

Wolfram Rollett · Hannah Bijlsma ·
Sebastian Röhl *Editors*

Student Feedback on Teaching in Schools

Using Student Perceptions for the
Development of Teaching and Teachers

OPEN ACCESS

 Springer

Student Feedback on Teaching in Schools

Wolfram Rollett · Hannah Bijlsma · Sebastian Röhl
Editors


Student Feedback on Teaching in Schools

Using Student Perceptions for
the Development of Teaching and Teachers

 Springer

Editors

Wolfram Rollett 
Institute for Educational Sciences
University of Education
Freiburg, Germany

Hannah Bijlsma 
Section of Teacher Professionalization
University of Twente
Enschede, The Netherlands

Sebastian Röhl 
Institute for Educational Sciences
University of Education
Freiburg, Germany



ISBN 978-3-030-75149-4 ISBN 978-3-030-75150-0 (eBook)
<https://doi.org/10.1007/978-3-030-75150-0>

© The Editor(s) (if applicable) and The Author(s) 2021. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Feedback is a hot topic—with so many studies, and now over 23 meta-analyses on the effect of feedback on achievement (and many more on other outcomes and in other disciplines). Most of this research is premised on feedback from teachers to students, whereas a critical missing link is the measurement, quality, and impact of feedback from students to teachers. This is a well-rehearsed topic in tertiary classes, but far less so in K-12. This book begins to add foundations to the debates about feedback to teachers in elementary, secondary, and high schools.

When I published *Visible Learning* in 2009, the average effect size from feedback to students was 0.78, and from feedback to teachers was 0.50 (nearly all of the latter studies being from university students). We recently revisited the meta-analyses and located almost every study in the 23 meta-analyses, and recalculated the individual and overall effect: the overall to students has reduced to 0.48 (Wisniewski et al. 2020, also see Chap. 8). One major reason is the over hype about feedback, the misplaced emphasis on increasing the quantity of feedback; the ignoring of the massive variability of feedback has led this to decrease. The variability is core to understanding the effect—This was seen in the major synthesis by Kluger and deNisi (1996), who observed that about one-third of feedback is negative, and who were careful to note that the search for these moderators is core. The same feedback may work for me but not you, the same feedback to me today works but not tomorrow. Understanding this variability is the core and this is so often forgotten.

The search for these critical moderators has been underway and gathering pace for many years. There are, from my research on feedback on learning, at least five moderators: Feedback is maximized when there is “where to next/improvement focused” information in the feedback; when feedback is aligned with the instructional cycle (about the task, process, self-regulation); that praise dilutes the effects (as students focus on and recall the praise over the information); when we overly focus on the giving compared to the reception of feedback; and, most critically, the effect on students is higher when teachers demonstrate they are willing to receive feedback about their impact.

Like students, teachers need to hear, understand, and action the feedback they receive. Some teachers are impervious to feedback, thinking that their task is to “give” feedback to students, not receive it themselves; some use the many cognitive

biases which make us humans to reinterpret student feedback (such as confirmation bias, where good feedback is about me and negative feedback is about the students); some are extremely good at selectively listening to student feedback; some dismiss student feedback as ill-informed and the whims of youngsters; and some collect the feedback too late, so it has little impact on improving the teaching for the students. In my own case as a University Professor, I have read student evaluations of my teaching for many decades and if they do not say “He speaks too fast,” then I question the validity of the other student responses. But the proper question should be: Why have I not improved my speaking skills? Such confirmation bias, dismissing of the value of student responses, and seeing evaluations as more worthwhile for promotion than improvement, means that I and my students are the losers.

