbezüglichen Erwartungen verfehlt (vgl. Anzahl der Teilnehmer_innen, die in Unterabschnitt 6.3.2.4 Muster (1b) und (1c) zugeordnet wurden).

Demgegenüber wurden sprachliche Schülerleistungen in einer tendenziell defizitorientierten Art und Weise festgestellt und beurteilt. Dies zeigte sich an den folgenden drei Teilbefunden: Ersten haben sich die Teilnehmer_innen, wenn sie sich beim lauten Denken zur sprachlichen Realisierung der vier Schülerlösungstexte äußerten, signifikant am häufigsten negativ wertend/ablehnend geäußert (im Median 69.6 % dementsprechender Äußerungen; vgl. Unterabschnitt 6.3.2.5). Zweitens hat eine Mehrheit der Teilnehmer_innen eine Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes nur dann vorgenommen, wenn dieser zu einem bestimmten Grad bezüglich sprachlicher Merkmale defizitär ist oder hat sich bei drei der vier Schülerlösungstexte überwiegend negativ wertend/ablehnend bezogen auf deren sprachliche Realisierung geäußert (vgl. Anzahl der Teilnehmer_innen, die in Unterabschnitt 6.3.2.4 Muster (1a), (1c) und (1d) zugeordnet wurden). Drittens berücksichtigte mehr als ein Drittel der Teilnehmer_innen sprachliche Merkmale der Schülerlösungstexten bei der Erwartungshorizonterstellung in einer auffällig defizitorientierten Art und Weise (z. B. durch expliziten Punkteabzug; vgl. Unterabschnitt 6.3.2.3; Hackemann, 2017, S. 59). Hinweise, die gegen eine Tendenz zur defizitorientierten Feststellung und Beurteilung sprachlicher Schülerleistungen sprechen, zeigten sich hingegen bei weniger als einem Drittel der Teilnehmer_innen bei ihrer laut-denkenden Erwartungshorizonterstellung (z. B. in einer expliziten Punktevergabe für sprachliche Merkmale; vgl. Unterabschnitt 6.3.2.3; Hackemann, 2017, S. 58) und/oder in qualitativen Prozessaspekten ihrer Leistungsurteilsgenese (z. B. in mehrheitlich positiv wertenden/akzeptierenden Äußerungen zur sprachlichen Realisierung von Schülerlösungstexten mit einer diesbezüglich hohen Qualität; vgl. Unterabschnitt 6.3.2.4).

6.5.2. Integration der Befunde zu Forschungsfrage (F2)

Sowohl die Analyse der Laut-Denk-Daten, als auch die der retrospektiven Befragungen lieferte Teilbefunde, die partikuläre Antworten auf die Frage darstellen, inwieweit Physiklehrkräfte fachlich-konzeptuelle und sprachliche Leistungsurteile miteinander konfundieren (Forschungsfrage (F2)). In den folgenden beiden Unterabschnitten erfolgt zunächst eine Zusammenführung der qualitativen und quantitativen Befunde zu Forschungsfrage (F2), bevor anschließend deren Quintessenz erörtert wird.

6.5.2.1. Qualitative Teilbefunde zur Konfundierung fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung

Im Rahmen der Extraktion und Interpretation qualitativer Prozessaspekte aus den Laut-Denk-Daten (vgl. Unterabschnitt 6.3.2.4) konnten insgesamt zwei Bewertungslogiken identifiziert werden, bei denen eine Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen stattfand. In der Abfolge mitvokalisierter Gedankenschritte von insgesamt...

... 7 Teilnehmer_innen konnte die Bewertungslogik identifiziert werden, fachlich-konzeptuell richtige oder anschlussfähige Denkfiguren in einem Schülerlösungstext zu relativieren, wenn in diesem die Sprachgebrauchserwartungen des_der Teilnehmers_-Teilnehmerin nicht hinreichend erfüllt wurden (vgl. Muster (2a) in Unterabschnitt 6.3.2.4).

... 4 Teilnehmer_innen konnte die Bewertungslogik identifiziert werden, sprachliche Mängel eines Schülerlösungstextes zu relativieren, wenn in diesem die fachlich-konzeptuellen Erwartungen des_der Teilnehmers_Teilnehmerin (in Teilen) erfüllt wurden (vgl. Muster (2c) in Unterabschnitt 6.3.2.4).

Daneben zeigte sich in den Laut-Denk-Daten von 7 Teilnehmer_innen die Bewertungslogik, fachlich-konzeptuelle Merkmale von Merkmalen der sprachlichen Realisierung zu trennen (vgl. Muster (2b) in Unterabschnitt 6.3.2.4). Insgesamt konnte also bei einigen Teilnehmer_innen zwar eine Herangehensweise bei der Leistungsurteilsgenese identifiziert werden, durch die eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile vermieden wird, gleichzeitig zeigte sich bei einer Vielzahl der Teilnehmer_innen, dass diese durch ein relativierendes Herangehen fachlich-konzeptuelle und sprachliche Schülerleistungen miteinander konfundieren.

Sehr deutliche qualitative Teilbefunde, aus denen eine Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen hervorgeht, offenbarten sich zudem in den Verbaldaten der retrospektiven Befragungen (vgl. Unterabschnitt 6.4.1.3). Es zeigte sich, dass die Teilnehmer_innen insgesamt 7 Beurteilungskriterien sowohl bei den fachlich-konzeptuellen, als auch bei den sprachlichen Paarvergleichen der Schülerlösungstexte eingesetzt haben. Von diesen 7 Beurteilungskriterien sind die Kriterien „Quantität an Fachwörtern“ und „Qualität der (Fach-)Sprache“ hervorzuheben. Grund hierfür ist, dass diese beiden Beurteilungskriterien eindeutig die sprachliche Realisierung eines Schülerlösungstextes betreffen, diese allerdings von mehreren (aber nicht allen) Teilnehmer_innen auch dazu verwendet wurden um zu begründen, inwieweit sich zwei Schülerlösungstexte hinsichtlich ihrer fachlich-konzeptuellen Qualität voneinander unterscheiden. Dies war insbesondere bei den fachlich-konzeptuellen Paarvergleichen der Schülerlösungstexte A und C der Fall, bei denen eine Mehrheit der Teilnehmer_innen unter anderem die „Qualität der (Fach-)Sprache“ als Beurteilungskriterium heranzog.

fachlich-konzeptuelle Qualität	Qualität der sprachlichen Realisierung	Schülerlösungstext	1. Äußerungen zur sprachlichen Realisierung (lautes Denken)				2. positiv wertende/ akzeptierende Äußerungen (lautes Denken)		3. negativ wertende/ ablehnende Äußerungen (lautes Denken)		4. Punkteverteilung (lautes Denken)		5. Einschätzungen in den Paarvergleichen (retrospektive Befragung)
			%	ES			%	ES	%	ES	Punkte	ES	τ^*
hoch	hoch	C	18.9	.62***	.55***	.15	66.7	.58***	14.3	.42**	5.0	.61***	.44**
	gering	A	19.1				27.8		29.5		2.5		
gering	hoch	B	0.0		.39**		10.5	.34*	45.7	.37*	1.0	.43**	
	gering	D	11.1				0.0		72.7		0.0		

Tabelle 6.35.: Zusammenfassung der Teilbefunde zu Forschungsfrage (F2), die aus der quantitativen Analyse der Laut-Denk-Daten, sowie jener der retrospektiven Befragungen hervorgingen (bei den angegebenen Prozent- und Punktwerten handelt es sich um die entsprechenden Stichprobenmediane; ***:$p \leq .001$; **:$p \leq .01$; *:$p \leq .05$).

6.5.2.2. Quantitative Teilbefunde zur Konfundierung fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung

Wie in Tabelle 6.35 dargestellt, konnten in den quantitativen Analysen der in der Laborsituation erhobenen Daten insgesamt 5 Teilbefunde gewonnen werden, die für eine Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen sprechen:

1. Bei der laut-denkenden Korrektur der Schülerlösungstexte, die sich durch eine hohe fachlich-konzeptuelle Qualität auszeichnen, äußerten sich die Teilnehmer_innen im Median (signifikant) häufiger zur sprachliche Realisierung, als bei den Schülerlösungstexten, die eine geringe fachlich-konzeptuelle Qualität aufweisen (vgl. Unterabschnitt 6.3.2.5). Hierin zeigt sich ein Hinweis, dass die Teilnehmer_innen eine Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes vor allem dann vornehmen, wenn dieser bis zu einem bestimmten Grad auch ihren fachlich-konzeptuellen Erwartungen entspricht.

2. Der Medianunterschied zwischen den prozentuellen Häufigkeiten positiv wertender/akzeptierender Äußerungen ist bei den Schülerlösungstexten mit geringer fachlich-konzeptueller Qualität deutlich geringer (moderater Effekt; $ES = .34$), als zwischen Schülerlösungstexten mit hoher fachlich-konzeptueller Qualität (starker Effekt; $ES = .58$) (vgl. Unterabschnitt 6.3.2.5). Dies deutet darauf hin, dass die Teilnehmer_innen während der laut-denkenden Korrektur der Schülerlösungstexte, die eine hohe fachlich-konzeptuelle Qualität aufweisen, sprachliche Qualitäten stärker positiv wertend/akzeptierend berücksichtigten, als bei den Schülerlösungstexten mit einer geringen fachlich-konzeptuellen Qualität.

3. Der Medianunterschied zwischen den prozentuellen Häufigkeiten negativ wertender/ablehnender Äußerungen ist bei den Schülerlösungstexten mit geringer fachlich-konzeptueller Qualität geringer (moderater Effekt; $ES = .37$), als zwischen Schülerlösungstexten mit hoher fachlich-konzeptueller Qualität (moderater Effekt; $ES = .42$) (vgl. Unterabschnitt 6.3.2.5). Dies deutet darauf hin, dass die Teilnehmer_innen während der laut-denkenden Korrektur der Schülerlösungstexte, die eine hohe fachlich-konzeptuelle Qualität aufweisen, sprachliche Mängel stärker negativ wertend/ablehnend berücksichtigten, als bei den Schülerlösungstexten mit einer geringen fachlich-konzeptuellen Qualität.
4. Der Unterschied zwischen den im Median vergebenen Punkten ist bei den Schülerlösungstexten mit geringer fachlich-konzeptueller Qualität deutlich geringer (moderater Effekt; $ES = .43$), als zwischen Schülerlösungstexten mit hoher fachlich-konzeptueller Qualität (starker Effekt; $ES = .61$) (vgl. Abschnitt 6.3.1). Hierin zeigt sich ein Hinweis, dass die Teilnehmer_innen bei den Schülerlösungstexten mit höherer fachlich-konzeptueller Qualität, sprachliche Merkmale in ihrer Punktevergabe stärker berücksichtigten, als bei den Schülerlösungstexten mit einer geringen fachlich-konzeptuellen Qualität.
5. Für die Einschätzungen der Teilnehmer_innen im Rahmen der fachlich-konzeptuellen und sprachlichen Paarvergleiche ergibt sich eine moderat ausgeprägte Rangkorrelation von $\tau^* = .44$, die signifikant verschieden von null ist (vgl. Unterabschnitt 6.4.2.2). Hierbei ist zu beachten, dass die vier Schülerlösungstexte im Rahmen der Entwicklungstudie derart ausgewählt wurden, dass ein Wert von $\tau^* = .00$ theoretisch zu erwarten gewesen wäre, wenn die Teilnehmer_innen bei den Paarvergleichen fachlich-konzeptuelle und sprachliche Schülerleistungen nicht miteinander konfundieren.

Zusammengefasst wurden bei der quantitativen Analyse der erhobenen Daten Kennwerte für unterschiedliche Facetten der von allen Teilnehmer_innen vorgenommenen Leistungsfeststellungen und -beurteilungen berechnet bzw. beim Vergleich von Kennwerten Muster identifiziert, die sich plausibel als Hinweise auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile durch die Teilnehmer_innen interpretieren lassen. Hinzu kommt, dass die in Tabelle 6.35 angegebenen Kennwerte (Effektstärkemaß ES; Rangkorrelationskoeffizient τ^*) bzw. Kennwertunterschied (Effektstärkemaß ES) überwiegend eine moderate Ausprägung aufweisen.

6.5.2.3. Quintessenz der Teilbefunde zur Konfundierung fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung

Aus einem unmittelbaren Vergleich der in Unterabschnitt 6.5.2.1 und 6.5.2.2 zusammengeführten qualitativen und quantitativen Teilbefunde zu Forschungsfrage (F2) geht folgendes hervor:

- Auf qualitativer Ebene konnten bei mehreren (aber nicht allen) Teilnehmer_innen verschiedene Auffälligkeiten bzw. Muster identifiziert werden, aus denen eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile im Rahmen der Laborsituation direkt hervorgeht. Mit Hilfe dieser qualitativen Teilbefunde lässt sich allerdings keine Aussage darüber treffen, wie ausgeprägt eine derartige Konfundierung ist.
- Bei der quantitativen Analyse wurden aus den erhobenen Daten Kennwerte berechnet bzw. Kennwertunterschiede identifiziert, die sich plausibel als indirekte Hinweise auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile durch die Teilnehmer_innen interpretieren lassen. Diese Kennwerte bzw. Kennwertunterschiede beziehen sich dabei auf verschiedene Facetten der fachlich-konzeptuellen und sprachlichen Leistungsurteilsgenese aller Teilnehmer_innen und sind überwiegend moderat ausgeprägt.

Insgesamt lieferte die qualitative und quantitative Analyse der erhobenen Daten also eine Vielzahl einander komplementärer Teilbefunde, die auf eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung durch Physiklehrkräfte schließen lassen. Die qualitativen Teilbefunde zeigten sich allerdings nicht bei allen Teilnehmer_innen und die quantitativen Teilbefunde bewegen sich überwiegend auf mittlerem Niveau. Aus den erhobenen Daten lässt sich daher weder auf eine nicht vorhandene noch auf eine stark ausgeprägte Konfundierung fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung schließen. Vielmehr kann auf Basis der in der Hauptstudie gewonnenen Teilbefunde auf eine moderaten Konfundierung der Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen geschlussfolgert werden. Auch dieser Befund lässt sich mit Hilfe des in Abschnitt 2.2.4 dargestellten Rahmenkonzepts einer Assessment Literacy interpretieren: Die vier Schülerlösungstexte, die sich kontrastierend bezüglich ihrer fachlich-konzeptuellen Qualität und/oder der Qualität ihrer sprachlichen Realisierung voneinander unterscheiden, stellten für die Teilnehmer_innen eine Kontextbedingung dar, die ihren tatsächlichen Handlungsspielraum in der Laborsituation mehr oder minder begrenzte. Dementsprechend weist eine moderate Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen durch die Teilnehmer_innen darauf hin, dass diese...

... entweder über kaum kompetenztheoretisch gedachtes Wissen und Können und/oder berufsbezogenen Überzeugungen darüber verfügen, wie bzw. dass fachlich-konzeptuelle und sprachliche Leistungen getrennt voneinander festgestellt und beurteilt werden können und/oder wie bzw. dass Konfundierungen fachlich-konzeptueller und sprachlicher Leistungsurteile vermieden werden können,

... oder sie zwar über derartiges kompetenztheoretisch gedachtes Wissen und Können und/oder berufsbezogenen Überzeugungen verfügen, es ihnen aber nicht gelingt Kontextbedingungen, die ihren Handlungsspielraum begrenzen, mit zu berücksichtigen, weswegen es ihnen in ihrer Logik des Handelns nicht gelingt, Konfundierungen fachlich-konzeptueller und sprachlicher Leistungsurteile zu vermeiden.

6.6. Zusammenfassung

Das vorangegangene Kapitel dient der umfassenden Darstellung der empirischen Hauptstudie der vorliegenden Arbeit. Dabei wurde sich den Fragen gewidmet, auf welche Ressourcen Physiklehrkräfte bei der Genese fachlich-konzeptueller und sprachlicher Leistungsurteile über schriftliche, aus einer Klassenarbeit stammenden Schülerleistungen zurückgreifen (Forschungsfrage (F1)) und inwieweit bei dieser Genese eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile stattfindet (Forschungsfrage (F2)).

Zunächst wurde in Unterkapitel 6.1 die heterogene Gelegenheitsstichprobe aus im Schuldienst aktiven Physiklehrkräften, die an der Hauptstudie der vorliegenden Arbeit teilgenommen haben, detailliert beschrieben. Dem folgte eine Erläuterung des Transkriptions- und Segmentierungssystems, mit dessen Hilfe die von den Teilnehmern_Teilnehmerinnen erhobenen Verbaldaten aufbereitet wurden (vgl. Unterkapitel 6.2).

Entsprechend dem Mixed-Methods-Triangulationsdesign zur geplanten Auswertung der in der Hauptstudie erhobenen Daten (vgl. Abschnitt 5.4.2) widmeten sich Unterkapitel 6.3 und 6.4 der qualitativen und quantitativen Analysen der von den Teilnehmer_innen erhobenen Laut-Denk-Daten, sowie der Daten, die von den Teilnehmern_Teilnehmerinnen in den retrospektiven Befragungen gewonnen wurden. Hierbei erfolgte eine umfassende Darstellung des methodischen Vorgehens, der Ergebnisse, der Interpretationen und der Limitationen der bei diesen Analysen gewonnenen Teilbefunde.

Schließlich wurden in Unterkapitel 6.5 die zuvor umfassend beschriebenen qualitativen und quantitativen Teilbefunde zu Forschungsfrage (F1) und (F2) in ein kaleidoskopartiges Gesamtbild zusammengeführt. Die zentralen Teilbefunde zu Forschungsfrage (F1) ließen sich dabei zu 6 Ressourcen zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen verdichten, die die Teilnehmer_innen als bis zu einem bestimmten Grad bezüglich fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung grundgebildet charakterisieren lassen. Bei den 21 Teilnehmer_innen der Hauptstudie zeigte sich...

... eine Nutzung von Lehrerwissen und -können, dass sich speziell auf die Feststellung und Beurteilung fachlich-konzeptueller und/oder sprachlicher Schülerleistungen bezieht (vgl. Unterabschnitt 6.5.1.1).

... in Teilen eine holistische Feststellung und Beurteilung schriftlicher Schülerleistungen (vgl. Unterabschnitt 6.5.1.2).

... ein Handeln auch auf Basis von generalisiertem Wissen und Können und berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung (vgl. Unterabschnitt 6.5.1.3).

... ein Handeln bei der Leistungsurteilsgenese, dass deutlich an der kriterialen Bezugsnorm orientiert ist (vgl. Unterabschnitt 6.5.1.4).

... vor allem eine Beachtung fachlich-konzeptueller Merkmale bei der Feststellung und Beurteilung schriftlicher Schülerleistungen (vgl. Unterabschnitt 6.5.1.5). Sprachliche Merkmale wurden von den Teilnehmer_innen (in unterschiedlichem Umfang) zwar ebenfalls beachtet, spielten bei der Leistungsfestellung und -beurteilung insgesamt aber eine deutlich geringere Rolle (vgl. Unterabschnitt 6.5.1.5).

... die Tendenz sprachliche Schülerleistungen defizitorientiert festzustellen und zu beurteilen (vgl. Unterabschnitt 6.5.1.6). Demgegenüber wurden fachlich-konzeptuelle Schülerleistungen von den Teilnehmer_innen zum Teil defizitorientiert, in Teilen aber auch fähigkeitsorientiert festgestellt und beurteilt (vgl. Unterabschnitt 6.5.1.6).

Ferner lieferten sowohl die qualitative, als auch die quantitative Analyse der erhobenen Daten, zahlreiche, einander komplementäre Teilbefunde, die dafür sprechen, dass die Teilnehmer_innen beim Feststellen und Beurteilen schriftlicher Schülerleistungen fachlich-konzeptuelle und sprachliche Schülerleistungen auf einem moderaten Niveau miteinander konfundieren (vgl. Abschnitt 6.5.2). Vor dem Hintergrund des Rahmenkonzepts einer Assessment Literacy (vgl. Abschnitt 2.2.4) spricht dieser Befund dafür, dass es den Teilnehmer_innen in der Laborsituation nur bedingt gelungen ist, das Spannungsverhältnis zwischen der kontrastierenden Auswahl der vier Schülerlösungstexte und ihrem Wissen und Können zur Feststellung und Beurteilung von (fachlich-konzeptuellen und/oder sprachlichen) Schülerleistungen und/oder ihren berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung aufzulösen. Auf qualitativer Ebene zeigte sich beispielsweise bei mehreren Teilnehmer_innen die Bewertungslogik fachlich-konzeptuelle Schülerleistungen zu relativieren, wenn gleichzeitig bestimmte Erwartungen bezüglich der sprachlichen Realisierung eines Schülerlösungstextes nicht erfüllt waren. Auf quantitativer Ebene zeigten sich z. B. Hinweise darauf, dass die Teilnehmer_innen eine Feststellung und Beurteilung der sprachlichen Realisierung eines Schülerlösungstextes vor allem dann vornehmen, wenn dieser bis zu einem bestimmten Grad auch ihren fachlich-konzeptuellen Erwartungen entspricht.

Alles in allem wurden im Rahmen der empirischen Hauptstudie also zum einen gänzlich neue Erkenntnisse gewonnen. Zum anderen wurde in vielerlei Hinsicht der in Teil I der vorliegenden Arbeit dargestellte Forschungsstand zu Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung und zur Bedeutung von Sprache im Rahmen von Leistungsfeststellung und -beurteilung im Physikunterricht weiter gestützt, was im folgenden Kapitel noch ausführlich dargestellt wird. Ferner werden im folgenden Kapitel die Befunde der empirischen Hauptstudie kritisch diskutiert, sowie Konsequenzen für zukünftige physikdidaktische Forschung und die Aus- und Weiterbildung von Physiklehrkräften abgeleitet.

7. Diskussion

Im empirischen Teil der vorliegenden Arbeit wurde den folgenden Forschungsfragen nachgegangen:

(F1) Welche Ressourcen werden von Physiklehrkräften bei schriftlichen, aus einer Klassenarbeit stammenden Schülerleistungen zur Genese fachlich-konzeptueller und sprachlicher Leistungsurteile eingesetzt?

(F2) Inwieweit findet im Rahmen einer Klassenarbeit bei der Feststellung und Beurteilung von schriftlichen Schülerleistungen durch Physiklehrkräfte eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile statt?

In Summe liefert die empirische Hauptstudie (vgl. Kapitel 6) reichhaltige Befunde zum bislang wenig erforschten Handeln von im Schuldienst aktiven Physiklehrkräften beim Feststellen und Beurteilen schriftlicher Schülerleistungen. Insbesondere ist es gelungen, empirische Evidenz zu gewinnen, die dafür spricht, dass Physiklehrkräfte bei der Genese von Leistungsurteilen fachlich-konzeptuelle und sprachliche Schülerleistungen auf einem moderaten Niveau miteinander konfundieren. Das folgende Kapitel hat zum Ziel, Limitationen des empirischen Teils der vorliegenden Arbeit zu diskutieren und Konsequenzen der gewonnenen Befunde für zukünftige physikdidaktische Forschung, sowie die Aus- und Weiterbildung von Physiklehrkräften abzuleiten.

7.1. Limitationen der empirischen Hauptstudie

Das Vorgehen in der empirischen Hauptstudie folgte einem deskriptiv-explorativen Studiendesign (vgl. Unterkapitel 4.2). Zuvor wurde in einer eigens hierfür angelegten Entwicklungsstudie geklärt, welches methodische Vorgehen sowohl zu den Forschungsfragen (F1) und (F2), als auch zu den Eigenschaften des Untersuchungsgegenstands passgenau ist und eine Verbindung zwischen Forschungsfragen und Untersuchungsgegenstand ermöglicht (vgl. Kapitel 5). Aus dem gewählten methodischen Vorgehen ergeben sich zwangsläufig auch Grenzen. Spezifische Limitationen der einzelnen qualitativen und quantitativen Teilanalysen wurden an entsprechender Stelle bereits erörtert (vgl. Unterabschnitt 6.3.1.4, 6.3.2.3, 6.3.2.4, 6.3.2.5, 6.4.1.4 und 6.4.2.2). Auf eine erneute Darstellung dieser Limitationen wird in diesem Unterkapitel daher verzichtet. Stattdessen beschränkt sich die Darstellung auf Limitationen, die den Untersuchungsgegenstand des empirischen Teils der vorliegenden Arbeit betreffen. Diese Limitationen sind für eine kritische Betrachtung des empirischen Teils der vorliegenden Arbeit von zentraler Bedeutung. Grund hierfür

ist, dass die analysierten Daten in einer vergleichsweise kleinen Gelegenheitsstichprobe erhoben wurden und damit die Übertragbarkeit der gewonnenen Befunde in besonderem Maße davon abhängig ist, inwieweit sich die Merkmale des Untersuchungsgegenstands der vorliegenden Arbeit auch im angestrebten Übertragungskontext wiederfinden.

Zunächst sind die Limitationen zu benennen, die aus der Stichprobenziehung resultieren. In der Hauptstudie wurde eine heterogene Gelegenheitsstichprobe von N = 21 Hamburger Physiklehrkräften gewonnen (vgl. Unterkapitel 6.1). Die Heterogenität dieser vergleichsweise kleinen Stichprobe begünstigte die Datenauswertung im Rahmen der Hauptstudie, da hierdurch unterschiedlichste Ressourcen von Physiklehrkräften zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen (Forschungsfrage (F1)), sowie zahlreiche empirische Hinweise, die für eine Konfundierung fachlich-konzeptueller und sprachlicher Leistungsurteile von Physiklehrkräften sprechen (Forschungsfrage (F2)), identifiziert werden konnten. Gleichzeitig weist die gewonnene Gelegenheitsstichrobe eine gewisse Homogenität auf, aus der sich die folgenden Limitationen ergeben:

1. Alle Teilnehmer_innen waren zum Erhebungszeitpunkt in der Freien und Hansestadt Hamburg im Schuldienst tätig. Die in der Hauptstudie gewonnenen Befunde beziehen sich also auf Physiklehrkräfte, deren tägliche Arbeit – insbesondere auch ihre Umsetzung von schulischer Leistungsfeststellung und -beurteilung – von den schulpolitischen Vorgaben, schulkulturellen Gepflogenheiten und gesellschaftlichen Rahmenbedingungen des Bildungssystems eines bestimmten Stadtstaates und Bundeslandes festgelegt und/oder beeinflusst wird. Die Befunde der Hauptstudie sind daher nicht zwangsläufig auf Physiklehrkräfte übertragbar, die beispielsweise in anderen Bundesländern oder in eher ländlichen Regionen im Schuldienst aktiv sind.
2. In Bezug auf Forschungsfrage (F2) spricht die Assessment Literacy Konzeption aus Abschnitt 2.2.4 dafür, dass es einen kritischen Variationsbereich gibt, innerhalb dessen sich das Wissen und Können, sowie die berufsbezogenen Überzeugungen von Physiklehrer_innen zu schulischer Leistungsfeststellung und -beurteilung in einer Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen niederschlagen (vgl. auch Unterabschnitt 6.5.2.3). Liegt daher das Wissen und Können, sowie die berufsbezogenen Überzeugungen von Physiklehrkräfte nicht (bzw. nur zum Teil) in diesem kritischen Bereich, so ist auch keine (bzw. nur teilweise eine) Konfundierung fachlich-konzeptueller und sprachlicher Teilleistungsurteile durch die Physiklehrkräfte zu erwarten (vgl. Renkl, 1993, S. 119). Da sich die Stichprobengewinnung in der Hauptstudie aus unterschiedlichen Gründen als recht schwierig darstellte (für Details vgl. Unterkapitel 6.1) und die befragten Physiklehrkräfte – wenn auch nur in sehr schwacher Tendenz – eine subjektiv eher hoch empfundene Diagnosesicherheit aufweisen (vgl. Abschnitt 6.1.2), handelt es sich bei der gewonnenen Gelegenheitsstichprobe möglicherweise jedoch um eine Positivauswahl von Physiklehrkräften in Bezug auf ihr Wissen und Können, sowie ihren berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung. Es kann daher nicht ausgeschlossen werden, dass der Befund einer Konfundierung

fachlich-konzeptueller und sprachlicher Schülerleistungen durch Physiklehrkräfte auf einem „lediglich“ moderaten Niveau auch darauf zurückgeführt werden kann, dass die Wissens- und Könnensbestände und die berufsbezogenen Überzeugungen der Teilnehmer_innen nur zum Teil in einem kritischen Bereich variieren.

Daneben sind Limitationen zu benennen, die aus der Ausgestaltung der Datenerhebung im Rahmen der Hauptstudie hervorgehen. Diese erfolgt in einer Laborsituation mit kontrolliertem Ablaufplan (vgl. Abschnitt 5.4.1). Das ökologische Validitätskriterium, dass die Laborsituation mit der Alltagspraxis von Physiklehrer_innen bei der Leistungsfeststellung und -beurteilung im Rahmen einer Klassenarbeit kongruieren soll, wurde in der Entwicklungsstudie in besonderem Maße berücksichtigt (vgl. Abschnitt 5.2.2). Allerdings ist es allein aufgrund des Umstands, dass die Befragung in einer Laborsituation stattfand, möglich, dass hierdurch das Handeln der Teilnehmer_innen beeinflusst und/oder verändert wurde (vgl. von Aufschnaiter, 2014, S. 85). Hinzu kommen die folgenden besonderen Merkmale der Laborsituation, die die Übertragbarkeit der in der Hauptstudie gewonnenen Befunde beschränken:

3. Die Teilnehmer_innen wurden in der Laborsituation gebeten zu einer bestimmten Klassenarbeitsaufgabe – die Aufgabe Weltraumspaziergang – entsprechend ihrer täglichen Berufspraxis einen Erwartungshorizont zu erstellen und zugehörige Schülerlösungstexte zu korrigieren (vgl. Abschnitt 5.4.1). Die in der Hauptstudie gewonnenen Befunde sind daher nicht zwangsläufig auf Klassenarbeitsaufgaben, die beispielsweise andersartige curriculare Inhalte und/oder Kompetenzbereiche von Schüler_innen abprüfen, übertragbar.
4. Klassenarbeiten bestehen im Schulalltag meist nicht aus einer einzigen sondern aus einer Vielzahl unterschiedlicher Aufgaben. In der Laborsituation werden die Teilnehmer_innen allerdings lediglich gebeten, sich in die Situation hineinzuversetzen, die Aufgabe Weltraumspaziergang in einer Klassenarbeit der 9. Jahrgangsstufe als Grundwissensaufgabe eingesetzt zu haben (vgl. Anhang C.1 Aufgabenheft für Physiklehrkräfte, S. 2). Die konkrete Ausgestaltung dieser Klassenarbeit, z. B. welche weiteren Aufgaben gestellt wurden, wurde nicht weiter spezifiziert. Damit sind aber die in der Hauptstudie gewonnenen Befunde insofern limitiert, als dass die Einbettung einer bestimmten Aufgabe in eine Klassenarbeit als mutmaßlicher Einflussfaktor auf die Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen durch Physiklehrkräfte in der Ausgestaltung der Laborsituation wenig berücksichtigt ist.
5. Wie in Unterabschnitt 6.3.2.1 dargestellt, zeigten sich in den Selbstauskünften der Teilnehmer_innen Hinweise auf reaktive Effekte durch die Erhebungsmethode des lauten Denkens. Auch wenn diese Hinweise geringfügig und nur bei vier Teilnehmer_innen anzutreffen waren, ist dennoch die Validität der durch diese Erhebungsmethode gewonnenen Daten bis zu einem bestimmten Grad kritisch zu hinterfragen. Hinzu kommt, dass auch die Instruktionen im Aufgabenheft (vgl. Anhang C.1 Aufgabenheft), sowie die Strukturierung der retrospektiven Befragung (vgl. Anhang

C.3 Durchführungsmanual, S. 8 u. f.) das Potenzial reaktiver Effekte in sich bergen. Diese weisen eine gewollt lenkende Wirkung auf das Handeln der Teilnehmer_innen in der Laborsituation auf. Gleichzeitig ist aber nicht auszuschließen, dass diese Lenkung zu einem Handeln der Teilnehmer_innen führte, dass sich lediglich eingeschränkt mit ihrer täglichen Berufspraxis bei der Feststellung und Beurteilung schriftlicher, aus einer Klassenarbeit stammender Schülerleistungen deckt. Dies gilt insbesondere für die stark lenkenden Instruktionen an die Teilnehmer_innen im Rahmen der fachlich-konzeptuellen und sprachlichen Paarvergleiche (vgl. Anhang C.3 Durchführungsmanual, S. 9).

6. Da die kriteriale Unterscheidbarkeit der Schülerlösungstexte A bis D bezüglich ihrer fachlich-konzeptuellen Qualität und der Qualität ihrer sprachlichen Realisierung in der Laborsituation im Vordergrund stehen sollte, wurden den Teilnehmer_innen im Rahmen der Laborsituation keine Informationen über Personenmerkmale der Schüler_innen, die die Schülerlösungstexte verfasst haben, zur Verfügung gestellt (vgl. Kapitel 5). Aus der bisherigen psychologischen Urteilsforschung zu schulischer Leistungsfeststellungen und -beurteilungen ist jedoch bekannt, dass Lehrkräfte bei ihrer Leistungsurteilsgenese für gewöhnlich auch Personenmerkmale von Schüler_innen berücksichtigen (vgl. Abschnitt 2.1.4). Es ist daher plausibel anzunehmen, dass dies auch auf die Genese fachlich-konzeptueller und sprachlicher Leistungsurteile durch Physiklehrkräfte zutrifft. Die vorliegende Arbeit kann zu dieser Annahme allerdings lediglich Hinweise liefern. So steht insbesondere der qualitative Befund, dass die Teilnehmer_innen vereinzelt auch auf mutmaßliche Personenmerkmale von Schüler_innen (z. B. ihr Geschlecht) zurückgegriffen haben (vgl. Unterabschnitt 6.3.2.3), mit dieser Annahme im Einklang. Über den Einfluss „tatsächlicher“ Personenmerkmalen von Schüler_innen auf die Genese fachlich-konzeptueller und sprachlicher Leistungsurteile von Physiklehrkräften lassen sich auf Grundlage des empirischen Teils der vorliegenden Arbeit jedoch keine Aussagen treffen.

7. Die Schülerlösungstexte A bis D wurden in der Entwicklungsstudie als Kontrastfälle ausgewählt (vgl. Kapitel 5). Diese kontrastierende Auswahl entspricht nicht notwendigerweise der Verteilung fachlich-konzeptueller und sprachlicher Schülerleistungen, wie sie Physiklehrkräfte in ihrer täglichen Berufspraxis vorfinden. Insbesondere bei Physiklehrkräften, die eine ausgeprägte soziale Bezugsnormorientierung aufweisen (vgl. Abschnitt 2.1.2), ist zu erwarten, dass diese in der Laborsituation der Hauptstudie deutlich anders handeln könnten, als in ihrem Berufsalltag. Bei den Teilnehmer_innen der Hauptstudie zeigte sich allerdings bei der Leistungsurteilsgenese ein Handeln, dass deutlich an der kriterialen Bezugsnorm orientiert ist (vgl. Unterabschnitt 6.5.1.4). Ferner weisen die Teilnehmer_innen ihrer Selbsteinschätzung nach in schwacher Tendenz eher eine kriteriale anstatt eine soziale Bezugsnormorientierung auf (vgl. Abschnitt 6.1.2). Es ist daher plausibel anzunehmen, dass die Befunde der empirischen Hauptstudie vor allem für ein an der kriterialen Bezugsnorm orientiertes Lehrerhandeln Gültigkeit beanspruchen können und dass diese beispielsweise

nicht zwangsläufig auf Lehrkräfte mit einer ausgeprägten sozialen Bezugsnormorientierung übertragbar sind.

Ziel des nun folgenden Unterkapitels ist eine Diskussion der Befunde der empirischen Hauptstudie, sowohl hinsichtlich bisheriger und zukünftiger physikdidaktischer Forschung, als auch im Hinblick auf Implikationen für die Aus- und Weiterbildung von Physiklehrkräften. Wie im Nachfolgenden deutlich wird, sind die im vorangegangenen Unterkapitel benannten Limitationen der empirischen Hauptstudie für diese Diskussion von zentraler Bedeutung.

7.2. Diskussion der Befunde der empirischen Hauptstudie

Die Befunde der empirischen Hauptstudie stützen in vielerlei Hinsicht den in Teil I der vorliegenden Arbeit dargestellten Forschungsstand zum Wissen und Können von Physiklehrkräften zu schulischer Leistungsfeststellung und -beurteilung. Des Weiteren ergeben sich Implikationen zum einen für zukünftige physikdidaktische Forschung und zum anderen für die Aus- und Weiterbildung von Physiklehrer_innen. In diesem Unterkapitel werden die Befunde der empirischen Hauptstudie daher sowohl hinsichtlich bisheriger und zukünftiger physikdidaktischer Forschung diskutiert, als auch im Hinblick auf Implikationen für die Lehrerbildung.

7.2.1. Diskussion der Befunde hinsichtlich bisheriger und zukünftiger physikdidaktischer Forschung

In Kapitel 2 wurde das weite Forschungsfeld um Lehrerwissen und -können zur schulischen Leistungsfeststellung und -beurteilung gesichtet und anschließend im Rahmenkonzept einer Assessment Literacy von Lehrkräften zusammengeführt, dass den theoretischen Konzeptionen der zuvor vorgestellten Forschungstraditionen erhaben ist. Auf Grundlage dieses Rahmenkonzepts erfolgte die Interpretation der Befunde im empirischen Teil der vorliegenden Arbeit. Wie bei der Integration der Befunde zu Forschungsfrage (F1) und (F2) ausführlich dargestellt (vgl. Unterkapitel 6.5), haben die Teilnehmer_innen in der Laborsituation für ihre fachlich-konzeptuellen und sprachlichen Leistungsfeststellungen und -beurteilungen auf Ressourcen zurückgegriffen, die im Konzept einer Assessment Literacy miteinander vereinbare und/oder zueinander in Konflikt stehende Faktoren darstellen und zwischen denen es für eine Lehrkraft im Rahmen einer schulische Leistungsfeststellung und -beurteilung betreffenden Handlungsepisode gilt, einen Kompromiss zu realisieren (Wissen und Können; berufsbezogene Überzeugungen zu (fachspezifischer) Leistungsfeststellung und -beurteilung; Kontextbedingungen der Handlungsepisode). Ergo: Die zentrale Grundannahme des heuristischen Referenzrahmens der vorliegenden Arbeit über die Logik des Handelns von Lehrkräften im Kontext von schulischer Leistungsfeststellung und -beurteilung hat im Lichte der in der vorliegenden Arbeit gewonnenen Erkenntnisse Be-

stand. Die Befunde der Hauptstudie bestärken damit also empirisch die Erklärmächtigkeit des Rahmenkonzepts einer Assessment Literacy von Lehrkräften (vgl. Abschnitt 2.2.4).

Darüber hinaus sind insbesondere, wie im folgenden zusammenfassend dargestellt, fünf Ergebnisse der empirischen Hauptstudie mit Befunden bisheriger Untersuchungen zum Umgang von Lehrkräften mit Sprache (in Leistungssituationen) im Physikunterricht vereinbar:

1. Im Handeln der Teilnehmer_innen zeigte sich eine Nutzung von Lehrerwissen und -können, dass sich speziell auch auf die Feststellung und Beurteilung sprachlicher Schülerleistungen bezieht (vgl. Unterabschnitt 6.5.1.1). Unter anderem fokussierten sie bei ihrer laut denkenden Korrekturarbeit explizit sprachliche Merkmale der Schülerlösungstexte, mitvokalisierten Beurteilungskriterien, die die sprachliche Realisierung eines Schülerlösungstextes betreffen und berücksichtigten bei der Erstellung ihrer Erwartungshorizonte (in unterschiedlichem Umfang) auch sprachliche Merkmale eines Schülerlösungstextes. Diese Befunde sind vergleichbar mit einem zentralen Ergebnis aus dem Fallbeispiel von Tajmel (vgl. Abschnitt 3.2.2). In diesem zeigte sich global zusammengefasst, dass (angehende) Naturwissenschaftslehrkräfte bei der Korrektur von Textprodukten „die sprachlichen Leistungen der Schülerinnen und Schüler mitbewerten“ (Tajmel, 2010, S. 174).
2. Sprachliche Merkmale der Schülerlösungstexte wurden von den Teilnehmer_innen bei ihrer Leistungsfeststellung und -beurteilung zwar berücksichtigt, spielten aber im Vergleich zu fachlich-konzeptuellen Merkmalen eine deutlich geringere Rolle (vgl. Unterabschnitt 6.5.1.5). Ein analoger Befund zeigte sich in der Untersuchung von Lyon zur Entwicklung der Expertise angehender Naturwissenschaftslehrkräfte bezüglich des Umgangs mit sprachlich-kultureller Heterogenität bei schulischen Leistungsfeststellungen und -beurteilungen (vgl. Abschnitt 3.2.1). In den hier vorgenommenen qualitativen Analysen der erhobenen Daten von insgesamt 3 Lehramtsstudierenden zeigte sich, dass diese ihre Leistungsfeststellungen und -beurteilungen überwiegend auf fachlich-konzeptuelles Wissen und Können von Schüler_innen beschränken und sprachliche Schülerleistungen, wenn überhaupt, nur marginal feststellten und beurteilten (vgl. Lyon, 2013d, S. 9).
3. Thieme & Mavruk (2018) zeigten, dass die von ihnen befragten Biologielehrkräfte bei Textprodukten von Schüler_innen feststellen und beurteilen, inwieweit diese hinreichend ausführlich, aber gleichzeitig nicht zu ausschweifend realisiert sind (vgl. ebd., S. 292). In den Beurteilungskriterien, die die Teilnehmer_innen der empirischen Hauptstudie im Rahmen der retrospektiven Befragung einsetzten, zeigte sich etwas Vergleichbares (vgl. Abschnitt 6.4.1). Etliche dieser Kriterien adressieren Ausführlichkeit bzw. die Knappheit eines Schülerlösungstextes. Dies gilt insbesondere für die Beurteilungskriterien „Differenziertheit/Komplexität des Textes“, „Verdichtungsgrad/Präzision des Textes“ und „Vorhandensein von Redundanz“, die die Teilnehmer_innen sowohl bei den fachlich-konzeptuellen, als auch bei den sprachlichen Paarvergleichen einsetzten. Es lässt sich daher vermuten, dass der Befund, den Thie-

me & Mavruk (2018) für Biologielehrkräfte gewonnen haben, sich möglicherweise auch auf Physiklehrkräfte übertragen lässt.

4. Sprachliche Schülerleistungen wurden von den Teilnehmer_innen in der Laborsituation in einer tendenziell defizitorientierten Art und Weise festgestellt und beurteilt (vgl. Unterabschnitt 6.5.1.5). Hinweise, die gegen eine Tendenz zur defizitorientierten Feststellung und Beurteilung sprachlicher Schülerleistungen sprechen, zeigten sich bei weniger als einem Drittel der Teilnehmer_innen. Vergleichbar hierzu sind Befunde aus dem Fallbeispiel von Tajmel (vgl. Abschnitt 3.2.2), sowie der Studie von Buxton et al. (2013). In diesen Untersuchungen zeigte sich analog, dass (angehende) Naturwissenschaftslehrkräfte bei der Leistungsfeststellung und -beurteilung die sprachlichen Defizite von Schülertexten in den Vordergrund rücken (vgl. Tajmel, 2017b, S. 260 u. f.), bzw. dass Naturwissenschaftslehrkräfte beim Unterrichten von sprachlich-kulturell heterogener Lerngruppen vorrangig die sprachlich-kulturell bedingten Defizite von Schüler_innen im Blick haben (Buxton et al., 2013).

5. Auch zu der Bewertungslogik einer Relativierung des fachlich-konzeptuellen Eindrucks eines Schülerlösungstextes aufgrund seiner sprachlichen Realisierung, bzw. umgekehrt, die bei der qualitativen Analyse der Laut-Denk-Daten der Teilnehmer_innen identifiziert werden konnte (vgl. Unterabschnitt 6.3.2.4), finden sich auffällig ähnliche Befunde in zwei bisherigen Untersuchungen zum Umgang von Lehrkräften mit Sprache in Leistungssituationen im Physikunterricht. Zum einen zeigten sich auch in den Leistungsurteilsbegründungen einiger Teilnehmer_innen der Untersuchung von (Tajmel, 2017b) Auffälligkeiten, die auf die Bewertungslogik hinweisen, fachlich-konzeptuelle Merkmale des Textproduktes eines_einer Schülers_Schülerin mehr oder weniger zu relativieren, wenn dieses zu viele Mängel bezüglich einer sprachlichen Norm aufweist (vgl. Abschnitt 3.2.2). Zum anderen stellte auch Willems (2007) in ihrer ethnographischen Untersuchung im Kontext von bilingualem Physikunterricht fest, „dass die sprachliche Ebene nur dann in die Bewertungen von schriftlichen Tests eingeht, wenn der physikalische Inhalt dadurch falsch oder unvollständig wird“ (ebd., S. 194). Sprachliche Schülerleistung konfundiert mit fachlich-konzeptueller Leistung und umgekehrt festzustellen und zu beurteilen, ist daher möglicherweise ein generelles Phänomen, dass sich im Handeln von Physiklehrkräften im Kontext von schulischer Leistungsfeststellung und -beurteilung zeigt.

Durch die vorgenommene Auflistung werden zum einen bereits bestehende (vor allem in Untersuchungen mit ausgeprägtem Fallstudiencharakter) Erkenntnisse zum Umgang von Lehrkräften mit Sprache (in Leistungssituationen) im Physikunterricht weiter gestützt. Zum anderen lässt sich dadurch, dass sich zwischen Ergebnissen bisheriger Untersuchungen und Befunden der empirischen Hauptstudie auffällige Ähnlichkeiten zeigen, begründet vermuten, dass besonders die oben aufgeführten fünf Ergebnisse der empirischen Hauptstudie ein Generalisierungspotenzial aufweisen.

Bei den Befunden der Hauptstudie handelt es sich um deskriptive und insbesondere explorative Ergebnisse, die aus erhobenen Daten in einer vergleichsweise kleinen Gelegen-

heitsstichprobe gewonnen wurden. Ferner ist die empirische Hauptstudie der vorliegenden Arbeit keine konfirmatorisch angelegte Untersuchung (vgl. Unterkapitel 4.2). Aus Sicht hypothesentestender erziehungswissenschaftlicher Forschung sind deshalb alle qualitativen und quantitativen Teilbefunde der Hauptstudie als Hypothesen zu verstehen, die sich in zukünftigen Untersuchungen bewähren müssen. Des Weiteren sind die Ergebnisse des empirischen Teils der vorliegenden Arbeit schon aufgrund des explorativen Charakters der Hauptstudie gewiss nicht allumfassend. Es erscheint daher sinnvoll zukünftig insbesondere den folgenden Fragen nachzugehen:

- Fragen in Bezug auf die sechs Ressourcen zur fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung, die aus der Integration der Befunde zu Forschungsfrage (F1) hervorgingen (vgl. Abschnitt 6.5.1):
    - Reproduzieren sich die beschriebenen Ressourcen auch in größeren und/oder konfirmatorisch angelegten Untersuchungen?
    - Finden sich die beschriebenen Ressourcen bei Physiklehrkräften generell? Inwieweit lassen sich, bezogen auf die Ressourcen, die Physiklehrkräfte zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen nutzen, empirisch und inhaltlich voneinander verschiedene Handlungstypen identifizieren?
    - Die beschriebenen Ressourcen stehen im Einklang mit dem Rahmenkonzept einer Assessment Literacy von Lehrkräften. Lassen sich diese Ressourcen für ein geschlossenes Instrument zur Erfassung (von Teilfacetten) der Assessment Literacy von Physiklehrkräften operationalisieren? Falls dies möglich ist: Welche empirischen Zusammenhänge zeigen sich zwischen dem (den) Konstrukt(en), die durch ein solches Instrument erfasst werden und beispielsweise dem Fachwissen-, dem fachdidaktischen oder dem pädagogischen Wissen von Physiklehrkräften?
    - Inwieweit werden die beschriebenen Ressourcen von Physiklehrkräften im Rahmen ihrer Aus- und Weiterbildung erworben? Inwieweit sind sie ein Produkt langjähriger beruflicher Praxis? Inwieweit sind sie Ausdruck einer Orientierung an dem Physikunterricht, den Physiklehrkräfte selbst als Schüler_innen erlebt haben (vgl. Klinghammer et al., 2016, S. 182)?
    - Sind die beschriebenen Ressourcen – insbesondere die einer Defizitorientierung bei der Feststellung und Beurteilung sprachlicher Schülerleistungen – Interventionsmaßnahmen zugänglich? Falls dies möglich ist: Wie sollten derartige Interventionen inhaltlich gestaltet sein und in welcher (welchen) Phase(n) der Lehrerbildung sollten sie angesiedelt sein, um wirksam zu werden?
    - Lassen sich neben den beschriebenen noch weitere Ressourcen von Physiklehrkräften zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen identifizieren?

- Inwieweit finden sich die beschriebenen Ressourcen auch im beruflichen Handeln von Physiklehrkräften, das nicht einem *Summative Assessment*, sondern einem *Initial* und/oder *Formative Assessment* dient (vgl. Abschnitt 1.1.2)?

- Fragen hinsichtlich einer Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen durch Physiklehrkräfte (vgl. Integration der Befunde zu Forschungsfrage (F2) in Abschnitt 6.5.2):

  - Lassen sich, die auf qualitativer Ebene identifizierten Bewertungslogiken,...

    ... sprachliche Schülerleistung konfundiert mit fachlicher Leistung festzustellen und zu beurteilen,...

    ... fachliche Schülerleistungen konfundiert mit sprachlichen Leistungen festzustellen und zu beurteilen und ...

    ... fachlich-konzeptuelle Merkmale einer Schülerleistung von Merkmalen der sprachlichen Realisierung zu trennen...

    ... auch in zukünftigen Untersuchungen mit Physiklehrkräften identifizieren? Lassen sich auf qualitativer Ebene weitere Bewertungslogiken ausfindig machen, bei denen fachlich-konzeptuelle und sprachliche Schülerleistungen miteinander konfundiert werden und/oder Bewertungslogiken, durch die eine solche Konfundierung vermieden wird?

  - Lassen sich die eben benannten Bewertungslogiken auch durch ein geschlossenes Testinstrument valide abbilden? Falls dies möglich ist: Finden sich in größer angelegten Studien bestimmte Handlungstypen von Physiklehrkräften, die gehäuft auf eine oder mehrere dieser Bewertungslogiken zurückgreifen?

  - Lässt sich der qualitative Befund, dass Physiklehrkräfte Beurteilungskriterien, die die sprachliche Realisierung des Textproduktes eines_einer Schülers_-Schülerin betreffen, auch zur Begründung von Unterschieden hinsichtlich der fachlich-konzeptuellen Qualität verwenden, in größeren und/oder konfirmatorisch angelegten Studien reproduzieren?

  - Lassen sich weitere quantitative empirische Befunde dafür gewinnen (vgl. Integration der quantitativen Befunde zu Forschungsfrage (F2) in Unterabschnitt 6.5.2.2), dass Physiklehrkräfte...

    ... vor allem dann die sprachliche Realisierung der Textprodukte von Schüler_innen feststellen und beurteilen, wenn diese auch fachlich-konzeptuelle Erwartungen bis zu einem bestimmten Grad erfüllen?

    ... bei Textprodukten von Schüler_innen, die eine hohe fachlich-konzeptuelle Qualität aufweisen, die sprachliche Realisierung stärker berücksichtigen, als bei Textprodukten mit einer geringen fachlich-konzeptuellen Qualität?

    ... beim Paarvergleich von Textprodukten von Schüler_innen fachlich-konzeptuelle und sprachliche Schülerleistungen miteinander konfundieren?

- Was sind die Ursprünge einer Logik des Handelns von Physiklehrkräften, aus der eine Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen hervorgeht? Wird eine solche Logik in bestimmten Phasen der Lehrerbildung erworben? Ist sie ein Produkt der eigenen beruflichen Praxis? Ist sie Ausdruck einer Orientierung an dem Physikunterricht, den Physiklehrkräfte selbst als Schüler_innen erlebt haben (vgl. Klinghammer et al., 2016, S. 182)?
- Sind die gewonnenen qualitativen und quantitativen Befunde, die für eine Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen durch Physiklehrkräfte sprechen, Interventionsmaßnahmen zugänglich? Falls dies möglich ist: Wie sollten derartige Interventionen inhaltlich gestaltet sein und in welcher (welchen) Phase(n) der Lehrerbildung sollten sie angesiedelt sein, um wirksam zu werden?
- Inwieweit zeigt sich eine Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen auch im beruflichen Handeln von Physiklehrkräften, das nicht einem *Summative Assessment*, sondern einem *Initial* und/oder *Formative Assessment* dient (vgl. Abschnitt 1.1.2)?

Berücksichtigt man die in Unterkapitel 7.1 dargestellten Limitationen der empirischen Hauptstudie, ergeben sich zudem folgende Fragen, denen sich zukünftige Forschung ebenfalls widmen sollte, um diese Limitationen zu überwinden:

- Zeigen sich die Ergebnisse der empirischen Hauptstudie auch außerhalb einer Laborsituation mit kontrolliertem Ablaufplan (z. B. in einer ethnographische Untersuchung)? Inwieweit decken sich die gewonnenen Befunde mit der tatsächlichen täglichen Berufspraxis von Physiklehrer_innen bei der Leistungsfeststellung und -beurteilung im Rahmen einer Klassenarbeit?
- Inwieweit sind die in der empirischen Hauptstudie gewonnenen Befunde auf Physiklehrkräfte, die in der Freien und Hansestadt Hamburg im Schuldienst aktiv sind, verallgemeinerbar? Lassen sie sich auf Physiklehrkräfte im Allgemeinen generalisieren? Sind sie auch in einer Stichprobe von Physiklehrkräfte, bei der eine Positivauswahl in Bezug auf ihr Wissen und Können, sowie ihre berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung ausgeschlossen werden kann, reproduzierbar?
- Lassen sich die Ergebnisse der empirischen Hauptstudie auf Klassenarbeitsaufgaben im Fach Physik verallgemeinern, in denen Schüler_innen Textprodukte anfertigen sollen, unabhängig von curricularen Inhalten und/oder Kompetenzbereichen, die diese Aufgaben abprüfen? Welchen Einfluss hat die Einbettung einer solchen Aufgabe in eine Klassenarbeit auf die Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen durch Physiklehrkräfte?
- Welchen Einfluss hat die Verteilung fachlich-konzeptueller und sprachlicher Schülerleistungen, die einer Physiklehrkraft zur Korrektur vorliegen, auf ihre Leistungsfeststellung und Beurteilung (Stichwort: soziale Bezugsnormorientierung)?

- Welchen Einfluss haben „tatsächliche“ Personenmerkmale von Schüler_innen auf die Genese fachlich-konzeptueller und sprachlicher Leistungsurteile von Physiklehrkräften? Inwieweit finden sich hierbei *kompensatorische*[156] und/oder *substitutive Mechanismen*[156]?

Neben einer Diskussion der Ergebnisse der empirischen Hauptstudie hinsichtlich bisheriger und zukünftiger physikdidaktischer Forschung lassen sich auf Grundlage der gewonnenen Befunde auch Konsequenzen im Hinblick auf die Aus- und Weiterbildung von Physiklehrkräften formulieren. Dies erfolgt im nun folgenden Abschnitt.

7.2.2. Diskussion der Befunde im Hinblick auf Implikationen für die Aus- und Weiterbildung von Physiklehrkräften

In diesem Abschnitt werden Konsequenzen des empirischen Teils der vorliegenden Arbeit hinsichtlich der Aus- und Weiterbildung von Physiklehrkräften erörtert. Die Ausgangsbasis für diese Erörterung bilden die in Unterabschnitt 2.2.5 formulierten theoretischen Implikationen des Konzepts einer Assessment Literacy für die Verbesserung der Lehrerbildung.

Um einen umfassenden Beitrag zur Grundbildung von Lehrkräften hinsichtlich schulischer Leistungsfeststellung und -beurteilung zu leisten, sollte in allen Phasen der Lehrerbildung ein diesbezüglich breitgefächertes Aus- bzw. Weiterbildungsangebot angesiedelt sein. Zum einen sollte ein am täglichen Unterrichtsgeschehen orientiertes Hintergrundwissen zu schulischer Leistungsfeststellung und -beurteilung vermittelt werden. Vor dem Hintergrund, dass alltägliche Leistungsfeststellung und -beurteilung durch Physiklehrkräfte einen Gegenstand dargestellt, mit dem sich erziehungswissenschaftliche Forschung bislang kaum auseinandergesetzt hat (vgl. Kapitel 2), kann vermutet werden, dass vielen (angehenden) Physiklehrkräften im Rahmen der Aus- und Weiterbildung (wenn überhaupt) lediglich individuell gewonnenes Erfahrungswissen ihrer Lehrerbildner_innen aus der Schulpraxis vermittelt wird (vgl. hierzu die in Unterabschnitt 2.1.1.3 dargestellten bisherigen Erkenntnisse zu diagnostischen Kompetenzen von Lehrkräften, die mit dem Selbstauskunftsansatz gewonnen wurden). Demgegenüber stellt die Hauptstudie der vorliegenden Arbeit empirisch fundiertes Hintergrundwissen darüber bereit, wie im Schuldienst aktive Physiklehrkräfte (in einer an der täglichen Berufspraxis orientierten Laborsituation) Textprodukte von Schüler_innen feststellen und beurteilen, welchen Logiken sie dabei folgen, welche Maßstäbe sie hierbei für angemessen halten, sowie inwieweit sie fachlich-konzeptuelle und sprachliche Schülerleistungen miteinander konfundieren (vgl. Kapitel 6). Lohnenswert erscheint deshalb, Befunde der empirischen Hauptstudie – zusammen mit ihren nicht zu vernachlässigenden Limitationen! – in Aus- bzw. Weiterbildungsangeboten für Physiklehrkräften zu thematisieren. Hierdurch kann ein Beitrag für eine *inhaltliche Evidenzbasierung*

[156] *Kompensatorische Mechanismen* sind solche, die (in empirischen Untersuchungen) zu einer Maskierung von Kausalzusammenhängen führen (vgl. Renkl, 1993, S. 120). *Substitutive Mechanismen* sind solche, in denen „das Fehlen eines Faktors [...] durch das Vorhandensein eines anderen ausgeglichen wird“ (vgl. ebd.).

von Lehrerbildung zu schulischer Leistungsfeststellung und -beurteilung geleistet werden. Es würde der Forderung nachgekommen, in der Lehrerbildung „(möglichst nur) „evidente“ Wissensbestände an [...] Lehrkräfte heran[zutragen], so dass diese ihr [...] Handeln auf empirisch [...] abgesichertes Wissen gründen und solcherart effektivieren“ (Neuweg, 2018, S.63). Offenkundig besonders geeignet hierfür sind die Befunde der empirischen Hauptstudie, die auffällige Ähnlichkeiten zu Ergebnissen bisheriger Untersuchungen zum Umgang von Lehrkräften mit Sprache (in Leistungssituationen) im Physikunterricht aufweisen und daher ein Generalisierungspotenzial in sich bergen (vgl. Abschnitt 7.2.1). Ergo: Es scheint gewinnbringend in der Lehreraus- und -weiterbildung zu thematisieren, dass der gegenwärtige Stand der Forschung in besonderem Maße dafür spricht, dass Physiklehrkräfte in ihrer täglichen Berufspraxis bei Korrektur der Textprodukte von Schüler_innen...

... nicht nur fachlich-konzeptuelle Leistungen feststellen und -beurteilen, sondern auch sprachliche Leistungen der Schüler_innen.

... überwiegend die fachlich-konzeptuellen Merkmale der Textprodukte im Blick haben, Merkmale, die die sprachliche Realisierung eines Textprodukts betreffen, hingegen eher weniger.

... deren sprachliche Realisierung mit Beurteilungskriterien feststellen und beurteilen, die unter anderem die Ausführlichkeit und gleichzeitig die Knappheit der Textprodukte adressieren.

... dazu tendieren sprachliche Schülerleistungen in einer defizitorientierten Art und Weise festzustellen und zu beurteilen.

... die Bewertungslogik zeigen, sprachliche Schülerleistung konfundiert mit fachlich-konzeptueller Leistung und umgekehrt festzustellen und zu beurteilen.

Ein Aus- und Weiterbildungsangebot, das für sich in Anspruch nimmt, einen umfassenden Beitrag zur Grundbildung von Physiklehrkräften bzgl. schulischer Leistungsfeststellung und -beurteilung zu leisten, darf sich allerdings nicht auf ein „bloßes Informieren“ über den gegenwärtigen Stand der Forschung zur von Lehrkräften praktizierten fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung beschränken. Erstens könnte hierdurch das Missverständnis aufkommen, dass professionelles Lehrerhandeln im Kontext schulischer Leistungsfeststellung und -beurteilung durch ein unhinterfragtes Agieren auf einer Art und Weise „wie sie im Lehrbuch steht“ auszeichnet (vgl. theoretische Ausprägungen der Assessment Literacy einer Lehrkraft in Tabelle 2.6). Zweitens würden derart konzipierte Angebote nicht der berufsbezogenen Überzeugung von Lehrkräften entgegenkommen, dass in der Lehrerbildung die erfolgreiche Integration von Hintergrundwissen zur Leistungsfeststellung und -beurteilung in das tägliche Unterrichtsgeschehen vermittelt werden sollte (vgl. Unterabschnitt 2.1.3.2). Drittens blieben Problematiken unthematisiert, die insbesondere eine Tendenz zur Defizitorientierung bei der Feststellung und Beurteilung sprachlicher Schülerleistungen (z. B. ein Übersehen der Lernfortschritte von Schüler_innen bezüglich eines schriftlichen Sprachgebrauchs, der den Normen der Schule gerecht wird; vgl. Unterkapitel 3.1), sowie eine Konfundierung fachlich-konzeptueller und

sprachlicher Schülerleistungen (z. B. ein Abwerten fachlich-konzeptuell richtiger Denkfiguren von Schüler_innen, anstatt diese wertzuschätzen) mit sich bringen. Dementsprechend bedarf es eines Aus- und Weiterbildungsangebots, das (angehenden) Lehrkräften evidenzbasiertes Wissen zu fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung im Physikunterricht bereitstellt, gleichzeitig aber auch...

... vergegenwärtigt, dass eine im Kontext von schulischer Leistungsfeststellung und -beurteilung professionell handelnde Lehrkraft nicht allein durch das Verfügen von evidenten Wissensbeständen charakterisiert wird. Vielmehr sollte im Sinne des Konzepts einer Assessment Litercy vermittelt werden, dass sich professionelles Lehrerhandeln beim Feststellen und Beurteilen von Schülerleistungen durch ein in verschiedenen Handlungsepisoden wiederkehrendes Kompromissfinden zwischen miteinander vereinbaren und/oder zueinander in Konflikt stehenden Faktoren auszeichnet (kompetenztheoretisch gedachtes Wissen und Können, berufsbezogene Überzeugungen und Bezugsnormorientierung, Kontextbedingungen).

... Gelegenheiten gibt, dieses Hintergrundwissen bezogen auf eigene berufsbezogene Überzeugungen zur Feststellung und Beurteilung von Schülerleistung zu reflektieren, sowie ferner Möglichkeiten zur Integration dieses Hintergrundwissens in die eigene berufliche Praxis eröffnet und auslotet. Lehrkräfte sollten dabei angeregt werden zu reflektieren, welche Zwecke bestimmte fachlich-konzeptuelle und sprachliche Leistungsfeststellungen und -beurteilungen in ihrem eigenen Physikunterricht (beispielsweise im Rahmen einer konkreten Klassenarbeit) erfüllen und erfüllen sollen, wie das im Rahmen des Aus- bzw. Weiterbildungsangebots vermittelte Hintergrundwissen für diese Zwecke angewendet werden kann und inwieweit dieses Hintergrundwissen mit eigenen berufsbezogenen Überzeugungen vereinbar ist und/oder mit diesen konfligiert.

... mögliche Begleiterscheinungen einer defizitorientierten Feststellung und Beurteilung sprachlicher Schülerleistungen, sowie einer Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen vor Augen führt und problematisiert. Ferner sollte thematisiert werden, welche Möglichkeiten es gibt, derartigen Begleiterscheinungen in der beruflichen Praxis entgegenzuwirken. Beispiele hierfür sind das Zurückgreifen auf eine Bewertungslogik des Trennens fachlich-konzeptueller und sprachlicher Schülerleistungen, wordurch einer Konfundierung dementsprechender Teilleistungsurteile entgegengewirkt wird (vgl. Abschnitt 6.5.2) oder eine Zuhilfenahme von Werkzeugen wie dem im Rahmen der Entwicklungsstudie entwickelten Kriterienraster zur Unterscheidung verschiedener Schülerlösungstexte bezüglich der Qualität ihrer sprachlichen Realisierung (vgl. Abschnitt 5.3.2), um hierdurch sowohl sprachliche Defizite als auch Fähigkeiten von Schüler_innen differenziert feststellen und beurteilen zu können.

Welche Lehr-Lernmethoden wirksame Aus- und Weiterbildungsangebote zum Thema fachlich-konzeptuelle und sprachliche Leistungsfeststellung und -beurteilung charakterisieren, ist kein Bestandteil des Erkenntnisinteresses, dem im empirischen Teil der vorliegenden Arbeit nachgegangen wurde. Die Befunde der empirische Hauptstudie können daher auch keine Antwort darauf liefern, wie ein Aus- und Weiterbildungsangebot für (angehende)

Physiklehrkräfte methodisch auszugestalten ist, das den eben aufgeführten inhaltlichen Anforderungen entspricht. Wie in Teil I der vorliegenden Arbeit dargestellt, finden sich in der Literatur allerdings bereits einzelne, hinsichtlich ihrer Wirksamkeit evaluierte Praxisbeispiele zu Aus- und Weiterbildungsmaßnahmen für Naturwissenschaftslehrkräfte zum Thema schulische Leistungsfeststellung und -beurteilung im Allgemeinen (z. B. M. A. Siegel & Wissehr, 2011), sowie zum Umgang mit Sprache (in Leistungssituationen) im naturwissenschaftlichen Fachunterricht (z. B. Lyon, 2013c; Tajmel, 2017b, S. 276 u. f.). Es scheint lohnenswert kritisch zu überprüfen, inwieweit die Lehr-Lernmethoden dieser Praxisbeispiele sich für die Entwicklung von konkreten Aus- und Weiterbildungsangeboten für Physiklehrkräfte, die den eben aufgeführten inhaltlichen Anforderungen genügen sollen, eignen. In jedem Fall sollten Aus- und Weiterbildungsangebote, die Befunde der empirischen Hauptstudie inhaltlich aufgreifen, hinsichtlich ihrer Wirksamkeit untersucht werden. Eine *methodisch evidenzbasierte*[157] Ausgestaltung derartiger Lehrerbildungsmaßnahmen stellt eine weiterführende Frage dar, der sich zukünftige physikdidaktische Forschung widmen sollte (vgl. Abschnitt 6.5.1).

Prägnant zusammengefasst ist damit festzuhalten: Hinsichtlich der Aus- und Weiterbildung von Physiklehrkräften sollten die Befunde der empirischen Hauptstudie als Grundsteine für die inhaltliche Ausgestaltung von Aus- und Weiterbildungsangeboten für (angehende) Lehrkräfte zu fachlich-konzeptueller und sprachlicher Leistungsfeststellung und -beurteilung im Physikunterricht aufgefasst werden, deren Wirksamkeit es zweifelsohne empirisch abzusichern gilt. Hierdurch könnte ein Beitrag für eine sowohl inhaltlich, wie auch methodisch evidenzbasierte Lehrerbildung geleistet werden.

[157] *Methodisch evidenzbasierte* Aus- und/oder Weiterbildungsangebote für Lehrkräfte zeichnen sich dadurch aus, dass diese „sich in der Ausgestaltung [...] der makrostrukturellen Didaktisierung von Lehrerbildungsgängen und in ihren mikrodidaktischen Inszenierungsformen an der „Evidenz“ zur Frage ausrichte[n], wie wirksame Lehrerbildung aussieht“ (Neuweg, 2018, S. 63-64).

Schlussbemerkung

> „Als ein sehr geeignetes Mittel der Wiederholung haben sich mir seit einer langen Reihe von Jahren auch schriftliche Klassenarbeiten besonderer Art bewährt[.] [...] Ich stelle [...] eine Anzahl Fragen, in der Regel sechs, die sich auf ein bestimmt begrenztes, nicht zu kleines Gebiet des durchgenommenen Pensums beziehen. [...] Die Antworten bestehen meist in wenigen kurzen Sätzen[.] [...] Man wird auch im Zensieren nicht zu streng sein dürfen, hat aber andererseits die Möglichkeit, alle Schüler mit gleichem Maße zu messen. In der Regel stellt sich bei der schriftlichen Bearbeitung heraus, daß ein oder der andere Gegenstand von einer Zahl von Schülern noch mangelhaft aufgefaßt worden ist, oder daß sich mit einer etwas abweichend vom Besprochenen abgefaßten Frage nicht alle haben zurechtfinden können. [...] Auch zeigt sich eine starke Verschiedenheit der Begabung darin, daß manche Schüler [...] mehr zur schriftlichen Form der Darstellung befähigt sind.“ (Poske, 1915, S. 84-87)

Obiges Zitat stammt aus der *Didaktik des Physikalischen Unterrichts* von Poske (1915). Dieses Lehrwerk stellt einen der ersten ernstzunehmenden Versuche zur Theoretisierung von Physikunterricht dar (vgl. Mikelskis, 1996, S. 26). Das Zitat illustriert, dass die Feststellung und Beurteilung schriftlicher Schülerleistungen im Allgemeinen, sowie speziell der Umgang von Physiklehrer_innen mit Sprache im Rahmen schulischer Leistungsfeststellungen und -beurteilungen Themen sind, mit denen sich die Physikdidaktik bereits in ihren Ursprüngen auseinandergesetzt hat. Im theoretischen Teil der vorliegenden Arbeit wurde der bisherige Forschungsstand über Lehrerwissen und -können zu schulischer Leistungsfeststellung und -beurteilung mit Schwerpunkt auf dem bisherigen Erkenntnisstand der naturwisssenschaftsdidaktischen, speziell der physikdidaktischen Forschung dargestellt. Dieser Forschungsstand wurde in einem Rahmenkonzept, das professionelles Lehrerhandeln im Kontext von schulischer Leistungsfeststellung und -beurteilung beschreibt, zusammengeführt (Assessment Literacy von Lehrkräften). Ferner erfolgte eine ausführliche Erörterung der Bedeutung von Sprache im Rahmen von Leistungsfeststellung und -beurteilung im Physikunterricht. Dabei zeigte sich, dass insbesondere alltägliche Leistungsfeststellung und -beurteilung durch Physiklehrkräfte, sowie die Rolle von Sprache in diesem Zusammenhang Gegenstände darstellen, mit dem sich die empirische erziehungswissenschaftliche Forschung bislang nur sehr wenig auseinandergesetzt hat.

Ziel des empirischen Teils der vorliegenden Arbeit war daher einen Beitrag zum Schließen dieser Forschungslücke zu leisten. Hierzu wurde einem deskriptiv-explorativen Studiendesign gefolgt und in einer empirischen Hauptstudie 21 im Schuldienst aktive Physiklehrkräfte befragt. Zentrale Teilbefunde der empirischen Hauptstudie ließen sich zu insgesamt sechs Ressourcen verdichten, die den Studienteilnehmer_innen zur Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen im Erhebungssetting dienten. Unter anderem zählen hierzu ein facettenreiches und überwiegend angemesse-

nes Wissen und Können zu (fachspezifischer) Leistungsfeststellung und -beurteilung, ein Handeln auch auf Basis von berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung, sowie eine defizit- und/oder fähigkeitsorientierte Herangehensweise bei der Feststellung und Beurteilung fachlich-konzeptueller und sprachlicher Schülerleistungen. Des Weiteren konnten zahlreiche einander komplementäre Teilbefunde identifiziert werden, die alles in allem dafür sprechen, dass die Teilnehmer_innen in der Erhebungssituation fachlich-konzeptuelle und sprachliche Schülerleistungen auf einem moderaten Niveau miteinander konfundierten. Vor allem dieser Befund eröffnet die weiterführende Frage, welche Aus- und Weiterbildungsmaßnahmen für (angehende) Physiklehrkräfte ergriffen werden können und sollten, um einer solchen Konfundierung entgegen zu wirken.

Summa summarum stellt die vorliegende Arbeit, sowohl auf theoretischer, wie auch empirischer Ebene, mannigfache Einsichten und Erkenntnisse über die fachlich-konzeptuelle und sprachliche Leistungsfeststellung und -beurteilung durch im Schuldienst aktive Physiklehrkräfte zur Verfügung. Schon aufgrund des explorativen Charakters der Hauptstudie, liefert der empirische Teil der vorliegenden Arbeit weder verallgemeinerte, noch allumfassende Ergebnisse. Vielmehr sind die gewonnenen Befunde als Motivation für zukünftige Untersuchungen zu verstehen. Insbesondere sollte in zukünftigen Studien der Frage nachgegangen werden, inwieweit zu den Erkenntnissen des empirischen Teils der vorliegenden Arbeit konvergierende, komplementäre oder auch divergierende Befunde gewonnen werden können.

Literaturverzeichnis

Abell, S. K. & Siegel, M. A. (2011). Assessment Literacy: What Science Teachers Need to Know and Be Able to Do. In D. Corrigan, J. Dillon, & R. Gunstone (Hrsg.), *The Professional Knowledge Base of Science Teaching* (S. 205–221). Heidelberg: Springer.

Abels, S. (2011). *LehrerInnen als „Reflective Practitioner". Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Abs, H. J. (2006). Zur Bildung diagnostischer Kompetenz in der zweiten Phase der Lehrerbildung. In C. Allemann-Ghionda & E. Terhart (Hrsg.), *Zeitschrift für Pädagogik 51. Beiheft. Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern: Ausbildung und Beruf* (S. 217–234). Weinheim: Beltz.

Abs, H. J., Peter, D., Gerlach-Jahn, A., & Klieme, E. (2009). *Pädagogische Entwicklungsbilanz an Studienseminaren (PEB-Sem). Auswahl und statistische Analyse der Erhebungsinstrumente.* Frankfurt am Main: Deutsches Institut für Internationale Pädagogische Forschung.

Adamzik, K. (1998). Fachsprachen als Varietäten. In L. Hoffmann, K. Hartwig, & H. E. Wiegand (Hrsg.), *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. 1. Halbband* (S. 181–189). Berlin: De Gruyter.

Airey, J. (2012). "I don't teach language". The linguistic attitudes of physics lectures in Sweden. *AILA Review*, *25*(1), 64–79.

Altheide, D. L. (1996). *Qualitative media analysis.* Thousand Oaks, Kalifornien, USA: Sage Publications Ltd.

Arras, U. (2007). *Wie beurteilen wir Leistung in der Fremdsprache? Strategien und Prozess bei der Beurteilung schriftlicher Leistungen in der Fremdsprache am Beispiel der Prüfung Test Deutsch als Fremdsprache (TestDaF).* Tübingen: Narr Francke Attempto Verlag.

Arras, U. (2013). Introspektive Verfahren in der Sprachtestforschung. In K. Aguado, K. Schramm, & L. Heine (Hrsg.), *Introspektive Verfahren und Qualitative Inhaltsanalyse in der Fremdsprachenforschung* (S. 74–91). Frankfurt am Main: Peter Lang.

Ashton, P. T. (2015). Historical overview and theoretical perspectives of research on teachers' beliefs. In H. Fives & M. G. Gill (Hrsg.), *International Handbook of Research on Teachers' Beliefs* (S. 31–47). New York, New York, USA: Routledge.

Bach, S. (1984). Systematische und empirische Untersuchung über das Verhältnis von Umgangssprache und Fachsprache im gymnasialen Physikunterricht. Dissertation. Universität Hamburg.

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' Beliefs About Assessment. In H. Fives & M. G. Gill (Hrsg.), *International Handbook of Research on Teachers' Beliefs* (S. 284–300). New York, New York, USA: Routledge.

Bates, C. & Nettelback, T. (2001). Primary School Teachers' Judgements of Reading Achievement. *Educational Psychology*, *21*(2), 177–187.

Baumann, K.-D. (1992). *Integrative Fachtextlinguistik*. Tübingen: Gunter Narr Verlag.

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., ... Tsai, Y.-M. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*(4), 469–520.

Bayerische Staatsministerium für Unterricht und Kultus (2007). Schulordnung für die Gymnasien in Bayern vom 23. Januar 2007. München.

Behörde für Schule und Berufsbildung Hamburg (2011). Bildungsplan. Gymnasium. Sekundarstufe I. Physik. Hamburg.

Behörde für Schule und Berufsbildung Hamburg (2014). Bildungsplan. Stadtteilschule. Jahrgangsstufe 7-11. Physik. Hamburg.

Bender, R., St., L., & Ziegler, A. (2002). Multiples Testen. *Deutsche Medizinische Wochenschrift*, *127*, T4–T7.

Benischek, I. (2006). *Leistungsbeurteilung im österreichischen Schulsystem*. Wien, Österreich: Lit.

Bergeler, E. (2009). Lernen durch eigenständiges Schreiben von sachbezogenen Texten im Physikunterricht. Eine Feldstudie zum Schreiben im Physikunterricht am Beispiel der Akustik. Dissertation. Technische Universität Dresden.

Beutel, S.-I. & Vollstädt, W. (2000). Leistung ermitteln und bewerten. Eine Einführung. In S.-I. Beutel & W. Vollstädt (Hrsg.), *Leistung ermitteln und bewerten* (S. 7–14). Hamburg: Bergmann + Helbig.

Biber, D. (2009). Multi-dimensional approaches. In A. Lüdeling & M. Kytö (Hrsg.), *Corpus Linguistics. An International Handbook. Volume 2* (S. 822–855). Berlin: De Gruyter.

Biggs, J. B. & Collis, K. F. (1982). *Evaluating the Quality of Learning. The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York, New York, USA: Academic Press Inc.

Bird, E. & Welford, G. (1995). The effect of language on the performance of second-language students in science examinations. *International Journal of Science Education*, *17*(3), 389–397.

Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, *49*(3), 219–224.

Birt, L., Scott, S., Cavers, D., Campbell, C., & Walter, F. (2016). Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? *Qualitative Health Research*, *26*(13), 1802–1811.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. Competence Viewed as a Continuum. *Zeitschrift für Psychologie*, *223*(1), 3–13.

Blömeke, S., Herzig, B., & Tulodziecki, G. (2007). *Gestaltung von Schule. Eine Einführung in Schultheorie und Schulentwicklung*. Bad Heilbrunn: Klinkhardt.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. New York, New York, USA: McGraw-Hill.

Böhmer, I., Gräsel, C., Krolak-Schwerdt, S., Höstermann, T., & Glock, S. (2017). Teachers' school tracking decisions. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Hrsg.), *Competence Assessment in Education. Research, Models and Instruments* (S. 131–147). Cham, Schweiz: Springer Nature.

Böhmer, M., Englich, B., & Böhmer, I. (2017). Schülerbeurteilungen aus der Perspektive dualer Prozessmodelle der sozialen Urteilsbildung. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 50–54). Münster: Waxmann.

Böhm-Kasper, O. (2004). *Schulische Belastung und Beanspruchung. Eine Untersuchung von Schülern und Lehrern am Gymnasium*. Münster: Waxmann.

Bortz, J. & Lienert, G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung. Leitfaden für die verteilungsfreie Analyse kleiner Stichproben* (3. Aufl.). Heidelberg: Springer.

Bortz, J., Lienert, G. A., & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik* (3. Aufl.). Heidelberg: Springer.

Bos, W., Pietsch, M., Gröhlich, C., & Janke, N. (2006). Ein Belastungsindex für Schulen als Grundlage der Ressourcenzuweisung am Beispiel von KESS 4. Versuch einer Klassifizierung von Schultypen. In W. Bos, H. G. Holtappels, H. Pfeiffer, H.-G. Rolff, & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung. Daten, Beispiele und Perspektiven. Band 14* (S. 149–160). Weinheim: Juventa.

Bos, W., Voss, A., & Goy, M. (2009). Leistung und Leistungsmessung. In S. Andersen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee, & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 563–576). Weinheim: Beltz.

Bourdieu, P. (1990). *Was heisst Sprechen?. Die Ökonomie des sprachlichen Tausches*. Wien, Österreich: Braumüller.

Bourdieu, P. (1993). *Soziologische Fragen*. Frankfurt am Main: Suhrkamp.

Boyatzis, R. E. (1998). *Transforming qualitative information. Thematic analysis and code development*. Thousand Oaks, Kalifornien, USA: Sage Publications Ltd.

Braaten, M. & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, *95*(4), 639–669.

Brennan, R. L. & Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, *41*, 687–699.

Briggs, D. (1970). The influence of handwriting on assessments. *Educational Research*, *13*(1), 50–55.

Bromme, R. (1992). *Der Lehrer als Experte. Zur Psychologie des professionellen Wissens*. Bern, Schweiz: Verlag Hans Huber.

Brookhart, S. M. (2003). Developing Measurement Theory for Classroom Assessment Purposes and Uses. *Educational Measurement*, *22*(4), 5–12.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., ... Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research*, *86*(4), 803–848.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education*, *11*(3), 301–318.

Brunner, M., Anders, Y., Hachfeld, A., & Krauss, S. (2011). Diagnostische Fähigkeiten von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 215–234). Münster: Waxmann.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2. Aufl.). Berkley und Los Angeles, Kalifornien, USA: University of California Press.

Bürgerschaft der Freien und Hansestadt Hamburg (2013). Schriftliche kleine Anfrage des Abgeordneten Robert Heinemann (CDU) vom 28.02.13 und Antwort des Senats. Drucksache 20/7094. Neufassung vom 03.05.2013. Hamburg.

Bürgerschaft der Freien und Hansestadt Hamburg (2017). Große Anfrage der Abgeordneten Karin Prien, Stephan Gamm, Philipp Heißner, Joachim Lenders, Richard Seelmaecker (CDU) und Fraktion vom 28.02.17 und Antwort des Senats. Drucksache 21/8179. Neufassung vom 28.03.2017. Hamburg.

Burke, P. J. (1991). Identity Processes and Social Stress. *American Sociological Review*, *56*(6), 836–849.

Bußmann, H. (2008). *Lexikon der Sprachwissenschaft* (4. Aufl.). Stuttgart: Alfred Kröner Verlag.

Buxton, C. A., Salinas, A., Mahotiere, M., Lee, O., & Secada, W. G. (2013). Leveraging cultural resources through teacher pedagogical reasoning: Elementary grade teachers analyze second language learners' science problem solving. *Teaching and Teacher Education*, *32*(1), 31–42.

Byers, J. L. & Evans, T. E. (1980). *Using a lens-model analysis to identify the factors in teacher judgment*. East Lansing, Michigan, USA: The Institute for Research on Teaching. Michigan State University.

Capelle, J. (1969). Sociological Aspects of Examinations. In J. A. Lauwerys & D. G. Scanlon (Hrsg.), *The World Year Book of Education 1969. Examinations* (S. 258–267). London, Großbritannien: Evan Brothers Limited.

Cassels, J. R. T. & Johnstone, A. H. (1984). The Effect of Language on Student Performance on Multiple Choice Tests in Chemistry. *Journal of Chemical Education*, *61*(7), 613–615.

Cauet, E. (2016). *Testen wir relevantes Wissen? Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*. Berlin: Logos.

Chi, M. T. H. (1997). Quantifying Qualitative Analyses of Verbal Data: A Practical Guide. *Journal of the Learning Sciences*, *6*(3), 271–315.

Cohen, A. (1996). Verbal Reports as a Source of Insights into Second Language Learner Strategies. *Applied Language Learning 7*, *7*(1 & 2), 5–24.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. Aufl.). Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates Publishers.

Coladarci, T. (1986). Accuracy of Teacher Judgements of Student Responses to Standardized Test Items. *Journal of Educational Psychology*, *78*(2), 141–146.

Comber, L. C. & Keeves, J. P. (1973). *Science Education in Nineteen Countries. An Empirical Study*. Stockholm: Almqvist & Wiksell.

Cooksey, R. W., Freebody, P., & Davidson, G. R. (1986). Teachers' Predictions of Children's Early Reading Achievement: An Application of Social Judgment Theory. *American Educational Research Journal*, *23*(1), 41–64.

Corbin, J. & Strauss, A. (1990). Grounded Theory Research: Procedures, Canons, and Evaluative Criteria. *Qualitative Sociology*, *13*(1), 3–21.

Corder, G. W. & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticans. A step-by-step approach*. Hoboken, New Jersey, USA: Wiley.

Creswell, J. W. & Plano Clark, V. L. (2007). *Designing and Conducting Mixed Methods Research*. London, Großbritannien: Sage Publications Ltd.

Criblez, L. (2001). Die Wirksamkeit der Lehrerbildungssysteme in der Schweiz: Forschungsfeld und Forschungskonzept. In F. Oser & J. Oelkers (Hrsg.), *Die Wirksamkeit der Lehrerbildungssysteme. Von der Allrounderbildung zur Ausbildung professioneller Standards* (S. 99–139). Zürich, Schweiz: Verlag Rüegger.

Cronbach, L. J. (1955). Processes affecting scores on "understanding others" and "assumed similarity". *Psychological Bulletin*, *52*(3), 177–193.

Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy. Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, *28*, 43–56.

Cummins, J. (1979). Cognitive-academic language proficiency. Linguistic Interdependence, the Optimum Age Question and some other Matters. *Working Papers on Bilingualism*, *19*, 198–205.

Cummins, J. (1981). The Role of Primary Language Development in Promoting Educational Success for Language Minority Students. In California State Department of Education (Hrsg.), *Schooling and language minority students* (S. 3–49). Los Angeles, Kalifornien, USA: Evaluation, Dissemination & Assessment Center.

Cummins, J. (2000). BICS and CALP. In M. Byram (Hrsg.), *Encyclopedia of Language Teaching and Learning* (S. 76–79). New York, New York, USA: Routledge.

Degen, M. (2015). Codierer-Effekte in Inhaltsanalysen - ein vernachlässigtes Forschungsfeld. In W. Wirth, K. Sommer, M. Wettsein, & J. Matthes (Hrsg.), *Qualitätskriterien in der Inhaltsanalyse* (S. 78–95). Köln: Herbert von Halem Verlag.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: a review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251–272.

Demaray, M. K. & Elliot, S. N. (1998). Teachers' Judgments of Students' Academic Functioning: A Comparison of Actual and Predicted Performances. *School Psychology Quarterly*, *13*(1), 8–24.

Demidow, I. (1999). Fachlernen in der Zweitsprache Deutsch: Wie zweisprachige Schüler(innen) Physik verstehen. *Zeitschrift für Didaktik der Naturwissenschaften*, *5*(2), 15–32.

Denzin, N. K. (1970). *The Research Act. A Theoretical Introduction to Sociological Methods*. Chicago, Illinois, USA: Aldine Publishing Company.

Deppner, J. (1989). *Fachsprache der Chemie in der Schule*. Heidelberg: Julius Groos Verlag.

Diebold, T. J. & Waldron, M. B. (1988). Designing Instructional Formats: The Effects of Verbal and Pictorial Components on Hearing-Impaired Students' Comprehension of Science Concepts. *American Annals of the Deaf*, *133*(1), 30–35.

Diekmann, A. (2013). *Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen*. Reinbek bei Hamburg: Rowohlt.

Dori, Y. J. & Avargil, S. (2016). Teachers' Understanding of Assessment. In R. Gunstone (Hrsg.), *Encyclopedia of Science Education. Volume 2. L-Z* (S. 1033–1035). Dordrecht, Niederlande: Springer.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Heidelberg: Springer.

Drach, E. (1928). Bildungssprache. In H. Schwartz (Hrsg.), *Pädagogisches Lexikon. Erster Band* (S. 665–674). Bielefeld: Verlag von Velhagen & Klasing.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science. Research into children's ideas*. New York, New York, USA: Routledge.

Duit, R. (1986a). *Der Energiebegriff im Physikunterricht*. Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

Duit, R. (1986b). Wärmevorstellungen. *Naturwissenschaften im Unterricht Physik/Chemie*, *34*, 30–33.

Duit, R. & Häußler, P. (1997). Unterricht vielfältig bewerten. Überlegungen und Vorschläge für die Unterrichtsbeurteilung und die Lernberatung. *Naturwissenschaften im Unterricht Physik*, *38*(8), 4–9.

Eckhardt, A. G. (2008). *Sprache als Barriere für den schulischen Erfolg. Potentielle Schwierigkeiten beim Erwerb schulbezogener Sprache für Kinder mit Migrationshintergrund*. Münster: Waxmann.

Einhaus, E. (2007). *Schülerkompetenzen im Bereich Wärmelehre. Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*. Berlin: Logos.

Erfurt, J. (1996). Sprachwandel und Schriftlichkeit. In H. Günther & O. Ludwig (Hrsg.), *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung. 2. Halbband* (S. 1387–1404). Berlin: De Gruyter.

Ergönenç, J., Neumann, K., & Fischer, H. E. (2014). The Impact of Pedagogical Content Knowledge on Cognitive Activation and Student Learning. In H. E. Fischer, P. La-

budde, K. Neumann, & J. Viiri (Hrsg.), *Quality of Instruction in Physics. Comparing Finland, Switzerland and Germany* (S. 145–160). Münster: Waxmann.

Erickson, G. & Tiberghien, A. (1985). Heat and Temperature. In R. Driver, G. Tiberghien, & A. Tiberghien (Hrsg.), *Children's Ideas in Science* (S. 52–84). Buckingham, Großbritannien: Open University Press.

Ericsson, K. A. & Simon, H. A. (1985). *Protocol Analysis. Verbal reports as data*. Cambridge, Massachusetts, USA: The MIT Press.

Fast, L. & Klein, H. (1998). *Notengebung - Beispiel Technikunterricht*. Bad Heilbrunn: Klinkhardt.

Fehm, L. & Fydrich, T. (2011). *Prüfungsangst*. Göttingen: Hogrefe.

Feilke, H. (2012). Schulsprache - Wie Schule Sprache macht. In G. Susanne, W. Imo, D. Meer, & J. G. Schneider (Hrsg.), *Kommunikation und Öffentlichkeit. Sprachwissenschaftliche Potenziale zwischen Empirie und Norm* (S. 151–175). Berlin: De Gruyter.

Fend, H. (1980). *Theorie der Schule*. München: Urban und Schwarzenberg.

Feser, M. S. & Höttecke, D. (2017a). How physics teachers assess students' texts in teacher-made tests. Paper presented at the ESERA conference 21st-25th August 2017 in Dublin. http://keynote.conference-services.net/resources/444/5233/pdf/ESERA2017_0099_paper.pdf. [Letzter Abruf: 5. November 2019].

Feser, M. S. & Höttecke, D. (2017b). Klassenarbeiten kriteriengeleitet korrigieren – Wie beurteile ich eine Schülererklärung? *Unterricht Physik*, *158*, 15–18.

Feser, M. S. & Höttecke, D. (2017c). Wie Physiklehrkräfte Schülertexte beurteilen – Instrumentenentwicklung. In C. Maurer (Hrsg.), *Implementation fachdidaktischer Innovation im Spiegel von Forschung und Praxis. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Zürich 2016* (S. 123–126).

Feser, M. S. & Höttecke, D. (2018). Physiklehrkräfte beurteilen Schülertexte – Eine Explorationsstudie. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht-normative und empirische Dimensionen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Regensburg 2017* (S. 218–221).

Feser, M. S. & Höttecke, D. (2019). Exploring physics teachers' assessment practices - a study about the role of language. Paper presented at the ESERA conference 26st-30th August 2019 in Bologna.

Feser, M. S. & Höttecke, D. (im Druck). Physiklehrkräfte beurteilen Schülertexte – Eine Explorationsstudie. In H. Sebastian (Hrsg.), *Naturwissenschaftliche Kompetenzen in der Gesellschaft von morgen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Wien 2019*.

Feser, M. S., Höttecke, D., & Ehmke, T. (2016). Testitems zur qualitativen Untersuchung der Ressourcen von Physiklehrkräften beim Bewerten schriftlicher Schülerleistungen. *PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung*, *o. V.*(o. N.), o. S.

Fiedler, K. & von Sydow, M. (2015). Heuristics and biases: Beyond Tversky and Kahneman's (1974) judgment under uncertainty. In M. W. Eysenck & D. Groome (Hrsg.),

Cognitive Psychology. Revisiting the Classical Studies (S. 146–161). London, Großbritannien: Sage Publications Ltd.

Fischer, R. & Malle, G. (1989). *Mensch und Mathematik. Eine Einführung in didaktisches Denken und Handeln.* Mannheim: BI Wissenschaftsverlag.

Fives, H. & Buehl, M. M. (2012). Spring cleaning fot the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, & T. Urdan (Hrsg.), *APA Educational Psychology Handbook. Volume 2. Individual Differences and Cultural and Contextual Factors* (S. 471–499). Washington, DC, USA: American Psychological Association.

Flick, U. (2004). Triangulation in Qualitative Research. In U. Flick, E. von Kardoff, & I. Steinke (Hrsg.), *A Companion to Qualitative Research* (S. 178–183). London, Großbritannien: Sage Publications Ltd.

Flick, U. (2011). *Triangulation. Eine Einführung* (3. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Fluck, H.-R. (1997). *Fachdeutsch in Naturwissenschaft und Technik. Einführung in die Fachsprachen und die Didaktik/Methodik des fachorientierten Fremdsprachenunterrichts (Deutsch als Fremdsprache).* Heidelberg: Julius Groos Verlag.

Follman, J. (1991). Teachers' Estimates of Pupils' IQs and Pupils' Tested IQs. *Psychological Reports*, *69*(1), 350.

Förster, N. & Böhmer, I. (2017). Das Linsenmodell - Grundlagen und exemplarische Anwendungen in der pädagogisch-psychologischen Diagnostik. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 46–50). Münster: Waxmann.

Förtsch, C., Sommerhoff, D., Fischer, F., Fischer, M. R., Girwidz, R., Obersteiner, A., ... Neuhaus, B. J. (2018). Systematizing Professional Knowledge of Medical Doctors and Teachers: Development of an Interdisciplinary Framework in the Context of Diagnostic Competences. *education sciences*, *8*(4), o. S.

Friege, G. (2017). Leistungsbewertung - eine ungeliebte Aufgabe. Formen, Probleme und Chancen. *Unterricht Physik*, *158*, 2–7.

Fulcher, G. (2012). Assessment Literacy for the Language Classroom. *Language Assessment Quarterly*, *9*(2), 113–132.

Füller, K. (1975). *Funktionen und Formen von Prüfungen.* Neuburgweier: Schindele.

Fuß, S. & Karbach, U. (2014). *Grundlagen der Transkription. Eine praktische Einführung.* Opladen: Verlag Barbara Budrich.

Geddis, A. N., Onslow, B., Beynon, C., & Oesch, J. (1993). Transforming content knowledge: Learning to teach about isotopes. *Science Education*, *77*(6), 575–591.

Georg, A. L. (1959). Quantitative and qualitative approaches to content analysis. In I. de Sola Pool (Hrsg.), *Trends in content analysis* (S. 7–32). Urbana, Illinois, USA: University of Illinois Press.

Gess-Newsome, J. (2015). A model of teacher professional knowledge and skill including PCK. Results of the thinking form the PCK Summit. In A. Berry, P. Friedrichsen, & J. Loughran (Hrsg.), *Re-examining Pedagogical Content Knowledge in Science Education* (S. 28–42). London, Großbritannien: Routledge.

Gilhooly, K. & Green, C. (2002). Protocol analysis: theoretical background. In J. T. E. Richardson (Hrsg.), *Handbook of Qualitative Research Methods for Psychology and the Social Sciences* (Neuauflage, S. 43–54). Oxford, Großbritannien: Blackwell Publishing.

Gläser, J. & Laudel, G. (2010). *Experteninterviews und qualitative Inhaltsanalyse. als Instrument rekonstruierender Untersuchungen* (4. Aufl.). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gläser-Zikuda, M. (2010). Leistungsvoraussetzungen diagnostizieren und Fördermaßnahmen realisieren. In T. Bohl, C. Schelle, & W. Helsper (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 369–376). Bad Heilbrunn: Klinkhardt.

Gogolin, I. (2007). Herausforderung Bildungssprache. Textkompetenz aus der Perspektive interkultureller Bildungsforschung. In K.-R. Bausch, E. Burwitz-Melzer, F. G. Königs, & H.-J. Krumm (Hrsg.), *Textkompetenzen. Arbeitspapiere der 27. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts* (S. 73–80). Tübingen: Narr Francke Attempto Verlag.

Gogolin, I. (2008). *Der monolinguale Habitus der multilingualen Schule*. Münster: Waxmann.

Gogolin, I. (2009). Zweisprachigkeit und die Entwicklung bildungssprachlicher Fähigkeiten. In I. Gogolin & U. Neumann (Hrsg.), *Streitfall Zweisprachigkeit - The Bilingual Controversy* (S. 263–290). Wiesbaden: VS Verlag für Sozialwissenschaften.

Gogolin, I. & Duarte, J. (2016). Bildungssprache. In J. Kilian, B. Brouër, & D. Lüttenberg (Hrsg.), *Handbuch der Sprache in der Bildung* (S. 478–499). Berlin: De Gruyter.

Gogolin, I. & Lange, I. (2011). Bindungssprache und Durchgängige Sprachbildung. In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Mehrsprachigkeit* (S. 107–127). Wiesbaden: Springer VS.

Göpferich, S. (1995). *Textsorten in Naturwissenschaften und Technik. Pragmatische Typologie - Kontrastierung - Translation*. Tübingen: Gunter Narr Verlag.

Gramzow, Y., Riese, J., & Reinhold, P. (2013). Modellierung fachdidaktischen Wissens angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 7–30.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge, Großbritannien: Cambridge University Press.

Green, C. & Gilhooly, K. (2002). Protocol analysis: practical implementation. In J. T. E. Richardson (Hrsg.), *Handbook of Qualitative Research Methods for Psychology and the Social Sciences* (Neuauflage, S. 55–74). Oxford, Großbritannien: Blackwell Publishing.

Green, J. A. (1975). *Teacher-Made Tests. Second Edition*. New York, New York, USA: Harper & Row.

Habermas, J. (1978). Umgangssprache, Wissenschaftssprache, Bildungssprache. *Merkur*, *359*, 327–342.

Hackemann, T. (2017). Was erwarten Physiklehrkräfte von ihren Schüler*innen? Analyse von Erwartungshorizonten zu physikalischen Leistungsaufgaben. unveröffentlichte Masterarbeit. Universität Hamburg.

Häder, M. (2015). *Empirische Sozialforschung. Eine Einführung* (3. Aufl.). Wiesbaden: Springer VS.

Halliday, M. A. K. (1993). Some Grammatical Problems in Scientific English. In M. A. K. Halliday & J. R. Martin (Hrsg.), *Writing Science. Literacy and Discursive Power* (S. 76–94). New York, New York, USA: Routledge.

Halliday, M. A. K., McIntosh, A., & Strevens, P. (1970). *The linguistic sciences and language teaching* (5. Aufl.). London, Großbritannien: Longman.