It does seem ironic that teachers will listen to feedback from external adult observers who come into their classes and conduct fleeting observations of their teaching. This feedback from observers is more often about *how* they teach, and usually not about the impact of their teaching on the students. There is a corpus of research, noted in many chapters, of the major issues with the reliability of these external observations. But there is already a plethora of studies showing the high reliabilities of student feedback to teachers, which is then often dismissed—Students could not possibly have worthwhile information, are biased (as they do not know good teaching!), and they are recipients not informers of teaching. Why do we (a) see unreliability as a major question for observational methods, but (b) when student feedback has been shown to meet these reliability criteria, we change the blame to other factors? This book reviews the evidence on how to maximize the reliability and validity of student feedback, outlines many of the more dependable measures, and invites deeper discussion on the informational value which can be derived from students’ feedback to teachers.

Students are not, however, “consumers”; they are learners, and learning is often messy. Failure, therefore, needs to be a learner’s best friend and welcomed by the teacher as opportunity to learn. Hearing from the students about their struggles, what they do not know, their ways of thinking about the content, their conceptions and misconceptions—surely this is the food for more effective teaching. Teachers who are impervious to this feedback are more likely to attribute failure to the students, devise explanations why the students cannot learn (they come from poor backgrounds, unsupportive families, are not well enough prepared for my class, they do not pay attention, are disruptive, have fixed mindsets, and so on)—when they themselves are the only people in the room paid to improve.

The editors set the scene by outlining the Process Model of Student Feedback on Teaching (SFT; Chap. 1 in this volume by Röhl, Bijlsma, and Rollett), which deals with ensuring a comprehensive, reliable, interpretable collection of data - but also has a major emphasis on how student feedback is understood and interpreted by the teacher. There is much emphasis on the teacher as receiver of the feedback—and this is investigated from cognitive as well as from affective perspectives. There is then much debate about the many instruments, their dimensionality, construction, and measurement properties. With the move in the 1990s such that validity now is

seen through the lens of whether the test information is correctly interpreted and has consequential impact, we do need to know more about the quality of the reports.

Van der Lans (Chap. 5) asks a core question about two teachers, John and Tess. Given that we have collected information from students about these teachers—What can we advise Tess and John? This is a reasonably undeveloped territory, and possibly could account for so much dismissal of student feedback—If we do not intend to use the information to improve, then what is its value, and why collect it. This question begs the next generation of researchers to devote more to the consequences of asking for student evaluation. When we give feedback to students, it is more often to improve their learning, and we are most creative and effective in framing feedback to engender improvement. Ditto, when we receive feedback from students. We are not there yet.

It is thus fascinating to find that the only study characteristic in Röhl's (Chap. 9) meta-analysis of longitudinal effects of receiving feedback from students, is the level of support to the teacher. Treatments with a high level of individual support for reflecting on feedback and teaching development showed a significantly higher effect size ($d = 0.52$) than studies with a medium ($d = -0.06$) or low ($d = 0.16$) supportive level. That is, feedback to teachers only made sense when there was support for subsequent teaching development, with ongoing advice on the subsequent development processes through individual or group consultations, counselling, or professional learning communities (also noted in Fleenor, Chap. 14). This means in the presence of greater professional learning about the consequences and actions from the feedback.

One of my concerns about many of the current student evaluation tools is that they focus on particular ways to teach, and then assume if you teach in these ways that there are positive effects on students. Once again, confusing correlation with causation. In most of the instruments outlined in these chapters, there are too few asking for feedback about the students' learning, and the impact teachers are having on students. We know, for example, if a student believes that the teacher has no credibility, then the teacher is unlikely to have much impact—even if they are using all the desirable teacher strategies with great classroom climates. This is a call for more about the students' conceptions of what it means to be a learner, whether the class is an inviting place to come and learn, and students' conceptions of their learning. Such feedback to teachers could be among the more powerful to improve teaching quality.