Hallwirth, U. (2015). *Schule auf dem Weg. Von der Haupt- und Realschule zur Stadtteilschule - eine Fallanalyse*. Münster: Waxmann.

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. In M. F. Kaplan & S. Schwartz (Hrsg.), *Human judgment and decision processes* (S. 271–312). New York, New York, USA: Academic Press Inc.

Harding, L. & Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari & J. Banerjee (Hrsg.), *Handbook of Second Language Assessment* (S. 413–427). Berlin: De Gruyter.

Harren, I. (2011). Die verborgene Arbeit der Fachlehrer - sprachliche Anforderungen im Fachunterricht. *Osnabrücker Beiträge zur Sprachtheorie*, *80*, 101–123.

Harren, I. (2015). *Fachliche Inhalte sprachlich ausdrücken lernen. Sprachliche Hürden und interaktive Vermittlungsverfahren im naturwissenschaftlichen Unterrichtsgespräch in der Mittel- und Oberstufe*. Mannheim: Verlag für Gesprächsforschung.

Härtig, H., Bernholt, S., Prechtl, H., & Retelsdorf, J. (2015). Unterrichtssprache im Fachunterricht – Stand der Forschung und Forschungsperspektiven am Beispiel des Textverständnisses. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 55–67.

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 128–143). Heidelberg: Springer.

Hartmann, P. (1991). *Wunsch und Wirklichkeit. Theorie und Empirie sozialer Erwünschtheit*. Wiesbaden: Deutscher Universitäts-Verlag.

Hashweh, M. Z. (2005). Teacher pedagogical constructions: a reconfiguration of pedagogical content knowledge. *Teachers and Teaching: theory and practice*, *11*(3), 273–292.

Häußler, P., Bünder, W., Duit, R., Gräber, W., & Mayer, J. (1998). *Naturwissenschaftsdidaktische Forschung. Perspektiven für die Unterrichtspraxis*. Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

Heine, L. (2010). *Problem solving in a foreign language. A study in content and language integrated learning*. Berlin: De Gruyter Mouton.

Heine, L. (2013). Introspektive Verfahren in der Fremdsprachenforschung: State-of-the-Art und Desiderata. In K. Aguado, K. Schramm, & L. Heine (Hrsg.), *Introspektive Verfahren und Qualitative Inhaltsanalyse in der Fremdsprachenforschung* (S. 13–30). Frankfurt am Main: Peter Lang.

Heine, L. & Schramm, K. (2007). Lautes Denken in der Fremdsprachenforschung: Eine Handreichung für die empirische Praxis. In H. J. Vollmer (Hrsg.), *Synergieeffekte in der Fremdsprachenforschung. Empirische Zugänge, Probleme, Ergebnisse* (S. 167–206). Frankfurt am Main: Peter Lang.

Heine, L. & Schramm, K. (2016). Introspektion. In D. Caspari, F. Klippel, M. K. Legutke, & K. Schramm (Hrsg.), *Forschungsmethoden in der Fremdsprachendidaktik. Ein Handbuch* (S. 173–181). Tübingen: Narr Francke Attempto Verlag.

Heitmann, P. (2013). *Bewertungskompetenz im Rahmen naturwissenschaftlicher Problemlöseprozesse. Modellierung und Diagnose der Kompetenzen Bewertung und analytisches Problemlösen für das Fach Chemie*. Berlin: Logos.

Heitmann, P., Hecht, M., Schwanewedel, J., & Schipolowski, S. (2014). Students' Argumentative Writing Skills in Science and First-Language Education: Commonalities and differences. *International Journal of Science Education*, *36*(18), 3148–3170.

Heitzmann, A. (2013). Von der Alltagssprache zur Fachsprache gelangen. In P. Labudde (Hrsg.), *Fachdidaktik Naturwissenschaft. 1.-9. Schuljahr* (2. Aufl., S. 73–86). Bern, Schweiz: Haupt Verlag.

Heller, K. A. & Hany, E. A. (2001). Standardisierte Schulleistungsmessungen. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 87–101). Weinheim: Beltz.

Helmke, A., Hosenfeld, I., & Schrader, F.-W. (2004). Vergleichsarbeiten als Instrument zur Verbesserung der Diagnosekompetenz von Lehrkräften. In R. Arnold & C. Griese (Hrsg.), *Schulleitung und Schulentwicklung: Voraussetzungen, Bedingungen, Erfahrungen* (S. 119–143). Baltmannsweiler: Schneider-Verlag Hohengehren.

Helmke, A. & van Anken, M. A. G. (1995). The causal ordering of academic achievement and selfconcept of ability during elementary school. A longitudinal study. *Journal of Educational Psychology*, *87*(4), 624–637.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (S. 71–176). Göttingen: Hogrefe.

Herppich, S., Praetorius, A.-K., Hetmanek, A., Glogger-Frey, I., Ufer, S., Leutner, D., … Südkamp, A. (2017). Ein Arbeitsmodell für die empirische Erforschung der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 75–93). Münster: Waxmann.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2011). Does Teaching Experience Help? Differences in the Assessment of Tutees' Understanding Between Teacher Tutors and Student Tutors. In L. Carlson, C. Hoelscher, & T. F. Shipley (Hrsg.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (S. 78–83). Austin, Texas, USA: Cognitive Science Society.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2013). Does it Make a Difference? Investigating the Assessment Accuracy of Teacher Tutors and Student Tutors. *The Journal of Experimental Education*, *81*(2), 242–260.

Hess-Lüttich, E. W. B. (1998). Fachsprachen als Register. In L. Hoffmann, K. Hartwig, & H. E. Wiegand (Hrsg.), *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. 1. Halbband* (S. 208–218). Berlin: De Gruyter.

Hinnenkamp, V. (1982). *Foreigner Talk und Tarzanisch. Eine vergleichende Studie über die Sprechweise gegenüber Ausländern am Beispiel des Deutschen und des Türkischen*. Hamburg: Buske Verlag.

Hoffmann, L. (1985). *Kommunikationsmittel Fachsprache. Eine Einführung* (2. Aufl.). Tübingen: Gunter Narr Verlag.

Hoge, R. D. & Coladaraci, T. (1989). Teacher-Based Judgements of Academic Achievement: A Review of Literature. *Review of Educational Research*, *59*(3), 297–313.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric Statistical Methods* (3. Aufl.). Hoboken, New Jersey, USA: Wiley.

Holmeier, M. (2013). *Leistungsbeurteilung im Zentralabitur*. Wiesbaden: Springer VS.

Holsti, O. R. (1969). *Content Analysis for the Social Sciences and Humanities*. Reading, Massachusetts, USA: Addison-Wesley Publishing Company.

Hopf, C. (2015). Qualitative Interviews - ein Überblick. In U. Flick, E. von Kardorff, & I. Steinke (Hrsg.), *Qualitative Forschung. Ein Handbuch* (11. Aufl., S. 349–360). Reinbek bei Hamburg: Rowohlt.

Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of educational measurement*, *22*(3), 177–182.

Hosenfeld, I., Helmke, A., & Schrader, F.-W. (2002). Diagnostische Kompetenz: Unterrichts- und lernrelevante Schülermerkmale und deren Einschätzung durch Lehrkräfte in der Unterrichtsstudie SALVE. In M. Prenzel & J. Doll (Hrsg.), *Zeitschrift für Pädagogik 45. Beiheft. Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen* (S. 65–82). Weinheim: Beltz.

Höttecke, D. (2015). Stolpersteine der Diagnose ... und wie man sie umgehen kann. *Unterricht Physik*, *147/148*, 11–13.

Höttecke, D. (2017). Naturwissenschaft und Sprache. In U. Gebhard, D. Höttecke, & M. Rehm (Hrsg.), *Pädagogik der Naturwissenschaften. Ein Studienbuch* (S. 107–124). Wiesbaden: Springer VS.

Höttecke, D., Ehmke, T., Krieger, K., & Kulik, M. A. (2017). Vergleichende Messung fachsprachlicher Fähigkeiten in den Domänen Physik und Sport. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*(1), 53–69.

Höttecke, D., Feser, M. S., Heine, L., & Ehmke, T. (2018). Do Linguistic Features Influence Item Difficulty in Physics Assessments? *Science Education Review Letters*, *2018*, 1–6.

Höttecke, D. & Wodzinski, R. (2015). Diagnostizieren und Fördern. Hintergründe, Ansätze und Probleme von Diagnostik im Physikunterricht. *Unterricht Physik*, *147/148*, 2–10.

Hughes, J. & Parkes, S. (2003). Trends in the use of verbal protocol analysis in software engineering research. *Behaviour & Information Technology*, *22*(2), 127–140.

Ingenkamp, K.-H. (1985). Erfassung und Rückmeldung des Lernerfolgs. In D. Lenzen (Hrsg.), *Enzyklopädie Erziehungswissenschaft. Band 4. Methoden und Medien der Erziehung und des Unterrichts* (S. 173–205). Stuttgart: Klett-Cotta.

Ingenkamp, K.-H. (1995). *Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte*. Weinheim: Beltz.

Ingenkamp, K.-H. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Auflage). Weinheim: Beltz.

Ischreyt, H. (1965). *Studien zum Verhältnis von Sprache und Technik. Institutionelle Sprachlenkung in der Terminologie der Technik.* Düsseldorf: Pädagogischer Verlag Schwann.

Jäger, R. S. (2007). *Beobachten, beurteilen und fördern! Lehrbuch für die Aus-, Fort- und Weiterbildung.* Landau: Verlag Empirische Pädagogik.

Jäger, R. S. (2009). Diagnostische Kompetenz und Urteilsbildung als Element von Lehrprofessionalität. In O. Zlatkin-Troitschanskaia, K. Beck, D. Sembill, R. Nickolaus, & R. Mulder (Hrsg.), *Lehrprofessionalität. Bedingungen, Genese, Wirkungen und ihre Messung* (S. 105–116). Weinheim: Beltz.

Jäger-Flor, D. & Jäger, R. S. (2008). *Bildungsbarometer zum Thema Förderung im Bildungssystem. Ergebnisse, Bewertungen und Perspektiven.* Landau: Verlag Empirische Pädagogik.

Janich, N. (1998). *Fachliche Information und inszenierte Wissenschaft. Fachlichkeitskonzepte in der Wirtschaftswerbung.* Tübingen: Gunter Narr Verlag.

Jatzwauk, P., Rumann, S., & Sandmann, A. (2008). Der Einfluss des Aufgabeneinsatzes im Biologieunterricht auf die Lernleistung der Schüler - Ergebnis einer Videostudie. *Zeitschrift für Didaktik der Naturwissenschaften*, *14*, 263–283.

Jucks, R. (2001). *Was verstehen Laien?. Die Verständlichkeit von Fachtexten aus der Sicht von Computer-Experten.* Münster: Waxmann.

Jung, W. (1983). *Anstöße. Ein Essay über die Didaktik der Physik und ihre Probleme.* Frankfurt am Main: Diesterweg.

Jung, W., Reul, H., & Schwedes, H. (1977). *Untersuchung zur Einführung in die Mechanik in den Klassen 3-6.* Frankfurt am Main: Diesterweg.

Jürgens, E. (1997). *Leistung und Bewertung in der Schule. Eine Einführung in Leistungs- und Bewertungsfragen aus pädagogischer Sicht* (3. Aufl.). Sankt Augustin: Academia.

Kaiser, J. & Möller, J. (2017). Diagnostische Kompetenz von Lehramtsstudierenden. In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals. Interdisziplinäre Betrachtungen, Befunde und Perspektiven* (S. 55–74). Wiesbaden: Springer VS.

Kalthoff, H. (1996). Das Zensurenpanoptikum. Eine ethnographische Studie zur schulischen Bewertungspraxis. *Zeitschrift für Soziologie*, *25*(2), 106–124.

Kalverkämper, H. (1992). Die kulturanthropologische Dimension von 'Fachlichkeit' im Handeln und Sprechen. Konstative Studie zum Deutschen, Englischen, Französischen, Italienischen und Spanischen. In J. Albrecht & R. Baum (Hrsg.), *Fachsprache und Terminologie in Geschichte und Gegenwart* (S. 31–58). Tübingen: Gunter Narr Verlag.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating Opportunities for Students to Show What They Know: The Role of Scaffolding in Assessment Tasks. *Science Education*, *98*(4), 674–704.

Karing, C. (2009). Diagnostische Kompetenz von Grundschul- und Gymnasiallehrkräften im Leistungsbereich und im Bereich Interessen. *Zeitschrift für Pädagogische Psychologie*, *23*(3-4), 197–209.

Karing, C., Matthäi, J., & Artelt, C. (2011). Genauigkeit von Lehrerurteilen über die Lesekompetenz ihrer Schülerinnen und Schüler in der Sekundarstufe I - Eine Frage der Spezifizität? *Zeitschrift für Pädagogische Psychologie*, *25*(3), 159–172.

Karst, K. & Förster, N. (2017). Ansätze zur Modellierung diagnostischer Kompetenz. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 19–20). Münster: Waxmann.

Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungsaufgaben*. Berlin: Logos.

Kelle, U. & Erzberger, C. (2015). Qualitative und quantiative Methoden: kein Gegensatz. In U. Flick, E. von Kardorff, & I. Steinke (Hrsg.), *Qualitative Forschung. Ein Handbuch* (11. Aufl., S. 299–309). Reinbek bei Hamburg: Rowohlt.

Kendall, M. & Gibbons, J. D. (1990). *Rank Correlation Methods* (5. Aufl.). London, Großbritannien: Edward Arnold.

Keupp, H., Ahbe, T., Gmür, W., Höfer, R., Mitzscherlich, B., Kraus, W., & Straus, F. (2002). *Identitätskonstruktionen. Das Patchwork der Identitäten in der Spätmoderne*. Reinbek bei Hamburg: Rowohlt.

Kiel, E. (2009). Literalität. In S. Andersen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee, & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft* (S. 592–605). Weinheim: Beltz.

Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften*. Berlin: Logos.

Klafki, W. (1995). Leistung. In D. Lenzen (Hrsg.), *Pädagogische Grundbegriffe. Band 2. Jugend bis Zeugnis* (S. 983–987). Reinbek bei Hamburg: Rowohlt.

Kleber, E. W. (1976). Tendenzen, die das Urteil des Lehrers beeinflussen. In E. W. Kleber, H. Meister, C. Schwarzer, & R. Schwarzer (Hrsg.), *Beurteilung und Beurteilungsprobleme. Eine Einführung in Beurteilungs- und Bewertungsfragen in der Schule* (S. 39–61). Weinheim: Beltz.

Kleber, E. W. (1992). *Diagnostik in pädagogischen Handlungsfeldern. Einführung in Bewertung, Beurteilung, Diagnose und Evaluation*. Weinheim und München: Jeventa.

Klemp, A. (2015). Das Verhältnis fachlicher und sprachlicher Kompetenzen beim Lösen physikalischer Leistungsaufgaben - eine explorative Untersuchung der Fähigkeiten von Schüler/innen. unveröffentlichte Masterarbeit. Universität Hamburg.

Klinghammer, J., Rabe, T., & Krey, O. (2016). Unterrichtsbezogene Vorstellungen von Lehramtsstudierenden der Physik. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*, 181–195.

Klug, J. (2017). Ein Prozessmodell zur Diagnostik und Förderung von selbstreguliertem Lernen. In A. Südkamp & A.-K. Praetorius (Hrsg.), *Diagnostische Kompetenz von Lehrkräften* (S. 54–59). Münster: Waxmann.

Knorr, P. & Schramm, K. (2012). Datenerhebung durch Lautes Denken und Lautes Erinnern in der fremdsprachendidaktischen Empirie. In S. Doff (Hrsg.), *Fremdspra-*

chenunterricht empirisch erforschen. Grundlagen - Methoden - Anwendung (S. 184–217). Tübingen: Narr Francke Attempto Verlag.

Koch, P. & Oesterreicher, W. (1985). Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In G. Folke, D. Jacob, A. Kablitz, B. König, K. Joachim, D. Nelting, ... S. Zepp (Hrsg.), *Romanistisches Jahrbuch. Band 36 - 1985* (S. 15–43). De Gruyter.

Köller, O. (2008). Bildungsstandards - Verfahren und Kriterien bei der Entwicklung von Messinstrumenten. *Zeitschrift für Pädagogik*, *54*(2), 163–173.

König, B. (2017). *Schriftliches Korrigieren im Schulalltag. Eine qualitative Analyse der Korrekturtätigkeit von Grundschullehrkräften.* Opladen: Verlag Barbara Budrich.

Konrad, K. (2010). Lautes Denken. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 476–490). Wiesbaden: VS Verlag für Sozialwissenschaften.

Krabbe, H. & Beese, M. (2015). Lesestrategien für Erklärungstexte in Physikbüchern. *MNU-Journal*, *68*(3), 148–155.

Kracauer, S. (1952). The Challenge of Qualitative Content Analysis. *The Public Opinion Quarterly*, *16*(4), 631–642.

Krauss, S., Lindl, A., Schilcher, A., Michael, F., Göhring, A., Hofmann, B., ... Mulder, R. H. (2017). *FALKO: Fachspezifischen Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik.* Münster: Waxmann.

Krippendorff, K. (2004). *Content Analysis. An Introduction to Its Methodology* (2. Aufl.). Thousand Oaks, California, USA: Sage Publications Ltd.

Krispenz, A., Dickhäuser, O., & Reinhard, M.-A. (2016). Assessing task difficulty for other people: when deeper evaluation means "it's more about me!" *Social Psychology of Education*, *19*(4), 865–877.

Kröger, J., Neumann, K., & Petersen, S. (2013). Messung professioneller Kompetenz im Fach Physik. In S. Bernholt (Hrsg.), *Inquiry-based Learning – Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012* (S. 533–535). Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

Krolak-Schwerdt, S., Böhmer, M., & Gräsel, C. (2012). Leistungsbeurteilung von Schulkindern. Welche Rolle spielen Ziele und Expertise der Lehrkraft? *Zeitschrift für Entwicklungsspsychologie und Pädagogische Psychologie*, *44*(3), 111–122.

Kroll, N. (2017). Rekonstruktion von Logiken und Ressourcen von Physiklehrkräften zum Beurteilen schriftlicher Schülerleistung. unveröffentlichte Masterarbeit. Universität Hamburg.

Krüger, H.-H. & Pfaff, N. (2004). Triangulation quantiativer und qualitativer Zugänge in der Schulforschung. In W. Helsper & J. Böhme (Hrsg.), *Handbuch der Schulforschung* (S. 159–182). Wiesbaden: Springer VS.

Krüger, J. (2017). *Schülerperspektiven auf die zeitliche Entwicklung der Naturwissenschaften. Theoretische Grundsatzüberlegungen und empirische Erkenntnisse.* Berlin: Logos.

Kuckartz, U. (2014). *Mixed Methods. Methodolgie, Foschungsdesigns und Analyseverfahren*. Wiesbaden: Springer VS.

Kuckartz, U. (2016). *Qualitative Inhaltsanalyse. Methoden, Praxis, Computerunterstützung* (3. Aufl.). Weinheim: Beltz Juventa.

Kühberger, C. (2014). *Leistungsfeststellung im Geschichtsunterricht. Diagnose - Bewertung - Beurteilung*. Schwalbach: Wochenschau Verlag.

Kühn, C. (2015). *Literacy in der Kita. Dialogische Bilderbuchbetrachtungen und deren Bedeutsamkeit für den Schriftspracherwerb*. Hamburg: disserta Verlag.

Kulgemeyer, C. (2010). *Physikalische Kommunikationskompetenz. Modellierung und Diagnostik*. Berlin: Logos.

Kulgemeyer, C. & Tomczyszyn, E. (2015). Physik erklären – Messung der Erklärensfähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation. *Zeitschrift für Didaktik der Naturwissenschaften*, *21*(1), 111–126.

Kultusministerkonferenz (1968). Erläuterungen der Notenstufen bei Schulzeugnissen und Einzelergebnissen in staatlichen Prüfungszeugnissen. Beschluss der Kultusministerkonferenz vom 03.10.1968.

Kultusministerkonferenz (2004a). Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss der Kultusministerkonferenz vom 16.12.2004.

Kultusministerkonferenz (2004b). Einheitliche Prüfungsanforderungen in der Abiturprüfung Physik. Beschluss der Kultusministerkonferenz vom 01.12.1989 in der Fassung vom 05.02.2004.

Kultusministerkonferenz (2004c). Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004.

Kultusministerkonferenz (2018). Ländergemeinsamen inhaltlichen Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 11.10.2018.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: Waxmann.

Kunz, R. (2015). Situative Kasuistik – Die Relationierung von Theorie und Praxis durch Schlüsselsituationen. In E. Bolay, A. Iser, & M. Weinhardt (Hrsg.), *Methodisches Handeln - Beiträge zu Maja Heiners Impulsen zur Professionalisierung der Sozialen Arbeit* (S. 77–89). Wiesbaden: Springer VS.

Lange, K. (2010). Zusammenhänge zwischen naturwissenschaftsbezogenem fachspezifischpädagogischem Wissen von Grundschullehrkräften und Fortschritten im Verständnis naturwissenschaftlicher Konzepte bei Grundschülerinnen und -schülern. Dissertation. Universität Münster.

Lee, E. & Luft, J. A. (2008). Experienced Secondary Science Teachers' Representation of Pedagogical Content Knowledge. *International Journal of Science Education*, *30*(10), 1343–1363.

Leisen, J. (1991). Über Sprachprobleme im deutschsprachigen Fachunterricht am Beispiel des Physikunterrichts. *Zielsprache Deutsch*, *22*(3), 143–151.

Leisen, J. (2005). Muss ich jetzt auch noch Sprache unterrichten? Sprache im Physikunterricht. *Unterricht Physik*, *3*(87), 4–9.

Leisen, J. (2017). *Handbuch Sprachförderung im Fach. Sprachsensibler Fachunterricht in der Praxis. Grundlagenteil*. Stuttgart: Klett.

Leisen, J. & Höttecke, D. (2011). Leistungsmessung und Schülerbeurteilung. In H. Wiesner, H. Schecker, & M. Hopf (Hrsg.), *Physikdidaktik kompakt* (S. 63–71). Hallbergmoos: Aulis.

Lemke, J. L. (1990). *Talking Science. Language, learning and values.* Norwood, New Jersey, USA: Ablex Publishing Corporation.

Lengyel, D., Heintze, A., Reich, H. H., Roth, H.-J., & Scheinhardt-Stettner, H. (2009). Prozessbegleitende Diagnose zur Schreibentwicklung: Beobachtung schriftlicher Sprachhandlungen in der Sekundarstufe I. In D. Lengyel, H. H. Reich, H.-J. Roth, & M. Döll (Hrsg.), *Von der Sprachdiagnose zur Sprachförderung* (S. 129–138). Münster: Waxmann.

Lengyel, D. & Roth, H.-J. (2012). Beobachtung der Schreibentwicklung in der Sekundarstufe I. In S. Fürstenau & M. Gomolla (Hrsg.), *Migration und schulischer Wandel: Leistungsbeurteilung* (S. 123–136). Wiesbaden: Springer VS.

Leuders, T., Leuders, J., & Philipp, K. (2014). Fachbezogene diagnostische Kompetenzen - Forschungsstand und Forschungsdesiderata. In J. Roth & J. Ames (Hrsg.), *Beiträge zum Mathematikunterricht 2014* (S. 731–734). Münster: WTM-Verlag.

Liebl, A., Haller, J., Jödicke, B., Baumgartner, H., Schlittmeier, S., & Hellbrück, J. (2012). Combined effects of acoustic and visual distraction on cognitive performance and well-being. *Applied Ergonomics*, *43*, 424–434.

Lienert, G. A. (1967). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, California, USA: Sage Publications Ltd.

Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D., … Kunter, M. (2015). Assessing Teachers' Educational Knowledge. Construct Specification and Validation Using Mixed Methods. *Zeitschrift für Entwicklungsspsychologie und Pädagogische Psychologie*, *47*(2), 72–83.

Lißman, U. & Paetzhold, B. (1982). Kriteriumsorientierte und sehr differenzierte Leistungsrückmeldung. Eine Längsschnittuntersuchung in Hauptschule. In F. Rheinberg (Hrsg.), *Bezugsnormen zur Schulleistungsbewertung: Analyse und Intervention* (S. 193–219). Düsseldorf: Schwann.

Lissmann, U. (2001). *Inhaltsanalyse von Texten. Ein Lehrbuch zur computergestützten und konventionellen Inhaltsanalyse*. Landau: Verlag Empirische Pädagogik.

Lomax, R. G. & Hahs-Vaughn, D. L. (2012). *An Introduction to Statistical Concepts* (3. Aufl.). New York, New York, USA: Routledge.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2018). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assessment in Education*, *25*(5), 442–467.

Loughran, J., Berry, A., & Mulhall, P. (2012). *Understanding and Developing Science Teachers' Pedagogical Content Knowledge* (2. Aufl.). Rotterdam, Niederlande: Sense Publishers.

Luchtenberg, S. (1989). Überlegungen zur Bedeutung von Fachsprache in Vorschule und Schule: Möglichkeiten und Schwierigkeiten. *Fachsprache. Internationale Zeitschrift für Fachsprachenforschung, -didaktik und Terminologie*, *2*, 153–171.

Lüders, C. (2015). Beobachtung im Feld und Ethnographie. In U. Flick, E. von Kardorff, & I. Steinke (Hrsg.), *Qualitative Forschung. Ein Handbuch* (11. Aufl., S. 384–401). Reinbek bei Hamburg: Rowohlt.

Ludwig, O. (1962). Über Kombination von Rangkorrelationskoeffizienten aus unabhängigen Meßreihen. *Biometrical Journal*, *4*(1), 40–50.

Lumley, T. (2005). *Assessing Second Langauge Writing. The Rater's Perspective*. Frankfurt am Main: Peter Lang.

Luykx, A., Lee, O., Hart, J., & Deaktor, R. (2007). Cultural and Home Language Influences on Children's Responses to Science Assessments. *Teachers College Record*, *109*(4), 897–926.

Lyon, E. G. (2013a). "Assessment as Discourse": A Pre-Service Physics Teacher's Evolving Capacity to Support an Equitable Padagogy. *education sciences*, *3*, 279–299.

Lyon, E. G. (2013b). Conceptualizing and Exemplifying Science Teachers' Assessment Expertise. *International Journal of Science Education*, *35*(7), 1208–1229.

Lyon, E. G. (2013c). Learning to Assess Science in Linguistically Diverse Classrooms: Tracking Growth in Secondary Science Preservice Teachers' Assessment Expertise. *Science Education*, *97*(3), 442–467.

Lyon, E. G. (2013d). What about language while equitably assessing science?: Case studies of preservice teachers' evolving expertise. *Teaching and Teacher Education*, *32*, 1–11.

Lyon, E. G., Tolbert, S., Stoddart, P., Solís, J., & Bunch, G. C. (2016). *Secondary Science Teaching for English Learners. Developing Supportive and Responsive Learning Contexts for Sense-Making and Language Development*. Lanham, Maryland, USA: Rowman & Littlefield.

Machts, N., Kaiser, J., Schmidt, F. T. C., & Möller, J. (2016). Accuracy of teachers' judgments of students' cognitive abilities: A meta-analysis. *Educational Research Review*, *19*, 85–103.

Magnusson, S., Krajcik, J., & Borko, H. (1999). Nature, Sources, and Development of Pedagogical Content Knowledge for Science Teaching. In J. Gess-Newsome & N. G. Lederman (Hrsg.), *Examining Pedagogical Content Knowledge. The Construct and its Implications for Science Education* (S. 95–132). New York, New York, USA: Kluwer Academic Publishers.

Maier, M. (2001). *Das Verbalzeugnis in der Grundschule: Anspruch und Wirklichkeit*. Landau: Verlag Empirische Pädagogik.

Maier, U. (2015). *Leistungsdiagnostik in Schule und Unterricht. Schülerleistungen messen, bewerten und fördern*. Bad Heilbrunn: Klinkhardt.

Marksteiner, T., Reinhard, M.-A., Dickhäuser, O., & Sporer, S. L. (2012). How do teachers perceive cheating students? Beliefs about cues to deception and detection accuracy in the educational field. *European Journal of Psychology of Education*, *27*(3), 329–350.

Marso, R. N. & Pigge, F. L. (1993). Teachers' Testing Knowledge, Skills, and Practices. In S. L. Wise (Hrsg.), *Teacher Training in Measurement and Assessment Skills* (S. 129–185). Lincoln, Nebraska, USA: Buros Institute of Mental Measurements.

Mathes, R. (1992). Hermeneutisch-klassifikatorische Inhaltsanalyse von Leitfadengesprächen. Über das Verhältnis von quantitativen und qualitativen Verfahren der Textanalyse und die Möglichkeit ihrer Kombination. In J. H. P. Hoffmeyer-Zlotnik (Hrsg.), *Analyse verbaler Daten. Über den Umgang mit qualitativen Daten* (S. 402–424). Opladen: Westdeutscher Verlag.

Mauermann, L. (1976). *Faktoren unterrichtlicher Kommunikation. Untersuchung im Rahmen eines Physik-Curriculums für den 5. Schülerjahrgang.* München: R. Oldenbourg Verlag.

Mayring, P. (2001). Kombination und Integration qualitativer und quantitativer Analyse. *Forum Qualitative Sozialforschung*, *2*(1), o. S.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Technicken* (12. Aufl.). Weinheim: Beltz.

Mbaye, M. (2018). *Eine rekonstruktive Studie zum Umgang mit Fehlern im Fremdsprachenunterricht in Deutschland und im Senegal.* Baden-Baden: Tectum.

Meier, S. (2014). *Physik Schulaufgaben, Übungen. 7.-10. Klasse Real- Mittel- und Gesamtschule.* North Charleston, South Carolina, USA: CreateSpace Independent Publishing Platform.

Meisel, J. M. (1975). Ausländerdeutsch und Deutsch ausländischer Arbeiter. *Zeitschrift für Literaturwissenschaft und Linguistik*, *5*(18), 9–53.

Merzyn, G. (2006). Zensuren, Lernerfolge und Schülereinstellungen. Besonderheiten von Chemie und Physik im Vergleich der Schulfächer. *MNU-Journal*, *65*(2), 116–120.

Meyer, H. (2016). *Unterrichtsmethoden I. Theorieband* (17. Aufl.). Berlin: Cornelsen.

Michigan Assessment Consortium (2017). *Assessment Literacy Standards. A national imperative.* Mason, Michigan, USA: Michigan Assessment Consortium.

Mikelskis, H. (1996). Didaktik der Physik. Selbstbesinnung über Stand und Perspektiven einer sich findenden Wissenschaftsdisziplin in Forschung und Lehre. *LLF-Berichte*, *15*, 23–42.

Mischo, C. & Rheinberg, F. (1995). Erziehungsziele von Lehrern und individuelle Bezugsnormen der Leistungsbewertung. *Zeitschrift für Pädagogische Psychologie*, *9*(3/4), 139–151.

Morek, M. & Heller, V. (2012). Bildungssprache - Kommunikative, epistemische, soziale und interaktive Aspekte ihres Gebrauchs. *Zeitschrift für Angewandte Linguistik*, *57*(1), 67–101.

Moss, P. A. (2003). Reconceptualizing Validity for Classroom Assessment. *Educational Measurement*, *22*(4), 13–25.

Muckenfuß, H. (1995). *Lernen im sinnstiftenden Kontext. Entwurf einer zeitgemäßen Didaktik des Physikunterrichts.* Berlin: Cornelsen.

Neumann, A. & Domenech, M. (2010). *Paradoxien des Schreibens in der Bildungssprache Deutsch. Befunde zu Schreibsozialisation, Schreibmotivation und Schreibfähigkeiten bei Schülerinnen und Schülern mit nichtdeutscher Muttersprache und zum Schreibunterricht im mehrsprachigen Kontext*. Hamburg: Verlag Dr. Kovač.

Neunhöffer, M. (1967). Physikunterricht und Spracherziehung. *Der Physikunterricht. Beiträge zu seiner Didaktik*, *1*(2).

Neuweg, G. H. (2000). *Schulische Leistungsbeurteilung. Rechtliche Grundlagen und pädagogische Hilfestellungen für die Schulpraxis* (1. Aufl.). Linz, Österreich: Trauner.

Neuweg, G. H. (2004). Die Beziehung zwischen Lehrerwissen und Lehrerkönnen: Zwölf Modellvorstellungen im Überblick. In M. Krainz-Dürr, H. Erzinger, & M. Schmoczer (Hrsg.), *Grenzen überschreiten in Bildung und Schule* (S. 74–82). Klagenfurt, Österreich: Drava.

Neuweg, G. H. (2009). *Schulische Leistungsbeurteilung. Rechtliche Grundlagen und pädagogische Hilfestellungen für die Schulpraxis* (4. Aufl.). Linz, Österreich: Trauner.

Neuweg, G. H. (2014). Das Wissen der Wissensvermittler. Problemstellungen, Befunde und Perspektiven der Forschung zum Lehrerwissen. In E. Terhart, H. Bennewitz, & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (2. Aufl., S. 583–614). Münster: Waxmann.

Neuweg, G. H. (2018). Evidenzbasierte Lehrerbildung. In G. H. Neuweg (Hrsg.), *Distanz und Einlassung. Gesammelte Schriften zur Lehrerbildung* (S. 63–67). Münster: Waxmann.

Nickel, S. (2007). Familienorientierte Grundbildung im Sozialraum als Schlüsselstrategie zur breiten Teilhabe an Literalität. In A. Grotlüsche & A. Linde (Hrsg.), *Literalität, Grundbildung oder Lesekompetenz?. Beiträge zu einer Theorie-Praxis-Diskussion* (S. 31–41). Münster: Waxmann.

Nickerson, R. S. (1999). How We Know — and Sometimes Misjudge — What Others Know: Imputing One's Own Knowledge to Others. *Psychological Bulletin*, *125*(6), 737–759.

Niederhaus, C. (2011). *Fachsprachlichkeit in Lehrbüchern. Korpuslinguistische Analysen von Fachtexten der beruflichen Bildung*. Münster: Waxmann.

Nitz, S., Nerdel, C., & Prechtl, H. (2012). Entwicklung eines Erhebungsinstruments zur Erfassung der Verwendung von Fachsprache im Biologieunterricht. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 117–139.

Nohl, A.-M. (2017). *Interview und Dokumentarische Methode. Anleitung für die Forschungspraxis* (5. Aufl.). Wiesbaden: Springer VS.

Oelkers, J. & Oser, F. (2000). *Die Wirksamkeit der Lehrerbildungssysteme in der Schweiz. Umsetzungsbericht*. Aarau, Schweiz: Fasler Druck AG.

Ogan-Bekiroglu, F. (2009). Assessing Assessment: Examination of pre-service physics teachers' attitudes towards assessment and factors affecting their attitudes. *International Journal of Science Education*, *31*(1), 1–39.

Ogborn, J., Kress, G., Martins, I., & McGillicuddy, K. (1996). *Explaining Science in the Classroom*. Buckingham, Großbritannien: Open University Press.

Opitz, R. (2002). Didaktische Aspekte der physikalischen Fachsprache. Dissertation. Gerhard-Mercator-Universität Duisburg.

Ortner, H. (2009). Rhetorisch-stilistische Eigenschaften der Bildungssprache. In U. Flix, A. Gardt, & J. Knape (Hrsg.), *Rhetorik und Stilistik. Ein internationales Handbuch historischer und systematischer Forschung. 2. Halbband* (S. 2227–2240). Berlin: De Gruyter.

Oser, F. (2001). Standards: Kompetenzen von Lehrpersonen. In F. Oser & J. Oelkers (Hrsg.), *Die Wirksamkeit der Lehrerbildungssysteme. Von der Allrounderbildung zur Ausbildung professioneller Standards* (S. 215–342). Zürich, Schweiz: Verlag Rüegger.

Ostermann, A., Leuders, T., & Nückles, M. (2015). Wissen, was Schülerinnen und Schülern schwer fällt. Welche Faktoren beeinflussen die Schwierigkeitsschätzung von Mathematikaufgaben. *Journal für Mathematik-Didaktik*, *36*(1), 45–76.

Özcan, N. (2013). *Zum Einfluss der Fachsprache auf die Leistung im Fach Chemie. Eine Förderstudie zur Fachsprache im Chemieunterricht.* Berlin: Logos.

Page, E. B. (1963). Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks. *Journal of the American Statistical Association*, *58*(301), 216–230.

Pajares, M. F. (1992). Teachers' Beliefs and Educational Research: Cleaning Up a Messy Construct. *Review of Educational Research*, *62*(3), 307–332.

Pallant, J. (2007). *SPSS Survival Manual. A Step by Step Guide to Data Analysis using SPSS for Windows* (3. Aufl.). Maidenhead, Großbritannien: Open University Press.

Park, S. & Oliver, J. S. (2008). Revisiting the Conceptualisation of Pedagogical Content Knowledge (PCK): PCK as a Conceptual Tool to Understand Teachers as Professionals. *Research in Science Education*, *38*(3), 261–284.

Paseka, A. (2010). Interviews „qualitativ" auswerten - ein Beispiel aus der Forschungspraxis. In C. Fridrich, M. Heissenberger, & A. Paseka (Hrsg.), *PH Wien. Forschungsperspektiven. Band 2* (S. 141–161). Wien, Österreich: LIT Verlag.

Pineker-Fischer, A. (2017). *Sprach- und Fachlernen im naturwissenschaftlichen Unterricht. Umgang von Lehrpersonen in soziokulturell heterogenen Klassen mit Bildungssprache.* Wiesbaden: Springer VS.

Pöhlmann-Lang, A. (2015). Bildungssprache. Nicht nur eine Herausforderung beim Zweitsprachlernen. In C. Kupfer-Schreiner & A. Pöhlmann-Lang (Hrsg.), *Didaktik des Deutschen als Zweitsprache - DiDaZ in Bamberg lehren und lernen. Eine Bilanz des Faches in Forschung und Lehre (2010 bis 2015)* (S. 103–113). Bamberg: University of Bamberg Press.

Popham, W. J. (2009). Assessment Literacy for Teachers: Faddish or Fundamental? *Theory Into Practice*, *48*(1), 4–11.

Popham, W. J. (2011). Assessment Literacy Overlooked: A Teacher Educator's Confession. *The Teacher Educator*, *46*(4), 265–273.

Poske, F. (1915). *Didaktik des physikalischen Unterrichts.* Leipzig: B. G. Teubner Verlag.

Praetorius, A.-K., Karst, K., & Lipowsky, F. (2011). Wie gut schätzen Lehrer die Fähigkeitsselbstkonzepte ihrer Schüler ein? Zur diagnostischen Kompetenz von Lehrkräften. *Psychologie in Erziehung und Unterricht. Zeitschrift für Forschung und Praxis*, *58*, 81–91.

Prophet, R. B. & Babede, N. B. (2009). Language and student performance in junior secondary science examinations: The case of second language learners in Botswana. *International Journal of Science and Mathematics Education*, *7*(2), 389–397.

Przyborski, A. & Wohlrab-Sahr, M. (2014). *Qualitative Sozialforschung. Ein Arbeitsbuch* (4. Aufl.). München: Oldenbourg Wissenschaftsverlag.

Ravitch, D. (2007). *EdSpeak. A Glossary of Education Terms, Phrases, Buzzwords, and Jargon.* Alexandria, Virginia, USA: Association for Supervision & Curriculum Development.

Rehm, M. (2012). Verstehen als Kompetenz. Wagenscheins Verstehen lernen in einem Kompetenzmodell. In N. Kruse, M. Rudolf, & W. Bernd (Hrsg.), *Martin Wagenschein - Faszination und Aktualität des Genetischen* (S. 119–136). Baltmannsweiler: Schneider-Verlag Hohengehren.

Reid, T. B. W. (1956). Linguistics, Structuralism and Philology. *Archivum Linguisticum*, *8*, 28–37.

Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: the Spanish instance, building upon Black and Wiliam (2005). *The Curriculum Journal*, *18*(1), 27–38.

Renkl, A. (1993). Korrelation und Kausalität. Ein ausreichend durchdachtes Problem in der pädagogisch-psychologischen Forschung? In C. Tarnai (Hrsg.), *Beiträge zur empirischen pädagogischen Forschung* (S. 115–123). Münster: Waxmann.

Reusser, K., Pauli, C., & Elmer, A. (2011). Berufsbezogene Überzeugungen von Lehrerinnen und Lehrern. In E. Terhart, H. Bennewitz, & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (1. Aufl., S. 478–495). Münster: Waxmann.

Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation.* Göttingen: Verlag für Psychologie Dr. C. J. Hogrefe.

Rheinberg, F. (1982). Bezugsnorm-Orientierung angehender Lehrer im Verlauf ihrer praktischen Ausbildung. In F. Rheinberg (Hrsg.), *Bezugsnormen zur Schulleistungsbewertung: Analyse und Intervention* (S. 235–247). Düsseldorf: Schwann.

Rheinberg, F. (1987). Soziale versus individuelle Leistungsvergleiche. In R. Olechowski & E. Persy (Hrsg.), *Fördernde Leistungsbeurteilung. Ein Symposium* (S. 81–115). Wien, Österreich: Jugend und Volk.

Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59–71). Weinheim: Beltz.

Rheinberg, F. & Fries, S. (2010). Bezugsnormorientierung. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (S. 61–67). Weinheim: Beltz.

Richter, D., Kuhl, P., Haag, N., & Pant, H. A. (2013). Aspekte der Aus- und Fortbildung von Mathematik- und Naturwissenschaftslehrkräften im Ländervergleich. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenz am Ende der Sekundarstufe I* (S. 367–390). Münster: Waxmann.

Riebling, L. (2013a). Heuristik der Bildungssprache. In I. Gogolin, I. Lange, U. Michel, & H. H. Reich (Hrsg.), *Herausforderung Bildungssprache - und wie man sie meistert* (S. 106–153). Münster: Waxmann.

Riebling, L. (2013b). *Sprachbildung im naturwissenschaftlichen Unterricht. Eine Studie im Kontext migrationsbedingter sprachlicher Heterogenität*. Münster: Waxmann.

Riese, J. (2009). *Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften*. Berlin: Logos.

Rincke, K. (2007). *Sprachentwicklung und Fachlernen im Mechanikunterricht. Sprache und Kommunikation bei der Einführung in den Kraftbegriff*. Berlin: Logos.

Rincke, K. (2010). Alltagssprache, Fachsprache und ihre besonderen Bedeutungen für das Lernen. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 235–260.

Rincke, K. & Leisen, J. (2015). Sprache im Physikunterricht. In E. Kircher, R. Girwidz, & P. Häußler (Hrsg.), *Physikdidaktik. Theorie und Praxis* (3. Aufl., S. 635–655). Berlin: Springer Spektrum.

Rincke, K. & Markic, S. (2018). Sprache und das Lernen von Naturwissenschaften. In D. Krüger, I. Parchmann, & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 31–48). Berlin: Springer Spektrum.

Roelcke, T. (2014). Zur Gliederung von Fachsprache und Fachkommunikation. *Fachsprache - International Journal of Specialized Communication*, (3-4), 154–178.

Roisch, H. (2003). Die horizontale und vertikale Geschlechterverteilung in der Schule. In M. Stürzer, H. Roisch, A. Hunze, & W. Cornelißen (Hrsg.), *Geschlechterverhältnisse in der Schule* (S. 21–52). Opladen: Leske + Budrich.

Rollnick, M., Bennett, J., Rhemtula, M., Dharsey, N., & Ndlovu, T. (2008). The Place of Subject Matter Knowledge in Pedagogical Content Knowledge: A case study of South African teachers teaching the amount of substance and chemical equilibrium. *International Journal of Science Education*, *30*(10), 1365–1387.

Rosenthal, J. W. (1996). *Teaching Science to Language Minority Students: Theory and Practice*. Clevedon, Philadelphia, USA: Multilingual Matters Ltd.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. London, Großbritannien: Sage Publications Ltd.

Rost, J. (1996). *Lehrbuch Testtheorie - Testkonstruktion*. Bern, Schweiz: Verlag Hans Huber.

Roth, H.-J., Reich, H., & Lengyel, D. (2011). FÖRMIG AG SEK I. Sprachhandlungen Erklären, Berichten und Argumentieren. Manual zu den Auswertungsrastern. unveröffentlichte Entwurfsfassung. Stand: Januar 2011.

Sacher, W. (1996). *Prüfen - Beurteilen - Benoten. Grundlagen, Hilfen und Denkanstöße für alle Schularten* (2. Aufl.). Bad Heilbrunn: Klinkhardt.

Sacher, W. (2009). *Leistungen entwickeln, überprüfen und beurteilen. Bewährte und neue Wege für die Primar- und Sekundarstufe* (5. Aufl.). Bad Heilbrunn: Klinkhardt.

Saldaña, J. (2016). *The Coding Manual for Qualitative Researchers* (3. Aufl.). London, Großbritannien: Sage Publications Ltd.

Salem, T., Neumann, M. U., Ursula, & Dobutowitsch, F. (2013). *Netzwerke für durchgängige Sprachbildung 1. Grundlagen und Fallbeispiele*. Münster: Waxmann.

Salzmann, C. (1971). Prüfen, Erproben. In H. Rombach (Hrsg.), *Lexikon der Pädagogik. Neue Ausgabe. Dritter Band. Kultur bis Schulbuch* (S. 358–359). Freiburg im Breisgau: Herder.

Sander, H. (2017). *Orientierungen von Jugendlichen beim Urteilen und Entscheiden in Kontexten nachhaltiger Entwicklung. Eine rekonstruktive Perspektive auf Bewertungskompetenz in der Didaktik der Naturwissenschaft*. Berlin: Logos.

Scarino, A. (2013). Language assessment literacy as self-awareness. Understanding the role of interpretation in assessment and in teacher learning. *Language Testing*, *30*(3), 309–327.

Schafer, W. D. (1993). Assessment Literacy for Teachers. *Theory into Practice*, *32*(2), 118–126.

Scherer, R. (2012). *Analyse der Struktur, Messinvarianz und Ausprägung komplexer Problemlösekompetenz im Fach Chemie. Eine Querschnittstudie in der Sekundarstufe I und am Übergang zur Sekundarstufe II*. Berlin: Logos.

Schleppegrell, M. J. (2001). Linguistic Features of the Language of Schooling. *Linguistics and Education*, *12*(4), 431–459.

Schleppegrell, M. J. (2004). *The Language of Schooling. A Functional Linguistics Perspective*. New York, New York, USA: Routledge.

Schmidt, M. (2008). *Kompetenzmodellierung und -diagnostik im Themengebiet Energie der Sekundarstufe I. Entwicklung und Erprobung eines Testinventars*. Berlin: Logos.

Schmidt, R. C. (1997). Managing Delphi Surveys Using Nonparametric Statistical Techniques. *Decision Sciences*, *28*(3), 763–774.

Schödl, A. & Göhring, A. (2017). FALKO-P: Fachspezifische Lehrerkompetenzen im Fach Physik. Entwicklung und Valididierung eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Physiklehrkräften. In S. Krauss, A. Lindl, A. Schilcher, F. Michael, A. Göhring, B. Hofmann, ... R. H. Mulder (Hrsg.), *FALKO: Fachspezifischen Lehrerkompetenzen. Konzeption von Professionswissenstests in den Fächern Deutsch, Englisch, Latein, Physik, Musik, Evangelische Religion und Pädagogik* (S. 201–242). Münster: Waxmann.

Schrader, F.-W. (2014). Lehrer als Diagnostiker. In E. Terhart, H. Bennewitz, & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (2. Aufl., S. 865–880). Münster: Waxmann.

Schrader, F.-W. & Helmke, A. (1987). Diagnostische Kompetenz von Lehrern: Komponenten und Wirkungen. *Empirische Pädagogik*, *1*(1), 27–52.

Schreiber, J.-R. & Siege, H. (2016). *Orientierungsrahmen für den Lernbereich Globale Entwicklung im Rahmen einer Bildung für nachhaltige Entwicklung* (2. Aufl.). Bonn: Engagement Global gGmbH.

Schreier, M. (2012). *Qualitative Content Analysis in Practice*. London, Großbritannien: Sage Publications Ltd.

Schreier, M. (2014a). Quantitative Content Analysis. In U. Flick (Hrsg.), *The SAGE Handbook of Qualitative Data Analysis* (S. 170–181). London, Großbritannien: Sage Publications Ltd.

Schreier, M. (2014b). Varianten qualitativer Inhaltsanalyse: Ein Wegweiser im Dickicht der Begrifflichkeiten. *Forum Qualitative Sozialforschung*, *15*(1), o. S.

Schröder, H. (1985). *Grundwortschatz Erziehungswissenschaft. Ein Wörterbuch der Fachbegriffe. Von «Abbilddidaktik» bis «Zielorientierung»*. München: Ehrenwirth.

Schulte, K., Hartig, J., & Pietsch, M. (2014). Der Sozialindex für Hamburger Schulen. In D. Fickermann & N. Maritzen (Hrsg.), *Grundlagen für eine daten- und theoriegestütze Schulentwicklung. Konzeption und Anspruch des Hamburger Instituts für Bildungsmonitoring und Qualitätsentwicklung (IfBQ)* (S. 67–80). Münster: Waxmann.

Shavelson, R. J. & Stern, P. (1981). Research on Teachers' Pedagogical Thoughts, Judgments, Decisions, and Behavior. *Review of Educational Research*, *51*(4), 455–498.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, *57*(1), 1–23.

Shulman, L. S. & Elstein, A. S. (1975). Studies of Problem Solving, Judgment, and Decision Making: Implications for Educational Research. *Review of Research in Education*, *3*, 3–42.

Siegel, M. A. & Wissehr, C. (2011). Preparing for the Plunge: Preservice Teachers' Assessment Literacy. *Journal of Science Teacher Education*, *22*(4), 371–391.

Siegel, S. (1976). *Nichtparametrische statistische Methoden.* Frankfurt am Main: Fachbuchhandlung für Psychologie Verlagsabteilung.

Siegel, S. & Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* New York, New York, USA: McGraw-Hill.

Simon, L. (1979). Über eine Möglichkeit der Bewertung schriftlicher Schülerleistungen im Physikunterricht. *Physik in der Schule*, *o. V.*(3), 90–96.

Skott, J. (2013). Understanding the role of the teacher in emerging classroom practices: searching for patterns of participation. *ZDM Mathematics Education*, *45*(4), 547–559.

Skott, J. (2015). The Promises, Problems, and Prospects of Research on Teachers' Beliefs. In H. Fives & M. G. Gill (Hrsg.), *International Handbook of Research on Teachers' Beliefs* (S. 13–30). New York, New York, USA: Routledge.

Smith, D. C. & Neale, D. C. (1989). The Construction of subject matter knowledge in primary science teaching. *Teaching & Teacher Education*, *5*(1), 1–20.

Smith, J. K. (2003). Reconsidering Reliability in Classroom Assessment and Grading. *Educational Measurement*, *22*(4), 26–33.

Snow, R. E. (1968). Brunswikian Approaches to Research on Teaching. *American Educational Research Journal*, *5*(4), 475–489.

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, *19*(1/2), 85–95.

Statistisches Bundesamt (2018). Schulen auf einen Blick. Ausgabe 2018.

Stefer, C. (2013). Die Gegenstandsangemessenheit empirischer Datenerhebungsmethoden im Kontext von Lehrevaluationen an Hochschulen. Dissertation. Philipps-Universität Marburg/Lahn.

Steinke, I. (1999). *Kriterien qualitativer Forschung. Ansätze zur Bewertung qualitativ-empirischer Sozialforschung.* Weinheim: Juventa.

Steinmüller, U. & Scharnhorst, U. (1987). Sprache im Fachunterricht. Ein Beitrag zur Diskussion über Fachsprache im Unterricht mit ausländischen Schülern. *Zielsprache Deutsch*, *18*(4), 3–12.

Stiggins, R. J. (1988). Revitalizing Classroom Assessment: The Highest Instructional Priority. *Phi Delta Kappan*, *69*(5), 363–368.

Stiggins, R. J. (1991). Assessment Literacy. *Phi Delta Kappan*, *72*(7), 534–539.

Stiggins, R. J. (1995). Assessment Literacy for the 21st Century. *Phi Delta Kappan*, *77*(3), 238–245.

Stiggins, R. J. (2014). Improve assessment literacy outside of schools too. *Phi Delta Kappan*, *96*(2), 67–72.

Strate, W. (2014). *Physik Schulaufgaben, Übungen. 7.-10. Klasse Gymnasium.* North Charleston, South Carolina, USA: CreateSpace Independent Publishing Platform.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of Teachers' Judgements of Students' Academic Achievement: A Meta-Analysis. *Journal of Educational Psychology*, *104*(3), 743–762.

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum. Eine experimentelle Untersuchung zur diagnostischen Kompetenz. *Zeitschrift für Pädagogische Psychologie*, *22*(3-4), 261–276.

Tajmel, T. (2010). DaZ-Förderung im naturwissenschaftlichen Fachunterricht. In B. Ahrenholz (Hrsg.), *Fachunterricht und Deutsch als Zweitsprache* (S. 167–184). Tübingen: Gunter Narr Verlag.

Tajmel, T. (2011). Wortschatzarbeit im mathematisch-naturwissenschaftlichen Unterricht. *ide. informationen zur deutschdidaktik. Zeitschrift für den Deutschunterricht in Wissenschaft und Schule*, *35*(1), 83–93.

Tajmel, T. (2013). Bildungssprache im Fach Physik. In I. Gogolin, I. Lange, U. Michel, & H. H. Reich (Hrsg.), *Herausforderung Bildungssprache - und wie man sie meistert* (S. 239–256). Münster: Waxmann.

Tajmel, T. (2017a). Die Bedeutung von ‚Alltagssprache' - eine physikdidaktische Betrachtung. In B. Lütke, I. Petersen, & T. Tajmel (Hrsg.), *Fachintegrierte Sprachbildung. Forschung, Theoriebildung und Konzepte für die Unterrichtspraxis* (S. 253–267). Berlin: De Gruyter.

Tajmel, T. (2017b). *Naturwissenschaftliche Bildung in der Migrationsgesellschaft. Grundzüge einer Reflexiven Physikdidaktik und kritisch-sprachbewussten Praxis.* Wiesbaden: Springer VS.

Tamir, P. (1988). Subject matter and related pedagogial knowledge in teacher education. *Teaching & Teacher Education*, *4*(2), 99–110.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders. Some reflections. *Language Testing*, *30*(3), 403–412.

Tenorth, H.-E. & Tippelt, R. (2007). *BELTZ Lexikon Pädagogik.* Weinheim: Beltz.

Tepner, O., Borowski, A., Dollny, S., Fischer, H. E., Jüttner, M., Kirschner, S., ... Wirth, J. (2012). Modell zur Entwicklung von Testitems zur Erfassung des Professionswissens von Lehrkräften in den Naturwissenschaften. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 7–28.

Terhart, E. (2000). Schüler beurteilen - Zensuren geben. Wie Lehrerinnen und Lehrer mit einem leidigen, aber unausweichlichen Element ihres Berufsalltags umgehen. In S.-I. Beutel & W. Vollstädt (Hrsg.), *Leistung ermitteln und bewerten* (S. 39–50). Hamburg: Bergmann + Helbig.

Terhart, E. (2011). Lehrerberuf und Professionalität: Gewandeltes Begriffsverständnis - neue Herausforderungen. In W. Helsper & R. Tippelt (Hrsg.), *Zeitschrift für Pädagogik 57. Beiheft. Pädagogische Professionalität* (S. 202–224). Weinheim: Beltz.

Thieme, C. & Mavruk, G. (2018). Merkmale der Bildungssprache im Fach Biologie. Bewertungskriterien für die schriftliche Darstellungsleistung in der gymnasialen Oberstufe. *MNU-Journal*, *71*(5), 292–297.

Thomas, S. (2010). Ethnografie. In G. Mey & K. Mruck (Hrsg.), *Handbuch Qualitative Forschung in der Psychologie* (S. 462–475). Wiesbaden: VS Verlag für Sozialwissenschaften.

Thonhauser, J. (2008). Warum (neues) Interesse am Thema ‚Aufgaben'? In J. Thonhauser (Hrsg.), *Aufgaben als Katalysatoren von Lernprozessen. Eine zentrale Komponente organisierten Lehrens und Lernens aus der Sicht von Lernforschung, Allgemeiner Didaktik und Fachdidaktik* (S. 13–26). Münster: Waxmann.

Tillmann, K.-J. & Vollstädt, W. (2000). Funktion der Leistungsbewertung. Eine Bestandsaufnahme. In S.-I. Beutel & W. Vollstädt (Hrsg.), *Leistung ermitteln und bewerten* (S. 27–38). Hamburg: Bergmann + Helbig.

Tomasevški, K. (2001). *Human rights obligations: making education available, accessible, acceptable and adaptable*. Göteborg, Schweden: Novum Grafiska AB.

Torgerson, W. S. (1956). A non-parametric test of correlation using rank orders within subgroups. *Psychometrika*, *21*(2), 145–152.

Trump, S. S. (2015). Mathematik in der Physik der Sekundarstufe II!?. Eine Benennung notwendiger mathematischer Fertigkeiten für einen flexiblen Umgang mit Mathematik beim Lösen physikalisch-mathematischer Probleme im Rahmen der Schul- und Hochschulbildung sowie eine systematische Analyse zur notwendigen Mathematik in der Physik der Sekundarstufe II. Dissertation. Universität Potsdam.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review*, *71*(6), 528–530.

Tversky, A. & Kahnemann, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

van de Wiel, M. A. & Di Bucchianico, A. (2001). Fast computation of the exact null distribution of Spearman's ρ and Page's L statistic for samples with and without ties. *Journal of Statistical Planning and Inference*, *92*, 133–145.

van Dijk, E. & Kattmann, U. (2010). Evolution im Unterricht: Eine Studie über fachdidaktisches Wissen von Lehrerinnen und Lehrern. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 7–21.

van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing Science Teachers' Pedagogical Content Knowledge. *Journal of Research in Science Teaching*, *35*(6), 673–695.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method. A practical guide to modelling cognitive processes*. London, Großbritannien: Academic Press.

VERBI Software. Consult. Sozialforschung. GmbH (2018). MAXQDA 12. Referenzhandbuch. www.maxqda.de/download/manuals/MAX12_manual_ger.pdf. [Letzter Abruf: 5. November 2019].

Victor, A., Elsäßer, A., Hommel, G., & Blettner, M. (2010). Wie bewertet man die p-Wert-Flut?. Hinweise zum Umgang mit dem multiplen Testen. *Deutsches Ärzteblatt*, *107*(4), 50–56.

Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*. Berlin: Logos.

Vogelsang, C. & Reinhold, P. (2013). Zur Handlungsvalidität von Tests zum professionellen Wissen von Lehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 103–128.

Vollmer, H. J. & Thürmann, E. (2010). Zur Sprachlichkeit des Fachlernens: Modellierung eines Referenzrahmens für Deutsch als Zeitsprache. In B. Ahrenholz (Hrsg.), *Fachunterricht und Deutsch als Zweitsprache* (S. 107–132). Tübingen: Gunter Narr Verlag.

von Aufschnaiter, C. (2014). Laborstudien zur Untersuchung von Lernprozessen. In D. Krüger, I. Parchmann, & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 81–94). Berlin: Springer Spektrum.

von Aufschnaiter, C., Cappell, J., Dübbelde, G., Ennemoser, M., Mayer, J., Stiensmeier-Pelster, J., ... Wolgast, A. (2015). Diagnostische Kompetenz. Theoretische Überlegungen zu einem zentralen Konstrukt der Lehrerbildung. *Zeitschrift für Pädagogik*, *61*(5), 738–758.

von Bargen, I. (2017). Formen der Leistungsbewertung im inklusiven Alltag. In B. Lütje-Klose, S. Miller, S. Schwab, & B. Streese (Hrsg.), *Inklusion: Profile für die Schul- und Unterrichtsentwicklung in Deutschland, Österreich und der Schweiz. Theoretisch Grundlagen – Empirische Befunde – Praxisbeispiele* (S. 141–151). Münster: Waxmann.

von Eye, A. (2006). An Alternative to Cohen's κ. *European Psychologist*, *11*(1), 12–24.

von Horváth, Ö. (2017). *Jugend ohne Gott* (10. Aufl.). Frankfurt am Main: Suhrkamp Verlag.

von Polenz, P. (1981). Über die Jargonisierung von Wissenschaftssprache und wider die Deagentivierung. In T. Bungarten (Hrsg.), *Wissenschaftssprache. Beiträge zur Methodologie, theoretischen Fundierung und Deskription* (S. 85–110). München: Fink.

Wagenschein, M. (1970a). Die Sprache im Physikunterricht. In M. Wagenschein (Hrsg.), *Ursprüngliches Verstehen und exaktes Denken II* (S. 158–173). Stuttgart: Klett.

Wagenschein, M. (1970b). Noten. In M. Wagenschein (Hrsg.), *Ursprüngliches Verstehen und exaktes Denken I* (S. 263–265). Stuttgart: Klett.

Wagenschein, M. (1970c). *Ursprüngliches Verstehen und exaktes Denken I*. Stuttgart: Klett.

Wagenschein, M. (1970d). *Ursprüngliches Verstehen und exaktes Denken II.* Stuttgart: Klett.

Wagenschein, M. (1976). *Die pädagogische Dimension der Physik* (4. Aufl.). Braunschweig: Westermann.

Wagenschein, M. (1983). *Erinnerungen für morgen. Eine pädagogische Autobiographie.* Weinheim: Beltz.

Wagenschein, M. (1986). *Die Sprache zwischen Natur und Naturwissenschaft.* Marburg: Jonas Verlag.

Wang, J.-R., Kao, H.-L., & Lin, S.-W. (2010). Preservice teachers' initial conceptions about assessment of science learning: The coherence with their views of learning science. *Teaching and Teacher Education*, *26*(3), 522–529.

Webb, N. L. (2002). Assessment Literacy in a Standards-Based Urban Education Setting. Paper presented at the 2002 Annual Meeting of the American Educational Research Association. New Orleans, Louisiana, USA.

Wegner, D. M. (1995). A computer network model of human transactive memory. *Social Cognition*, *13*(3), 319–339.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, *92*(4), 548–573.

Weinert, F. E. (2001a). Concept of competence. A conceptual clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–65). Seattle, Washington, USA: Hogrefe & Huber.

Weinert, F. E. (2001b). Vergleichende Leistungsmessungen in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–32). Weinheim: Beltz.

Weinert, F. E. & Schrader, F.-W. (1986). Diagnose des Lehrers als Diagnostiker. In H. Petillon, J. W. L. Wagner, & B. Wolf (Hrsg.), *Schülergerechte Diagnose. Theoretische und empirische Beiträge zur Pädagogischen Diagnostik* (S. 11–29). Weinheim: Beltz.

Wellington, J. & Osborne, J. (2001). *Language and Literacy in Science Education.* Buckingham, Großbritannien: Open University Press.

White, E. (2009). Are you assessment literate? Some fundamental questions regarding effective classroom-based assessment. *OnCUE Journal*, *3*(1), 3–25.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2016). *Engineering Psychology and Human Performance* (4. Aufl.). New York, New York, USA: Routledge.

Wierzioch, K.-D. (2018). www.mathe-physik-aufgaben.de. [Letzter Abruf: 5. November 2019].

Willems, K. (2007). *Schulische Fachkultur und Geschlecht. Physik und Deutsch - natürliche Gegenpole?* Bielefeld: transcript Verlag.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen.* Göttingen: Hogrefe.

Wodzinski, C. T. (2007). Differenzierte Leistungsbewertung. Grundlegende lnformationen und pratkische Vorschläge. *Unterricht Physik*, *99/100*(18), 70–77.

Woolfolk Hoy, A., Davis, H. A., & Pape, S. J. (2006). Teacher knowledge and beliefs. In P. A. Alexander & H. Winne Philip (Hrsg.), *Handbook of educational psychology* (2. Aufl., S. 715–737). New York, New York, USA: Routledge.

Wright, P. (1974). *The language of British industry*. London, Großbritannien: Macmillan.

Würffel, N. (2001). Protokolle Lauten Denkens als Grundlage für die Erforschung von hypertextgeleiteten Lernprozessen im Fremdsprachenunterricht. In A. Müller-Hartmann & M. Schocker-v. Ditfurth (Hrsg.), *Qualitative Forschung im Bereich Fremdsprachen lehren und lernen* (S. 163–186). Tübingen: Narr Francke Attempto Verlag.

Xu, Y. & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149–162.

Yore, L. D., Pimm, D., & Tuan, H.-L. (2007). The Literacy Component of Mathematical and Scientific Literacy. *International Journal of Science and Mathematics Education*, *5*(4), 559–589.

Appendix

A. Kriterienraster der Entwicklungsstudie

A.1. Kriterienraster für die fachlich-konzeptuelle Qualität eines Schülerlösungstextes

Im Folgenden findet sich das Kriterienraster, das im Rahmen der Entwicklungsstudie entwickelt wurde, um Schülerlösungstexte zur Klassenarbeitsaufgabe Weltraumspaziergang bezüglich ihrer fachlich-konzeptuellen Qualität voneinander unterscheiden zu können. Wie in Abschnitt 5.3.2 beschrieben, wurde dieses Kriterienraster in einem deduktiv-induktiven Vorgehen entwickelt. Als Grundlage hierfür dienten die Arbeiten von Kang et al. (2014) und Braaten & Windschitl (2011), sowie die „SOLO-Taxonomie“ von Biggs & Collis (1982).

Für die Anwendung des Kriterienrasters gelten folgende Regeln:

1. Jedem Schülerlösungstext wird in jedem der vier Kriterien („Rahmenbau der Erklärung“, „Rolle von Evidenzbezügen in der Erklärung“, „Tiefe der Erklärung“, „Konsistenz der Erklärung“) genau eine Ausprägung zugewiesen.
2. Das gesamte Material ist kriterienweise zu codieren. Das heißt allen zu codierenden Schülerlösungstexten wird zunächst jeweils eine Ausprägung des Kriterium „Rahmenbau der Erklärung“ zugewiesen, danach wird allen Schülerlösungstexten jeweils eine Ausprägung des Kriteriums „Rolle von Evidenzbezügen in der Erklärung“ zugewiesen, und so weiter.
3. Bei jedem der vier Kriterien ist jeder Ausprägung ein Score zugewiesen. Die Gesamteinschätzung der fachlich-konzeptuellen Qualität eines Schülertextes ergibt sich als Summenscore, in dem der Score für das Kriterium „Konsistenz“ mit dem Faktor 3 gewichtet ist. Für die Bildung des Summenscores gilt also folgende Formel:

$$\underset{\text{fachlich}-\text{konzeptuelle Qualität}}{\text{Summenscore}} = \underset{\text{Rahmenbau}}{\text{Score}} + \underset{\text{Evidenzbezüge}}{\text{Score}} + \underset{\text{Tiefe}}{\text{Score}} + 3 \cdot \underset{\text{Konsistenz}}{\text{Score}}$$

Rahmenbau der Erklärung

Allgemein	Ausprägungen	Beispiele
Der Rahmenbau ist ein Indikator für das Denkmuster, das der Bearbeitung einer Aufgabe zugrunde gelegen hat. Es wird **nicht** berücksichtigt, ob der Schülerlösungstext fachlich konsistent ist (wird in der Kategorie „Konsistenz“ berücksichtigt). Ausprägung **(R1) Erzählend:** Erklärung in Form einer Geschichte oder Nacherzählung über ein Phänomen und / oder Erklärung, die nicht / kaum empirie- und / oder theoriegeleitet erscheint. Ausprägung **(R2) Konstruierend**: Erklärung, die empirie- und / oder theoriegeleitet erscheint und im argumentierenden, generalisierenden Duktus vorgetragen wird.	**(R1) Erzählend (Score = 0)** • Der Schülerlösungstext muss nicht konsistent sein (wird in der Kategorie „Konsistenz“ berücksichtigt). • 1. Fall: **(mindestens 1 Kriterium muss erfüllt sein)** ○ Erklärgerüst basierend auf umformulierten Teilen des Aufgabenstamms, evtl. auch Bruchstücke memorierter Merksätze (z. B. „Energie ist die Fähigkeit Arbeit zu verrichten“). ○ Erzählender Charakter ist dominant. • 2. Fall: **(mindestens 1 Kriterium muss erfüllt sein)** ○ Text lässt sich nicht / kaum als empirie- noch theoriegeleitet einstufen. Z. T. werden unsichtbare Mechanismen oder physikalische Konzepte als etwas Wahrnehmbares dargestellt. ○ Beobachtungen und messbare Größen werden nicht / kaum unterschieden.	*„Der Kollege drückt den Helm näher an den Kopf des Astronauten. Dadurch hört der Astronaut seinen anderen Astronauten leise aber er hört ihn.“* *„Die Schallwellen werden durch den Helm übertragen. so kann man leise hören, weil die Helme sich berühren. Der eine Astronaut kann nicht gehört werden, weil sich die Helme nicht berühren.“*
	(R2) Konstruiert (Score = 2) • Der Schülerlösungstext muss nicht konsistent sein (wird in der Kategorie „Konsistenz“ berücksichtigt). • 1. Fall: Empiriegeleitete Erklärung **(mindestens das erste Kriterium muss erfüllt sein)** ○ Argumentation auf der Grundlage empirischer Daten, Erfahrungen oder naturwissenschaftlichen Sachwissens, das dem Aufgabenstamm entnommen oder darüber hinausgeht und sinnvoll mit der Aufgabe in Verbindung gebracht worden ist. ○ Generalisierender Charakter ist dominant. • 2. Fall: Theoriegeleitete Erklärung **(mindestens das erste Kriterium muss erfüllt sein)** ○ Erklärung eines Sachverhalts durch Anwendung physikalischen Sachwissens. ○ Generalisierender Charakter ist dominant.	*„Im Universum gibt es keinen Sauerstoff, somit auch keine Geräusche. Im Helm schon.“* *„Es liegt an den Schallwellen weil im Weltraum keine Luft ist können keine Schallwellen geleitet werden“* *„Ton wird durch Schall weitergegeben. Teilchen schlagen gegeneinander und leiten den Ton weiter. Je dichter das Material desto besser wird es geleitet. Im All ist ein Vakuum und Ton kann nicht geleitet werden. Da sich die Helme berühren kann der Schall durch das Glas geleitet werden.“*

Rolle von Evidenzbezügen in der Erklärung

Allgemein	Ausprägungen	Beispiele
Erklärungen können durch Bezüge auf Evidenzmittel gestützt werden. Evidenzmittel sind Messdaten oder naturwissenschaftliches Sachwissen, aber auch Alltagsbeispiele, Meinungen oder eine anekdotenhafte Wiedergabe von Einzelbeispielen. Evidenzmittel sind **angemessen**, wenn sie für die genannten Fakten oder Befunde **relevant** sind. Evidenzmittel sind **hinlänglich**, wenn sie **zur Rechtfertigung** einer Aussage **hinreichend und glaubwürdig** erscheinen. Es werden drei Ausprägungen unterschieden. **Dabei ist zu beachten**, dass auch Sachverhalte erklärt worden sein können, die **nicht oder kaum** im Fokus der Aufgabenstellung stehen (wird in der Kategorie „Konsistenz“ berücksichtigt). **Entscheidend ist** also nur die Angemessenheit und Hinlänglichkeit der Evidenzmittel, selbst wenn sie eine fachlich problematische Erklärung stützen sollten.	**(E1) Erklärung ohne Evidenzbezüge (Score = 0)** • Es werden keine Bezüge auf Evidenzmittel zur Stützung von Aussagen eingesetzt.	*„Schall ist, wenn sich in der Luft Töne ausbreiten.“* *„Das macht keinen Sinn. Das zweite Phänomen sehe ich nicht.“*
	(E2) Erklärung mit leichten Evidenzbezügen (Score = 1) • Evidenzmittel werden verwendet, sind aber entweder nicht/kaum angemessen oder nicht/kaum hinlänglich oder beides nicht/kaum. • **Mindestens 1 Kriterium muss erfüllt sein:** ○ Es wird lediglich auf Vorgänge, Fakten oder Befunde verwiesen, die vorwiegend aus dem Alltagskontext stammen. ○ Die Verbindung verwendeter Evidenzmittel zur Erklärung ist schwach, z. B. wenn lediglich auf Tätigkeiten/Fakten/Befunde hingewiesen wird, ohne dabei zentrale Muster in den Tätigkeiten/Fakten/Befunden und deren Zusammenhang mit dem erklärten Sachverhalt darzustellen.	*„Weil man durch die Schwerkraft im Weltraum sich vom Weiten nicht mehr richtig hört aber wenn man näher rankommt kann man sich ein bisschen hören, weil man dann auch sehr laut spricht.“* *„Wegen den Schallwellen. Wenn Sie sich berühren gelangen die Schallwellen den anderen.“* *„Der Kollege drückt den Helm näher an den Kopf des Astronauten. Dadurch hört der Astronaut seinen anderen Astronauten leise, aber er hört ihn.“*
	(E3) Erklärung mit starken Evidenzbezügen (Score = 2) • Angemessene und hinlängliche Verwendung von Evidenzmitteln. • **Mindestens 1 Kriterium muss erfüllt sein:** ○ Verwendung von Daten-, Beobachtungs- oder Tätigkeitsmustern zur Erklärung, die vorwiegend aus fachlichen Kontexten stammen und / oder generalisierten Charakter haben. ○ Evidenzmittel stehen in unmittelbarem Bezug zum Sachverhalt, der erklärt wird.	*„Er hört ihn nicht, weil er einen Helm auf hat und draußen keine Luft ist, wodurch der Schrei getragen werden kann. Wenn sie aneinander halten können sie die Vibrationen spüren und leise etwas hören.“* *„Im Universum gibt es keinen Sauerstoff , somit auch keine Geräusche. Im Helm schon.“*

Tiefe der Erklärung

Allgemein	Ausprägungen	Beispiele
Die Tiefe der Erklärung gibt an, ob ein Sachverhalt strukturiert und detailliert dargestellt wird und ob eine Erklärung auch tatsächlich erklärt, wie und warum etwas geschieht. Für die Tiefe einer Erklärung wird **nicht** berücksichtigt, ob der Schülerlösungstext konsistent ist (wird in der Kategorie „Konsistenz“ berücksichtigt). Es werden drei Ausprägungen unterschieden.	**(T1) „Was“-Erklärung (Score = 0)** • 1. Fall: Ein Sachverhalt wird lediglich beschrieben, ohne dass Ursachen angegeben oder angedeutet werden. • 2. Fall: „Erklärung per Erläuterung“ **(mindestens 1 Kriterium muss erfüllt sein)** ○ Es werden lediglich Begriffe oder Konzepte genannt und evtl. erläutert, ohne dass ein Bezug zu einem Sachverhalt hergestellt wird. ○ Es werden lediglich Erklärungs- oder Problemlösestrategien genannt oder beschrieben, die zur Lösung der Aufgabe herangezogen worden sind.	*„Weil diese zwei Helme so nah gegeneinander kommen.“* *„Schall ist, wenn sich in der Luft Töne ausbreiten.“* *„Das macht keinen Sinn. Das zweite Phänomen sehe ich nicht.“*
	(T2) Einfache Kausalerklärung oder „Wie“-Erklärung (Score = 1) • Die hergestellten Ursache-Wirkungs-Zusammenhänge müssen nicht konsistent sein (wird in der Kategorie „Konsistenz“ berücksichtig). • 1. Fall: „unistrukturaler Fall“ **(alle Kriterien müssen erfüllt sein)** ○ Die Erklärung basiert auf einem kausalen Zusammenhang (z. B. aus Phänomen 1 folgt Phänomen 2). ○ Auf zugrundeliegende Mechanismen oder Prinzipien (z. B. was die Beobachtung fundamental beeinflusst) wird nicht/kaum eingegangen. ○ Keine probabilistischen Aussagen in der Erklärung (z. B. ein Phänomen tritt mit einer bestimmten Wahrscheinlichkeit ein). • 2. Fall: „multistrukturaler Fall“: **(alle Kriterien müssen erfüllt sein)** ○ Analog zum ersten Fall, allerdings fokussiert die Erklärung zwei (oder mehrere) kausale Zusammenhänge. ○ Jeder dieser Zusammenhänge beschreibt die Beziehung zwischen zwei beobachtbaren Teilphänomenen des Sachverhalts. Dies geschieht allerdings isoliert voneinander, d. h. die hergestellten Zusammenhänge bleiben unverknüpft.	*„Der Schall überträgt sich über die Helme“* *„Im Weltraum wird kein Schall weitergeleitet. Als sie die Helme aneinanderdrücken, leiten die Helme die Schallwellen weiter“*
	(T3) „Warum“-Erklärung (Score = 2) • Die Erklärung muss nicht konsistent sein (wird in der Kategorie „Konsistenz“ berücksichtig). • **Alle Kriterien müssen erfüllt sein:** ○ Es wird ein vollständiger Kausalzusammenhang angegeben. Es werden zur Erklärung modellbezogene, theoretische und/oder probabilistische Ereignisse/Prozesse angeführt. ○ Zusammenhänge zwischen Teilphänomenen werden strukturiert und nicht isoliert voneinander dargestellt.	*„Ton wird über Schall weitergegeben. Teilchen schlagen gegeneinander und leiten den Ton weiter. Je dichter das Material desto besser wird es geleitet. Im All ist ein Vakuum und Ton kann nicht geleitet werden. Da sich die Helme berühren kann der Schall durch das Glas geleitet werden.“*

Konsistenz der Erklärung

Allgemein	Ausprägungen	Beispiele
Erklärungen werden hinsichtlich ihrer internalen (Zusammenhalt betreffend) und externalen Konsistenz (Widerspruchsfreiheit betreffend) eingeschätzt. Eine Erklärung ist **internal konsistent**, wenn Ursache-Wirkungsbeziehungen lückenlos sind und sie Phänomene oder **Sachverhalte, die im Fokus der Aufgabenstellung stehen**, erklärt. Eine Erklärung ist **external konsistent**, wenn sie mit naturwissenschaftlichen Begriffen und Konzepten nicht im Widerspruch steht. Es werden drei Ausprägungen unterschieden.	**(K1) inkonsistente Erklärung (Score = 0)** **(Mindestens 1 Kriterium muss erfüllt sein)** ○ Die Erklärung bezieht sich auf Phänomene / Sachverhalte, die nicht/kaum im Fokus der Aufgabenstellung stehen. Die Antwort ist **vor allem internal inkonsistent**. Sie kann **zusätzlich external inkonsistent** sein. ○ Die Erklärung umfasst Begriffe und Konzepte, die in keinem fachlich richtigen Zusammenhang mit dem Sachverhalt der Aufgabenstellung stehen. Die Antwort ist **external inkonsistent** (z.B. einen Regenbogen mit Gravitation erklären). ○ Die Erklärung wirkt unzusammenhängend und bruchstückhaft und / oder es werden (Teil-)Informationen über den zu erklärenden Sachverhalt aufgelistet.	*„Im Weltraum werden die Schallwellen anders geleitet Dadurch dass die Funkverbindung besser wird desto näher man einander ist und im Weltraum ist die Funkverbindung sehr schlecht“* *„Weil man durch die Schwerkraft im Weltraum sich von Weitem nicht mehr richtig hört aber wenn man dann näher rankommt kann man sich ein bisschen hören, weil man dann auch sehr laut spricht.“* *„Weil diese Helme so nah gegeneinander kommen.“*
	(K2) partiell konsistente Erklärung (Score = 1) **(Mindestens 1 Kriterium muss erfüllt sein)** ○ Die Erklärung ist **teilweise korrekt**. **Entweder...** a) Teile der Erklärung sind **external inkonsistent**. Sie sind **übergeneralisiert** und daher fachlich falsch. **und / oder** b) Teile der Erklärung sind **external konsistent aber internal inkonsistent**: Sie sind i. Allg. fachlich richtige Zusammenhänge zwischen Phänomenen / Sachverhalten, die im Fokus der Aufgabenstellung stehen. Diese Zusammenhänge erklären aber nicht das, was in der Aufgabe erklärt werden soll. Sie sind **irrelevant** für die Aufgabenstellung (z. B. hängt Schallausbreitung i. Allg. mit dem Abstand zwischen Sender und Empfänger zusammen, erklärt aber nicht, warum Schallausbreitung im Vakuum nicht möglich ist). ○ Die Erklärung ist **unvollständig**. Teile des zu erklärenden Sachverhalts werden nicht aufgeklärt. Die Erklärung ist **internal inkonsistent**.	*„Im Universum gibt es keinen Sauerstoff, somit auch keine Geräusche. Im Helm schon.“* *„Der Schall breitet sich aus und je näher man aneinander ist desto einen kürzeren weg hat der Schall, er breitet sich über das Glas des Helmes aus.“* *„Der Schall überträgt sich über die Helme“*
	(K3) konsistente Erklärung (Score = 2) **(Beide Kriterien müssen erfüllt sein)** ○ Die Erklärung ist lückenlos und umfassend (**internal konsistent**). Sie bezieht sich auf Phänomene / Sachverhalte die im Fokus der Aufgabenstellung stehen. ○ Die Erklärung steht nicht im Konflikt mit relevanten naturwissenschaftlichen Begriffen und Konzepten (**externale Konsistenz**).	*„Im Weltraum gibt es keine Luft die den Schall transportiert. Legt man aber die Helme aneinander, übertragen diese den Schall und werden in Schwingung versetzt.“*

A.2. Kriterienraster für die Qualität der sprachlichen Realisierung eines Schülerlösungstextes

Im Folgenden findet sich das Kriterienraster, das im Rahmen der Entwicklungsstudie entwickelt wurde, um Schülerlösungstexte zur Klassenarbeitsaufgabe Weltraumspaziergang bezüglich der Qualität ihrer sprachlichen Realisierung voneinander unterscheiden zu können. Wie in Abschnitt 5.3.2 beschrieben, wurde dieses Kriterienraster in einem deduktiv-induktiven Vorgehen entwickelt. Als Grundlage hierfür diente das von der FörMig-Initiative entwickelte Raster zur Beobachtung und Analyse bildungssprachlicher Fähigkeiten von Schüler_innen im natur- und sozialwissenschaftlichen Unterricht für die Sprachhandlung „Erklären“ (vgl. Lengyel et al., 2009, S. 135 u. f.; Roth et al., 2011, S. 8 u. f.; Lengyel & Roth, 2012), das um Aspekte aus den Arbeiten von Koch & Oesterreicher (1985), sowie weiteren linguistischen Merkmalen für Sprachgebrauch im Physikunterricht (vgl. Fluck, 1997, S. 35 u. f.; Tajmel, 2011) ergänzt wurde.

Für die Anwendung des Kriterienrasters gelten folgende Regeln:

1. Jedem Schülerlösungstext wird in jedem der beiden Kriterien („Lexik/Semantik“, „Syntax/Stilistik“) genau eine Ausprägung zugewiesen.
2. Für die Anwendung des Kriterienrasters spielen Rechtschreib- und Satzzeichenfehler keine Rolle.
3. Das gesamte Material ist kriterienweise zu codieren. Das heißt allen zu codierenden Schülerlösungstexte wird zunächst jeweils eine Ausprägung des Kriterium „Lexik/Semantik“ zugewiesen, danach werden allen Schülerlösungstexten jeweils eine Ausprägung des Kriteriums „Syntax/Stilistik“ zugewiesen.
4. Bei jedem der beiden Kriterien ist jeder Ausprägung ein Score zugewiesen. Die Gesamteinschätzung der Qualität der sprachlichen Realisierung eines Schülertextes ergibt sich als Summenscore:

$$\underset{\textbf{Qualität der sprachlichen Realisierung}}{\textbf{Summenscore}} = \underset{\textbf{Lexik / Semantik}}{\textbf{Score}} + \underset{\textbf{Syntax / Stilistik}}{\textbf{Score}}$$

Lexik/Semantik

Allgemein	Ausprägungen	Beispiele
In 3 Ausprägungen wird unterschieden, inwieweit die Ausdrucksweisen in einem Schülerlösungstext dem Sprachgebrauch im Physikunterricht entsprechen. Besonders berücksichtigt werden dabei folgende Kriterien: Eine präzise und angemessene sprachliche Darstellung von Phänomenen / Sachverhalten besteht aus verständlichen und treffenden (fachsprachlichen) Begriffen und Formulierungen. Hierzu werden neben geeigneten **Nomen** (Nennwörter) und **Verben** (Tätigkeitswörter) auch geeignete **Adjektive** (Eigenschaftswörter), **Adverbien** (Umstandswörter), **Präpositionen** (Verhältniswörter) und **Nominalisierungen** (zu Nomen umgeformte Wörter) benötigt. Im fachlich adäquaten Sprachgebrauch kommt es weniger darauf an, Fachnomen zu kennen, sondern diese im Sinne einer fachsprachlichen Norm durch weitere normadäquate Wörter als **Nomen-Wort-Verbindungen** in ein Satzgefüge einbetten zu können. Deshalb orientieren sich die drei Ausprägungen dieser Kategorie in großen Teilen daran, ob die von der Schüler_in verwendeten Nomen-Wort-Verbindungen Ausdrucksweisen im Sinne einer fachsprachlichen Norm des Physikunterrichts sind oder nicht sind. **Stets entscheidend ist, welche Kriterien den Schülerlösungstext dominieren**.	**(L1) Alltagswortschatz mit mündlichem Charakter (Score = 0)** • **Alle folgenden Kriterien müssen erfüllt sein:** ○ **Geläufiger Alltagswortschatz**: Es werden verständliche und treffende Begriffe und Formulierungen verwendet, wie sie aus dem alltäglichen Sprachgebrauch bekannt sind. ○ Die Ausdrucksweisen haben einen **mündlichen Charakter**, z. B. *„Er hört ihn nicht weil er einen Helm auf hat...“*. • **Folgende Kriterien können erfüllt sein. Wenn sie auftreten, sind sie starke Indikatoren für Ausprägung (L1):** ○ Verwendung von **Näherungsbergriffen** um Dinge und / oder Vorgänge andeutungsweise zu bezeichnen. Durch diese Näherungsbegriffe wird ein Sachverhalt, nur angenähert bzw. allgemein bezeichnet. Näherungsbegriffe sind z. B. Nomen in Verbindung mit häufig verwendeten, in ihrer Bedeutung wenig spezifische Verben, Adjektiven, Adverbien oder Präpositionen, z. B. *„... weil der Schall nicht weit kommt“*, *„...weil der Gehörgang anders ist“*, *„Da Glas gegen Glas ist...“*. ○ Verwendung alltagssprachlicher Ausdrücke, die wie **mündliche Einsprengsel** in einen schriftlichen Text wirken, z. B. *„... weil die Stimme schallt“*, *„Wir haben in ein Glas Calciumkörner reingeschmissen“*. ○ Dinge und Vorgänge werden **umschrieben**, anstatt konkret bezeichnen zu werden, z. B. *„...weil dort keine Luft ist“* anstatt *„im Weltraum herrscht Vakuum“*. • **Folgende Kriterien können erfüllt sein. Im Vergleich zu den übrigen Indikatoren dürfen diese (vor allem in längeren Schülerlösungstexten) nur vereinzelt auftreten:** ○ Verwendung spezifischer Begriffe, die aus dem Aufgabenstamm übernommen worden sind, z. B. Nomen wie „*Funkverbindung*“, „*Astronaut*“, „*Kollege*“ oder Verben wie „*pressen*“, „*abbrechen*“. ○ Verwendung allgemein **gebräuchlicher Fachnomen**, die in den Alltagswortschatz eingegangen sind, wie z. B. „*Schall*“, „*Funk*“, „*Schallwellen*“.	*„Der eine kann den anderen durch die Schallwellen hören die durch den Helm kommen, wenn der eine schreit halt das durch die Helme wieder und wenn es dann an den Helm des anderen Astronauten kommt schallt es durch die Helme wieder“* *„Er hört ihn erst nicht, weil der Schall im Weltraum gestoppt wird. Wenn die Helme ganz dicht aneinder sind, dann wird der Schall noch nicht vollständig gestoppt, und er versteht ihn.“*
	(L2) „elaborierter“ Alltagswortschatz (Score = 1) • **Alle folgenden Kriterien müssen erfüllt sein:** ○ Verwendung eines **Alltagswortschatzes**, der aber **deutlich differenzierter und anspruchsvoller ist als der in Ausprägung (L1)**, z. B. *„...die Stimme kann sich ausbreiten“*, *„...und Schall kann sich verbreiten“*. ○ Die Ausdrücke sind präzise und kontextspezifisch. ○ Die Ausdrucksweise **setzt sich deutlich von einer Mündlichkeit ab**. ○ Verwendung **allgemein gebräuchliche Fachnomen**, die in den Alltagswortschatz eingegangen sind, z. B. „*Schall*“, „*Funk*“, „*Schallwellen*“, „*Weltall*“, „*Töne*“. • **Folgende Kriterien können erfüllt sein. Sie beschreiben Übergangsphänomene zur Ausprägung (L3). Sie sind starke Indikatoren für Ausprägung (L2), wenn sie häufiger auftreten (zählen!) als Kollokationen, Funktionsverbgefüge und Komposita (siehe letztes Kriterium):** ○ Verwendung von Verben, Adjektiven, Adverbien, Präpositionen oder Nominalisierungen in einer **Nomen-Wort-Verbindung**. Diese Worte sind aus fachsprachlicher Sicht präzise und kontextspezifisch angemessen. Allerdings sind diese Worte mit einem Nomen oder einer Nominalisierungsgruppe aus dem Alltagswortschatz verbunden. **Es entsteht eine Nomen-Wort Verbindung, die aus Gründen einer fachsprachlichen Norm nicht gängig ist**. Auf den_die Leser_in wirkt sie daher „**schief**“. Sinnentstellungen sind möglich. Substituiert man das Alltagsnomen / die Nominalisierungsgruppe durch ein geeignetes Fachnomen, entsteht eine im Sinne einer fachsprachlichen Norm adäquate Kollokation / ein Funktions-Verb-Gefüge z. B. wenn im Nebensatz *„...wodurch der Schrei getragen werden kann.“* das Nomen „*Schrei*“ durch „*Schall*“ ersetzt.	*„Der jüngere kann den älteren hören, da im Weltraum Vakuum enthalten ist. Im Helm ist Luft und durch das Vakuum kann das Geräusch hörbar sein.“* *„Die Astronauten können sich nicht hören, weil sie beide einen Helm tragen und nur in dem Helm Sauerstoff ist. Wenn die Helme aneinander gepresst sind ist nur noch die Hülle des Helms dazwischen und nichts was die Geräusche komplett blocken würde“*

Lexik/Semantik (Fortsetzung)

Allgemein	Ausprägungen	Beispiele
	○ Verbindung eines allgemein gebräuchlichen Fachnomens mit Verben, Adjektiven, Adverbien, Präpositionen oder Nominalisierungen zu einer **Nomen-Wort-Verbindung**. Die einzelnen Worte sind aus fachsprachlicher Sicht präzise und kontextspezifisch angemessen. **Es entsteht eine Nomen-Wort-Verbindung, die aus Gründen einer fachsprachlichen Norm nicht gängig ist**. Auf den_die Leser_in wirken sie „**schief**“ oder „**überkorrekt**“. Auch hier sind Sinnentstellungen möglich. Analog zum ersten Fall entsteht durch Substitution geeigneter Wörter eine im Sinne einer fachsprachlichen Norm adäquate Kollokation / ein Funktions-Verb-Gefüge, z. B. wenn im Nebensatz *„... da im All Vakuum enthalten ist“* *„enthalten ist“* durch *„herrscht“* ersetzt wird. • **Folgendes Kriterium kann erfüllt sein. Es ist ein starker Indikator für Ausprägung (L2), wenn es seltener auftritt (zählen!) als die oben beschriebenen Übergangsphänomen:** ○ Verwendung von bezüglich einer fachsprachlichen Norm gängigen **Kollokationen**, **Funktionsverbgefügen** und **Komposita** (siehe Ausprägung (L3)).	
	(L3)Fachwortschatz des Physikunterrichts (Score = 2) • **Alle folgenden Kriterien müssen erfüllt sein:** ○ Verwendung **spezifischer fachsprachlicher Bezeichnungen oder Begriffe aus dem Physikunterricht**. ○ Die Ausdrucksweise zeigt, dass die fachliche Sprache zielgerichtet eingesetzt werden kann. • **Folgende Kriterien können erfüllt sein. Sie sind starke Indikatoren für Ausprägung (L3), wenn sie häufiger auftreten (zählen!) als die in Ausprägung (L2) beschriebenen Übergangsphänomenen:** ○ Es wird ein Fachnomen oder Nomen mit Verben, Adjektiven, Adverbien, Präpositionen oder Nominalisierungen, verbunden. Die einzelnen Worte sind aus fachsprachlicher Sicht präzise und kontextspezifisch angemessen. Bei diesen Nomen-Wort-Verbindungen handelt es sich um im Sinne einer fachsprachlichen Norm gängige **Kollokationen**, z. B. *„die Schallwellen werden geleitet“*, *„die Ausbreitung des Schalls“*, *„die Helme sind elektrisch aufgeladen“*. ○ Analog zum ersten Fall werden sog. **Funktionsverbgefüge** gebildet. Im Gegensatz zu Kollokationen sind diese durch ein Verb substituierbar, z. B. *„eine Forderung stellen – fordern“* ○ Bildung aus fachsprachlicher Sicht präziser und kontextspezifisch angemessener **Komposita**. Hierbei handelt es sich um zusammengesetzte Nomen, Verben, Adjektive, Adverbien oder Präpositionen, z. B. „*Weltraumvakuum*“, „*Reibungsenergie*“, „*luftleer*“. • **Folgende Kriterien können erfüllt sein. Sie sind starke Indikatoren für Ausprägung (L3), wenn sie in dem Schülerlösungstext seltener auftreten (zählen!) als Kollokationen, Funktionsverbgefüge und Komposita (siehe vorletztes Kriterium):** ○ Die Ausdrucksweise entspricht nicht durchgängig einer Fachsprache, wie sie in Fachtexten eingesetzt wird. ○ Auftreten von **Übergangsphänomenen** (beschrieben in Ausprägung (L2)).	*„Das erste Phänomen ist so zu erklären: In dem All herrscht Leere und so auch keine Luft die die Schallwellen leiten könnte. Das zweite Phänomen auch: Dadurch dass die Astronauten ihre Helme aneinanderhalten leitet das Glas die Schwingungen. Dies erfolgt zwar nur Schwach, aber es erfolgt.“* *„Da im Weltraum ein Vakuum herrscht entsteht kein Schall und man kann nichts hören, berühren sich jedoch beide Helme welche mit Luft gefüllt sind kann die Vibration des Schalls über die Helme übertragen werden.“*

Syntax/Stilistik

Allgemein	Ausprägungen	Beispiele
Es werden 3 Ausprägungen unterschieden. **Stets entscheidend ist, welche Kriterien den Schülerlösungstext dominieren**. Auf syntaktischer Ebene steht im Fokus, ob Aussagen sprachlich miteinander verknüpft sind und welche Arten von **satzverbindenden Elementen** hierzu eingesetzt werden. Hinzu kommt die Verwendung sprachliche Mittel, die einer **syntaktische Komplexitätsreduktion** dienen. Auf stilistischer Ebene wird zwischen einer **Sprache der Nähe** und einer **Sprache der Distanz** unterschieden. Hierbei sind Sprache der Nähe und Sprache der Distanz **idealisierte Gegenpole mit fließendem Übergang**.	**(S1) unvollständig/einfache/Reihungen von Sätzen; Sprache der Nähe (Score = 0)** • **Sind syntaktische und stilistische Ebene erfüllt, entspricht ein Schülerlösungstext Ausprägung (S1).** **SYNTAKTISCHE EBENE** • **Mindestens eines der folgenden Kriterien muss erfüllt sein:** ○ Verwendung einfacher Sätze, die nicht mit sprachlichen Mitteln verbunden sind. Es bleibt der Aktivität des_der Leser_in überlassen, diese Verbindung herzustellen. ○ Es werden **nur** isolierte Nebensätze ohne Hauptsätze gebildet. Diese Nebensätze können syntaktische und stilistische Merkmale der Ausprägungen (S2) und (S3) tragen. ○ Schematische Reihung von Sätzen: Gedankliche Verbindungen werden nur angedeutet aber nicht genauer bestimmt. Hierzu werden **satzverbindende Elemente** verwendet, die im alltäglichen Sprachgebrauch häufig auftreten. **Hierzu gezählt werden...** ...**häufig auftretende nebenordnende Konjunktionen** (Bindewort), vor allem „*und*“, „*und dann*“, „*und wenn*“ , seltener auch „*oder*“, „*aber*“, **...und häufig auftretende Konjunktionaladverbien** (Umstandswort mit Bindewortfunktion), vor allem „*dann*“, „*da*“, „*wo*“. **STILISTISCHE EBENE** • **Folgendes Kriterium muss erfüllt sein:** ○ Der Schülerlösungstext entspricht tendenziell einer **Sprache der Nähe**. Eine Sprache der Nähe zeichnet sich durch eine **persönliche face-to-face Interaktion**, sowie eine **expressive und affektive Teilnahme der Kommunikationspartner** aus. **Es zeigen sich Indikator a) und / oder b):** **a) Persönliche Wendungen mit Pronomen (Fürwörter):** Verwendung von Pronomen im Rahmen einer persönlichen Wendung, z B. „*...sie können sich...*“, „*...ich glaube...*“. Diese Pronomen stehen stellvertretend für den_die Schüler_in (als Schreiber_in), der_die Leser_in oder beide oder für in der Erklärung thematisierte Personen mit einer eindeutig bestimmten Identität. Alternativ drücken diese Pronomen die Zugehörigkeit / Verbundenheit zwischen mehreren Personen mit eindeutig bestimmter Identität oder einer oder mehrere Personen mit eindeutig bestimmter Identität und einem Objekt aus. **Hierzu gezählt werden...** ...**Personalpronomen** (persönliche Fürwörter), wie „*ich*“, „*mich*“, „*du*“, „*Sie*“, „*dein*“, „*wir*“, „*uns*“, „*er*“, „*sie*“, usw. ...**Possessivpronomen** (besitzanzeigende Fürwörter), wie „*mein*“, „*dein*“, „*seinen*“ „*unseren*“, usw. ...und **Reflexivpronomen** (rückbezügliche Fürwörter), wie „*mich*“, „*dich*“, „*dir*“, „*euch*“, „*sich*“, usw. **Ausgenommen** sind hierbei Pronomen, die das neutrale Geschlecht/Neutrum betreffen, vor allem „*es*“. **b) Zeigewörter und Pronomen ohne explizit im Text versprachlichte Bedeutung:** Verwendung von Zeigewörtern und Pronomen, die auf **personelle**, **temporale** oder **lokale Charakteristika eines Gegenstands** hinweisen. Ihre Bedeutung wird nur, wenn der zugehörige Kontext (z. B. der Aufgabenstamm) bekannt ist, deutlich. Als Beispiel hierfür kann folgende Schülerlösungstext zur Aufgabe Weltraumspaziergang gelten: *„Er hört ihn nicht, weil er einen Helm auf hat und draußen keine Luft ist wodurch der Schrei getragen werden kann. Wenn sie aneinander halten können sie die Vibrationen spüren und leise etwas hören.“* Hier wird erst durch gedankliche Hinzunahme des Aufgabenstamms deutlich, dass mit „*er*“, „*ihn*“ und „*sie*“ die beiden Astronauten und mit „*draußen*“ der Weltraum gemeint ist.	*„Er hört seinen Freund nicht. Er presst die Helme zusammen. Er kann ihn wieder hören.“* *„Weil diese zwei Helme so nah gegeneinander kommen.“* *„Da glas gegen glas ist, können sie sich besser hören.“* *„Da die Helme sich berühren werden die Schallwellen übertragen.“* *„Im All ist nichts und wenn sie sich nicht berühren und zwischen ihnen nichts ist hören sie sich nicht. Aber dann ist zwischen ihnen was und sie hören sich.“*

Syntax/Stilistik (Fortsetztung)

Allgemein	Ausprägungen	Beispiele
	(S2) verbundene Sätze; Sprache der Nähe (Score = 1) • **Sind syntaktische und stilistische Ebene erfüllt, entspricht ein Schülerlösungstext Ausprägung (S2).** **SYNTAKTISCHE EBENE** • **Die letzten beiden Kriterien müssen erfüllt sein. Das erste Kriterium kann zusätzlich erfüllt sein:** ○ Verwendung von **satzverbindenden Elementen**, wie sie in Ausprägung (S1) beschriebenen sind. ○ Es treten keine / nur einzelne Nominalisierungen, Nominalisierungsgruppen oder Komposita auf. ○ Verwendung **weiterer satzverbindender Elemente**. Durch diese sind Aussagen über eine einfache Reihung hinaus miteinander verbunden. Diese satzverbindenden Elemente können sehr **vielfältig** sein. **Hierzu zählen im alltäglichen Sprachgebrauch seltener auftretende...** **...nebenordnende Konjunktionen** (Bindewort), wie „*jedoch*“, „*beziehungsweise*“, usw. ...und **Konjunktionaladverbien** (Umstandswort mit Bindewortfunktion), wie „*somit*“, „*folglich*“, „*deswegen*“, „*deshalb*“, usw. **Hinzu kommen...** ... **kausale** und **konditionale Konjunktionen** (Bindewort), wie „*weil*“, „*wenn*“, „*falls*“, *während*“, „*sofern*“, „*als*“, usw. **...weitere unterordnende Konjunktionen** (Bindewort), wie „*damit*“, „*während*“, „*dass*“, „*sodass*“, „*obwohl*“, „*sobald*“, „*sowie*, „*indem*“, usw. **...Relativpronomen** (bezügliches Fürwort), wie „*..., das...*“, „*..., welches...*“, „*..., was...*“, „*..., dessen...*“, usw. **...mehrteilige Konjunktionen**, wie „*entweder... oder*“, „*je... desto*“, „*sowohl... als auch*“, usw. ...und **Infinitivkonjunktionen,** wie „*um... zu*“, „*anstatt... zu*“, „*außer... zu*“, usw. **STILISTISCHE EBENE** • **Beide Kriterien müssen erfüllt sein:** ○ Der Schülerlösungstext entspricht tendenziell einer **Sprache der Nähe** (siehe Ausprägung (S1)). ○ **Indikatoren** für eine **Sprache der Distanz** (siehe Ausprägung (S3)) treten **nicht / kaum** auf.	*„Er hört ihn nicht, weil er einen Helm auf hat und draußen keine Luft ist wodurch der Schrei getragen werden kann. Wenn sie aneinander halten können sie die Vibrationen spüren und leise etwas hören.“* *„Ich glaube das die beiden Helme schalldämpfen und deswegen können sie sich nicht hören ich meine hinter einer geschlossenen Glaswand kann mich auch niemand hören.“*

Syntax/Stilistik (Fortsetztung)

Allgemein	Ausprägungen	Beispiele
	(S3) Sprachliche Verdichtung; Sprache der Distanz (Score = 2) • **Damit ein Schülerlösungstext Ausprägung (S3) entspricht, müssen...** ○ **... entweder syntaktische und stilistische Ebene erfüllt sein,** ○ **... oder nur das erste Kriterium auf syntaktischer Ebene und die stilistische Ebene erfüllt sein,** ○ **... oder das letzte Kriterium auf syntaktischer Ebene und das erste Kriterium auf stilistischer Ebene erfüllt sein.** **SYNTAKTISCHE EBENE** • **Mindestens das letzte Kriterium muss erfüllt sein:** ○ Aussagen sind sprachlich miteinander verbunden. Verwendet werden hierzu **satzverbindende Elemente**, wie sie in Ausprägung (S1) und (S2) beschriebenen sind. ○ **Mehrfach** finden sich **Nominalisierungen**, z. B. „*Schwingung*“, „Reibung“, **Nominalisierungsgruppen**, z. B. „*die Ausbreitung des Schalls*“ oder **Komposita** (zusammengesetzte Fachbegriffe), z. B. „*Weltraumvakuum*“, „*Schallschwingungen*“, „Schallwellen“. Durch diese wird eine **syntaktische Komplexitätsreduktion** erreicht, da durch sie Sachzusammenhänge, die sonst nur mit Hilfe eines Verbs oder eines Nebensatzes ausgedrückt werden können, in einem Nomen bzw. einer Aneinanderreihung von Nomen zusammengefasst dargestellt werden. **STILISTISCHE EBENE** • **Folgende Kriterien müssen erfüllt sein:** ○ Indikatoren für eine **Sprache der Nähe** treten **nicht / kaum** auf. ○ Der Schülerlösungstext entspricht tendenziell einer **Sprache der Distanz**. Eine Sprache der Distanz charakterisiert sich durch ihre **Anonymität und die Teilnahmslosigkeit der Kommunikationspartner**. **Es zeigen sich Indikator a) und / oder b) und / oder c):** a) **Generalisierte Wendungen mit Indefinitpronomen (unbestimmte Fürwörter):** Verwendung von **Indefinitpronomen** im Rahmen einer unpersönlichen Wendung, z. B. „*...man kann...*“ . **Hierzu gehören** „*man*“, „*jemand*“, „*keine*“, „*alle*“, „*jede*“, „*eine*“, „*andere*“, usw. Diese Pronomen stehen stellvertretend für eine nicht genauer bezeichnete Person oder Personengruppe. Alternativ können diese Pronomen auch eine Zugehörigkeit oder Verbundenheit zwischen mehreren allgemein gehaltenen Personen oder einer oder mehrerer allgemein gehaltenen Personen und einem oder mehreren Objekten ausdrücken. b) **Zeigewörter und Pronomen mit explizit im Text versprachlichter Bedeutung:** Verwendung von Zeigewörtern und Pronomen, die auf personelle, temporale oder lokale Charakteristika eines versprachlichten Gegenstandes hinweisen. Diese **Charakteristika** wurden **im Voraus bzw. werden nachträglich explizit im Text versprachlicht**. Ihre Bedeutung ist daher im Gegensatz zu den Zeigewörtern und Pronomen aus Stufe (S1), ohne dass Rückschlusse auf den Kontext notwendig sind, eindeutig ersichtlich. Ein Beispiele hierfür sind die folgenden: „*Das erste Phänomen ist das die Astronauten sich nicht hören, dass liegt dahran, dass sie sich nicht berühren. Das 2 Phönomen ist das sie sich wenn sie sich berühren hören. Das liegt an den Schallwellen.*“ „*Es liegt an der Schwerkraft, da die keine Luft ist hört man vom weiten nichts, denn im Weltall gibt es keine Wellen die das Geräusch oder die Stimme ins Ohr schallen lassen.*“ c) **Personale, temporale oder lokale Charakteristika werden durchgehend explizit versprachlicht**. z. T. werden Indefinitpronomen als Artikelwörter eingesetzt, z. B. „*...andere Person...*“.	„*Ich glaube, dass die Töne beim schreien also aus einer größeren entfernung auf dem Weg zur anderen Person verfligen. Sie verteilen sich so schnell das kein Ton mehr bei der anderen Person ankommt.* *Wenn die beiden Astronauten aber ganz dicht aneinander sind haben die Töne einen kürzeren Weg und können sich nicht so schnell verteilen deshalb kommen mehr Töne in dem Ohr des anderen an.* *Das ist wie bei einen Parfum. Wenn man es aus einer großen entfernung sprüht verteilen sich die Teilchen schnell im Raum. Wenn die Entfernung kleiner ist kommen mehr Teilchen auf der Haut an.*“ „*Da im Weltraum ein Vakuum herrscht entsteht kein Schall und man kann nichts hören, berühren sich jedoch beide Helme welche mit Luft gefüllt sind kann die Vibration des Schalls über die Helme übertragen werden.*“

B. Konsenskoeffizient Ξ

Im Folgenden werden theoretische Überlegungen zum *Konsenskoeffizient* Ξ dargestellt, der im Rahmen der Entwicklungsstudie als Maß für die Codiererübereinstimmung bei einem bestimmten Schülerlösungstext hinsichtlich einer bestimmten Ausprägung eines Kriteriums eingesetzt wurde (vgl. Abschnitt 5.3.4).

B.1. Problemaufriss

Einhaus (2007) entwickelte im Rahmen seines Dissertationsprojekts ein Verfahren zur Einschätzung von Itemmerkmalen[158] mittels Expertenbefragung, das in der naturwissenschaftsdidaktischen Forschung inzwischen auch von anderen Autor_innen eingesetzt wurde, um die Operationalisierung theoretischer Konstrukte, sowie die Konstruktion von Testaufgaben empirisch abzusichern (z. B. M. Schmidt, 2008, S. 74 u. f.; Scherer, 2012, S. 89 u. f.; Heitmann, 2013, S. 69 u. f.). In diesem Verfahren wird aus den im Rahmen eines Expertenratings gewonnenen Daten ein Koeffizient errechnet, der die „Beurteilungsübereinstimmung bei einem bestimmten Item hinsichtlich eines bestimmten Merkmals [beschreibt]“ (Einhaus, 2007, S. 33). Bei der Berechnung dieses Koeffizient wird mathematische quantifiziert, inwieweit bei einem Expertenrating bei einem einzelnen Item ein Konsens zwischen den befragten Experten_Expertinnen darüber besteht, ob diesem Item die Ausprägung eines bestimmten Itemmerkmals zuzuweisen ist (vgl. ebd., S. 25 u. f.). Dieses Itemmerkmal wurde dabei von den befragten Experten_Expertinnen auf einer Ratingskala eingeschätzt (vgl. ebd.). Die Übereinstimmung unter den Experten_Expertinnen bezeichnet Einhaus als *Einigkeit* und seinen vorgeschlagenen Koeffizienten dementsprechend als *Einigkeitskoeffizienten* η (vgl. ebd.). Der entscheidende Gedanke hinter diesem Verfahren ist, ein Expertenrating nicht durch die Übereinstimmung, sondern durch die Nichtübereinstimmung der Experten_Expertinnen untereinander zu beschreiben, sowie Kriterien festzulegen, ab wann diese Nichtübereinstimmung bei einem bestimmten Item „gering genug“ ist, um von einem Konsens unter den Experten_Expertinnen zu sprechen (vgl. ebd.).

Am von Einhaus (2007) vorgeschlagenen Verfahren lassen sich allerdings folgende Punkte kritisieren:

[158] In diesem Kontext ist unter dem Begriff „Item“ ein einzelnes, einer Gruppe von Experten_Expertinnen zur Einschätzung vorlegte Untersuchungsobjekt zu verstehen (z. B. eine Testaufgabe oder ein einzelnes *Can do Statement* (vgl. Köller, 2008, S. 164) aus der Operationalisierung einer bestimmten Kompetenz).

1. Das Verfahren von Einhaus (2007) ist nur für den Spezialfall, dass genau 6 Experten_Expertinnen ein Item auf einer 4-stufigen Ratingskala einschätzen, anwendbar. Im Umkehrschluss ist es daher nicht möglich, dieses Verfahren in einem Expertenrating anzuwenden, in dem sich die Anzahl der Experten_Expertinnen von 6 unterscheidet und/oder für ein Rating mit beliebig langer Ratingskala.
2. Das Vorgehen von Einhaus (2007) weist mathematische Lücken auf. Der Autor beschreibt zwar das Vorgehen, wie man in einem konkreten Fall eines Expertenratings zu dem von ihm vorgeschlagenen Einigkeitskoeffizienten gelangt (vgl. ebd., S. 26 u. f.). Was jedoch fehlt ist eine geschlossene Formel zur Berechnung von η.
3. Der Einigkeitskoeffizienten η ist zwar an die „Standardnormierung von Übereinstimmungsmaßen an[gepasst] (0: schlechte Übereinstimmung, 1: gute Übereinstimmung)“ (ebd., S. 29), allerdings gilt für diesen $0 < \eta \leq 1$ (vgl. ebd., S. 30). Wünschenswert wäre hingegen ein Koeffizient, der genau dann den Wert 0 annimmt, wenn das zu betrachtende Expertenrating die geringstmögliche Übereinstimmung aufweist.

Im weiteren Verlauf wird daher eine Verallgemeinerung des von Einhaus (2007) entwickelten Koeffizienten vorgeschlagen, in der zudem die eben benannten mathematischen Lücken geschlossen und Problematiken beseitigt werden. Dieser verallgemeinerte Koeffizient wird, um ihn vom Einigkeitskoeffizienten η nach Einhaus (2007) besser unterscheiden zu können, als *Konsenskoeffizient* Ξ bezeichnet.

B.2. Grundbegriffe und -annahmen

Ein Item soll ein Expertenrating durchlaufen, in dem ihm ein bestimmtes Merkmal von m Experten_Expertinnen zugeordnet werden soll. Für diese Zuordnung steht den Experten_Expertinnen eine w-stufige Ratingskala mit geordneten Skalenwerten $1, 2, ..., w$ zur Verfügung. Um die Experten_Expertinnen voneinander unterscheiden zu können, ist es sinnvoll, jede_n der m Experten_Expertinnen mit einer Nummer von 1 bis m zu versehen. Ferner soll der Skalenwert, den der_die i-te Experte_Expertin (mit $1 \leq i \leq m$) dem Item zuordnet, mit b_i bezeichnet werden. Durch diese Setzungen lässt sich das gesamte Expertenrating als Zeilenvektor $[b_1, b_2, ..., b_m]$ in einer kompakten Form darstellen.

In Anlehnung an Einhaus (2007) sollen in dem im weiteren Verlauf vorgestellten Verfahren zwei Kennwerte für die Nichtübereinstimmung im Expertenrating $[b_1, b_2, ..., b_m]$ berücksichtigt werden. Diese Kennwerte werden in den folgenden beiden Abschnitten vorgestellt.

B.2.1. Anzahl der paarweise nicht übereinstimmenden Einschätzungen

Der erste Kennwert, der berücksichtigt werden soll, ist die Anzahl der Expertenpaaren, die in ihrer Merkmalseinschätzung des Items nicht übereinstimmen. Für alle $i, j \in \{1, 2, ...m\}$ mit $i \neq j$ wird also überprüft, ob $b_i = b_j$ oder $b_i \neq b_j$ gilt. Mit Hilfe der Setzung...

$$\delta_{ij} := \begin{cases} 1, & \text{wenn } b_i \neq b_j \\ 0, & \text{wenn } b_i = b_j \end{cases}$$

... erhält man die Anzahl der paarweise nichtübereinstimmenden Einschätzungen N via...

$$N = \frac{1}{2} \cdot \sum_{i=1}^{m} \sum_{j=1}^{m} \delta_{ij}. \tag{B.1}$$

Man beachte, dass sich der Faktor $\frac{1}{2}$ daraus ergibt, dass in Gleichung B.1 jedes Expertenpaar doppelt gezählt wird, sowie aus der Symmetrieeigenschaft $\delta_{ij} = \delta_{ji}$. Ferner gilt für alle $i \in \{1, 2, ...m\}$ $\delta_{ii} = 0$, weswegen in Gleichung B.1 die Fälle, in denen $i = j$ gilt, nicht zu einer fälschlichen Erhöhung von N führen.

N ist sinnvollerweise ins Verhältnis zur maximal möglichen Anzahl von paarweisen Nichtübereinstimmungen $N_{max(m,w)}$ zu setzen (vgl. Einhaus, 2007, S. 27). Der Quotient $\frac{N}{N_{max(m,w)}}$ liefert dann eine Maßzahl, die Werte zwischen 0 und 1 annimmt, wobei der Wert 1 genau dann angenommen wird, wenn $N = N_{max(m,w)}$ gilt. Der Wert 0 wird genau dann angenommen, wenn der Fall $N = 0$ eintritt, also genau dann, wenn alle m Experten_Expertinnen dem Item einstimmig dieselbe Ausprägung eines bestimmtes Merkmals zugewiesen haben.

Wie die Notation bereits andeutet, hängt $N_{max(m,w)}$ sowohl von der Anzahl an befragten Experten_Expertinnen, als auch von der Länge der Ratingskala ab:

Ist $w \geq m$, so ist es möglich, dass alle befragten Experten_Expertinnen unterschiedliche Einschätzungen abgeben. $N_{max(m,w)}$ ist dann gerade $\frac{1}{2} \cdot m \cdot (m-1)$, also die Anzahl an verschiedenen Expertenpaaren (vgl. Einhaus, 2007, S. 27).

Im Fall $w < m$ können nicht alle Experten_Expertinnen unterschiedliche Einschätzungen abgeben (vgl. ebd.). Die maximale Anzahl der paarweisen Nichtübereinstimmungenen lässt sich in diesem Fall nur schwer direkt bestimmen, ergibt sich aber indirekt aus der Anzahl an verschiedenen Expertenpaaren abzüglich der minimalen Anzahl übereinstimmender Expertenpaare, die als N^* bezeichnet werden soll. N^* kann durch folgende Gleichung berechnet werden:

$$
\begin{aligned}
N^* =& \frac{w - \left(m - w \cdot \lfloor \frac{m}{w} \rfloor\right)}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot \left(\lfloor \frac{m}{w} \rfloor - 1 \right) \\
&+ \frac{m - w \cdot \lfloor \frac{m}{w} \rfloor}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot \left(\lfloor \frac{m}{w} \rfloor + 1 \right)
\end{aligned}
\tag{B.2}
$$

$N_{max(m,w)}$ kann also über folgende Gleichung bestimmt werden:

$$
\begin{aligned}
N_{max(m,w)} =& \frac{1}{2} \cdot m \cdot (m-1) - N^* \qquad \text{(B.3)} \\
=& \frac{1}{2} \cdot m \cdot (m-1) \\
&- \frac{w - \left(m - w \cdot \lfloor \frac{m}{w} \rfloor\right)}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot \left(\lfloor \frac{m}{w} \rfloor - 1 \right) \\
&- \frac{m - w \cdot \lfloor \frac{m}{w} \rfloor}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot \left(\lfloor \frac{m}{w} \rfloor + 1 \right)
\end{aligned}
\tag{B.4}
$$

Die Gleichung B.2 zur Berechnung von N^* ergibt sich aus elementaren kombinatorischen Überlegungen. Für diese Überlegungen sei $[b_1, b_2, ..., b_m]$ ein Expertenrating über ein bestimmtes Merkmal eines Items auf einer w-stufigen Ratingskala, für das $N = N_{max(m,w)}$ gilt. Man nehme an, man besitze beliebig viele Kugeln, die mit den Ziffern 1 bis w beschriftet sind (man besitzt also w verschiedene Kugelsorten) und beliebig viele Urnen. Die Urnen werden allmählich mit Kugeln gefüllt. Dabei wird genau eine Kugel mit der Beschriftung b in eine Urne gelegt, wenn ein_eine Experte_Expertin dem Item den Skalenwert $b \in \{1, 2, ..., w\}$ zugeordnet hat. Zusätzlich gilt die Regel, dass in einer Urne maximal eine Kugel einer bestimmten Sorte platziert werden darf. In jeder Urne haben also maximal w Kugeln Platz und zwei Kugeln mit derselben Beschriftung b liegen in verschiedenen Urnen.

Da in jeder Urne nur eine Kugel jeder Sorte platziert werden darf, werden umso mehr Urnen benötigt, je mehr Experten_Expertinnen dem Item denselben Skalenwert zugeordnet haben. Je größer die paarweise Übereinstimmung der m befragten Experten_Expertinnen ist, desto mehr Urnen werden also insgesamt benötigt. Die minimalen Anzahl von Urnen, die benötigt wird, um m Kugeln nach obigen Regeln zu platzieren, ergibt sich daher genau dann, wenn in dem Expertenrating, auf dessen Grundlage die Kugeln auf Urnen verteilt werden, die minimale Anzahl übereinstimmender Expertenpaare N^* auftritt. Da zu Beginn der Überlegungen das zu betrachtende Expertenrating so gewählt wurde, dass $N = N_{max(m,w)}$ gilt, ist diese Bedingung erfüllt, denn es gilt $N_{max(m,w)} = \frac{1}{2} \cdot m \cdot (m-1) - N^*$.

Die minimale Anzahl von Urnen für m Kugeln, von denen w Sorten zur Verfügung stehen, erhält man, indem $\lfloor \frac{m}{w} \rfloor$ Urnen mit je w verschiedenen Kugeln befüllt werden[159]. Zusätzlich wird eine weitere Urne mit $m - w \cdot \lfloor \frac{m}{w} \rfloor$ verschiedenen Kugeln gefüllt. Ist $\frac{m}{w}$ eine ganze Zahl, bleibt die letzte Urne leer, da in diesem Fall $m - w \cdot \lfloor \frac{m}{w} \rfloor = 0$ gilt.

Die Anzahl an Paaren, die aus den verteilten Kugeln pro Kugelsorte gebildet werden können, entspricht der Anzahl von übereinstimmenden Expertenpaaren pro Skalenwert. Die Summe dieser Anzahlen für alle w Kugelsorten ergibt daher die minimale Anzahl übereinstimmender Expertenpaare N^*. In der letzten Urne befinden sich $m - w \cdot \lfloor \frac{m}{w} \rfloor$ verschiedene Kugeln und von jeder dieser Kugeln wurden insgesamt genau $\lfloor \frac{m}{w} \rfloor + 1$ Kugeln der gleichen Sorte auf die insgesamt $\lfloor \frac{m}{w} \rfloor + 1$ Urnen verteilt. Für jede dieser Kugelsorten ergeben sich daher $\frac{1}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot (\lfloor \frac{m}{w} \rfloor + 1)$ Paare. Analog wurden von den $w - \left(m - w \cdot \lfloor \frac{m}{w} \rfloor\right)$ Kugelsorten, die sich nicht in der letzten Urne befinden, jeweils genau $\lfloor \frac{m}{w} \rfloor$ Kugeln auf die insgesamt $\lfloor \frac{m}{w} \rfloor$ Urnen verteilt. Für jede dieser Kugelsorten ergeben sich daher $\frac{1}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot (\lfloor \frac{m}{w} \rfloor - 1)$ Paare. Für alle w Kugelsorten ergeben sich also in Summe $\frac{w - \left(m - w \cdot \lfloor \frac{m}{w} \rfloor\right)}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot \left(\lfloor \frac{m}{w} \rfloor - 1\right) + \frac{m - w \cdot \lfloor \frac{m}{w} \rfloor}{2} \cdot \lfloor \frac{m}{w} \rfloor \cdot \left(\lfloor \frac{m}{w} \rfloor + 1\right)$ Kugelpaare, was der rechten Seite von Gleichung B.2 entspricht.

B.2.2. Abstand zwischen den Einschätzungen

Der Quotient $\frac{N}{N_{max(m,w)}}$ ist ein Maß, das beschreibt, ob und wenn ja wie häufig die befragten Experten_Expertinnen ein Merkmal des betreffenden Item unterschiedlich einschätzen. Die Summe der Abstände der Einschätzungen aller Expertenpaare untereinander hingegen beschreibt, wie weit (unterschiedliche) Ratings der Experten_Expertinnen auf der eingesetzten Ratingskala auseinanderliegen. Sie charakterisiert also das Ausmaß der Unterschiedlichkeit der einzelnen Ratings und liefert damit eine Maßzahl, die das Expertenrating $[b_1, b_2, ..., b_m]$ auf eine prinzipiell andere Art und Weise beschreibt als $\frac{N}{N_{max(m,w)}}$ (vgl. Einhaus, 2007, S. 28 u. f.). Sie soll daher als zweiter Kennwert für die Nichtübereinstimmung im Expertenrating berücksichtigt werden.

Sind b_i und b_j die Einschätzungen des_der i-ten bzw. j-ten Experten_Expertin (für alle $i, j \in \{1, 2, ...m\}$), so ergibt sich analog zum Vorgehen zur Berechnung von N der Gesamtabstand A der Einschätzungen aller Experten_Expertinnen als Summe der Beträge der Differenz der einzelnen Expertenpaareinschätzungen:

$$A = \frac{1}{2} \cdot \sum_{i=1}^{m} \sum_{j=1}^{m} |b_i - b_j| \tag{B.5}$$

Ebenso wie für N, ist es auch für den Gesamtabstand A sinnvoll diesen im Verhältnis zum maximal möglichen Abstand $A_{max(m,w)}$ zu betrachten. Der Quotient $\frac{A}{A_{max(m,w)}}$ liefert analog eine Maßzahl, die Werte zwischen 0 und 1 annimmt, wobei der Wert 1 genau dann angenommen wird, wenn der Abstand zwischen den Ratings maximal ist und der

[159] $\lfloor \cdot \rfloor$ symbolisiert die *Abrundungsfunktion*. Für diese gilt beispielsweise $\lfloor 2.1 \rfloor = \lfloor 2.9 \rfloor = 2$.

Wert 0 genau dann, wenn alle m Experten_Expertinnen, dem Item denselben Skalenwerte zugeordnet haben (vgl. Einhaus, 2007, S. 28 u. f.).

Auch $A_{max(m,w)}$ hängt sowohl von der Anzahl befragten Experten_Expertinnen, als auch von der Länge der Ratingskala ab. Sowohl für den Fall $w \geq m$, als auch für den Fall $w < m$ gilt:

$$A_{max(m,w)} = \begin{cases} \frac{m}{2} \cdot \frac{m}{2} \cdot (w-1)\,, & \text{wenn } m \equiv 0 \mod 2 \\ \frac{m-1}{2} \cdot \frac{m+1}{2} \cdot (w-1)\,, & \text{wenn } m \equiv 1 \mod 2 \end{cases} \tag{B.6}$$

Auch Gleichung B.6 ergibt sich aus elementaren kombinatorischen Überlegungen, bei denen, analog zu denen aus Abschnitt B.2.1 die Einschätzung von m Experten_Expertinnen über ein bestimmtes Merkmal eines Items auf einer w-stufigen Ratingskala in einem Gedankenexperiment betrachtet wird. In dem für diese Überlegungen zu betrachtenden Expertenrating soll allerdings gelten $A = A_{max(m,w)}$.

Erneut stehen beliebig viele Kugeln, die mit den Ziffern 1 bis w beschriftet sind, zur Verfügung. Man gehe zusätzlich davon aus, man besitze genau w Schubläden, die mit den Ziffern 1 bis w beschriftet sind. Man legt nun genau dann eine Kugel mit der Beschriftung b in die Schublade mit der Beschriftung b, wenn ein_eine Experte_Expertin dem Item den Skalenwert $b \in \{1, 2, ..., w\}$ zugeordnet hat. Der Betrag der Differenz aus den Ziffern von zwei Schubläden, in denen jeweils genau eine Kugel platziert wurden, entspricht dann gerade dem Abstand der Einschätzungen zweier Experten_Expertinnen.

Der Abstand zwischen den Einschätzungen aller m Experten_Expertinnen ist umso größer, je mehr Kugelpaare mit Kugeln aus Schubladen mit möglichst weit auseinanderliegenden Beschriftungen gebildet werden können. $A_{max(m,w)}$ ergibt sich also aus der maximalen Anzahl von Kugelpaaren, die mit Kugeln aus Schubladen mit maximalem weit auseinanderliegenden Beschriftungen gebildet werden können. Ist m gerade ($m \equiv 0 \mod 2$), wird diese maximale Anzahl von Kugelpaaren genau dann erreicht, wenn sich $\frac{m}{2}$ Kugeln in der Schublade mit der Beschriftung 1 und $\frac{m}{2}$ Kugeln in der Schublade mit der Beschriftung w befinden. Die Gleichung zur Berechnung von $A_{max(m,w)}$ lautet in diesem Fall daher $\frac{m}{2} \cdot \frac{m}{2} \cdot (w-1)$. Ist m ungerade ($m \equiv 1 \mod 2$), wird diese maximale Anzahl von Kugelpaaren genau dann erreicht, wenn sich $\frac{m+1}{2}$ Kugeln in der Schublade mit der Beschriftung 1 und $\frac{m-1}{2}$ Kugeln in der Schublade mit der Beschriftung w befinden (oder umgekehrt). Die Gleichung zur Berechnung von $A_{max(m,w)}$ lautet in diesem Fall daher $\frac{m-1}{2} \cdot \frac{m+1}{2} \cdot (w-1)$.

B.3. Verknüpfung zum Konsenskoeffizienten Ξ

Aus den Quotienten $\frac{N}{N_{max(m,w)}}$ und $\frac{A}{A_{max(m,w)}}$ lässt sich zunächst der von Einhaus (2007) vorgeschlagene Einigkeitskoeffizient η bilden. Dieser berechnet sich, indem das geometri-

sche Mittel aus $\frac{N}{N_{max(m,w)}}$ und $\frac{A}{A_{max(m,w)}}$ gebildet und dieser Mittelwert von 1 abgezogen wird (vgl. Einhaus, 2007, S. 29):

$$\eta := 1 - \sqrt{\frac{N}{N_{max(m,w)}} \cdot \frac{A}{A_{max(m,w)}}} \tag{B.7}$$

Ein in dieser Art und Weise definierter Koeffizient ist allerdings problematisch. Da beim Einigkeitskoeffizienten nach Einhaus (2007) $w = 4$ fest gewählt ist, gilt für diesen $0 < \eta \leq 1$ (vgl. ebd., S. 30). $\eta = 1$ wird genau dann erreicht, wenn die befragten Experten_Expertinnen sich einstimmig für einen Skalenwert entscheiden (es gilt dann $N = A = 0$). In den Fällen $N = N_{max(m,w)}$ oder $A = A_{max(m,w)}$ werden Werte nahe 0 allerdings nicht exakt 0 angenommen. Grund hierfür ist, dass nur im Fall einer dichotomen Ratingskala ($w = 2$) der Fall $N = N_{max(m,w)}$ genau dann eintritt, wenn $A = A_{max(m,w)}$ gilt[160], was in Gleichung B.7 zu $\eta = 0$ führt.

Dieses Problem soll bei der Konstruktion des Konsenskoeffizienten Ξ beseitigt werden. Es soll also $0 \leq \Xi \leq 1$ gelten, wobei $\Xi = 0$ genau dann eintreten soll, wenn $0 = N = A$ gilt und gleichzeitig soll $\Xi = 1$ genau dann eintreten, wenn $N = N_{max(m,w)}$ oder $A = A_{max(m,w)}$ gilt.

Dass im Allgemeinen für die Fälle $N = N_{max(m,w)}$ oder $A = A_{max(m,w)}$ $\eta \neq 0$ gilt, begründet sich aus dem funktionalen Zusammenhang, der für die Berechnung von η gewählt wurde. Aus mathematischer Sicht handelt es sich hierbei um eine Funktion f in den Variablen N, A, m und w. Dabei ist es nicht möglich, die Variablen N und A zu separieren, d. h. f in die Form $f(N, A, m, w) = f_1(N, m, w) \cdot f_2(A, m, w)$ zu überführen. Dies stellt aber ein notwendiges Kriterium für die eben geforderte Oder-Bedingung für den Fall $\Xi = 0$ dar. Der Funktionale Zusammenhang für die Berechnung des Koeffizienten Ξ wird daher ausgehend von diesem Separationsansatz konstruiert. Es soll also gelten

$$\Xi = \Xi(N, A, m, w) = \Xi_1(N, m, w) \cdot \Xi_2(A, m, w),$$

wobei Ξ_1 und Ξ_2 Funktionen sind, die folgenden Forderungen genügen:

1. Ξ_1 und Ξ_2 hängen jeweils nicht von A bzw. N ab.
2. Der Wertebereich von Ξ_1 und Ξ_2 ist das abgeschlossene Intervall zwischen 0 und 1.
3. $N_{max(m,w)}$ und $A_{max(m,w)}$ sind die einzigen Nullstellen von Ξ_1 und Ξ_2. Es soll also $\Xi_1(N_{max(m,w)}, m, w) = 0$ und $\Xi_1(N, m, w) \neq 0$ für alle $N \neq N_{max(m,w)}$, sowie $\Xi_2(A_{max(m,w)}, m, w) = 0$ und $\Xi_2(A, m, w) \neq 0$ für alle $A \neq A_{max(m,w)}$ gelten.

[160] Nur für den Fall $w = 2$ werden in den kombinatorischen Überlegungen in Abschnitt B.2.1 und B.2.2 genau 2 Kugelsorten in gleicher Anzahl auf Urnen und Schubläden verteilt: Für $m \equiv 0 \mod 2$ werden genau $\frac{m}{2}$ Kugeln einer Sorte auf je $\frac{m}{2}$ Urnen (die dritte Urne bleibt leer) und genau $\frac{m}{2}$ Kugeln auf 2 Schubläden verteilt. Für $m \equiv 1 \mod 2$ werden genau $\frac{m-1}{2}$ Kugeln einer Sorte auf je $\frac{m-1}{2}$ Urnen und 1 Kugel auf die letzte Urne verteilt, sowie genau $\frac{m+1}{2}$ bzw. $\frac{m-1}{2}$ Kugeln auf 2 Schubläden verteilt. Für $w > 2$ werden $w > 2$ Kugelsorten auf Urnen, jedoch 2 Kugelsorten auf 2 Schubläden verteilt.

4. Analog zur dritten Forderung soll der Wert $0 = N = A$ der einzige sein, an dem die Funktionen Ξ_1 und Ξ_2 den Wert 1 annehmen. Es soll also $\Xi_1(0, m, w) = 1$ und $\Xi_1(N, m, w) \neq 1$ für alle $N \neq 0$, sowie $\Xi_2(0, m, w) = 1$ und $\Xi_2(A, m, w) \neq 1$ für alle $A \neq 0$ gelten.

Wie durch Einsetzen unmittelbar folgt, werden alle vier aufgeführten Forderungen an Ξ_1 und Ξ_2 von den Termen $(1 - \frac{N}{N_{max(m,w)}})$ und $(1 - \frac{A}{A_{max(m,w)}})$ erfüllt. Der Konsenskoeffizient wird daher definiert als geometrisches Mittel aus diesen beiden Termen:

$$\Xi := \sqrt{\left(1 - \frac{N}{N_{max(m,w)}}\right) \cdot \left(1 - \frac{A}{A_{max(m,w)}}\right)} \tag{B.8}$$

B.4. Kritische Werte für den Konsenskoeffizienten Ξ

Nachdem im vorherigen Unterkapitel Ξ definiert wurde, stellt sich die Frage, ab welchem kritischen Wert des Konsenskoeffizienten von einem hinreichenden Konsens unter den befragten Experten_Expertinnen gesprochen werden kann. Genauer ausgedrückt: Was ist eine sinnvolle Wahl für einen Wert Ξ_{krit}, für den gilt:

1. Für alle $\Xi < \Xi_{krit}$ kann nicht von einem Konsens im Expertenrating $[b_1, b_2, ..., b_m]$ gesprochen werden.
2. Für alle $\Xi \geq \Xi_{krit}$ kann angenommen werden, dass es unter den Experten_Expertinnen einen Konsens über die Ausprägung eines bestimmten Merkmals bei einem bestimmten Item gibt.

Für den Einigkeitskoeffizienten führt Einhaus (2007) eine heuristische Analyse aller möglichen Ratings für den Fall $m = 6$ und $w = 4$ durch, was für η zu einem kritischen Wert von $\eta \approx .40$ führt (vgl. ebd., S. 30 u. f.). Diese heuristische Analyse stützt sich auf eine Betrachtung der Mehrheitsverhältnisse der Experten_Expertinnen untereinander und lässt sich für den Konsenskoeffizienten Ξ wie folgt verallgemeinern und in einen allgemeinen Formelzusammenhang überführen:

Der kritische Wert Ξ_{krit} wird durch drei Bedingungen, die gleichzeitig erfüllt sein müssen, charakterisiert:

1. Für den Faktor $h \in [\frac{1}{2}, 1]$, der die Mehrheitsverhältnisse der Experten_Expertinnen untereinander beschreibt, ist es sinnvoll mindestens den Wert $h = \frac{1}{2}$ (absolute Mehrheit) zu wählen. Für ein Expertenrating, das besonders strengen Auswahlregeln genügen soll, ist es sinnvoll mindestens den Wert $h = \frac{2}{3}$ (Zweidrittelmehrheit) zu wählen.
2. Mindestens $k_{krit} = \lceil h \cdot m \rceil$ (mit $h \in [\frac{1}{2}, 1]$) der befragten Experten_Expertinnen hat dem Item denselben Skalenwert $b \in \{1, 2, ..., w\}$ zugeordnet[161]. Ist $h = \frac{1}{2}$ und $h \cdot m$

[161] $\lceil \cdot \rceil$ symbolisiert die *Aufrundungsfunktion*. Für diese gilt beispielsweise $\lceil 2.1 \rceil = \lceil 2.9 \rceil = 3$.

eine ganze Zahl, müssen mindestens $k_{krit} = \lceil h \cdot m \rceil + 1$ der befragten Experten_-Expertinnen dem Item denselben Skalenwert $b \in \{1, 2, ..., w\}$ zugeordnet haben (ansonsten wäre ein Patt zwischen den befragten Experten_Expertinnen möglich).

3. Alle übrigen $m - \lceil h \cdot m \rceil$ Experten_Expertinnen (bzw. $m - \lceil h \cdot m \rceil - 1$ Experten_-Expertinnen) sollten den Item Skalenwerte zugeordnet haben, die „in der Nähe von" b liegen. Für den kritischen Wert ist es sinnvoll einen Maximalabstand von $\lfloor \frac{w}{2} \rfloor$ zu setzen. A_{krit} ergibt sich also für den Fall, dass alle übrigen Experten_Expertinnen dem Item denselben Skalenwert $c \in \{1, 2, ..., w\}$ zugeordnet haben, für den gilt $|c - b| = \lfloor \frac{w}{2} \rfloor$.

Aus diesen drei Bedingungen lassen sich zunächst kritische Werte für N und A bestimmen. Durch einfaches Abzählen ergeben sich folgende Gleichungen:

$$N_{krit} = \begin{cases} (\lceil h \cdot m \rceil + 1) \cdot (m - \lceil h \cdot m \rceil - 1), & \text{wenn } h = \frac{1}{2} \text{ und } m \equiv 0 \mod 2 \\ \lceil h \cdot m \rceil \cdot (m - \lceil h \cdot m \rceil), & \text{sonst} \end{cases} \tag{B.9}$$

$$A_{krit} = \lfloor \frac{w}{2} \rfloor \cdot \begin{cases} \lceil (h \cdot m \rceil + 1) \cdot (m - \lceil h \cdot m \rceil - 1), & \text{wenn } h = \frac{1}{2} \text{ und } m \equiv 0 \mod 2 \\ \lceil h \cdot m \rceil \cdot (m - \lceil h \cdot m \rceil), & \text{sonst} \end{cases} \tag{B.10}$$

Durch Einsetzen von Gleichung B.9 und B.10 in Gleichung B.8 ergibt sich dann unmittelbar der kritische Wert des Konsenskoeffizienten:

$$\Xi_{krit} = \Xi(N_{krit}, A_{krit}, m, w) \tag{B.11}$$

Tritt der Fall $\Xi \geq \Xi_{krit}$ in einem Expertenrating $[b_1, b_2, ..., b_m]$ ein, ist es allerdings nicht nur von Interesse, dass es unter den Experten_Expertinnen einen Konsens über einen Skalenwert $d \in \{1, 2, ..., w\}$ gibt, sondern auch auf welchen der w Skalenwerte sich die Experten_Expertinnen verständigt haben. Dieser Wert lässt sich mathematisch wie folgt bestimmen: Sind $b_1, b_2, ..., b_m$ die Einschätzungen der Experten_Expertinnen in einem Expertenrating mit $\Xi \geq \Xi_{krit}$, so ist der Konsensskalenwert d der *Modalwert* von $\{b_1, b_2, ..., b_m\}$. Für die absolute Häufigkeit k_d, mit der d in $\{b_1, b_2, ..., b_m\}$ auftritt ist, gilt also:

$$k_d = \max_{1 \leq j \leq w} |\{b \mid b \in \{b_1, b_2, ..., b_m\} \wedge b = j\}| \tag{B.12}$$

Ferner ist im Fall $\Xi \geq \Xi_{krit}$ von Interesse, ob beim Konsensskalenwert d eine zufällige Übereinstimmung der befragten Experten_Expertinnen ausgeschlossen werden kann. Zufällige Übereinstimmung meint dabei, dass die Experten_Expertinnen dem Item jede der w Ausprägungen eines Merkmals durch „bloßes Raten“, also mit der Wahrscheinlichkeit $\frac{1}{w}$ zugeordnet haben. Insbesondere für den Konsensskalenwerkt d gilt dann $p_d = \frac{1}{w}$ (Nullhypothese). Für die Häufigkeit des Auftretens des Konsensskalenwerts k_d lässt sich daher ein rechtsseitiger Binomialtests ($H_0 : p_d = \frac{1}{w}$; $H_1 : p_d > \frac{1}{w}$) als Signifikanztest durchführen (vgl. S. Siegel, 1976, S. 36 u. f.). Hierzu ist der Funktionswert P der oberen kumulierten Binomialverteilung für m, k_d und $p_d = \frac{1}{w}$ zu berechnen und mit einem vorgegebenen Signifikanzniveau α zu vergleichen. Ist $P(m, k_d, \frac{1}{w}) \leq \alpha$, dann ist H_0 zu verwerfen. Für $\Xi \geq \Xi_{krit}$ kann also plausibel ausgeschlossen werden, dass es sich beim Konsensskalenwert d um eine zufällige Übereinstimmung der befragten Experten_Expertinnen handelt, wenn gilt:

$$P(m, k_d, \tfrac{1}{w}) = \sum_{i=k_d}^{m} \frac{m!}{i! \cdot (m-i)!} \cdot \left(\frac{1}{w}\right)^i \cdot \left(1 - \frac{1}{w}\right)^{m-i} \leq \alpha \tag{B.13}$$

Tabelle B.1 gibt einen Überblick den kritischen Wert des Konsenskoeffizienten für $h = \frac{1}{2}$ und $h = \frac{2}{3}$ in Abhängigkeit von m und w. Zudem sind hier jeweils in Klammern die Werte der kumulierten Binomialverteilung für m, k_{krit} und $p_d = \frac{1}{w}$ angegeben (die Wahrscheinlichkeit, dass $\Xi \geq \Xi_{krit}$ gilt, obwohl alle befragten Experten_Expertinnen ihre Einschätzung durch „bloßes Raten“ vorgenommen haben). Wie aus Tabelle B.1 hervorgeht, unterliegt Ξ_{krit} deutlichen Schwankungen. Diese Schwankungen lassen sich darauf zurückführen, dass in Gleichung B.9 und B.10 berücksichtigt ist, dass nur „ganze Personen“ am Expertenrating teilnehmen können und bei einer w-stufigen Ratingskala sinnvollerweise nur von ganzen Skalenschritten gesprochen werden kann. Vor allem nimmt in einigen Fällen in Tabelle B.1 Ξ_{krit} den Wert .00 an, was damit zusammenhängt, dass in diesen Fällen $N_{max(m,w)} = N_{krit} = A_{max(m,w)} = A_{krit}$ gilt.

m:	h: \ w:	2	3	4	5	6	7	8
2	½	1.00 (.250)	1.00 (.111)	1.00 (.063)	1.00 (.040)	1.00 (.028)	1.00 (.020)	1.00 (.016)
	⅔	1.00 (.250)	1.00 (.111)	1.00 (.063)	1.00 (.040)	1.00 (.028)	1.00 (.020)	1.00 (.016)
3	½	.00 (.500)	.41 (.259)	.33 (.156)	.41 (.104)	.37 (.074)	.41 (.055)	.38 (.043)
	⅔	.00 (.500)	.41 (.259)	.33 (.156)	.41 (.104)	.37 (.074)	.41 (.055)	.38 (.043)
4	½	.25 (.313)	.53 (.111)	.50 (.051)	.57 (.027)	.53 (.016)	.57 (.010)	.54 (.007)
	⅔	.25 (.313)	.50 (.111)	.50 (.051)	.56 (.027)	.52 (.016)	.56 (.010)	.53 (.007)
5	½	.00 (.500)	.35 (.210)	.33 (.104)	.45 (.058)	.40 (.035)	.45 (.023)	.41 (.016)
	⅔	.33 (.188)	.58 (.045)	.56 (.016)	.63 (.007)	.60 (.003)	.63 (.002)	.61 (.001)
6	½	.11 (.344)	.46 (.100)	.40 (.038)	.50 (.017)	.47 (.009)	.51 (.005)	.48 (.003)
	⅔	.11 (.344)	.43 (.100)	.40 (.038)	.49 (.017)	.47 (.009)	.51 (.005)	.48 (.003)
7	½	.00 (.500)	.35 (.173)	.33 (.071)	.43 (.033)	.40 (.018)	.46 (.010)	.43 (.006)
	⅔	.17 (.227)	.47 (.045)	.44 (.013)	.53 (.005)	.50 (.002)	.55 (.001)	.52 (.001)
8	½	.06 (.363)	.42 (.088)	.38 (.027)	.47 (.010)	.43 (.005)	.49 (.002)	.46 (.001)
	⅔	.25 (.145)	.52 (.020)	.50 (.004)	.57 (.001)	.54 (.000)	.59 (.000)	.57 (.000)
9	½	.00 (.500)	.36 (.145)	.33 (.049)	.43 (.020)	.40 (.009)	.45 (.005)	.43 (.002)
	⅔	.10 (.254)	.43 (.042)	.40 (.010)	.49 (.003)	.46 (.001)	.51 (.000)	.49 (.000)
10	½	.04 (.377)	.41 (.077)	.36 (.020)	.46 (.006)	.42 (.002)	.48 (.001)	.45 (.001)
	⅔	.16 (.172)	.46 (.020)	.44 (.004)	.52 (.001)	.49 (.000)	.54 (.000)	.52 (.000)
11	½	.00 (.500)	.35 (.122)	.33 (.034)	.43 (.012)	.40 (.005)	.45 (.002)	.43 (.001)
	⅔	.20 (.113)	.49 (.009)	.47 (.001)	.55 (.000)	.52 (.000)	.56 (.000)	.54 (.000)
12	½	.03 (.387)	.40 (.066)	.35 (.014)	.45 (.004)	.42 (.001)	.47 (.000)	.44 (.000)
	⅔	.11 (.194)	.43 (.019)	.41 (.003)	.49 (.001)	.47 (.000)	.51 (.000)	.49 (.000)
13	½	.00 (.500)	.35 (.104)	.33 (.024)	.43 (.007)	.40 (.002)	.46 (.001)	.43 (.000)
	⅔	.14 (.133)	.45 (.009)	.43 (.001)	.51 (.000)	.49 (.000)	.53 (.000)	.51 (.000)
14	½	.02 (.395)	.40 (.058)	.34 (.010)	.45 (.002)	.41 (.001)	.47 (.000)	.44 (.000)
	⅔	.18 (.090)	.48 (.004)	.45 (.000)	.54 (.000)	.51 (.000)	.56 (.000)	.53 (.000)
15	½	.00 (.500)	.36 (.088)	.33 (.017)	.43 (.004)	.40 (.001)	.46 (.000)	.43 (.000)
	⅔	.11 (.151)	.43 (.009)	.40 (.001)	.50 (.000)	.46 (.000)	.52 (.000)	.49 (.000)
16	½	.02 (.402)	.40 (.050)	.34 (.007)	.45 (.001)	.41 (.000)	.47 (.000)	.44 (.000)
	⅔	.14 (.105)	.45 (.004)	.43 (.000)	.51 (.000)	.48 (.000)	.53 (.000)	.51 (.000)
17	½	.00 (.500)	.35 (.075)	.33 (.012)	.43 (.003)	.40 (.001)	.46 (.000)	.43 (.000)
	⅔	.17 (.072)	.47 (.002)	.44 (.000)	.53 (.000)	.50 (.000)	.55 (.000)	.52 (.000)
18	½	.01 (.407)	.40 (.043)	.34 (.005)	.45 (.001)	.41 (.000)	.46 (.000)	.43 (.000)
	⅔	.11 (.119)	.43 (.004)	.41 (.000)	.50 (.000)	.47 (.000)	.52 (.000)	.49 (.000)
19	½	.00 (.500)	.35 (.065)	.33 (.009)	.43 (.002)	.40 (.000)	.46 (.000)	.43 (.000)
	⅔	.13 (.084)	.45 (.002)	.42 (.000)	.51 (.000)	.48 (.000)	.53 (.000)	.50 (.000)
20	½	.01 (.412)	.39 (.038)	.34 (.004)	.45 (.001)	.40 (.000)	.46 (.000)	.43 (.000)
	⅔	.16 (.058)	.46 (.001)	.44 (.000)	.52 (.000)	.49 (.000)	.54 (.000)	.52 (.000)

Tabelle B.1.: Kritische Werte des Konsenskoeffizienten Ξ für $h = \frac{1}{2}$ und $h = \frac{2}{3}$, $2 \leq m \leq 20$ und $2 \leq w \leq 8$. In Klammern ist jeweils der zugehörige Funktionswerte der oberen kumulierten Binomialverteilung für m, k_{krit} und $p_d = \frac{1}{w}$ angegeben.

C. Materialien für die Laborsituation der Hauptstudie

Im Folgenden werden die Materialien aufgeführt, die ihm Rahmen der Entwicklungsstudie entwickelt (vgl. Abschnitt 5.4) und in der Laborsituation der Hauptstudie eingesetzt wurden (vgl. Kapitel 6). Diese Materialien sind:

- Das Aufgabenheft für Physiklehrkräfte (vgl. Anhang C.1)
- Der Lehrkräftefragebogen (vgl. Anhang C.2)
- Das Manual zur Durchführung der Erhebung (vgl. Anhang C.3)

Aus der Dokumentation der Erhebungsinstrumente der *COACTIV*-Studie (vgl. Baumert et al., 2009) wurden die Fragen 8 und 9 des Lehrkräftefragebogen übernommen und bilden die Skalen „Bewertung nach sozialer Bezugsnorm versus kriterialer Norm“ (vgl. ebd., S. 169) und „Diagnose im Leistungsbereich“ (vgl. ebd., S. 172). Die Frage 10 ist aus dem Fragebogen der Lehrkräftebefragung zu Sprachbildung im naturwissenschaftlichen Unterricht von Riebling (2013b, S. 108 u. f.) übernommen und bildet die Skala „Vermittlung der Domänenspezifischen Bildungssprache“ (vgl. ebd. S. 116).

Die Instruktionen und Übungsaufgaben im Manual zur Durchführung der Erhebung sind zum Teil in gekürzter und/oder geänderter Form aus den Arbeiten von Arras (2007, S. 499) und van Someren et al. (1994, S. 174) übernommen (für Details siehe Fußnoten im Manual).

C.1. Aufgabenheft für Physiklehrkräfte

Physikdidaktik

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Think-Aloud-Aufgaben für Physiklehrkräfte

Vielen Dank für die Teilnahme an unserer Untersuchung. Die Bearbeitung dieser Aufgaben ist freiwillig. Wenn Sie die Aufgaben nicht bearbeiten, wird dies keine Nachteile für Sie haben.

Bitte erstellen Sie Ihren anonymisierten Code.
Notieren Sie hierzu die ersten drei Buchstaben des Vornamens Ihres Vaters, Ihre eigene Körpergröße in cm und eine beliebige Zahl zwischen 0 und 9 (z. B. Ber/176/2 oder Mar/182/7).

___ ___ ___ / ___ ___ ___ / ___
Vorname **Körpergröße** **Zahl**

1

Lesen Sie die Instruktion genau durch. Blättern Sie anschließend auf Seite 2.

Instruktion

1. Ziel der folgenden Aufgaben ist es, möglichst viel darüber herauszufinden, wie Sie bei der Bewertung einer Klassenarbeit vorgehen.

2. Die Aufgaben haben eine feste Reihenfolge und sollen nur mit Hilfe bestimmter Materialien bearbeitet werden. Dies wird Ihnen in den Aufgaben genau beschrieben. Weichen Sie hiervon nicht ab.

3. Wenden Sie das sogenannte laute Denken an, während Sie die Aufgaben bearbeiten. Ein Diktiergerät wird dabei Ihre Äußerungen aufzeichnen.

4. Sie sollten die Texte auch laut vorlesen, damit deutlich wird, an welchen Stellen Sie ggf. Schwierigkeiten haben.

5. Da keine Videokamera mitläuft, ist es auch wichtig, dass Sie jeweils verbalisieren, was Sie gerade tun.

6. Während Sie die Aufgaben bearbeiten, werde ich anwesend sein. Ggf. werde ich etwas sagen, wenn Sie den Redefluss zu lange unterbrechen. Meine Anwesenheit sollte Sie jedoch nicht irritieren.

2

Für Aufgabe 1 benötigen Sie nur die Seiten 2 und 3.

Stellen Sie sich folgende Situation vor:

- Sie unterrichten eine 9. Klasse in Physik und haben eine Klassenarbeit geschrieben.
- In dieser Klassenarbeit haben Sie die Aufgabe „Weltraumspaziergang“ als Grundwissensaufgabe eingesetzt (siehe Seite 3).
- Sie haben Ihren Schülerinnen und Schülern zusätzlich folgende Anweisung gegeben:
  1. *„Schreibt eure Antwort in ganzen Sätzen auf.“*
  2. *„Skizzen oder Zeichnungen können bei dieser Aufgabe nicht gezählt werden.“*
- Für die Aufgabe „Weltraumspaziergang“ möchten Sie 0 bis maximal 5 Punkte vergeben.

Aufgabe 1:
Erstellen Sie für die oben beschriebene Situation einen geeigneten Erwartungshorizont für die Aufgabe „Weltraumspaziergang“.

Verwenden Sie hierzu das karierte Papier auf Seite 3.
Gehen Sie dabei so vor, wie Sie dies unter normalen Umständen auch tun würden. Geben Sie die Bedeutung spezieller Abkürzungen mit an (z. B. „✓“ für 1 Punkt).

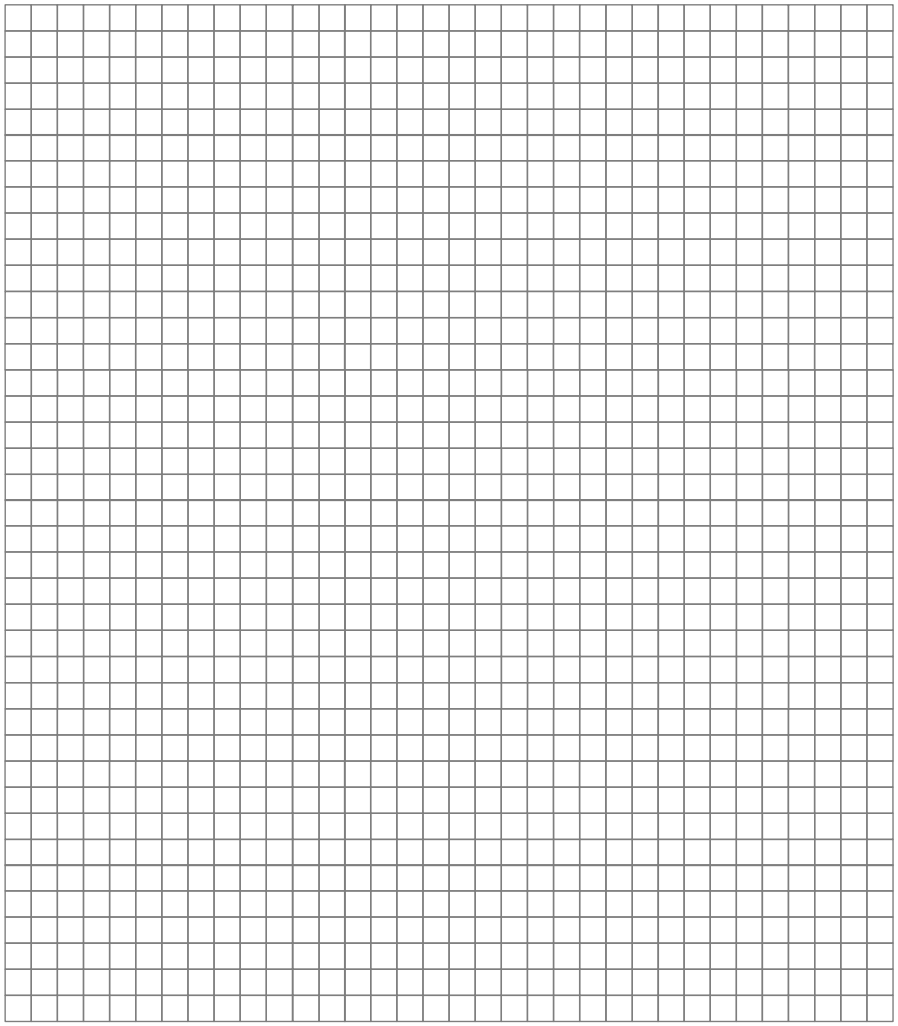
Wenn Sie mit Aufgabe 1 fertig sind, blättern Sie weiter zu Seite 4.

3

Weltraumspaziergang (5 Punkte)

Bei einem Weltraumspaziergang reißt zwischen zwei Astronauten die Funkverbindung ab. Obwohl der eine Astronaut aus Leibeskräften schreit, hört ihn sein Kamerad nicht. Der ältere Astronaut hält seinen in Panik geratenen Kollegen fest und presst seinen Helm an den des Kollegen. Plötzlich kann der jüngere den älteren leise hören. Erkläre beide Phänomene genau!

Erwartungshorizont:



4

Für Aufgabe 2 benötigen Sie nur die Seiten 3 bis 8.

- Die Antworten A, B, C, D sind Antworten von Schülerinnen und Schülern auf die Aufgabe „Weltraumspaziergang“ (siehe Seite 5 bis 8).
- Zur besseren Lesbarkeit sind die Antworten in einer einheitlichen Handschrift geschrieben und von Rechtschreibfehlern bereinigt worden.
- Die Zeilen der Antworten sind nachträglich nummeriert.

Aufgabe 2:
Bewerten Sie die Antworten A, B, C, D mit 0 bis maximal 5 Punkten. Verwenden Sie hierzu Ihren Erwartungshorizont auf Seite 3 und einen Rotstift. Gehen Sie dabei so vor, wie Sie dies unter normalen Umständen auch tun würden.

Wenn Ihnen beim Korrigieren einer Antwort etwas auffällt, ist es sinnvoll, die Zeilennummer zu verbalisieren. Hierdurch wird in der Tonbandaufnahme klar, welche Stelle der Antwort Sie gemeint haben.

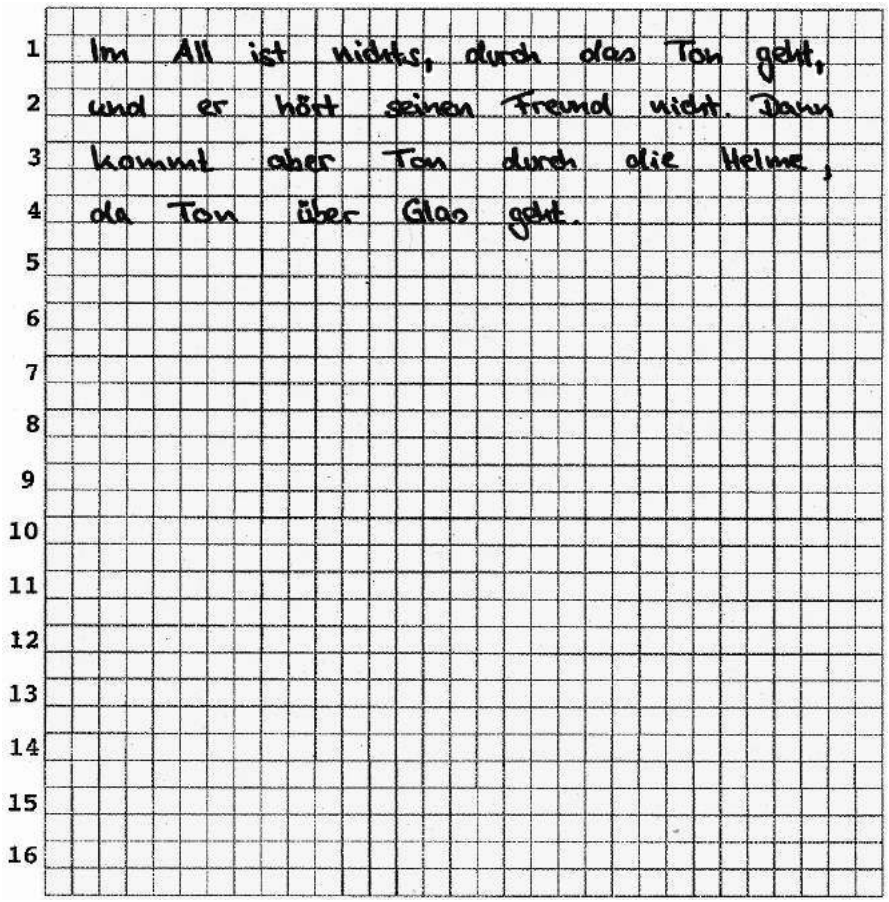
Wenn Sie mit Aufgabe 2 fertig sind, geben Sie dem Versuchsleiter ein Zeichen.

5

Antwort A

Weltraumspaziergang (5 Punkte)

Bei einem Weltraumspaziergang reißt zwischen zwei Astronauten die Funkverbindung ab. Obwohl der eine Astronaut aus Leibeskräften schreit, hört ihn sein Kamerad nicht. Der ältere Astronaut hält seinen in Panik geratenen Kollegen fest und presst seinen Helm an den des Kollegen. Plötzlich kann der jüngere den älteren leise hören. Erkläre beide Phänomene genau!

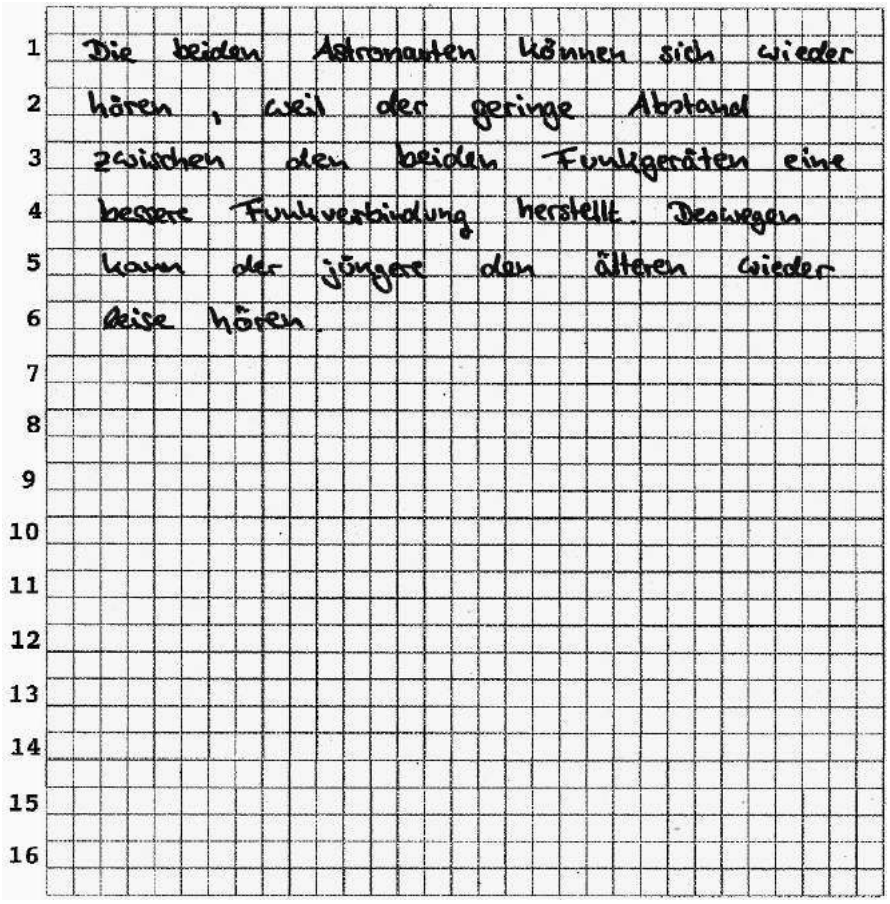


6

Antwort B

Weltraumspaziergang (5 Punkte)

Bei einem Weltraumspaziergang reißt zwischen zwei Astronauten die Funkverbindung ab. Obwohl der eine Astronaut aus Leibeskräften schreit, hört ihn sein Kamerad nicht. Der ältere Astronaut hält seinen in Panik geratenen Kollegen fest und presst seinen Helm an den des Kollegen. Plötzlich kann der jüngere den älteren leise hören. Erkläre beide Phänomene genau!

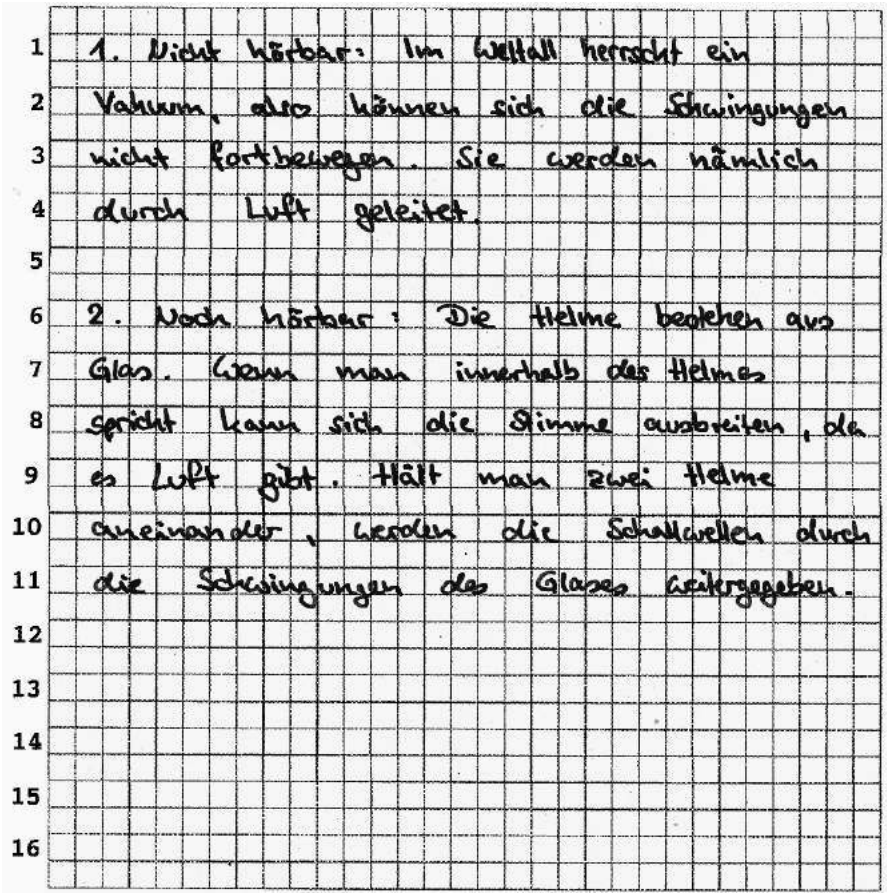


7

Antwort C

Weltraumspaziergang (5 Punkte)

Bei einem Weltraumspaziergang reißt zwischen zwei Astronauten die Funkverbindung ab. Obwohl der eine Astronaut aus Leibeskräften schreit, hört ihn sein Kamerad nicht. Der ältere Astronaut hält seinen in Panik geratenen Kollegen fest und presst seinen Helm an den des Kollegen. Plötzlich kann der jüngere den älteren leise hören. Erkläre beide Phänomene genau!


1. Nicht hörbar: Im Weltall herrscht ein Vakuum, also können sich die Schwingungen nicht fortbewegen. Sie werden nämlich durch Luft geleitet.

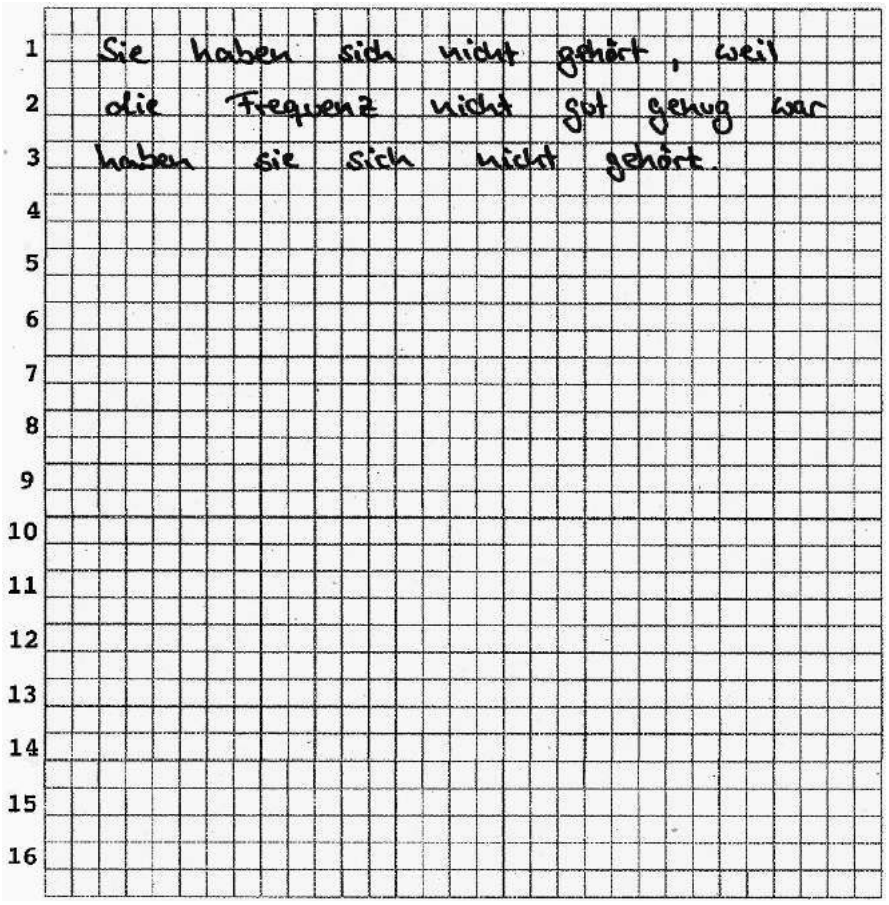
2. Noch hörbar: Die Helme bestehen aus Glas. Wenn man innerhalb des Helmes spricht kann sich die Stimme ausbreiten, da es Luft gibt. Hält man zwei Helme aneinander, werden die Schallwellen durch die Schwingungen des Glases weitergegeben.

8

Antwort D

Weltraumspaziergang (5 Punkte)

Bei einem Weltraumspaziergang reißt zwischen zwei Astronauten die Funkverbindung ab. Obwohl der eine Astronaut aus Leibeskräften schreit, hört ihn sein Kamerad nicht. Der ältere Astronaut hält seinen in Panik geratenen Kollegen fest und presst seinen Helm an den des Kollegen. Plötzlich kann der jüngere den älteren leise hören. Erkläre beide Phänomene genau!



C.2. Lehrkräftefragebogen

Bitte erstellen Sie Ihren anonymisierten Fragebogen-Code.
Notieren Sie hierzu die ersten drei Buchstaben des Vornamens Ihres Vaters, Ihre eigene Körpergröße in cm und eine beliebige Zahl zwischen 0 und 9 (z. B. Ber/176/2 oder Mar/182/7).

___ ___ ___ / ___ ___ ___ / ___
Vorname **Körpergröße** **Zahl**

I. Allgemeine Angaben

1. Welches Geschlecht haben Sie?

❍ männlich ❍ weiblich

2. In welchem Jahr sind Sie geboren?

___ ___ ___ ___

3. Welches Studium haben Sie absolviert?

Lehramt für _____

Unterrichtsfächer: _____

Anderes Studium
(„Quereinstieg“) : _____

4. Wie viele Jahre umfasst Ihre Berufserfahrung als Lehrerin oder Lehrer?

___ ___ Jahre

5. Welche Fächer unterrichten Sie zurzeit?

1. _____ 2. _____

3. _____ 4. _____

6. An welcher Schulform unterrichten Sie zurzeit?

❍ Gymnasium ❍ Stadtteilschule ❍ _____

7. Wie viele Unterrichtsstunden pro Woche erteilen Sie zurzeit in den folgenden Stufen?

Sekundarstufe I insgesamt: ___ ___ Stunden; Physik: ___ ___ Stunden

Sekundarstufe II insgesamt: ___ ___ Stunden; Physik: ___ ___ Stunden

II. Fragen zu Bewertung und Diagnose

8. Wie bewerten Sie Ihre Schüler im Einzelnen?	**trifft nicht zu**	**trifft eher nicht zu**	**trifft eher zu**	**trifft zu**
a) Ich lege den Bewertungsschlüssel vor der Klassenarbeit/Schulaufgabe fest.	❍1	❍2	❍3	❍4
b) Die Anforderungen einer Notenstufe lege ich vor der Klassenarbeit/Schulaufgabe fest und ändere daran nichts, auch wenn dann relativ viele Arbeiten gut oder schlecht ausfallen.	❍1	❍2	❍3	❍4
c) Die Noten ermittle ich im Vergleich der Schüler/innen meiner Klasse.	❍1	❍2	❍3	❍4
d) Ich orientiere meine Noten am Durchschnitt der Klasse.	❍1	❍2	❍3	❍4

9. Wie sicher sind Sie in Ihrer Schülerdiagnose?	**trifft nicht zu**	**trifft eher nicht zu**	**trifft eher zu**	**trifft zu**
a) Es fällt mir leicht festzustellen, ob ein/eine Schüler/in eine Aufgabe verstanden hat.	❍1	❍2	❍3	❍4
b) Ich merke sehr schnell, wenn jemand etwas nicht verstanden hat.	❍1	❍2	❍3	❍4
c) Ich weiß, bei welchen Aufgaben die einzelnen Schüler/innen Schwierigkeiten haben.	❍1	❍2	❍3	❍4
d) Ich kenne die Stärken und Schwächen der einzelnen Schüler/innen.	❍1	❍2	❍3	❍4
e) Ich merke sofort, wenn ein/eine Schüler/in im Unterricht nicht mitkommt.	❍1	❍2	❍3	❍4

III. Fragen zur Gestaltung Ihres Fachunterrichts (Biologie/Chemie/Physik/Naturwissenschaften)

10. In welchem Maße spielt <u>Sprache</u> in ihrem Fachunterricht eine Rolle? **Im Fachunterricht...**	**fast nie**				**fast immer**
a) lege ich mit den Schülern eine Wortschatzliste/ein Glossar an, das fortlaufend erweitert wird.	❍1	❍2	❍3	❍4	❍5
b) führe ich den neuen Wortschatz ausführlich ein (Tafel, OH-Folie etc.).	❍1	❍2	❍3	❍4	❍5
c) stelle ich den Schülern Aufgaben, die explizit der Einübung des Fachwortschatzes dienen (Skizzen beschriften, Diagramme ergänzen, Lückentexte bearbeiten etc.).	❍1	❍2	❍3	❍4	❍5
d) führe ich fachtypische Wort- und Satzkonstruktionen ein.	❍1	❍2	❍3	❍4	❍5
e) bespreche ich die zentralen grammatikalischen Merkmale der Fachsprache (z. B. die Passivverwendung).	❍1	❍2	❍3	❍4	❍5

C.3. Durchführungsmanual

Physikdidaktik

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Manual

Think-Aloud-Aufgaben für Physiklehrkräfte

Legen Sie im Anschluss an die Durchführung das Manual zusammen mit den bearbeiteten Aufgaben in den Umschlag.

1

Basisinformationen

Notieren Sie hier Basisinformationen über die Durchführung:

anonymer Code der Lehrkraft	___ ___ ___ / ___ ___ ___ / ___  Vorname Körpergröße Zahl
Dauer (Uhrzeit von bis)	___ ___ : ___ ___ Uhr bis ___ ___ : ___ ___ Uhr
Datum	___ ___ . ___ ___ . 2016
Ort	
Besondere Vorkommnisse während der Durchführung	

2

Materialcheckliste für die Durchführung

Legen Sie folgende Materialien **vor der Durchführung** bereit:

Material	Anmerkung	bereit gelegt?
Manual	Seite 1, 2, 7, 9 sind von der Leitung **auszufüllen**!	❍ ja
Lernvideo	Auf einem Laptop oder Tablet-PC; Ton überprüfen; Kopfhörer bereithalten	❍ ja
Übungsaufgabe	Übungsmaterial für die Think-Aloud-Methode	❍ ja
Aufgabenheft für Physiklehrkräfte	8 Seiten + Titelblatt ; Seiten in vorgegebener Reihenfolge anordnen ; Wasserflasche ; Stifte (blau und rot)	❍ ja
Lehrkräftefragebogen	2 Seiten	❍ ja
Diktiergeräte	2 Geräte laufen lassen; Batterien überprüfen; ggf. Handy als Zweitaufnahmegerät mitlaufen lassen.	❍ ja
Uhr	zur Zeitkontrolle	❍ ja
1 Tisch, 3 Stühle	im Raum anordnen (siehe unten)	❍ ja

Der Raum besteht aus einem Tisch mit 2 Stühlen. Auf einem nimmt die Lehrkraft platz, auf einem Sie als Leitung. Der 3. Stuhl wird so positioniert, dass er sich außerhalb des Blickfeldes der Lehrkraft befindet.

Zeitplan

Die Untersuchung erfolgt in mehreren Schritten. Es ergibt sich der folgende Zeitplan, der eingehalten werden sollte:

Zeitplan	
Einführung und allg. Instruktion	ca. 5 min
Instruktion und Üben der Think-Aloud-Methode	15 min
Instruktion Think-Aloud-Aufgaben für Physiklehrkräfte	5 min
Think-Aloud-Aufgaben für Physiklehrkräfte	40 min
Postinterview	10 min
Lehrkräftefragebogen	5 min
Schluss + Puffer	5 min
Total	**90 min**

3

Einführung und allg. Instruktion (5 min)

- Warten Sie bis die teilnehmende Lehrkraft ihren Platz eingenommen hat.
- **Informieren Sie die Lehrkraft[1] und generieren Sie ihren anonymen Code:**
  Zuerst möchte ich mich bei Ihnen für Ihre Bereitschaft bedanken, bei dieser Untersuchung mitzumachen. Ihre Angaben und alle Daten aus der Untersuchung werden selbstverständlich absolut anonym gehalten. Hierzu generieren wir zunächst Ihren anonymen Code:

Bitte erstellen Sie Ihren anonymen Code.
Notieren Sie hierzu die ersten drei Buchstaben des Vornamens Ihres Vaters, Ihre eigene Körpergröße in cm und eine beliebige Zahl zwischen 0 und 9 (z. B. Ber/176/2 oder Mar/182/7).

___ ___ ___ / ___ ___ ___ / ___
Vorname **Körpergröße** **Zahl**

- **Übertragen Sie** diesen Code auf das Manual, die Think-Aloud-Aufgaben für Physiklehrkräfte und den Lehrkräftefragebogen.
- **Starten Sie die Tonbandaufzeichnung. Sagen Sie zu Beginn den Anonymen Code und das Datum.**
- **Lesen Sie der Lehrkraft folgende Instruktion vor:**
  *Ziel dieser Studie ist es, möglichst viel darüber herauszufinden, wie Sie bei der Bewertung einer Klassenarbeit vorgehen. Es geht also **nicht** darum, Ihre Arbeit zu kontrollieren. Vielmehr geht es darum, mehr über die Strategien herauszufinden, die Sie bei der Beurteilung anwenden. Die Bewertungsarbeit sollte deshalb möglichst so ablaufen, wie Sie dies unter normalen Umständen auch tun würden. Alle hierfür notwendigen Materialien stehen Ihnen zur Verfügung.*

[1] Gekürzt und geändert aus Arras, U. (2007). Wie beurteilen wir Leistung in der Fremdsprache? Strategien und Prozess bei der Beurteilung schriftlicher Leistungen in der Fremdsprache am Beispielder Prüfung Test Deutsch als Fremdsprache (TestDaF). Tübingen: Narr Francke Attempto Verlag.

4

Wir verwenden die sog. Think-Aloud-Methode (Methode des lauten Denkens). Das bedeutet: Während Sie bewerten, sollen Sie all das laut äußern, was sie gerade denken und machen. Stellen Sie sich am besten vor, dass Sie alleine im Raum sind und mit sich selbst sprechen. Ein Diktiergerät wird dabei Ihre Äußerungen aufzeichnen. Das ist nicht ganz einfach, weil man sehr viel schneller denkt, als man verbalisieren kann. Deshalb werde ich Ihnen bevor wir mit der eigentlichen Untersuchung beginnen ein Lernvideo zur Think-Aloud-Methode zeigen. Anschließend werden wir eine kurze Übungssequenz machen, bei der Sie sich auf die Methode einstellen können. Bei dieser Übungsphase werde ich Ihnen ggf. noch ein paar Hinweise geben.
Haben Sie soweit Fragen?

- Beantworten Sie alle Fragen der Lehrkraft. **Hilfreich ist noch einmal zu wiederholen,** dass die Vorstellung hilft, alleine im Raum zu sein und mit sich selbst zu sprechen.

5

Instruktion und Üben der Think-Aloud-Methode (15 min)

- **Zeigen Sie der Lehrkraft das Lernvideo zur Think-Aloud-Methode.** Bei schlechten Klangverhältnissen in den Räumlichkeiten verwenden Sie Kopfhörer.

- *Haben Sie bisher Fragen zu dem was Sie im Video gesehen haben?* (**Beantworten Sie alle Fragen der Lehrkraft**)

- **Beginnen Sie anschließend mit der Übungssequenz:**
  Wie angekündigt möchte ich nun mit Ihnen die Think-Aloud-Methode üben. Ich werde dabei auf dem Stuhl hinter Ihnen Platz nehmen, damit Sie meine Anwesenheit nicht irritiert, während Sie laut denken.

- **Legen Sie die gekürzte Aufgabenstellung verdeckt vor die Lehrkraft** (siehe Seite 6). Nehmen Sie Ihren Platz ein. Beginnen Sie anschließend mit der Übungssequenz:
  Ich möchte Sie bitten sich mit folgendem Problem zu beschäftigen und dabei das laute Denken anzuwenden (**lesen Sie der Lehrkraft die Aufgabenstellung vor**):

Verbesserung technischer Geräte[2]
Ihre Aufgabe besteht darin ein technisches Gerät zu verbessern. Ich werde Ihnen ein technisches Gerät nennen und Ihre Aufgabe ist es, **fünf** Verbesserungen für dieses Gerät zu erfinden. Denken Sie dabei laut.

Name des Gerätes: Ihr Handy

Auf dem Zettel, steht noch einmal die Aufgabenstellung. Sie können nun mit der Aufgabe beginnen.

[2] Übersetzt, gekürzt und geändert aus van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). The think aloud method. A practical guide to modelling cognitive processes. London: Academic Press.

6

Gekürzte Aufgabenstellung:

Verbesserung technischer Geräte

Erfinden Sie **fünf** Verbesserungen für Ihr Handy.

- Geben Sie der Lehrkraft ggf. folgende Hinweise aus dem Lernvideo:
  - *„Es soll alles laut gesagt werden, was man denkt und macht!"*
  - *„Es sollte gut hörbar und deutlich gesprochen werden!"*
  - *„Lieber zu viel, als zu wenig sagen!"*
  - *„Es sollte jede noch so kleine Tätigkeit kommentiert werden."*
  - *„Es sollte sehr detailliert darauf eingegangen werden, was man denkt."*
  - *„Stellen Sie sich vor, ich bin gar nicht anwesend und Sie führen ein Selbstgespräch."*

- **Vermeiden Sie Hinweise** wie z. B. *„Sagen Sie mir, was Sie denken".* Die Lehrkraft könnte dies so verstehen, dass nach Ihrer Meinung oder einer Bewertung ihrer Gedanken gefragt ist.

- Beenden Sie die Übungssequenz, sobald die Lehrkraft eine Lösung gefunden hat. **Die Bearbeitungszeit sollte 5 Minuten nicht überschreiten.** Gehen Sie abschließend auf weitere Fragen der Lehrkraft ein.

7

Think-Aloud-Aufgaben für Lehrkräfte (5 min + 40 min)

- Legen Sie folgende Materialien der Lehrkraft vor:
  1. Think-Aloud-Aufgaben (anonymer Code eingetragen?)
  2. Stifte (blau und rot)
  3. Wasserflasche

- **Lesen Sie der Lehrkraft Aufgaben-Instruktionen vor (Think-Aloud-Aufgaben Seite 1). Weisen Sie auf die fett gedruckten Instruktionen im Aufgabenheft hin (diese weisen darauf hin, welche Seiten für eine Aufgabe benötigt werden und was zu tun ist wenn die Aufgabe abgeschlossen ist). Gehen Sie auf Fragen der Lehrkraft ein.**

- Nehmen Sie Ihren Platz im Raum ein. Dieser sollte außerhalb des Blickfeldes der Lehrkraft sein (die Lehrkraft soll das Gefühl haben „alleine im Raum zu sein"). **Geben Sie der Lehrkraft nun das Kommando zu beginnen.**

- **Notieren Sie hier die Reihenfolge, in der die Lehrkraft die Antworten korrigiert (sofern dies möglich ist):**

_____

(z. B. A, B, D, A, C, B, A, ...)

- **Wichtige Hinweise für die Leitung während der Aufzeichnung:**

  - **Nehmen Sie eine zurückhaltende Rolle ein**, während die Lehrkraft die Aufgabe bearbeitet. Greifen Sie nur ein, sollte die Lehrkraft aufhören zu sprechen. Geben Sie in diesem Fall **nur** den Impuls „*Reden Sie weiter" oder „Denken Sie auch daran zu verbalisieren, was Sie gerade tun".*

  - Da Sie die Aufgaben kennen, sind Sie evtl. dazu geneigt, die Lehrkraft, während Sie die Aufgabe bearbeitet, zu korrigieren oder zu unterstützen. **Vermeiden Sie dies unter allen Umständen!**

  - Evtl. wird die Lehrkraft Fragen bzgl. der Aufgabenstellung an Sie richten. Beantworten Sie diese **möglichst kurz** und **ohne** der Lehrkraft **zusätzliche Hinweise** zu geben, z. B. durch die Impulse „*Ja und denken Sie weiterhin laut*" oder „*Ja, blättern Sie auf die nächste Seite*".

8

Postinterview (10 min)

- **Stellen Sie der Lehrkraft folgende Fragen** (stellen Sie ggf. Nachfragen):
  Wir sind jetzt mit dem Teil fertig, in dem Sie laut Denken sollen.
  - *Wie geht es Ihnen denn jetzt gerade?*
  - *Wie sind Sie denn mit dem lauten Denken zurecht gekommen?*
  - *Was ist Ihnen leicht bzw. was ist Ihnen schwer gefallen?*

- Anschließend wird die fachlich-konzeptuelle Qualität und die Qualität der sprachlichen Realisierung der Schülerlösungstexte getrennt voneinander aus Sicht der Lehrkraft noch einmal beleuchtet. **Geben Sie hierzu folgende Instruktion:**
  Wir wollen nun noch einmal gemeinsam die Schülerantworten A bis D betrachten. Ich werde Ihnen nun je zwei Schülerantworten (z. B. Schülerantwort A und B) vorlegen und möchte Sie bitten folgende Frage zu beantworten:

- **Lesen Sie hierzu der Lehrkraft Instruktion 1 vor (siehe Seite 9).** Legen Sie ihr anschließend die Antwortpaare aus Tabelle 1 nacheinander vor (siehe Spalte „Antworten zur Auswahl").

- **Vermerken Sie das Urteil der Lehrkraft in Tabelle 1.**

- Fordern Sie die Lehrkraft ggf. noch einmal dazu auf ihre Entscheidung zu begründen. (z. B. wenn die Lehrkraft nur die Antwort „*A ist besser*" gibt).

- **Verfahren Sie anschließend in analoger Art und Weise mit Instruktion 2 und Tabelle 2 (siehe Seite 9).**

9

Instruktion 1:
*Beurteilen Sie, ob eine der beiden Antworten **fachlich besser ist, oder ob sie fachlich gleich gut sind**. Ob evtl. eine der beiden Antworten sprachlich besser ist, soll hierbei komplett unberücksichtigt bleiben. **Bitte begründen Sie Ihre Entscheidung**.*

Antworten zur Auswahl	Urteil der Lehrkraft?		
A und B	❍A	❍B	❍gleich
A und C	❍A	❍C	❍gleich
A und D	❍A	❍D	❍gleich
B und C	❍B	❍C	❍gleich
B und D	❍B	❍D	❍gleich
C und D	❍C	❍D	❍gleich

Tabelle 1

Instruktion 2:
*Beurteilen Sie, ob eine der beiden Antworten **sprachlich besser ist, oder ob sie sprachlich gleich gut sind**. Ob evtl. eine der beiden Antworten fachlich besser ist, soll hierbei komplett unberücksichtigt bleiben. **Bitte begründen Sie Ihre Entscheidung**.*

Antworten zur Auswahl	Urteil der Lehrkraft?		
A und B	❍A	❍B	❍gleich
A und C	❍A	❍C	❍gleich
A und D	❍A	❍D	❍gleich
B und C	❍B	❍C	❍gleich
B und D	❍B	❍D	❍gleich
C und D	❍C	❍D	❍gleich

Tabelle 2

10

Fragebogen; Schluss + Puffer (10 min)

- **Geben Sie folgende Instruktion:**
  Wir sind nun am Ende der Untersuchung angekommen. Sie erhalten nun noch einen kurzen Fragebogen, den ich Sie bitten würde auszufüllen. Vorab: Haben Sie noch irgendwelche Fragen oder möchten Sie etwas ergänzen, was sie gerne ansprechen würden?

- **Beenden Sie nach den Äußerungen der Lehrkraft die Aufnahme der Diktiergeräte.**

- Die Lehrkraft bekommt ca. 5 Minuten Zeit den Fragebogen auszufüllen.

- Bereiten Sie währenddessen ggf. den Raum für die nächste Lehrkraft vor.

- Legen Sie das ausgefüllte Manual, die bearbeiteten Aufgaben und den ausgefüllten Fragebogen in den Umschlag.

- Verabschieden Sie sich und bedanken Sie sich noch einmal für die Teilnahme der Lehrkraft.

D. Transkriptions- und Segmentierungsystem der Hauptstudie

D.1. Transkiptionssystem der Hauptstudie

Im Folgenden findet sich das Transkiptionssystem, das im Rahmen der Hauptstudie entwickelt wurde. Wie in Unterkapitel 6.2 beschrieben, diente es der Transkiption der in der Hauptstudie erhobenen Verbaldaten von im Schuldienst aktiven Physiklehrkräften. Als Grundlage für dieses Transkiptionssystem dienten die Arbeit von Fuß & Karbach (2014), sowie das Transkriptionssystem von Arras (2007, S. 191 u. f.). Zur besseren Übersicht ist das Transkriptionssystem in Anlehnung an Fuß & Karbach (2014) in Module unterteilt.

Modul Sprachglättung, Lautäußerungen, Interaktion

Bezeichnung	Beispiel(e)	Beschreibung
keine Glättung	(.) Okay, jetzt denk- mer erst ma-grad darüber nach, worauf die Aufgabe hinaus will. Ich hätte gesagt, es geht um (.) Schallübertragung, die ja da durch ein Medium geht. (.) So, 0 zwischen 5 Punkte.	Anwendung der literarischen Umschrift. Das heißt: • Beibehaltung von Dialekt und umgangssprachlicher Ausdrucksweisen • Beibehaltung fehlerhafter Ausdrücke • Beibehaltung eines fehlerhaften Satzbaus • Beibehaltung mundartlicher Ausdrücke
Lautäußerungen	Ich bin jetzt ähm in Zeile 5. (zustimmend) mhm (+) (verneinend) mhm (+)	Planungsäußerung werden transkribiert (z. B. ähm, mhm, öhm). Eindeutig zustimmende / ablehnende Lautäußerungen im Sinne von „ja“ / „nein“ werden kommentiert mit runden Klammern am Anfang und (+) am Ende transkribiert.
Interaktion	Also die Erklärung find- ich halt besser **(Leitung: Okay)**. Das wirkt fachlich sicherer.	Zuhörsignale und kurze Anmerkungen der Leitung werden **fett** und in runden Klammern im Fließtext gekennzeichnet.

Modul Pausen, Wortabbrüche, Satzabbrüche

Bezeichnung	Beispiel(e)	Beschreibung
Pausen	(.)	Pausen von bis zu einer Sekunde.
	(..)	Pausen zwischen einer und zwei Sekunden.
	(...)	Pausen von drei und mehr Sekunden.
Wortabbrüche	einf-	abgebrochenes Wort.
	Arbeits- äh –amt	Wiederaufnahme eines abgebrochenen Worts.
Satzabbrüche	Ich arbeite also in der... Ich hab zwei Arbeitsplätze.	unvollendete bzw. auslaufende Sätze (Fade-out) werden mit drei Auslassungspunkten gekennzeichnet.

Modul nicht-sprachliche Ereignisse, Sprachklang

Bezeichnung	Beispiel(e)	Beschreibung
non-verbale Äußerungen	(räuspert sich), (seufz), (lacht)	parasprachliche Äußerungen werden in Klammern im Fließtext durch runde Klammern vermerkt.
Lachende Betonung	(lachend) Mensch, so was habe ich noch nie gehört. (+)	Am Beginn einer lachend ausgesprochenen Sequenz wird (lachend) im Fließtext vermerkt. Das Ende der lachenden Aussprache wird mit (+) gekennzeichnet.
ironische Betonung	(ironisch) Mensch, so was habe ich noch nie gehört. (+)	Am Beginn einer ironisch ausgesprochenen Sequenz wird (ironisch) im Fließtext vermerkt. Das Ende der ironischen Aussprache wird mit (+) gekennzeichnet.
abwertende Betonung	(abwertend) Naja. (+)	Am Beginn einer abwertend ausgesprochenen Sequenz wird (abwertend) im Fließtext vermerkt. Das Ende der abwertenden Aussprache wird mit (+) gekennzeichnet.
Handlungen I	(schreibt „Vakuum“) (streicht „Vakuum“ durch)	Hörbare Handlungen werden als Kommentar im Fließtext durch runde Klammern vermerkt.
Handlungen II	Das ist aber schön. Schön erklärt. Dann schreiben wird noch dazu (schreibt) schöne Erklärung bzw. Vergleich. (+)	Am Beginn einer Handlungen (z. B. schreiben) die eindeutig während einer ausgesprochen Sequenz abläuft, wird die Handlung in runden Klammern im Fließtext vermerkt. Das Ende der Handlung, wird mit (+) gekennzeichnet.
Geräusche	(Telefon klingelt)	Hörbare Hintergrundgeräusche werden als Kommentar im Fließtext durch runde Klammern vermerkt.

Modul Unsicherheit, Unterbrechung und Auslassung

Bezeichnung	Beispiel(e)	Beschreibung
Unsicherheit in der Transkription	(...?)	unverständliches Wort
	(...??)	mehrere unverständliche Wörter
	(mein?)	unverständliches Wort mit vermutetem Wortlaut
	(mein?/dein?)	unverständliches Wort mit vermutetem unter alternativem Wortlaut
Auslassung	[...]	nicht transkribierte Gesprächssequenz

Modul Zeichensetzung

Bezeichnung	Beispiel(e)	Beschreibung
Zeichensetzung in Anlehnung an die grammatikalische Zeichensetzung	.	Satzende
	,	Aufzählen, Nebensätze, etc.
	?	Frage
	:	Ankündigung einer Aufzählung
	: „ “	wörtliche Rede

D.2. Segmentierungssystem der Hauptstudie

Im Folgenden findet sich das Segmentierungssystem, das im Rahmen der Hauptstudie entwickelt wurde. Wie in Unterkapitel 6.2 beschrieben, diente es der Segmentierung und Feintranskribierung des lauten Denkens von im Schuldienst aktiven Physiklehrkräften im Rahmen der Hauptstudie der vorliegenden Arbeit. Als Grundlage für dieses System dienten die Segmentierungssysteme von Ericsson & Simon (1985, S. 299 u. f.), van Someren et al. (1994, S. 117 u. f.), Chi (1997, S. 284 u. f.), A. Green (1998, S. 75 u. f.), C. Green & Gilhooly (2002, S. 60 u. f.), Hughes & Parkes (2003, S. 129) und Arras (2007, S. 194 u. f.). Zur besseren Übersicht ist das Transkriptionssystem in Anlehnung an Fuß & Karbach (2014) in Module unterteilt.

Modul Sonstige

Bezeichnung	Beispiel(e)	Beschreibung
Handlungen und Geräusche	(..) Mhm (..) *eine sogenannte Brücke haben* na gut unterringeln ma- das Brücke (unterringelt „Brücke“ in Zeile 9)	Handlungen und Geräusche, die unkommentiert sind oder zu denen sich nicht parallel geäußert wird, sind eigene Segmente (z. B. Umblättern).
Äußerungen der Leitung	(...) *da Ton über Glas geht.* (...) Warum geht Ton über Glas? (lacht) **Leitung: Bitte daran denken auch auszusprechen was man macht, nicht nur was man denkt.** Ach so, ja, okay.	Äußerungen der Leitung sind eigene Segmente. Sie werden nicht weiter zerlegt.
Sonstige	(..) *durch* (.) *das Ton geht* (lacht) Ahhhhh (...???) (schreibt) Warum (.) ? (+)	Äußerungen, die nicht klar zugeordnet werden können sind (vorläufig) eigene Segmente.

Modul Äußerungen zum Gelesen, zu Handlungen oder zu eigenen Gedanken

Bezeichnung	Beispiel(e)	Beschreibung
Äußerungen zu Gelesenem	Ähm (.) *für die Aufgabe (.) Weltraumspaziergang möchten Sie 0 bis maximal 5 Punkte vergeben.* Okay 5 Punkte. Is- schon mal -ne schwierige Aufgabe.	Äußerungen zu eben gelesen Textpassagen sind eigene Segmente.
Äußerungen zu eigenen Handlungen	(..) Mhm (..) *eine sogenannte Brücke haben* na gut unterringeln ma- das Brücke (unterringelt „Brücke“ in Zeile 9) und schreiben hin: (schreibt) direkter Kontakt (+)	Äußerungen, die das eigene Handeln betreffen sind eigene Segmente (z. B. Kommentare). Läuft eine Handlung parallel zum Sprechen ab, sind Handlung und Äußerung ein Segment.
Äußerungen zu eigenen Gedanken	(..) Gut. (..) Also erstelle ich da jetzt einen Erwartungshorizont. (..) Ich bin bloß grad die ganze Zeit am Überlegen, ob überhaupt Weltraumspaziergang. Na gut, ich bin jetzt im bayrischen Lehrplan in der neunten Klasse, ob sowas überhaupt kommt (.) im bayrischen Lehrplan is, glaub' ich, in der 9. Klasse des gar nicht drin (.) aber äh egal. (.) Ähm	Äußerungen, die eigene Gedanken betreffen sind eigene Segmente (z. B. Kommentare).

Modul besondere sprachliche Ereignisse

Bezeichnung	Beispiel(e)	Beschreibung
Schülerlösungstext / Aufgabenheft wird gelesen	*Im All ist nichts durch das Ton geht und er hört seinen Freund nicht.*	Sequenz, in denen Texte gelesen werden kursiv gesetzt.
Schülerlösungstext / Aufgabenheft wird anders als im Original gelesen	*Im Weltall ist nichts ... und er hört seinen Freund nicht.*	Sequenzen, in denen Texte gelesen werden, aber Wörter korrigiert oder Ausgelassen werden *kursiv* gesetzt und zusätzlich unterstrichen.
Paraphrasierung eines Schülerlösungstextes / des Aufgabenheftes	Er schreibt, dass sich die Astronauten dadurch nicht hören, weil im Weltraum nichts ist.	paraphrasierte Teile eines Textes werden durch Unterstreichung gekennzeichnet.
Eigener Erwartungshorizont wird aufgeschrieben, gelesen, zitiert oder paraphrasiert	Also da fällt weder Vakuum noch... Hah (.) Medium könnte man sagen.	Aufschreiben, lesen, zitieren und paraphrasieren des Erwartungshorizonts wird durch doppelte Unterstreichung gekennzeichnet.
Äußerungen der Leitung	(...) Warum geht Ton über Glas? (lacht) **Leitung: Bitte daran denken auch auszusprechen was man macht, nicht nur was man denkt.** Ach so, ja, okay.	Äußerungen der Leitung werden **fett** hervorgehoben.

Modul Lesen der Materialien

Bezeichnung	Beispiel(e)	Beschreibung
Lesen	Ähm (.) *für die Aufgabe (.) Weltraumspaziergang möchten Sie 0 bis maximal 5 Punkte vergeben.* Okay 5 Punkte. Is- schon mal -ne schwierige Aufgabe. *Aufgabe 1: Erstellen Sie für die oben beschriebene Situation geeigneten Erwartungshorizont für die Aufgabe Weltraumspaziergang.* (lachend) Herrje, kann ich die selber? (+)	Zusammenhängendes Lesen des Materials (Schülerlösungstexte, eigener Erwartungshorizont, Instruktionen im Aufgabenheft, usw.) sind eigene Segmente. Planungsäußerungen, Pause, Wortabbrüche, usw. sind nur Segmentgrenzen, wenn durch sie das zusammenhängende Lesen eindeutig unterbrochen wird (z. B. weil ein Äußerung bzgl. des eben Gelesen folgt).
wiederholtes Lesen	(...) Und (...) *weil der geringe Abstand zwischen den beiden eine bessere Funkverbindung herstellt*, (...) *weil der geringe Abstand zwischen den beiden eine bessere Funkverbindung herstellt* (..) find- ich, vom Satzbau her natürlich auch etwas ähm unglücklich. (...) Ich muss nach wie vor sagen, es steht ja eigentlich nichts in der Aufgabenstellung, dass die Funkverbindung komplett gekappt sein sollte.	Wird eine Textpassage erneut gelesen, ist dies ein eigenes Segment.
suchendes Lesen	(..) *Dann kommt aber Ton durch die Helme, da Ton über Glas geht.* (..) Des ist jetzt also der zweite Fall mit dem <u>Kontakt</u>. (..) Hat er jetzt eigentlich hier auch nicht wirklich geschrieben, dass des daran liegt, dass es einen <u>direkten Kontakt</u> gibt, sondern er schreibt nur, *da Ton über Glas geht.* Na, in der vierten Zeile ist auch noch mal <u>Ton</u>. Das unterstreich- mer auch noch gleich mal (unterstreicht „Ton“ in Zeile 4) (...) Also *über Glas geht* <u>geht</u> find ich auch gar nicht.	Das Scannen des Materials zur Suche nach Belegen wird in Segmente zerlegt. Jeder Scanschritt stellt ein eigenes Segment dar.

Modul Gedanken und Übergänge

Bezeichnung	Beispiel(e)	Beschreibung
Einzelgedanken	(...) Wenn sich keine Schallwellen ausbreiten können dann hört der eine Kamerad den andern nicht. Aber das würd- ich jetzt in meinen Erwartungshorizont nich- schreiben, weil das find ich jetzt einfach einfach dann einfach -ne ganz logische Schlussfolgerung. (...) Dann, ähm der 2. Fall (.) ist dann, als der ältere Astronaut den Kollegen festhält und den Helm an den anderen presst und dann kann der jüngere den älteren hören. Das ist dann natürlich so, dass wir eine Schallübertragung über die Helme haben.	Äußerung, die einen Einzelgedanken darstellen sind eigene Segmente (z. B. Nennung eines Phänomens). Planungsäußerungen, kurze Pause, Wortabbrüche, etc. sind nur Segmentgrenzen, wenn durch sie ein Einzelgedanke beendet wird. Einzelgedanken, die als miteinander verbunden erscheinen, werden voneinander getrennt (z. B. Nennung eines Phänomens und einer zugehörigen Wertung).
Übergänge	(schreibt) die Schallwelle (schmatzt) vom älteren (.) zum jüngeren Astronauten (..) übertragen werden. (+) (...) Mhm. (.) Jetz- ich überleg- grad wie ich's denn noch besser formuliere, warum das jetzt (.) wirklich funktioniert. (..) Mhm. (...) Würd- ich sagen es funktioniert hier, weil ja (.) wir kein Vakuum zwischendrin haben, sondern der Kontakt oder der Weg des Schalls nur über Medien geht, die Schall auch übertragen können. Also eventuell die Luft im Helm oder halt dann die Hülle des Helms.	Äußerungen, die einen Übergang von einem Gedanken einem zweiten Gedanken darstellen, werden dem Segment des zweiten Gedankens zugeschlagen (z. B. Planungsäußerung).

E. Kategoriensystem zur Analyse der Laut-Denk-Protokolle

Im Folgenden findet sich das Kategoriensystem, das im Rahmen der vorliegenden Arbeit entwickelt wurde, um die Laut-Denk-Protokolle der Physiklehrkräfte, die an der Hauptstudie teilgenommen haben, inhaltsanalytisch auszuwerten. Wie in Unterabschnitt 6.3.2.2 beschrieben, wurde dieses Kategoriensystem in einem deduktiv-induktiven Vorgehen entwickelt und basiert in Teilen auf den Kategoriensystemen, die Lumley (2005, S. 140 u. f.) und Arras (2007, S. 200 u. f.) in ihren Laut-Denk-Studien zur Feststellung und Beurteilung schriftlicher Leistungen in einer Fremdsprache entwickelten. Für die Anwendung des Kategoriensystems gelten folgende Regeln:

1. Die Laut-Denk-Protokolle liegen in segmentierter Form vor. Beim Codieren gilt die Regel, dass jedes Segment genau eine Codiereinheit (vgl. Schreier, 2012, S. 131 u. f.) darstellt.
2. Bis zu der Stelle, an der die Lehrkraft der Leitung ein Zeichen gibt, dass sie ihre Korrekturarbeit abgeschlossen hat, wird jedes Segment eines Protokolls codiert.
3. Da nicht jeder Kategorie in jedem Protokoll auftritt, wird segmentweise nacheinander und nicht kategorienweise codiert.
4. Einige Subkategorien des Kategoriensystems sind mit (Sub-)Subsubkategorien versehen. Je nach entsprechender Subkategorie sind einem Segement eine oder drei (Sub-)Subsubkategorien zuzuweisen. Details hierzu sind den entsprechenden Kategoriendefinitionen zu entnehmen.
5. Oft ist der Kontext ausschlaggebend, um entscheiden zu können welche Kategorie einem Segment zuzuordnen ist. Im Zweifelsfall sind daher auch die zum entsprechenden Segment unmittelbar benachbarten Segmente bei der Codierung zu beachten.
6. Bei der Codierung sind zudem die Regeln des Transkriptions- und Segemetierungssystems, mit dessen Hilfe die Laut-Denk-Protokolle angefertigt wurden, zu beachten (vgl. Anhang D). Oftmals schränkt die Transkription die Anzahl möglicher Codes bereits erheblich ein (z. B. bedeutet *kursive Schrift*, dass eine Lehrkraft etwas liest, weshalb hier nur die Kategorien mit der Codenummer 0.0 und 0.1 in Frage kommen).
7. In einigen Beispielen im Kategoriensystem sind Segmente in **grauer Schrift** gehalten. Dies bedeutet, dass das entsprechende Segment nicht der Kategorie entspricht, für das hier ein Beispiel gegeben werden soll, sondern nur der Anschaulichkeit dient.

1. lesen/erfassen eines Textes

Codenr.	Definition	Beispiele
1.0	zusammenhängendes Lesen einzelner Textabschnitte **Subsubkategorien (genau 1 ist zuzuweisen)**: 1.0.1 Anweisungen im Aufgabenheft 1.0.2 Aufgabe „Weltraumspaziergang“ 1.0.3 Schülerlösungstext	BEISPIEL 1 (Subsubkategorie 1.0.1): *Für die Aufgabe 2 benötigen Sie nur die Seiten 3 bis 8.* BEISPIEL 2 (Subsubkategorie 1.0.3) *Im All ist nichts (.) durch das Ton geht und er hört seinen Freund nicht.*
1.1	fragmentarisches Lesen (z. B. Scannen eines Textes/Search Reading zur Suche nach Belegen) **Subsubkategorien (genau 1 ist zuzuweisen)**: 1.1.1 Anweisungen im Aufgabenheft 1.1.2 Aufgabe „Weltraumspaziergang“ 1.1.3 Schülerlösungstext	Also: *… d- weil es Luft gibt.* Da würd- ich mal so’n Haken hinter setzen.
1.2	Paraphrasierung eines Textes **Subsubkategorien (genau 1 ist zuzuweisen)**: 1.2.1 Anweisungen im Aufgabenheft 1.2.2 Aufgabe „Weltraumspaziergang“ 1.2.3 Schülerlösungstext	Okay. 0 bis 5 Punkte. Und ich soll die... den Erwartungshorizont sozusagen (.) ähm für mich erstellen. Okay.

2. erstellen des Erwartungshorizonts zur Aufgabe Weltraumspaziergang

Codenr.	Definition	Beispiele
2.0	Beurteilungskriterien werden benannt/kommentiert/abgewogen/festgelegt (Handlung, Kommentar oder selbstreflektierte Äußerung, KEINE „activity descriptions“)	Also: (schreibt) 1. (+) Schreib- ich schonmal hin. Ähm (schreibt) Astronaut (.) 1 (.) schreit (+) Und wird nicht gehört. Das schreib- ich hier einmal hin. (schreibt „, wird nicht gehört“)
2.1	Zuweisung von Punkten im Erwartungshorizont (Handlung, Kommentar oder selbstreflektierte Äußerung, KEINE „activity descriptions“)	(.) Ähm ich würd- mal sagen, (.) von der Gewichtung her is- das obere 2, das untere is- deutlich mehr wert. [...] (liest) … Schall direkt von einem Helm zum anderen übertragen werden. (+) Das is- schonma-... das is- 1 Punkt wert. (zeichnet Haken hinter „kann der Schall direkt…“)

3. Korrektur der Schülerlösungstexte

Codenr.	Definition und Codierregel	Beispiele
3.0	Feststellung und Beurteilung eines Schülerlösungstextes (Handlung, Kommentar oder selbstreflektierte Äußerung, KEINE „activity descriptions“). Hierzu zählen: mündliche Kommentare, fällen/aushandeln einer endgültigen Entscheidung, Punkte (nicht) vergeben, schriftliche Beurteilungskommentare anfertigen, Schülerlösungstexte verbessern, usw. **Hinweis: Punkte zusammenzählen oder Vergeben der Gesamtpunkzahl wird mit den Codes „3.0.1.4“, „3.0.2.5“ versehen. Bei 5 Punkten wird der Code „3.0.3.1“ vergeben, bei 0 Punkte „3.0.3.2“ und in allen anderen Fällen der Code „3.0.3.3“ (vgl. BEISPIEL 3).** **Je Subsubkategorie ist genau 1 Subsubsubkategorie zuzuweisen:**	BEISPIEL 1 (Subsubsubkategorie 3.0.1.1, 3.0.2.1, 3.0.3.1): Also: *Im All ist nichts, durch das Ton geht…* Das wäre ja so (.) mein (.) Medium. Auch wenn dann jetz- hier als äh Fachbegriff nich- kommt. BEISPIEL 2 (Subsubsubkategorie 3.0.1.2, 3.0.2.1, 3.0.3.2): Also: *Im All ist nichts, durch das Ton geht…* Das wäre ja so (.) mein (.) Medium. Auch wenn dann jetz- hier als äh Fachbegriff nich- kommt. BEISPIEL 3 (Subsubsubkategorie 3.0.1.4, 3.0.2.5, 3.0.3.3): Also von den 3 Punkten is- es maximal die Hälfte. Joar doch. 1,5. (schreibt „1,5“ hinter Zeile 3) Dann komm- ich auf 3 von 5. BEISPIEL 4 (Subsubsubkategorie 3.0.1.4, 3.0.2.4, 3.0.3.2): *… weil die Frequenz nicht gut genug war.* Uiuiui! (..) Da würd- ich mal meinen… (..) da geht… da is- nicht viel beim Schüler angekommen.

Subsubkategorie	Subsubsubkategorien
3.0.1 Welches Merkmal wird fokussiert?	3.0.1.1 fachlich-konzeptueller Eindruck
	3.0.1.2 sprachliche Realisierung
	3.0.1.3 sonstiges Merkmal (Textlänge, Rechtschreibung, Zeichensetzung, Handschrift, Gliederung, usw.)
	3.0.1.4 mehrere Merkmale/nicht eindeutig/allgemeiner Eindruck
3.0.2 In Bezug zu was wird der Schülerlösungstext verortet?	3.0.2.1 sachliches Kriterium
	3.0.2.2 andere Schülerlösungstexte
	3.0.2.3 Erfahrungen mit Physiklernenden allgemein
	3.0.2.4 mutmaßliches Personmerkmal des_der Schülers_Schülerin (Geschlecht, Herkunft, Alter, intellektuelle Reife, usw.)
	3.0.2.5 Mehrere Bezugspunkte/Bezug nicht eindeutig
3.0.3 Wie wird sich geäußert?	3.0.3.1 positiv wertend/akzeptierend
	3.0.3.2 negativ wertend/ablehnend
	3.0.3.3 neutral/gemischt/sonstig

3. Korrektur der Schülerlösungstexte

Codenr.		Definition und Codierregel	Beispiele
		sachliche Kriterien zur Beurteilung eines Schülerlösungstextes werden **ad hoc benannt/abgewogen** oder Kriterien aus dem Erwartungshorizont werden **paraphrasiert/zitiert/abgewogen**.	*Die beiden Astronauten können sich wieder hören, weil der geringe Abstand zwischen den beiden Funkgeräten eine bessere Funkverbindung herstellt. Deswegen kann der jüngere den älteren wieder leise hören.*
		Subsubkategorien (genau 1 ist zuzuweisen)::	Äh.
3.1	3.1.1	fachlich-konzeptueller Eindruck	(...) Is- die Frage:
	3.1.2	sprachliche Realisierung	Kann das tatsächlich... kann das technisch, physikalisch genauso sein?
	3.1.3	sonstiges Merkmal (Textlänge, Rechtschreibung, Zeichensetzung, Handschrift, Gliederung/strukturelle Ordnung, usw.)	(.) Ist das überhaupt denkbar, dass es so is-?
	3.1.4	mehrere Merkmale/nicht eindeutig/allgemeiner Eindruck	(..) Äh (..) das is-... grundsätzlich kann es sein, dass die... (.) wenn die Funkverbindung abreißt, dass das am Funkgerät...
			Natürlich, das liecht am Funkgerät.

4. Äußerungen außerhalb der eigentlichen Korrektur und Erwartungshorizonterstellung

Codenr.	Definition und Codierregel	Beispiele
4.0	Beschreibung allgemeiner Handlungsstrategien beim Beurteilen oder Äußerungen zum allgemeinen Vorgehen/zu allgemeinen Erfahrungen beim Beurteilen (Kommentar, Kritik oder selbstreflektierte Äußerung).	BEISPIEL 1: So dann fang- ich... wie üblich fang- ich bei dem an, der (.) aus meiner Sicht die beste Lösung hat. BEISPIEL 2: Ich markier- mir das mal so 1 Punkt. Ich möchte aber gleich nochma- vergleichen mit äh Antwort B, C, D.
4.1	Kommentar zur, Kritik an, Fragen zur oder Interpretation der Aufgabe „Weltraumspaziergang“. **Ausgenommen sind Äußerungen, die der Definition eines Codes mit Nummer 2.X oder 3.X entsprechen.**	Und jetz- gehts um mein Erwartungshorizont: Ähm (.) mhm dahinter steckt ja ähm des Thema Akustik. Is- eigentlich schon ähm in Klasse 7 bei uns dran gewesen. Oder (.) nochmal dran gewesen, weil wir hier so’n Projekt haben. Ähm das heißt (.) eigentlich wissen die Schüler wie das äh funktioniert die Schallübertragung.
4.2	Kommentar zu, Kritik an, Fragen zu oder Interpretation von Anweisungen in oder zum Aufbau des Aufgabenhefts.	Okay. Antwort in ganze Sätzen. Schreib- ich auch immer auf meine Klassenarbeiten. Das passt zu mir.
4.3	Beschreibung allgemeiner Handlungsstrategien zum Erstellen eines Erwartungshorizonts oder Kommentar zum, Kritik oder selbstreflektierende Äußerung am eigenen Erwartungshorizont allgemein.	Also: Kariertes Papier verwenden, unter normalen Umständen... Ja unter normalen Umständen (..) äh habe ich ein- kompletten Text oder Stichworte dazu. [...] Natürlich (.) sollte sowas da auch erscheinen, welchen Weg der Schall insgesamt nimmt. Aber wie gesacht, dass würd- ich gegebenenfalls nach Durchsicht der Lösungen, würd- ich das nochmal vielleicht umstricken. Aber so zunächst Mal würd- ich das als Erwartungshorizont so stehen lassen.

5. emotionale Äußerungen und nichtsprachliche Ereignisse

Codenr.	Definition und Codierregel	Beispiele
5.0	Lachen, Stöhnen, Ausdruckspartikel (So!, Gut!, usw.), Planungs- und Verzögerungslaute (ähm, mhm, tk- tk- tk-, usw.), usw.	Also ich bin da schon auch dabei 0 Punkte zu geben. Ähm. (...) (fragend) Mhm. (+) [...] (.) Die Frequenz spielt hier ja garkeine Rolle. (..) Äh. (..) (schnallst mit der Zunge) Joar.
5.1	Klanggesten (Fingerschnippen, Klatschen, Klopfen, usw.)	(..) (kopft mit dem Stift auf den Tisch)

6. sonstige Äußerungen/Artefakte des lauten Denkens/sonstige nichtsprachliche Ereignisse

Codenr.	Definition und Codierregel	Beispiele
6.0	„activity descriptions“, sonstige Handlungen (blättern, trinken,...) oder Äußerungen zu sonstigen eigenen Handlungen/Verhaltensweisen/Gedanken während der Erhebung.	(..) Dann blätter- ich mal um. (legt Stift aus der Hand) Stift aus der Hand gelegt. (blättert auf Seite 4) [...] *Im Weltall herrscht ein Vakuum...* Da würd- ich mir so -ne Markierung ran machen für 1 Punkt. (zeichnet Punkt über „Vakuum“ in Zeile 2) Oder'n Haken.
6.1	Ankündigungen (also:, hier:,...), weitere Äußerungen, weitere Transkriptsegmente	Also (.) die Aufgabe heißt: *Erkläre beide Phänomene.* [...] (Schulglocke klingelt)

F. Einschätzungen der Teilnehmer_innen im Rahmen der Paarvergleiche der retrospektiven Befragung

In Tabelle F.1 und F.2 sind die Einschätzungen aller 21 Teilnehmer_innen der vier Schülerlösungstexte in den Paarvergleichen der retrospektiven Befragung zusammengefasst. Die dargestellte Übersicht basiert auf den Mitschriften der Leitung zur retrospektiven Befragung, deren Korrektheit im Rahmen der inhaltlich strukturierenden qualitativen Inhaltsanalyse der retrospektiven Befragung überprüft wurde (vgl. Abschnitt 6.4.1). Zudem sind in der letzten Spalte von Tabelle F.1 und F.2 die einzelnen Einschätzungen, die ein_e Teilnehmer_in im Rahmen der Paarvergleiche vorgenommen hat, zu einer Rangreihe der vier Schülerlösungstexte bezüglich ihrer fachlich-konzeptuellen Qualität bzw. der Qualität ihrer sprachlichen Realisierung zusammenfasst. Diese Rangreihen bildeten die Grundlage für die quantitative Analyse der Einschätzungen der 21 Teilnehmer_innen im Rahmen der Paarvergleiche (vgl. Abschnitt 6.4.2).

Pseudonym	*„Beurteilen Sie, ob eine der beiden Antworten **fachlich besser ist, oder ob sie fachlich gleich gut sind.** Ob evtl. eine der beiden Antworten sprachlich besser ist, soll hierbei komplett unberücksichtigt bleiben. **Bitte begründen Sie Ihre Entscheidung.**“* (Anhang C.3 Durchführungsmanual, S. 9, Hervorhebung im Original)						**Zusammenfassung der Einschätzungen in eine Rangreihe**
	Schülerlösungstext A und B	**Schülerlösungstext A und C**	**Schülerlösungstext A und D**	**Schülerlösungstext B und C**	**Schülerlösungstext B und D**	**Schülerlösungstext C und D**	
Herr Abney	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Balke	B < A	A < C	D < A	B < C	D = B	D < C	D = B < A < C
Herr Carboni	B < A	A < C	D < A	B < C	D = B	D < C	D = B < A < C
Herr Dassow	B < A	A < C	D < A	B < C	D = B	D < C	D = B < A < C
Herr Einert	B < A	A < C	D < A	B < C	D = B	D < C	D = B < A < C
Herr Feldner	B < A	A < C	D < A	B < C	D = B	D < C	D = B < A < C
Herr Geppert	B = A	A < C	D < A	B < C	D < B	D < C	D < B = A < C
Herr Hastedt	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Iezzi	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Jonuzi	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Frau Kirik	B < A	A < C	D < A	B < C	D = B	D < C	D = B < A < C
Herr Lemos	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Mehlert	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Frau Novack	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Onne	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Frau Pinna	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Quezada	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Rittershaus	B < A	A < C	D < A	B < C	D = B	D < C	D = B < A < C
Frau Sohm	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Trummer	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C
Herr Uckermark	B < A	A < C	D < A	B < C	D < B	D < C	D < B < A < C

Tabelle F.1.: Einschätzungen aller 21 Teilnehmer_innen bezüglich der fachlich-konzeptuellen Qualität der vier Schülerlösungstexte in den Paarvergleichen der retrospektiven Befragung.

Pseudonym	*„Beurteilen Sie, ob eine der beiden Antworten* ***sprachlich besser ist, oder ob sie sprachlich gleich gut sind****. Ob evtl. eine der beiden Antworten fachlich besser ist, soll hierbei komplett unberücksichtigt bleiben.* ***Bitte begründen Sie Ihre Entscheidung.****“* (Anhang C.3 Durchführungsmanual, S. 9, Hervorhebung im Original)						**Zusammenfassung der Einschätzungen in eine Rangreihe**
	Schülerlösungstext A und B	**Schülerlösungstext A und C**	**Schülerlösungstext A und D**	**Schülerlösungstext B und C**	**Schülerlösungstext B und D**	**Schülerlösungstext C und D**	
Herr Abney	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Herr Balke	A < B	A < C	D < A	C < B	D < B	D < C	D < A < C < B
Herr Carboni	A < B	A < C	D = A	B = C	D < B	D < C	D = A < B = C
Herr Dassow	A < B	A < C	D = A	B = C	D < B	D < C	D = A < B = C
Herr Einert	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Herr Feldner	A < B	A < C	D < A	B = C	D < B	D < C	D < A < B = C
Herr Geppert	A < B	A < C	D = A	B = C	D < B	D < C	D = A < B = C
Herr Hastedt	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Herr Iezzi	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Herr Jonuzi	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Frau Kirik	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Herr Lemos	A < B	A < C	A < D	B = C	D < B	D < C	A < D < B = C
Herr Mehlert	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Frau Novack	A < B	A < C	D < A	B = C	D < B	D < C	D < A < B = C
Herr Onne	A < B	A < C	A < D	B = C	D < B	D < C	A < D < B = C
Frau Pinna	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Herr Quezada	A = B	A < C	D < A	B < C	D < B	D < C	D < A = B < C
Herr Rittershaus	A < B	A < C	D = A	B < C	D < B	D < C	D = A < B < C
Frau Sohm	A < B	A < C	D < A	B < C	D < B	D < C	D < A < B < C
Herr Trummer	A < B	A < C	D = A	C < B	D < B	D < C	D = A < C < B
Herr Uckermark	A < B	A < C	D = A	B = C	D < B	D < C	D = A < B = C

Tabelle F.2.: Einschätzungen aller 21 Teilnehmer_innen bezüglich der Qualität der sprachlichen Realisierung der vier Schülerlösungstexte in den Paarvergleichen der retrospektiven Befragung.

Abstract

Im Physikunterricht sind neben fachlichen Inhalten viele Anforderungen an Schüler_innen sprachbezogen. Zugleich wird von (Physik-)Lehrkräften Sprache oft nicht als Lerngegenstand angesehen, sondern es als selbstverständlich vorausgesetzt, dass sich Schüler_innen versiert sprachlich ausdrücken können. Eine naheliegende Vermutung ist, dass sich diese Erwartungshaltung auch auf die Korrektur von Klassenarbeiten niederschlägt. Darüber, wie Physiklehrkräfte in ihrer täglichen Berufspraxis bei der Feststellung und Beurteilung schriftlicher, aus einer Klassenarbeit stammender Schülerleistungen tatsächlich vorgehen, welchen Logiken sie dabei folgen und welche Maßstäbe sie für angemessen halten, liegt allerdings bis dato keine belastbare empirische Evidenz vor (Forschungsfrage 1). Selbiges gilt für die Frage, inwieweit Physiklehrkräfte bei der Korrektur einer Klassenarbeit fachlich-konzeptuelle und sprachliche Schülerleistungen miteinander konfundieren (Forschungsfrage 2). Die vorliegende Arbeit exploriert diese beiden Fragen. Den theoretischen Referenzrahmen stellt dabei das Konzept einer Assessment Literacy von Lehrkräften dar, das professionelles Lehrerhandeln im Kontext von schulischer Leistungsfeststellung und -beurteilung beschreibt.

In einer Entwicklungsstudie wurde zunächst ein gegenstandsangemessenes Erhebungssetting entwickelt. Bei diesem werden im Schuldienst aktive Physiklehrkräfte gebeten, lautdenkend einen Erwartungshorizont zu einer Klassenarbeitsaufgabe so zu erstellen, wie sie dies in ihrer täglichen Berufspraxis auch tun würden. Mit Hilfe ihres Erwartungshorizonts korrigieren die Lehrkräfte anschließend (ebenfalls laut-denkend) vier auf sprachlicher und fachlich-konzeptueller Ebene stark unterschiedliche Schülerlösungstexte. Schließlich werden die Schülerlösungstexte den Lehrkräften in einer retrospektiven Befragung erneut vorgelegt. Dabei werden sie gebeten, die fachlich-konzeptuelle Qualität und die Qualität der sprachlichen Realisierung der vier Schülerlösungstexte getrennt voneinander einzuschätzen. Die Klassenarbeitsaufgabe, mit der die Lehrkräfte konfrontiert werden, wurde aus einem Pool von Aufgaben, die Physiklehrkräfte tatsächlich in Klassenarbeiten eingesetzt haben, ausgewählt. Die vier Schülerlösungstexte wurden mit Hilfe eines mehrschrittigen Codierverfahrens aus insgesamt 116 Texten ausgewählt, die von Schüler_innen der neunten Jahrgangsstufe verfasst wurden.

Die Erhebung der Hauptstudie, in der das entwickelte Erhebungssetting eingesetzt wurde, fand von April bis September 2016 statt. Es wurde eine heterogene Gelegenheitsstichprobe von N = 21 Hamburger Physiklehrkräften gewonnen. Die erhobenen Daten wurden in einem Mixed-Methods-Triangulationsdesign zunächst sowohl qualitativen als auch quantitativen Analysen unterzogen. Anschließend wurden die zuvor gewonnenen qualitativen

und quantitativen Teilbefunde zu beiden Forschungsfragen in ein Gesamtbild zusammengeführt.

Bezüglich Forschungsfrage 1 zeigte sich, dass die Teilnehmer_innen im Erhebungssetting ein facettenreiches und überwiegend angemessenes Wissen und Können zu (fachspezifischer) Leistungsfeststellung und -beurteilung nutzten. Des Weiteren handelten sie auch auf Basis ihrer berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung, sowie ihrer Bezugsnormorientierungen. Wie sich allerdings auch zeigte, wurden von den Teilnehmer_innen vor allem fachlich-konzeptuelle Merkmale der Schülerlösungstexte beachtet. Sprachliche Merkmale spielten bei der Leistungsfeststellung und -beurteilung hingegen eine untergeordnete Rolle. Ferner erfolgte die Feststellung und Beurteilung sprachlicher Schülerleistungen in einer tendenziell defizitorientierten Art und Weise. Fachlich-konzeptuelle Schülerleistungen wurden hingegen zum Teil defizitorientiert, in Teilen aber auch fähigkeitsorientiert festgestellt und beurteilt. Hinsichtlich Forschungsfrage 2 lieferten sowohl die qualitativen, als auch die quantitativen Analysen der erhobenen Daten zahlreiche einander komplementäre Teilbefunde. In ihrer Gesamtheit sprechen diese Teilbefunde dafür, dass die Teilnehmer_innen im Erhebungssetting fachlich-konzeptuelle und sprachliche Schülerleistungen auf einem moderaten Niveau miteinander konfundierten. Im Sinne des Konzepts einer Assessment Literacy von Lehrkräften bedeutet dies, dass es den Teilnehmer_innen nur bedingt gelungen ist, das Spannungsverhältnis zwischen ihrem Wissen und Können, sowie ihren berufsbezogenen Überzeugungen zu schulischer Leistungsfeststellung und -beurteilung und der Kontextbedingung des Erhebungssettings (fachlich-konzeptuell und sprachlich stark unterschiedliche Schülerlösungstexte korrigieren zu müssen) aufzulösen.

In Summe liefert die vorliegende Arbeit reichhaltige Einblicke in die bislang wenig erforschte fachlich-konzeptuelle und sprachliche Leistungsfeststellung und -beurteilung durch im Schuldienst aktive Physiklehrkräfte. Vor allem der Befund einer moderaten Konfundierung fachlich-konzeptueller und sprachlicher Schülerleistungen eröffnet die weiterführende Frage, welche Aus- und Weiterbildungsmaßnahmen für (angehende) Physiklehrkräfte ergriffen werden können und sollten, um dem entgegen zu wirken. Da die analysierten Daten allerdings in einer vergleichsweise kleinen Gelegenheitsstichprobe erhoben wurden, liefert die vorliegende Arbeit keine verallgemeinerbaren Erkenntnisse. Vielmehr ist die Übertragbarkeit der gewonnenen Befunde davon abhängig, inwieweit sich die Merkmale des Untersuchungskontextes der vorliegenden Arbeit auch im angestrebten Übertragungskontext wiederfinden. Insbesondere sind die Ergebnisse dieser Arbeit als Motivation für zukünftige, dann vielleicht auch konfirmatorisch angelegte Studien zu verstehen.

Abstract in englischer Sprache

Many requirements for students in physics classes, besides the subject-specific content, are language related. At the same time, (physics) teachers often do not regard language as something that should be taught but take it for granted that students are adept in applying adequate language repertoires on a high level. An obvious assumption is that this expectation is also reflected in how physics teachers assess students' performances in teacher-made tests. To date, however, there is no reliable empirical evidence as to how physics teachers, in their daily professional practice, actually proceed, the logic that they follow, and the standards they deem appropriate when assessing written student performances derived from a teacher-made test (research question 1). The same applies to the question as to what extent physics teachers confound subject-specific conceptual and linguistic student achievement when they assess the quality of a student's performance in a teacher-made test (research question 2). The present study explores both these questions. The theoretical frame of reference presents the concept of teachers' assessment literacy, which characterizes a teacher who acts professionally in the context of academic performance assessment.

In an initial study, the following survey setting was developed: In-service physics teachers create an answer key for a short essay question thinking out loud, as they would in their daily professional practice. With the help of their answer key, teachers then assess (also thinking out loud) four student texts, which differ substantially at a linguistic and subject-specific conceptual level. Finally, the student texts are re-submitted to the teachers in a retrospective interview. Thereby, they separately rate the conceptual quality and the quality of the linguistic realization of the four student texts. In the initial study, the short essay question was selected from a pool of essay questions, that actually have been used in teacher-made-tests. The four student texts were selected from a total of 116 texts written by 9th-grade students, applying a multi-step coding procedure.

The survey of the main study, in which the developed survey setting was used, took place from April to September 2016. In total, a convenience sample of 21 German secondary school physics teachers from Hamburg participated. The Data analysis followed a mixed-method triangulation design. Collected data were first subjected to both qualitative and quantitative analyses. Subsequently, the previously obtained qualitative and quantitative findings on both research questions were merged into the overall findings.

Regarding research question 1, it was shown that participants used a multi-faceted and predominantly adequate knowledge and ability for (subject-specific) performance assessment. Furthermore, they also acted on the basis of their beliefs about academic perfor-

mance assessments, as well as their reference norm orientations. However, as it turned out, the participants considered primarily the subject-specific conceptual features of the student texts. Linguistic characteristics played a minor role in their performance assessment. In addition, the assessment of student linguistic achievement showed a tendency of a deficit-orientation. In contrast, subject-specific conceptual student achievement was assessed partly deficit-oriented, but also partly in an ability-oriented manner. Regarding research question 2, both the qualitative and the quantitative data analyses yielded partial findings complementary to each other. In their entirety, these partial findings suggest that participants confound at a moderate level subject-specific conceptual and linguistic student achievement. In the sense of the concept of teachers' assessment literacy, this means that participants have only partially succeeded in balancing tensions between their knowledge and abilities, as well as their beliefs on academic performance assessment and the condition set by the context of the survey setting (to assess student texts, which differ substantially on a linguistic and subject-specific conceptual level).

In sum, the present study provides rich insights into the hitherto poorly researched assessment of subject-specific conceptual and linguistic achievement by in-service physics teachers. Above all the finding of a moderate confounding of subject-specific conceptual and linguistic student achievement opens the further question as to which education and training measures for (prospective) physics teachers can and should be taken to counteract this situation. However, since data analyzed were collected from a relatively small convenience sample, the present study does not provide findings which can be generalized. Rather, the transferability of the findings depends on the extent to which the characteristics of the survey setting are also reflected in the desired transference context. In particular, the results of this study should be understood as motivation for future research, and perhaps also conformational studies.

Bisher erschienene Bände der Reihe *„Studien zum Physik- und Chemielernen“*

ISSN 1614-8967 (vormals *Studien zum Physiklernen* ISSN 1435-5280)

1 Helmut Fischler, Jochen Peuckert (Hrsg.): Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie
ISBN 978-3-89722-256-4 40.50 EUR

2 Anja Schoster: Bedeutungsentwicklungsprozesse beim Lösen algorithmischer Physikaufgaben. *Eine Fallstudie zu Lernprozessen von Schülern im Physiknachhilfeunterricht während der Bearbeitung algorithmischer Physikaufgaben*
ISBN 978-3-89722-045-4 40.50 EUR

3 Claudia von Aufschnaiter: Bedeutungsentwicklungen, Interaktionen und situatives Erleben beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-89722-143-7 40.50 EUR

4 Susanne Haeberlen: Lernprozesse im Unterricht mit Wasserstromkreisen. *Eine Fallstudie in der Sekundarstufe I*
ISBN 978-3-89722-172-7 40.50 EUR

5 Kerstin Haller: Über den Zusammenhang von Handlungen und Zielen. *Eine empirische Untersuchung zu Lernprozessen im physikalischen Praktikum*
ISBN 978-3-89722-242-7 40.50 EUR

6 Michaela Horstendahl: Motivationale Orientierungen im Physikunterricht
ISBN 978-3-89722-227-4 50.00 EUR

7 Stefan Deylitz: Lernergebnisse in der Quanten-Atomphysik. *Evaluation des Bremer Unterrichtskonzepts*
ISBN 978-3-89722-291-5 40.50 EUR

8 Lorenz Hucke: Handlungsregulation und Wissenserwerb in traditionellen und computergestützten Experimenten des physikalischen Praktikums
ISBN 978-3-89722-316-5 50.00 EUR

9 Heike Theyßen: Ein Physikpraktikum für Studierende der Medizin. *Darstellung der Entwicklung und Evaluation eines adressatenspezifischen Praktikums nach dem Modell der Didaktischen Rekonstruktion*
ISBN 978-3-89722-334-9 40.50 EUR

10 Annette Schick: Der Einfluß von Interesse und anderen selbstbezogenen Kognitionen auf Handlungen im Physikunterricht. *Fallstudien zu Interessenhandlungen im Physikunterricht*
ISBN 978-3-89722-380-6 40.50 EUR

11 Roland Berger: Moderne bildgebende Verfahren der medizinischen Diagnostik. *Ein Weg zu interessanterem Physikunterricht*
ISBN 978-3-89722-445-2 40.50 EUR

12 Johannes Werner: Vom Licht zum Atom. *Ein Unterrichtskonzept zur Quantenphysik unter Nutzung des Zeigermodells*
ISBN 978-3-89722-471-1 40.50 EUR

13 Florian Sander: Verbindung von Theorie und Experiment im physikalischen Praktikum. *Eine empirische Untersuchung zum handlungsbezogenen Vorverständnis und dem Einsatz grafikorientierter Modellbildung im Praktikum*
ISBN 978-3-89722-482-7 40.50 EUR

14 Jörn Gerdes: Der Begriff der physikalischen Kompetenz. *Zur Validierung eines Konstruktes*
ISBN 978-3-89722-510-7 40.50 EUR

15 Malte Meyer-Arndt: Interaktionen im Physikpraktikum zwischen Studierenden und Betreuern. *Feldstudie zu Bedeutungsentwicklungsprozessen im physikalischen Praktikum*
ISBN 978-3-89722-541-1 40.50 EUR

16 Dietmar Höttecke: Die Natur der Naturwissenschaften historisch verstehen. *Fachdidaktische und wissenschaftshistorische Untersuchungen*
ISBN 978-3-89722-607-4 40.50 EUR

17 Gil Gabriel Mavanga: Entwicklung und Evaluation eines experimentell- und phänomenorientierten Optikcurriculums. *Untersuchung zu Schülervorstellungen in der Sekundarstufe I in Mosambik und Deutschland*
ISBN 978-3-89722-721-7 40.50 EUR

18 Meike Ute Zastrow: Interaktive Experimentieranleitungen. *Entwicklung und Evaluation eines Konzeptes zur Vorbereitung auf das Experimentieren mit Messgeräten im Physikalischen Praktikum*
ISBN 978-3-89722-802-3 40.50 EUR

19 Gunnar Friege: Wissen und Problemlösen. *Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs*
ISBN 978-3-89722-809-2 40.50 EUR

20 Erich Starauschek: Physikunterricht nach dem Karlsruher Physikkurs. *Ergebnisse einer Evaluationsstudie*
ISBN 978-3-89722-823-8 40.50 EUR

21 Roland Paatz: Charakteristika analogiebasierten Denkens. *Vergleich von Lernprozessen in Basis- und Zielbereich*
ISBN 978-3-89722-944-0 40.50 EUR

22 Silke Mikelskis-Seifert: Die Entwicklung von Metakonzepten zur Teilchenvorstellung bei Schülern. *Untersuchung eines Unterrichts über Modelle mithilfe eines Systems multipler Repräsentationsebenen*
ISBN 978-3-8325-0013-9 40.50 EUR

23 Brunhild Landwehr: Distanzen von Lehrkräften und Studierenden des Sachunterrichts zur Physik. *Eine qualitativ-empirische Studie zu den Ursachen*
ISBN 978-3-8325-0044-3 40.50 EUR

24 Lydia Murmann: Physiklernen zu Licht, Schatten und Sehen. *Eine phänomenografische Untersuchung in der Primarstufe*
ISBN 978-3-8325-0060-3 40.50 EUR

25 Thorsten Bell: Strukturprinzipien der Selbstregulation. *Komplexe Systeme, Elementarisierungen und Lernprozessstudien für den Unterricht der Sekundarstufe II*
ISBN 978-3-8325-0134-1 40.50 EUR

26 Rainer Müller: Quantenphysik in der Schule
ISBN 978-3-8325-0186-0 40.50 EUR

27 Jutta Roth: Bedeutungsentwicklungsprozesse von Physikerinnen und Physikern in den Dimensionen Komplexität, Zeit und Inhalt
ISBN 978-3-8325-0183-9 40.50 EUR

28 Andreas Saniter: Spezifika der Verhaltensmuster fortgeschrittener Studierender der Physik
ISBN 978-3-8325-0292-8 40.50 EUR

29 Thomas Weber: Kumulatives Lernen im Physikunterricht. *Eine vergleichende Untersuchung in Unterrichtsgängen zur geometrischen Optik*
ISBN 978-3-8325-0316-1 40.50 EUR

30 Markus Rehm: Über die Chancen und Grenzen moralischer Erziehung im naturwissenschaftlichen Unterricht
ISBN 978-3-8325-0368-0 40.50 EUR

31 Marion Budde: Lernwirkungen in der Quanten-Atom-Physik. *Fallstudien über Resonanzen zwischen Lernangeboten und SchülerInnen-Vorstellungen*
ISBN 978-3-8325-0483-0 40.50 EUR

32 Thomas Reyer: Oberflächenmerkmale und Tiefenstrukturen im Unterricht. *Exemplarische Analysen im Physikunterricht der gymnasialen Sekundarstufe*
ISBN 978-3-8325-0488-5 40.50 EUR

33 Christoph Thomas Müller: Subjektive Theorien und handlungsleitende Kognitionen von Lehrern als Determinanten schulischer Lehr-Lern-Prozesse im Physikunterricht
ISBN 978-3-8325-0543-1 40.50 EUR

34 Gabriela Jonas-Ahrend: Physiklehrervorstellungen zum Experiment im Physikunterricht
ISBN 978-3-8325-0576-9 40.50 EUR

35 Dimitrios Stavrou: Das Zusammenspiel von Zufall und Gesetzmäßigkeiten in der nichtlinearen Dynamik. *Didaktische Analyse und Lernprozesse*
ISBN 978-3-8325-0609-4 40.50 EUR

36 Katrin Engeln: Schülerlabors: authentische, aktivierende Lernumgebungen als Möglichkeit, Interesse an Naturwissenschaften und Technik zu wecken
ISBN 978-3-8325-0689-6 40.50 EUR

37 Susann Hartmann: Erklärungsvielfalt
ISBN 978-3-8325-0730-5 40.50 EUR

38 Knut Neumann: Didaktische Rekonstruktion eines physikalischen Praktikums für Physiker
ISBN 978-3-8325-0762-6 40.50 EUR

39 Michael Späth: Kontextbedingungen für Physikunterricht an der Hauptschule. *Möglichkeiten und Ansatzpunkte für einen fachübergreifenden, handlungsorientierten und berufsorientierten Unterricht*
ISBN 978-3-8325-0827-2 40.50 EUR

40 Jörg Hirsch: Interesse, Handlungen und situatives Erleben von Schülerinnen und Schülern beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-8325-0875-3 40.50 EUR

41 Monika Hüther: Evaluation einer hypermedialen Lernumgebung zum Thema Gasgesetze. *Eine Studie im Rahmen des Physikpraktikums für Studierende der Medizin*
ISBN 978-3-8325-0911-8 40.50 EUR

42 Maike Tesch: Das Experiment im Physikunterricht. *Didaktische Konzepte und Ergebnisse einer Videostudie*
ISBN 978-3-8325-0975-0 40.50 EUR

43 Nina Nicolai: Skriptgeleitete Eltern-Kind-Interaktion bei Chemiehausaufgaben. *Eine Evaluationsstudie im Themenbereich Säure-Base*
ISBN 978-3-8325-1013-8 40.50 EUR

44 Antje Leisner: Entwicklung von Modellkompetenz im Physikunterricht
ISBN 978-3-8325-1020-6 40.50 EUR

45 Stefan Rumann: Evaluation einer Interventionsstudie zur Säure-Base-Thematik
ISBN 978-3-8325-1027-5 40.50 EUR

46 Thomas Wilhelm: Konzeption und Evaluation eines Kinematik/Dynamik-Lehrgangs zur Veränderung von Schülervorstellungen mit Hilfe dynamisch ikonischer Repräsentationen und graphischer Modellbildung – mit CD-ROM
ISBN 978-3-8325-1046-6 45.50 EUR

47 Andrea Maier-Richter: Computerunterstütztes Lernen mit Lösungsbeispielen in der Chemie. *Eine Evaluationsstudie im Themenbereich Löslichkeit*
ISBN 978-3-8325-1046-6 40.50 EUR

48 Jochen Peuckert: Stabilität und Ausprägung kognitiver Strukturen zum Atombegriff
ISBN 978-3-8325-1104-3 40.50 EUR

49 Maik Walpuski: Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback
ISBN 978-3-8325-1184-5 40.50 EUR

50 Helmut Fischler, Christiane S. Reiners (Hrsg.): Die Teilchenstruktur der Materie im Physik- und Chemieunterricht
ISBN 978-3-8325-1225-5 34.90 EUR

51 Claudia Eysel: Interdisziplinäres Lehren und Lernen in der Lehrerbildung. *Eine empirische Studie zum Kompetenzerwerb in einer komplexen Lernumgebung*
ISBN 978-3-8325-1238-5 40.50 EUR

52 Johannes Günther: Lehrerfortbildung über die Natur der Naturwissenschaften. *Studien über das Wissenschaftsverständnis von Grundschullehrkräften*
ISBN 978-3-8325-1287-3 40.50 EUR

53 Christoph Neugebauer: Lernen mit Simulationen und der Einfluss auf das Problemlösen in der Physik
ISBN 978-3-8325-1300-9 40.50 EUR

54 Andreas Schnirch: Gendergerechte Interessen- und Motivationsförderung im Kontext naturwissenschaftlicher Grundbildung. *Konzeption, Entwicklung und Evaluation einer multimedial unterstützten Lernumgebung*
ISBN 978-3-8325-1334-4 40.50 EUR

55 Hilde Köster: Freies Explorieren und Experimentieren. *Eine Untersuchung zur selbstbestimmten Gewinnung von Erfahrungen mit physikalischen Phänomenen im Sachunterricht*
ISBN 978-3-8325-1348-1 40.50 EUR

56 Eva Heran-Dörr: Entwicklung und Evaluation einer Lehrerfortbildung zur Förderung der physikdidaktischen Kompetenz von Sachunterrichtslehrkräften
ISBN 978-3-8325-1377-1 40.50 EUR

57 Agnes Szabone Varnai: Unterstützung des Problemlösens in Physik durch den Einsatz von Simulationen und die Vorgabe eines strukturierten Kooperationsformats
ISBN 978-3-8325-1403-7 40.50 EUR

58 Johannes Rethfeld: Aufgabenbasierte Lernprozesse in selbstorganisationsoffenem Unterricht der Sekundarstufe I zum Themengebiet ELEKTROSTATIK. *Eine Feldstudie in vier 10. Klassen zu einer kartenbasierten Lernumgebung mit Aufgaben aus der Elektrostatik*
ISBN 978-3-8325-1416-7 40.50 EUR

59 Christian Henke: Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe. *Untersuchung am Beispiel des HIGHSEA-Projekts in Bremerhaven*
ISBN 978-3-8325-1515-7 40.50 EUR

60 Lutz Kasper: Diskursiv-narrative Elemente für den Physikunterricht. *Entwicklung und Evaluation einer multimedialen Lernumgebung zum Erdmagnetismus*
ISBN 978-3-8325-1537-9 40.50 EUR

61 Thorid Rabe: Textgestaltung und Aufforderung zu Selbsterklärungen beim Physiklernen mit Multimedia
ISBN 978-3-8325-1539-3 40.50 EUR

62 Ina Glemnitz: Vertikale Vernetzung im Chemieunterricht. *Ein Vergleich von traditionellem Unterricht mit Unterricht nach Chemie im Kontext*
ISBN 978-3-8325-1628-4 40.50 EUR

63 Erik Einhaus: Schülerkompetenzen im Bereich Wärmelehre. *Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*
ISBN 978-3-8325-1630-7 40.50 EUR

64 Jasmin Neuroth: Concept Mapping als Lernstrategie. *Eine Interventionsstudie zum Chemielernen aus Texten*
ISBN 978-3-8325-1659-8 40.50 EUR

65 Hans Gerd Hegeler-Burkhart: Zur Kommunikation von Hauptschülerinnen und Hauptschülern in einem handlungsorientierten und fächerübergreifenden Unterricht mit physikalischen und technischen Inhalten
ISBN 978-3-8325-1667-3 40.50 EUR

66 Karsten Rincke: Sprachentwicklung und Fachlernen im Mechanikunterricht. *Sprache und Kommunikation bei der Einführung in den Kraftbegriff*
ISBN 978-3-8325-1699-4 40.50 EUR

67 Nina Strehle: Das Ion im Chemieunterricht. *Alternative Schülervorstellungen und curriculare Konsequenzen*
ISBN 978-3-8325-1710-6 40.50 EUR

68 Martin Hopf: Problemorientierte Schülerexperimente
ISBN 978-3-8325-1711-3 40.50 EUR

69 Anne Beerenwinkel: Fostering conceptual change in chemistry classes using expository texts
ISBN 978-3-8325-1721-2 40.50 EUR

70 Roland Berger: Das Gruppenpuzzle im Physikunterricht der Sekundarstufe II. *Eine empirische Untersuchung auf der Grundlage der Selbstbestimmungstheorie der Motivation*
ISBN 978-3-8325-1732-8 40.50 EUR

71 Giuseppe Colicchia: Physikunterricht im Kontext von Medizin und Biologie. *Entwicklung und Erprobung von Unterrichtseinheiten*
ISBN 978-3-8325-1746-5 40.50 EUR

72 Sandra Winheller: Geschlechtsspezifische Auswirkungen der Lehrer-Schüler-Interaktion im Chemieanfangsunterricht
ISBN 978-3-8325-1757-1 40.50 EUR

73 Isabel Wahser: Training von naturwissenschaftlichen Arbeitsweisen zur Unterstützung experimenteller Kleingruppenarbeit im Fach Chemie
ISBN 978-3-8325-1815-8 40.50 EUR

74 Claus Brell: Lernmedien und Lernerfolg - reale und virtuelle Materialien im Physikunterricht. *Empirische Untersuchungen in achten Klassen an Gymnasien (Laborstudie) zum Computereinsatz mit Simulation und IBE*
ISBN 978-3-8325-1829-5 40.50 EUR

75 Rainer Wackermann: Überprüfung der Wirksamkeit eines Basismodell-Trainings für Physiklehrer
ISBN 978-3-8325-1882-0 40.50 EUR

76 Oliver Tepner: Effektivität von Aufgaben im Chemieunterricht der Sekundarstufe I
ISBN 978-3-8325-1919-3 40.50 EUR

77 Claudia Geyer: Museums- und Science-Center-Besuche im naturwissenschaftlichen Unterricht aus einer motivationalen Perspektive. *Die Sicht von Lehrkräften und Schülerinnen und Schülern*
ISBN 978-3-8325-1922-3 40.50 EUR

78 Tobias Leonhard: Professionalisierung in der Lehrerbildung. *Eine explorative Studie zur Entwicklung professioneller Kompetenzen in der Lehrererstausbildung*
ISBN 978-3-8325-1924-7 40.50 EUR

79 Alexander Kauertz: Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben
ISBN 978-3-8325-1925-4 40.50 EUR

80 Regina Hübinger: Schüler auf Weltreise. *Entwicklung und Evaluation von Lehr-/ Lernmaterialien zur Förderung experimentell-naturwissenschaftlicher Kompetenzen für die Jahrgangsstufen 5 und 6*
ISBN 978-3-8325-1932-2 40.50 EUR

81 Christine Waltner: Physik lernen im Deutschen Museum
ISBN 978-3-8325-1933-9 40.50 EUR

82 Torsten Fischer: Handlungsmuster von Physiklehrkräften beim Einsatz neuer Medien. *Fallstudien zur Unterrichtspraxis*
ISBN 978-3-8325-1948-3 42.00 EUR

83 Corinna Kieren: Chemiehausaufgaben in der Sekundarstufe I des Gymnasiums. *Fragebogenerhebung zur gegenwärtigen Praxis und Entwicklung eines optimierten Hausaufgabendesigns im Themenbereich Säure-Base*
978-3-8325-1975-9 37.00 EUR

84 Marco Thiele: Modelle der Thermohalinen Zirkulation im Unterricht. *Eine empirische Studie zur Förderung des Modellverständnisses*
ISBN 978-3-8325-1982-7 40.50 EUR

85 Bernd Zinn: Physik lernen, um Physik zu lehren. *Eine Möglichkeit für interessanteren Physikunterricht*
ISBN 978-3-8325-1995-7 39.50 EUR

86 Esther Klaes: Außerschulische Lernorte im naturwissenschaftlichen Unterricht. *Die Perspektive der Lehrkraft*
ISBN 978-3-8325-2006-9 43.00 EUR

87 Marita Schmidt: Kompetenzmodellierung und -diagnostik im Themengebiet Energie der Sekundarstufe I. *Entwicklung und Erprobung eines Testinventars*
ISBN 978-3-8325-2024-3 37.00 EUR

88 Gudrun Franke-Braun: Aufgaben mit gestuften Lernhilfen. *Ein Aufgabenformat zur Förderung der sachbezogenen Kommunikation und Lernleistung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-2026-7 38.00 EUR

89 Silke Klos: Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht. *Der Einfluss eines integrierten Unterrichtskonzepts*
ISBN 978-3-8325-2133-2 37.00 EUR

90 Ulrike Elisabeth Burkard: Quantenphysik in der Schule. *Bestandsaufnahme, Perspektiven und Weiterentwicklungsmöglichkeiten durch die Implementation eines Medienservers*
ISBN 978-3-8325-2215-5 43.00 EUR

91 Ulrike Gromadecki: Argumente in physikalischen Kontexten. *Welche Geltungsgründe halten Physikanfänger für überzeugend?*
ISBN 978-3-8325-2250-6 41.50 EUR

92 Jürgen Bruns: Auf dem Weg zur Förderung naturwissenschaftsspezifischer Vorstellungen von zukünftigen Chemie-Lehrenden
ISBN 978-3-8325-2257-5 43.50 EUR

93 Cornelius Marsch: Räumliche Atomvorstellung. *Entwicklung und Erprobung eines Unterrichtskonzeptes mit Hilfe des Computers*
ISBN 978-3-8325-2293-3 82.50 EUR

94 Maja Brückmann: Sachstrukturen im Physikunterricht. *Ergebnisse einer Videostudie*
ISBN 978-3-8325-2272-8 39.50 EUR

95 Sabine Fechner: Effects of Context-oriented Learning on Student Interest and Achievement in Chemistry Education
ISBN 978-3-8325-2343-5 36.50 EUR

96 Clemens Nagel: eLearning im Physikalischen Anfängerpraktikum
ISBN 978-3-8325-2355-8 39.50 EUR

97 Josef Riese: Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften
ISBN 978-3-8325-2376-3 39.00 EUR

98 Sascha Bernholt: Kompetenzmodellierung in der Chemie. *Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität*
ISBN 978-3-8325-2447-0 40.00 EUR

99 Holger Christoph Stawitz: Auswirkung unterschiedlicher Aufgabenprofile auf die Schülerleistung. *Vergleich von Naturwissenschafts- und Problemlöseaufgaben der PISA 2003-Studie*
ISBN 978-3-8325-2451-7 37.50 EUR

100 Hans Ernst Fischer, Elke Sumfleth (Hrsg.): nwu-essen – 10 Jahre Essener Forschung zum naturwissenschaftlichen Unterricht
ISBN 978-3-8325-3331-1 40.00 EUR

101 Hendrik Härtig: Sachstrukturen von Physikschulbüchern als Grundlage zur Bestimmung der Inhaltsvalidität eines Tests
ISBN 978-3-8325-2512-5 34.00 EUR

102 Thomas Grüß-Niehaus: Zum Verständnis des Löslichkeitskonzeptes im Chemieunterricht. *Der Effekt von Methoden progressiver und kollaborativer Reflexion*
ISBN 978-3-8325-2537-8 40.50 EUR

103 Patrick Bronner: Quantenoptische Experimente als Grundlage eines Curriculums zur Quantenphysik des Photons
ISBN 978-3-8325-2540-8 36.00 EUR

104 Adrian Voßkühler: Blickbewegungsmessung an Versuchsaufbauten. *Studien zur Wahrnehmung, Verarbeitung und Usability von physikbezogenen Experimenten am Bildschirm und in der Realität*
ISBN 978-3-8325-2548-4 47.50 EUR

105 Verena Tobias: Newton'sche Mechanik im Anfangsunterricht. *Die Wirksamkeit einer Einführung über die zweidimensionale Dynamik auf das Lehren und Lernen*
ISBN 978-3-8325-2558-3 54.00 EUR

106 Christian Rogge: Entwicklung physikalischer Konzepte in aufgabenbasierten Lernumgebungen
ISBN 978-3-8325-2574-3 45.00 EUR

107 Mathias Ropohl: Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion. *Entwicklung und Analyse von Testaufgaben*
ISBN 978-3-8325-2609-2 36.50 EUR

108 Christoph Kulgemeyer: Physikalische Kommunikationskompetenz. *Modellierung und Diagnostik*
ISBN 978-3-8325-2674-0 44.50 EUR

109 Jennifer Olszewski: The Impact of Physics Teachers' Pedagogical Content Knowledge on Teacher Actions and Student Outcomes
ISBN 978-3-8325-2680-1 33.50 EUR

110 Annika Ohle: Primary School Teachers' Content Knowledge in Physics and its Impact on Teaching and Students' Achievement
ISBN 978-3-8325-2684-9 36.50 EUR

111 Susanne Mannel: Assessing scientific inquiry. *Development and evaluation of a test for the low-performing stage*
ISBN 978-3-8325-2761-7 40.00 EUR

112 Michael Plomer: Physik physiologisch passend praktiziert. *Eine Studie zur Lernwirksamkeit von traditionellen und adressatenspezifischen Physikpraktika für die Physiologie*
ISBN 978-3-8325-2804-1 34.50 EUR

113 Alexandra Schulz: Experimentierspezifische Qualitätsmerkmale im Chemieunterricht. *Eine Videostudie*
ISBN 978-3-8325-2817-1 40.00 EUR

114 Franz Boczianowski: Eine empirische Untersuchung zu Vektoren im Physikunterricht der Mittelstufe
ISBN 978-3-8325-2843-0 39.50 EUR

115 Maria Ploog: Internetbasiertes Lernen durch Textproduktion im Fach Physik
ISBN 978-3-8325-2853-9 39.50 EUR

116 Anja Dhein: Lernen in Explorier- und Experimentiersituationen. *Eine explorative Studie zu Bedeutungsentwicklungsprozessen bei Kindern im Alter zwischen 4 und 6 Jahren*
ISBN 978-3-8325-2859-1 45.50 EUR

117 Irene Neumann: Beyond Physics Content Knowledge. *Modeling Competence Regarding Nature of Scientific Inquiry and Nature of Scientific Knowledge*
ISBN 978-3-8325-2880-5 37.00 EUR

118 Markus Emden: Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. *Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I*
ISBN 978-3-8325-2867-6 38.00 EUR

119 Birgit Hofmann: Analyse von Blickbewegungen von Schülern beim Lesen von physikbezogenen Texten mit Bildern. *Eye Tracking als Methodenwerkzeug in der physikdidaktischen Forschung*
ISBN 978-3-8325-2925-3 59.00 EUR

120 Rebecca Knobloch: Analyse der fachinhaltlichen Qualität von Schüleräußerungen und deren Einfluss auf den Lernerfolg. *Eine Videostudie zu kooperativer Kleingruppenarbeit*
ISBN 978-3-8325-3006-8 36.50 EUR

121 Julia Hostenbach: Entwicklung und Prüfung eines Modells zur Beschreibung der Bewertungskompetenz im Chemieunterricht
ISBN 978-3-8325-3013-6 38.00 EUR

122 Anna Windt: Naturwissenschaftliches Experimentieren im Elementarbereich. *Evaluation verschiedener Lernsituationen*
ISBN 978-3-8325-3020-4 43.50 EUR

123 Eva Kölbach: Kontexteinflüsse beim Lernen mit Lösungsbeispielen
ISBN 978-3-8325-3025-9 38.50 EUR

124 Anna Lau: Passung und vertikale Vernetzung im Chemie- und Physikunterricht
ISBN 978-3-8325-3021-1 36.00 EUR

125 Jan Lamprecht: Ausbildungswege und Komponenten professioneller Handlungskompetenz. *Vergleich von Quereinsteigern mit Lehramtsabsolventen für Gymnasien im Fach Physik*
ISBN 978-3-8325-3035-8 38.50 EUR

126 Ulrike Böhm: Förderung von Verstehensprozessen unter Einsatz von Modellen
ISBN 978-3-8325-3042-6 41.00 EUR

127 Sabrina Dollny: Entwicklung und Evaluation eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Chemielehrkräften
ISBN 978-3-8325-3046-4 37.00 EUR

128 Monika Zimmermann: Naturwissenschaftliche Bildung im Kindergarten. *Eine integrative Längsschnittstudie zur Kompetenzentwicklung von Erzieherinnen*
ISBN 978-3-8325-3053-2 54.00 EUR

129 Ulf Saballus: Über das Schlussfolgern von Schülerinnen und Schülern zu öffentlichen Kontroversen mit naturwissenschaftlichem Hintergrund. *Eine Fallstudie*
ISBN 978-3-8325-3086-0 39.50 EUR

130 Olaf Krey: Zur Rolle der Mathematik in der Physik. *Wissenschaftstheoretische Aspekte und Vorstellungen Physiklernender*
ISBN 978-3-8325-3101-0 46.00 EUR

131 Angelika Wolf: Zusammenhänge zwischen der Eigenständigkeit im Physikunterricht, der Motivation, den Grundbedürfnissen und dem Lernerfolg von Schülern
ISBN 978-3-8325-3161-4 45.00 EUR

132 Johannes Börlin: Das Experiment als Lerngelegenheit. *Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität*
ISBN 978-3-8325-3170-6 45.00 EUR

133 Olaf Uhden: Mathematisches Denken im Physikunterricht. *Theorieentwicklung und Problemanalyse*
ISBN 978-3-8325-3170-6 45.00 EUR

134 Christoph Gut: Modellierung und Messung experimenteller Kompetenz. *Analyse eines large-scale Experimentiertests*
ISBN 978-3-8325-3213-0 40.00 EUR

135 Antonio Rueda: Lernen mit ExploMultimedial in kolumbianischen Schulen. *Analyse von kurzzeitigen Lernprozessen und der Motivation beim länderübergreifenden Einsatz einer deutschen computergestützten multimedialen Lernumgebung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-3218-5 45.50 EUR

136 Krisztina Berger: Bilder, Animationen und Notizen. *Empirische Untersuchung zur Wirkung einfacher visueller Repräsentationen und Notizen auf den Wissenserwerb in der Optik*
ISBN 978-3-8325-3238-3 41.50 EUR

137 Antony Crossley: Untersuchung des Einflusses unterschiedlicher physikalischer Konzepte auf den Wissenserwerb in der Thermodynamik der Sekundarstufe I
ISBN 978-3-8325-3275-8 40.00 EUR

138 Tobias Viering: Entwicklung physikalischer Kompetenz in der Sekundarstufe I. *Validierung eines Kompetenzentwicklungsmodells für das Energiekonzept im Bereich Fachwissen*
ISBN 978-3-8325-3277-2 37.00 EUR

139 Nico Schreiber: Diagnostik experimenteller Kompetenz. *Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*
ISBN 978-3-8325-3284-0 39.00 EUR

140 Sarah Hundertmark: Einblicke in kollaborative Lernprozesse. *Eine Fallstudie zur reflektierenden Zusammenarbeit unterstützt durch die Methoden Concept Mapping und Lernbegleitbogen*
ISBN 978-3-8325-3251-2 43.00 EUR

141 Ronny Scherer: Analyse der Struktur, Messinvarianz und Ausprägung komplexer Problemlösekompetenz im Fach Chemie. *Eine Querschnittstudie in der Sekundarstufe I und am Übergang zur Sekundarstufe II*
ISBN 978-3-8325-3312-0 43.00 EUR

142 Patricia Heitmann: Bewertungskompetenz im Rahmen naturwissenschaftlicher Problemlöseprozesse. *Modellierung und Diagnose der Kompetenzen Bewertung und analytisches Problemlösen für das Fach Chemie*
ISBN 978-3-8325-3314-4 37.00 EUR

143 Jan Fleischhauer: Wissenschaftliches Argumentieren und Entwicklung von Konzepten beim Lernen von Physik
ISBN 978-3-8325-3325-0 35.00 EUR

144 Nermin Özcan: Zum Einfluss der Fachsprache auf die Leistung im Fach Chemie. *Eine Förderstudie zur Fachsprache im Chemieunterricht*
ISBN 978-3-8325-3328-1 36.50 EUR

145 Helena van Vorst: Kontextmerkmale und ihr Einfluss auf das Schülerinteresse im Fach Chemie
ISBN 978-3-8325-3321-2 38.50 EUR

146 Janine Cappell: Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase
ISBN 978-3-8325-3356-4 38.50 EUR

147 Susanne Bley: Förderung von Transferprozessen im Chemieunterricht
ISBN 978-3-8325-3407-3 40.50 EUR

148 Cathrin Blaes: Die übungsgestützte Lehrerpräsentation im Chemieunterricht der Sekundarstufe I. *Evaluation der Effektivität*
ISBN 978-3-8325-3409-7 43.50 EUR

149 Julia Suckut: Die Wirksamkeit von piko-OWL als Lehrerfortbildung. Eine Evaluation zum Projekt *Physik im Kontext* in Fallstudien
ISBN 978-3-8325-3440-0 45.00 EUR

150 Alexandra Dorschu: Die Wirkung von Kontexten in Physikkompetenztestaufgaben
ISBN 978-3-8325-3446-2 37.00 EUR

151 Jochen Scheid: Multiple Repräsentationen, Verständnis physikalischer Experimente und kognitive Aktivierung: *Ein Beitrag zur Entwicklung der Aufgabenkultur*
ISBN 978-3-8325-3449-3 49.00 EUR

152 Tim Plasa: Die Wahrnehmung von Schülerlaboren und Schülerforschungszentren
ISBN 978-3-8325-3483-7 35.50 EUR

153 Felix Schoppmeier: Physikkompetenz in der gymnasialen Oberstufe.*Entwicklung und Validierung eines Kompetenzstrukturmodells für den Kompetenzbereich Umgang mit Fachwissen*
ISBN 978-3-8325-3502-5 36.00 EUR

154 Katharina Groß: Experimente alternativ dokumentieren. *Eine qualitative Studie zur Förderung der Diagnose- und Differenzierungskompetenz in der Chemielehrerbildung*
ISBN 978-3-8325-3508-7 43.50 EUR

155 Barbara Hank: Konzeptwandelprozesse im Anfangsunterricht Chemie. *Eine quasiexperimentelle Längsschnittstudie*
ISBN 978-3-8325-3519-3 38.50 EUR

156 Katja Freyer: Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie
ISBN 978-3-8325-3544-5 38.00 EUR

157 Alexander Rachel: Auswirkungen instruktionaler Hilfen bei der Einführung des (Ferro-)Magnetismus. *Eine Vergleichsstudie in der Primar- und Sekundarstufe*
ISBN 978-3-8325-3548-3 43.50 EUR

158 Sebastian Ritter: Einfluss des Lerninhalts Nanogrößeneffekte auf Teilchen- und Teilchenmodellvorstellungen von Schülerinnen und Schülern
ISBN 978-3-8325-3558-2 36.00 EUR

159 Andrea Harbach: Problemorientierung und Vernetzung in kontextbasierten Lernaufgaben
ISBN 978-3-8325-3564-3 39.00 EUR

160 David Obst: Interaktive Tafeln im Physikunterricht. *Entwicklung und Evaluation einer Lehrerfortbildung*
ISBN 978-3-8325-3582-7 40.50 EUR

161 Sophie Kirschner: Modellierung und Analyse des Professionswissens von Physiklehrkräften
ISBN 978-3-8325-3601-5 35.00 EUR

162 Katja Stief: Selbstregulationsprozesse und Hausaufgabenmotivation im Chemieunterricht
ISBN 978-3-8325-3631-2 34.00 EUR

163 Nicola Meschede: Professionelle Wahrnehmung der inhaltlichen Strukturierung im naturwissenschaftlichen Grundschulunterricht. *Theoretische Beschreibung und empirische Erfassung*
ISBN 978-3-8325-3668-8 37.00 EUR

164 Johannes Maximilian Barth: Experimentieren im Physikunterricht der gymnasialen Oberstufe. *Eine Rekonstruktion übergeordneter Einbettungsstrategien*
ISBN 978-3-8325-3681-7 39.00 EUR

165 Sandra Lein: Das Betriebspraktikum in der Lehrerbildung. *Eine Untersuchung zur Förderung der Wissenschafts- und Technikbildung im allgemeinbildenden Unterricht*
ISBN 978-3-8325-3698-5 40.00 EUR

166 Veranika Maiseyenka: Modellbasiertes Experimentieren im Unterricht. *Praxistauglichkeit und Lernwirkungen*
ISBN 978-3-8325-3708-1 38.00 EUR

167 Christoph Stolzenberger: Der Einfluss der didaktischen Lernumgebung auf das Erreichen geforderter Bildungsziele am Beispiel der W- und P-Seminare im Fach Physik
ISBN 978-3-8325-3708-1 38.00 EUR

168 Pia Altenburger: Mehrebenenregressionsanalysen zum Physiklernen im Sachunterricht der Primarstufe. *Ergebnisse einer Evaluationsstudie.*
ISBN 978-3-8325-3717-3 37.50 EUR

169 Nora Ferber: Entwicklung und Validierung eines Testinstruments zur Erfassung von Kompetenzentwicklung im Fach Chemie in der Sekundarstufe I
ISBN 978-3-8325-3727-2 39.50 EUR

170 Anita Stender: Unterrichtsplanung: Vom Wissen zum Handeln.
Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung
ISBN 978-3-8325-3750-0 41.50 EUR

171 Jenna Koenen: Entwicklung und Evaluation von experimentunterstützten Lösungsbeispielen zur Förderung naturwissenschaftlich-experimenteller Arbeitsweisen
ISBN 978-3-8325-3785-2 43.00 EUR

172 Teresa Henning: Empirische Untersuchung kontextorientierter Lernumgebungen in der Hochschuldidaktik. *Entwicklung und Evaluation kontextorientierter Aufgaben in der Studieneingangsphase für Fach- und Nebenfachstudierende der Physik*
ISBN 978-3-8325-3801-9 43.00 EUR

173 Alexander Pusch: Fachspezifische Instrumente zur Diagnose und individuellen Förderung von Lehramtsstudierenden der Physik
ISBN 978-3-8325-3829-3 38.00 EUR

174 Christoph Vogelsang: Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. *Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*
ISBN 978-3-8325-3846-0 50.50 EUR

175 Ingo Brebeck: Selbstreguliertes Lernen in der Studieneingangsphase im Fach Chemie
ISBN 978-3-8325-3859-0 37.00 EUR

176 Axel Eghtessad: Merkmale und Strukturen von Professionalisierungsprozessen in der ersten und zweiten Phase der Chemielehrerbildung. *Eine empirisch-qualitative Studie mit niedersächsischen Fachleiter_innen der Sekundarstufenlehrämter*
ISBN 978-3-8325-3861-3 45.00 EUR

177 Andreas Nehring: Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie. Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung
ISBN 978-3-8325-3872-9 39.50 EUR

178 Maike Schmidt: Professionswissen von Sachunterrichtslehrkräften. Zusammenhangsanalyse zur Wirkung von Ausbildungshintergrund und Unterrichtserfahrung auf das fachspezifische Professionswissen im Unterrichtsinhalt „Verbrennung“
ISBN 978-3-8325-3907-8 38.50 EUR

179 Jan Winkelmann: Auswirkungen auf den Fachwissenszuwachs und auf affektive Schülermerkmale durch Schüler- und Demonstrationsexperimente im Physikunterricht
ISBN 978-3-8325-3915-3 41.00 EUR

180 Iwen Kobow: Entwicklung und Validierung eines Testinstrumentes zur Erfassung der Kommunikationskompetenz im Fach Chemie
ISBN 978-3-8325-3927-6 34.50 EUR

181 Yvonne Gramzow: Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion
ISBN 978-3-8325-3931-3 42.50 EUR

182 Evelin Schröter: Entwicklung der Kompetenzerwartung durch Lösen physikalischer Aufgaben einer multimedialen Lernumgebung
ISBN 978-3-8325-3975-7 54.50 EUR

183 Inga Kallweit: Effektivität des Einsatzes von Selbsteinschätzungsbögen im Chemieunterricht der Sekundarstufe I. *Individuelle Förderung durch selbstreguliertes Lernen*
ISBN 978-3-8325-3965-8 44.00 EUR

184 Andrea Schumacher: Paving the way towards authentic chemistry teaching. *A contribution to teachers' professional development*
ISBN 978-3-8325-3976-4 48.50 EUR

185 David Woitkowski: Fachliches Wissen Physik in der Hochschulausbildung. *Konzeptualisierung, Messung, Niveaubildung*
ISBN 978-3-8325-3988-7 53.00 EUR

186 Marianne Korner: Cross-Age Peer Tutoring in Physik. *Evaluation einer Unterrichtsmethode*
ISBN 978-3-8325-3979-5 38.50 EUR

187 Simone Nakoinz: Untersuchung zur Verknüpfung submikroskopischer und makroskopischer Konzepte im Fach Chemie
ISBN 978-3-8325-4057-9 38.50 EUR

188 Sandra Anus: Evaluation individueller Förderung im Chemieunterricht.*Adaptivität von Lerninhalten an das Vorwissen von Lernenden am Beispiel des Basiskonzeptes Chemische Reaktion*
ISBN 978-3-8325-4059-3 43.50 EUR

189 Thomas Roßbegalle: Fachdidaktische Entwicklungsforschung zum besseren Verständnis atmosphärischer Phänomene. *Treibhauseffekt, saurer Regen und stratosphärischer Ozonabbau als Kontexte zur Vermittlung von Basiskonzepten der Chemie*
ISBN 978-3-8325-4059-3 45.50 EUR

190 Kathrin Steckenmesser-Sander: Gemeinsamkeiten und Unterschiede physikbezogener Handlungs-, Denk- und Lernprozesse von Mädchen und Jungen
ISBN 978-3-8325-4066-1 38.50 EUR

191 Cornelia Geller: Lernprozessorientierte Sequenzierung des Physikunterrichts im Zusammenhang mit Fachwissenserwerb. *Eine Videostudie in Finnland, Deutschland und der Schweiz*
ISBN 978-3-8325-4082-1 35.50 EUR

192 Jan Hofmann: Untersuchung des Kompetenzaufbaus von Physiklehrkräften während einer Fortbildungsmaßnahme
ISBN 978-3-8325-4104-0 38.50 EUR

193 Andreas Dickhäuser: Chemiespezifischer Humor. *Theoriebildung, Materialentwicklung, Evaluation*
ISBN 978-3-8325-4108-8 37.00 EUR

194 Stefan Korte: Die Grenzen der Naturwissenschaft als Thema des Physikunterrichts
ISBN 978-3-8325-4112-5 57.50 EUR

195 Carolin Hülsmann: Kurswahlmotive im Fach Chemie. Eine Studie zum Wahlverhalten und Erfolg von Schülerinnen und Schülern in der gymnasialen Oberstufe
ISBN 978-3-8325-4144-6 49.00 EUR

196 Caroline Körbs: Mindeststandards im Fach Chemie am Ende der Pflichtschulzeit
ISBN 978-3-8325-4148-4 34.00 EUR

197 Andreas Vorholzer: Wie lassen sich Kompetenzen des experimentellen Denkens und Arbeitens fördern? *Eine empirische Untersuchung der Wirkung eines expliziten und eines impliziten Instruktionsansatzes*
ISBN 978-3-8325-4194-1 37.50 EUR

198 Anna Katharina Schmitt: Entwicklung und Evaluation einer Chemielehrerfortbildung zum Kompetenzbereich Erkenntnisgewinnung
ISBN 978-3-8325-4228-3 39.50 EUR

199 Christian Maurer: Strukturierung von Lehr-Lern-Sequenzen
ISBN 978-3-8325-4247-4 36.50 EUR

200 Helmut Fischler, Elke Sumfleth (Hrsg.): Professionelle Kompetenz von Lehrkräften der Chemie und Physik
ISBN 978-3-8325-4523-9 34.00 EUR

201 Simon Zander: Lehrerfortbildung zu Basismodellen und Zusammenhänge zum Fachwissen
ISBN 978-3-8325-4248-1 35.00 EUR

202 Kerstin Arndt: Experimentierkompetenz erfassen.
Analyse von Prozessen und Mustern am Beispiel von Lehramtsstudierenden der Chemie
ISBN 978-3-8325-4266-5 45.00 EUR

203 Christian Lang: Kompetenzorientierung im Rahmen experimentalchemischer Praktika
ISBN 978-3-8325-4268-9 42.50 EUR

204 Eva Cauet: Testen wir relevantes Wissen? *Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*
ISBN 978-3-8325-4276-4 39.50 EUR

205 Patrick Löffler: Modellanwendung in Problemlöseaufgaben. *Wie wirkt Kontext?*
ISBN 978-3-8325-4303-7 35.00 EUR

206 Carina Gehlen: Kompetenzstruktur naturwissenschaftlicher Erkenntnisgewinnung im Fach Chemie
ISBN 978-3-8325-4318-1 43.00 EUR

207 Lars Oettinghaus: Lehrerüberzeugungen und physikbezogenes Professionswissen. *Vergleich von Absolventinnen und Absolventen verschiedener Ausbildungswege im Physikreferendariat*
ISBN 978-3-8325-4319-8 38.50 EUR

208 Jennifer Petersen: Zum Einfluss des Merkmals Humor auf die Gesundheitsförderung im Chemieunterricht der Sekundarstufe I.
Eine Interventionsstudie zum Thema Sonnenschutz
ISBN 978-3-8325-4348-8 40.00 EUR

209 Philipp Straube: Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik
ISBN 978-3-8325-4351-8 35.50 EUR

210 Martin Dickmann: Messung von Experimentierfähigkeiten. *Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*
ISBN 978-3-8325-4356-3 41.00 EUR

211 Markus Bohlmann: Science Education. Empirie, Kulturen und Mechanismen der Didaktik der Naturwissenschaften
ISBN 978-3-8325-4377-8 44.00 EUR

212 Martin Draude: Die Kompetenz von Physiklehrkräften, Schwierigkeiten von Schülerinnen und Schülern beim eigenständigen Experimentieren zu diagnostizieren
ISBN 978-3-8325-4382-2 37.50 EUR

213 Henning Rode: Prototypen evidenzbasierten Physikunterrichts. *Zwei empirische Studien zum Einsatz von Feedback und Blackboxes in der Sekundarstufe*
ISBN 978-3-8325-4389-1 42.00 EUR

214 Jan-Henrik Kechel: Schülerschwierigkeiten beim eigenständigen Experimentieren. *Eine qualitative Studie am Beispiel einer Experimentieraufgabe zum Hooke'schen Gesetz*
ISBN 978-3-8325-4392-1 55.00 EUR

215 Katharina Fricke: Classroom Management and its Impact on Lesson Outcomes in Physics. *A multi-perspective comparison of teaching practices in primary and secondary schools*
ISBN 978-3-8325-4394-5 40.00 EUR

216 Hannes Sander: Orientierungen von Jugendlichen beim Urteilen und Entscheiden in Kontexten nachhaltiger Entwicklung. *Eine rekonstruktive Perspektive auf Bewertungskompetenz in der Didaktik der Naturwissenschaft*
ISBN 978-3-8325-4434-8 46.00 EUR

217 Inka Haak: Maßnahmen zur Unterstützung kognitiver und metakognitiver Prozesse in der Studieneingangsphase. *Eine Design-Based-Research-Studie zum universitären Lernzentrum Physiktreff*
ISBN 978-3-8325-4437-9 46.50 EUR

218 Martina Brandenburger: Was beeinflusst den Erfolg beim Problemlösen in der Physik? *Eine Untersuchung mit Studierenden*
ISBN 978-3-8325-4409-6 42.50 EUR

219 Corinna Helms: Entwicklung und Evaluation eines Trainings zur Verbesserung der Erklärqualität von Schülerinnen und Schülern im Gruppenpuzzle
ISBN 978-3-8325-4454-6 42.50 EUR

220 Viktoria Rath: Diagnostische Kompetenz von angehenden Physiklehrkräften. *Modellierung, Testinstrumentenentwicklung und Erhebung der Performanz bei der Diagnose von Schülervorstellungen in der Mechanik*
ISBN 978-3-8325-4456-0 42.50 EUR

221 Janne Krüger: Schülerperspektiven auf die zeitliche Entwicklung der Naturwissenschaften
ISBN 978-3-8325-4457-7 45.50 EUR

222 Stefan Mutke: Das Professionswissen von Chemiereferendarinnen und -referendaren in Nordrhein-Westfalen. *Eine Längsschnittstudie*
ISBN 978-3-8325-4458-4 37.50 EUR

223 Sebastian Habig: Systematisch variierte Kontextaufgaben und ihr Einfluss auf kognitive und affektive Schülerfaktoren
ISBN 978-3-8325-4467-6 40.50 EUR

224 Sven Liepertz: Zusammenhang zwischen dem Professionswissen von Physiklehrkräften, dem sachstrukturellen Angebot des Unterrichts und der Schülerleistung
ISBN 978-3-8325-4480-5 34.00 EUR

225 Elina Platova: Optimierung eines Laborpraktikums durch kognitive Aktivierung
ISBN 978-3-8325-4481-2 39.00 EUR

226 Tim Reschke: Lesegeschichten im Chemieunterricht der Sekundarstufe I zur Unterstützung von situationalem Interesse und Lernerfolg
ISBN 978-3-8325-4487-4 41.00 EUR

227 Lena Mareike Walper: Entwicklung der physikbezogenen Interessen und selbstbezogenen Kognitionen von Schülerinnen und Schülern in der Übergangsphase von der Primar- in die Sekundarstufe. *Eine Längsschnittanalyse vom vierten bis zum siebten Schuljahr*
ISBN 978-3-8325-4495-9 43.00 EUR

228 Stefan Anthofer: Förderung des fachspezifischen Professionswissens von Chemielehramtsstudierenden
ISBN 978-3-8325-4498-0 39.50 EUR

229 Marcel Bullinger: Handlungsorientiertes Physiklernen mit instruierten Selbsterklärungen in der Primarstufe. *Eine experimentelle Laborstudie*
ISBN 978-3-8325-4504-8 44.00 EUR

230 Thomas Amenda: Bedeutung fachlicher Elementarisierungen für das Verständnis der Kinematik
ISBN 978-3-8325-4531-4 43.50 EUR

231 Sabrina Milke: Beeinflusst *Priming* das Physiklernen?
Eine empirische Studie zum Dritten Newtonschen Axiom
ISBN 978-3-8325-4549-4 42.00 EUR

232 Corinna Erfmann: Ein anschaulicher Weg zum Verständnis der elektromagnetischen Induktion. *Evaluation eines Unterrichtsvorschlags und Validierung eines Leistungsdiagnoseinstruments*
ISBN 978-3-8325-4550-5 49.50 EUR

233 Hanne Rautenstrauch: Erhebung des (Fach-)Sprachstandes bei Lehramtsstudierenden im Kontext des Faches Chemie
ISBN 978-3-8325-4556-7 40.50 EUR

234 Tobias Klug: Wirkung kontextorientierter physikalischer Praktikumsversuche auf Lernprozesse von Studierenden der Medizin
ISBN 978-3-8325-4558-1 37.00 EUR

235 Mareike Bohrmann: Zur Förderung des Verständnisses der Variablenkontrolle im naturwissenschaftlichen Sachunterricht
ISBN 978-3-8325-4559-8 52.00 EUR

236 Anja Schödl: FALKO-Physik – Fachspezifische Lehrerkompetenzen im Fach Physik. *Entwicklung und Validierung eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Physiklehrkräften*
ISBN 978-3-8325-4553-6 40.50 EUR

237 Hilda Scheuermann: Entwicklung und Evaluation von Unterstützungsmaßnahmen zur Förderung der Variablenkontrollstrategie beim Planen von Experimenten
ISBN 978-3-8325-4568-0 39.00 EUR

238 Christian G. Strippel: Naturwissenschaftliche Erkenntnisgewinnung an chemischen Inhalten vermitteln. *Konzeption und empirische Untersuchung einer Ausstellung mit Experimentierstation*
ISBN 978-3-8325-4577-2 41.50 EUR

239 Sarah Rau: Durchführung von Sachunterricht im Vorbereitungsdienst. *Eine längsschnittliche, videobasierte Unterrichtsanalyse*
ISBN 978-3-8325-4579-6 46.00 EUR

240 Thomas Plotz: Lernprozesse zu nicht-sichtbarer Strahlung. *Empirische Untersuchungen in der Sekundarstufe 2*
ISBN 978-3-8325-4624-3 39.50 EUR

241 Wolfgang Aschauer: Elektrische und magnetische Felder. *Eine empirische Studie zu Lernprozessen in der Sekundarstufe II*
ISBN 978-3-8325-4625-0 50.00 EUR

242 Anna Donhauser: Didaktisch rekonstruierte Materialwissenschaft. *Aufbau und Konzeption eines Schülerlabors für den Exzellenzcluster Engineering of Advanced Materials*
ISBN 978-3-8325-4636-6 39.00 EUR

243 Katrin Schüßler: Lernen mit Lösungsbeispielen im Chemieunterricht. *Einflüsse auf Lernerfolg, kognitive Belastung und Motivation*
ISBN 978-3-8325-4640-3 42.50 EUR

244 Timo Fleischer: Untersuchung der chemischen Fachsprache unter besonderer Berücksichtigung chemischer Repräsentationen
ISBN 978-3-8325-4642-7 46.50 EUR

245 Rosina Steininger: Concept Cartoons als Stimuli für Kleingruppendiskussionen im Chemieunterricht. *Beschreibung und Analyse einer komplexen Lerngelegenheit*
ISBN 978-3-8325-4647-2 39.00 EUR

246 Daniel Rehfeldt: Erfassung der Lehrqualität naturwissenschaftlicher Experimentalpraktika
ISBN 978-3-8325-4590-1 40.00 EUR

247 Sandra Puddu: Implementing Inquiry-based Learning in a Diverse Classroom: Investigating Strategies of Scaffolding and Students' Views of Scientific Inquiry
ISBN 978-3-8325-4591-8 35.50 EUR

248 Markus Bliersbach: Kreativität in der Chemie. *Erhebung und Förderung der Vorstellungen von Chemielehramtsstudierenden*
ISBN 978-3-8325-4593-2 44.00 EUR

249 Lennart Kimpel: Aufgaben in der Allgemeinen Chemie. *Zum Zusammenspiel von chemischem Verständnis und Rechenfähigkeit*
ISBN 978-3-8325-4618-2 36.00 EUR

250 Louise Bindel: Effects of integrated learning: explicating a mathematical concept in inquiry-based science camps
ISBN 978-3-8325-4655-7 37.50 EUR

251 Michael Wenzel: Computereinsatz in Schule und Schülerlabor. *Einstellung von Physiklehrkräften zu Neuen Medien*
ISBN 978-3-8325-4659-5 38.50 EUR

252 Laura Muth: Einfluss der Auswertephase von Experimenten im Physikunterricht. *Ergebnisse einer Interventionsstudie zum Zuwachs von Fachwissen und experimenteller Kompetenz von Schülerinnen und Schülern*
ISBN 978-3-8325-4675-5 36.50 EUR

253 Annika Fricke: Interaktive Skripte im Physikalischen Praktikum. *Entwicklung und Evaluation von Hypermedien für die Nebenfachausbildung*
ISBN 978-3-8325-4676-2 41.00 EUR

254 Julia Haase: Selbstbestimmtes Lernen im naturwissenschaftlichen Sachunterricht. *Eine empirische Interventionsstudie mit Fokus auf Feedback und Kompetenzerleben*
ISBN 978-3-8325-4685-4 38.50 EUR

255 Antje J. Heine: Was ist Theoretische Physik? *Eine wissenschaftstheoretische Betrachtung und Rekonstruktion von Vorstellungen von Studierenden und Dozenten über das Wesen der Theoretischen Physik*
ISBN 978-3-8325-4691-5 46.50 EUR

256 Claudia Meinhardt: Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern
ISBN 978-3-8325-4712-7 47.00 EUR

257 Ann-Kathrin Schlüter: Professionalisierung angehender Chemielehrkräfte für einen Gemeinsamen Unterricht
ISBN 978-3-8325-4713-4 53.50 EUR

258 Stefan Richtberg: Elektronenbahnen in Feldern. Konzeption und Evaluation einer webbasierten Lernumgebung
ISBN 978-3-8325-4723-3 49.00 EUR

259 Jan-Philipp Burde: Konzeption und Evaluation eines Unterrichtskonzepts zu einfachen Stromkreisen auf Basis des Elektronengasmodells
ISBN 978-3-8325-4726-4 57.50 EUR

260 Frank Finkenberg: Flipped Classroom im Physikunterricht
ISBN 978-3-8325-4737-4 42.50 EUR

261 Florian Treisch: Die Entwicklung der Professionellen Unterrichtswahrnehmung im Lehr-Lern-Labor Seminar
ISBN 978-3-8325-4741-4 41.50 EUR

262 Desiree Mayr: Strukturiertheit des experimentellen naturwissenschaftlichen Problemlöseprozesses
ISBN 978-3-8325-4757-8 37.00 EUR

263 Katrin Weber: Entwicklung und Validierung einer Learning Progression für das Konzept der chemischen Reaktion in der Sekundarstufe I
ISBN 978-3-8325-4762-2 48.50 EUR

264 Hauke Bartels: Entwicklung und Bewertung eines performanznahen Videovignettentests zur Messung der Erklärfähigkeit von Physiklehrkräften
ISBN 978-3-8325-4804-9 37.00 EUR

265 Karl Marniok: Zum Wesen von Theorien und Gesetzen in der Chemie. *Begriffsanalyse und Förderung der Vorstellungen von Lehramtsstudierenden*
ISBN 978-3-8325-4805-6 42.00 EUR

266 Marisa Holzapfel: Fachspezifischer Humor als Methode in der Gesundheitsbildung im Übergang von der Primarstufe zur Sekundarstufe I
ISBN 978-3-8325-4808-7 50.00 EUR

267 Anna Stolz: Die Auswirkungen von Experimentiersituationen mit unterschiedlichem Öffnungsgrad auf Leistung und Motivation der Schülerinnen und Schüler
ISBN 978-3-8325-4781-3 38.00 EUR

268 Nina Ulrich: Interaktive Lernaufgaben in dem digitalen Schulbuch eChemBook. *Einfluss des Interaktivitätsgrads der Lernaufgaben und des Vorwissens der Lernenden auf den Lernerfolg*
ISBN 978-3-8325-4814-8 43.50 EUR

269 Kim-Alessandro Weber: Quantenoptik in der Lehrerfortbildung. *Ein bedarfsgeprägtes Fortbildungskonzept zum Quantenobjekt Photon mit Realexperimenten*
ISBN 978-3-8325-4792-9 55.00 EUR

270 Nina Skorsetz: Empathisierer und Systematisierer im Vorschulalter. *Eine Fragebogen- und Videostudie zur Motivation, sich mit Naturphänomenen zu beschäftigen*
ISBN 978-3-8325-4825-4 43.50 EUR

271 Franziska Kehne: Analyse des Transfers von kontextualisiert erworbenem Wissen im Fach Chemie
ISBN 978-3-8325-4846-9 45.00 EUR

272 Markus Elsholz: Das akademische Selbstkonzept angehender Physiklehrkräfte als Teil ihrer professionellen Identität. *Dimensionalität und Veränderung während einer zentralen Praxisphase*
ISBN 978-3-8325-4857-5 37.50 EUR

273 Joachim Müller: Studienerfolg in der Physik. *Zusammenhang zwischen Modellierungskompetenz und Studienerfolg*
ISBN 978-3-8325-4859-9 35.00 EUR

274 Jennifer Dörschelln: Organische Leuchtdioden. *Implementation eines innovativen Themas in den Chemieunterricht*
ISBN 978-3-8325-4865-0 59.00 EUR

275 Stephanie Strelow: Beliefs von Studienanfängern des Kombi-Bachelors Physik über die Natur der Naturwissenschaften
ISBN 978-3-8325-4881-0 40.50 EUR

276 Dennis Jaeger: Kognitive Belastung und aufgabenspezifische sowie personenspezifische Einflussfaktoren beim Lösen von Physikaufgaben
ISBN 978-3-8325-4928-2 50.50 EUR

277 Vanessa Fischer: Der Einfluss von Interesse und Motivation auf die Messung von Fach- und Bewertungskompetenz im Fach Chemie
ISBN 978-3-8325-4933-6 39.00 EUR

278 René Dohrmann: Professionsbezogene Wirkungen einer Lehr-Lern-Labor-Veranstaltung. *Eine multimethodische Studie zu den professionsbezogenen Wirkungen einer Lehr-Lern-Labor-Blockveranstaltung auf Studierende der Bachelorstudiengänge Lehramt Physik und Grundschulpädagogik (Sachunterricht)*
ISBN 978-3-8325-4958-9 40.00 EUR

279 Meike Bergs: Can We Make Them Use These Strategies? *Fostering Inquiry-Based Science Learning Skills with Physical and Virtual Experimentation Environments*
ISBN 978-3-8325-4962-6 39.50 EUR

280 Marie-Therese Hauerstein: Untersuchung zur Effektivität von Strukturierung und Binnendifferenzierung im Chemieunterricht der Sekundarstufe I. *Evaluation der Strukturierungshilfe Lernleiter*
ISBN 978-3-8325-4982-4 42.50 EUR

281 Verena Zucker: Erkennen und Beschreiben von formativem Assessment im naturwissenschaftlichen Grundschulunterricht. *Entwicklung eines Instruments zur Erfassung von Teilfähigkeiten der professionellen Wahrnehmung von Lehramtsstudierenden*
ISBN 978-3-8325-4991-6 38.00 EUR

282 Victoria Telser: Erfassung und Förderung experimenteller Kompetenz von Lehrkräften im Fach Chemie
ISBN 978-3-8325-4996-1 50.50 EUR

283 Kristine Tschirschky: Entwicklung und Evaluation eines gedächtnisorientierten Aufgabendesigns für Physikaufgaben
ISBN 978-3-8325-5002-8 42.50 EUR

284 Thomas Elert: Course Success in the Undergraduate General Chemistry Lab
ISBN 978-3-8325-5004-2 41.50 EUR

285 Britta Kalthoff: Explizit oder implizit? *Untersuchung der Lernwirksamkeit verschiedener fachmethodischer Instruktionen im Hinblick auf fachmethodische und fachinhaltliche Fähigkeiten von Sachunterrichtsstudierenden*
ISBN 978-3-8325-5013-4 37.50 EUR

286 Thomas Dickmann: Visuelles Modellverständnis und Studienerfolg in der Chemie. *Zwei Seiten einer Medaille*
ISBN 978-3-8325-5016-5 44.00 EUR

287 Markus Sebastian Feser: Physiklehrkräfte korrigieren Schülertexte. *Eine Explorationsstudie zur fachlich-konzeptuellen und sprachlichen Leistungsfeststellung und -beurteilung im Physikunterricht*
ISBN 978-3-8325-5020-2 49.00 EUR

Studien zum Physik- und Chemielernen

Herausgegeben von Hans Niedderer, Helmut Fischler und Elke Sumfleth

Die Reihe umfasst inzwischen eine große Zahl von wissenschaftlichen Arbeiten aus vielen Arbeitsgruppen der Physik- und Chemiedidaktik und zeichnet damit ein gültiges Bild der empirischen physik- und chemiedidaktischen Forschung in Deutschland.
Die Herausgeber laden daher Interessenten zu neuen Beiträgen ein und bitten sie, sich im Bedarfsfall an den Logos-Verlag oder an ein Mitglied des Herausgeberteams zu wenden.

Kontaktadressen:

Prof. Dr. Hans Niedderer
Institut für Didaktik der Naturwissenschaften,
Abt. Physikdidaktik, FB Physik/Elektrotechnik,
Universität Bremen,
Postfach 33 04 40, 28334 Bremen
Tel. 0421-218 2484/4695, e-mail:
niedderer@physik.uni-bremen.de

Prof. Dr. Helmut Fischler
Didaktik der Physik, FB Physik, Freie Universität Berlin,
Arnimallee 14, 14195 Berlin
Tel. 030-838 56712/55966, e-mail:
fischler@physik.fu-berlin.de

Prof. Dr. Elke Sumfleth
Didaktik der Chemie,
Fachbereich Chemie,
Universität Duisburg-Essen,
Schützenbahn 70, 45127 Essen
Tel. 0201-183 3757/3761, e-mail:
elke.sumfleth@uni-essen.de

Im Physikunterricht spielen sprachbezogene Anforderungen oft eine Rolle. Zugleich setzen viele Physiklehrkräfte es als selbstverständlich voraus, dass sich Lernende sprachlich versiert ausdrücken können. Eine naheliegende Vermutung ist, dass sich diese Erwartungshaltung auch auf die Korrektur von Klassenarbeiten niederschlägt. Darüber, wie Physiklehrkräfte bei der Feststellung und Beurteilung von Schülerleistungen in einer Klassenarbeit tatsächlich vorgehen, liegt allerdings bislang kaum belastbare empirische Evidenz vor (Forschungsfrage 1). Es stellt sich auch die Frage, inwieweit Physiklehrkräfte bei der Korrektur einer Klassenarbeit fachlich-konzeptuelle und sprachliche Schülerleistungen miteinander konfundieren (Forschungsfrage 2). Die vorliegende Arbeit exploriert diese beiden Fragen im Rahmen einer Laut-Denk-Studie mit 21 im Schuldienst aktiven Physiklehrkräften.

Bezüglich Forschungsfrage 1 zeigte sich – neben weiteren Befunden –, dass die befragten Lehrkräfte sprachliche Schülerleistungen in einer tendenziell defizitorientierten Art und Weise feststellen und beurteilen. Die Feststellung und Beurteilung fachlich-konzeptueller Schülerleistungen erfolgt hingegen zum Teil defizitorientiert, aber auch fähigkeitsorientiert. Hinsichtlich Forschungsfrage 2 konnten zahlreiche komplementäre Teilbefunde gewonnen werden. Sie sprechen in ihrer Gesamtheit dafür, dass die befragten Lehrkräfte fachlich-konzeptuelle und sprachliche Schülerleistungen auf einem moderaten Niveau miteinander konfundieren.

Logos Verlag Berlin

ISBN 978-3-8325-5020-2