It seems a cogent discovery that much of the use of student evaluations, and probably their subsequent impact on improving the quality of the teaching, starts with a high sense of teacher self-efficacy. Röhl and Gärtner (Chap. 10) noted that this impact is based on the teacher's attitude towards considering students as trustworthy or competent as feedback providers. So often I have seen teachers dismiss student evaluation, whereas building confidence in the informational value, the benefits to the teacher to thence learn how to improve, may be a critical first step; this once again highlights the importance of the value of the reports to help the teacher interpret and take actions. Also note the critical comments by Schweig and Martínez (Chap. 6) about the information in the variances. I was taught by my mentor, Rod McDonald (a famous psychometrician), that the answers often lie in the residuals—the detail

is in the variances. No teacher would proclaim that all students think and learn alike—so attending to the variances (often missing in some current measures and reports) seems well worth pursuing. When you see variance, this is a great moment to triangulate with other information, and so any deviations or surprises can be more closely investigated.

This book provides a line in the sand. It reviews what we know about student feedback to teachers, it makes the powerful point that most teachers have positive attitudes towards receiving such feedback (see Göbel et al. in Chap. 11), outlines many of the measures and measurement issues, and raises the more important questions still to be resolved. This makes this book timely. It is detailed and it is a pleasure to read. To have these chapters in one place—and from those most up to date with the research literature and doing the research—is a gift.

John Hattie
Melbourne Graduate School of Education
Melbourne, Australia

References

- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>.
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>.

Contents

| | | |
|--|--|------------|
| 1 | The Process Model of Student Feedback on Teaching (SFT): A Theoretical Framework and Introductory Remarks | 1 |
| | Sebastian Röhl, Hannah Bijlsma, and Wolfram Rollett | |
| Part I Measuring Student Perceptions of Teaching: Reliability, Validity, and Theoretical Considerations | | |
| 2 | A Reflection on Student Perceptions of Teaching Quality from Three Psychometric Perspectives: CCT, IRT and GT | 15 |
| | Hannah Bijlsma, Rikkert van der Lans, Tim Mainhard, and Perry den Brok | |
| 3 | Student Perceptions of Teaching Quality: Dimensionality and Halo Effects | 31 |
| | Sebastian Röhl and Wolfram Rollett | |
| 4 | The Quality of Student Perception Questionnaires: A Systematic Review | 47 |
| | Hannah Bijlsma | |
| 5 | A Probabilistic Model for Feedback on Teachers' Instructional Effectiveness: Its Potential and the Challenge of Combining Multiple Perspectives | 73 |
| | Rikkert van der Lans | |
| 6 | Understanding (Dis)Agreement in Student Ratings of Teaching and the Quality of the Learning Environment | 91 |
| | Jonathan D. Schweig and José Felipe Martínez | |
| 7 | Student Ratings of Teaching Quality Dimensions: Empirical Findings and Future Directions | 111 |
| | Richard Göllner, Benjamin Fauth, and Wolfgang Wagner | |

Part II Using Student Feedback for the Development of Teaching and Teachers

- 8 Functions and Success Conditions of Student Feedback in the Development of Teaching and Teachers** 125
Benedikt Wisniewski and Klaus Zierer
- 9 Effects of Student Feedback on Teaching and Classes: An Overview and Meta-Analysis of Intervention Studies** 139
Sebastian Röhl
- 10 Relevant Conditions for Teachers' Use of Student Feedback** 157
Sebastian Röhl and Holger Gärtner
- 11 Student Feedback as a Source for Reflection in Practical Phases of Teacher Education** 173
Kerstin Göbel, Corinne Wyss, Katharina Neuber, and Meike Raaflaub
- 12 Reciprocal Student–Teacher Feedback: Effects on Perceived Quality of Cooperation and Teacher Health** 191
Jan-Erik Schmidt and Caterina Gawrilow

Part III Relating to Other Fields of Research

- 13 Student Voice and Student Feedback: How Critical Pragmatism Can Reframe Research and Practice** 209
Mari-Ana Jones and Valerie Hall
- 14 What Can We Learn from Research on Multisource Feedback in Organizations?** 221
John W. Fleenor
- 15 Lessons Learned from Research on Student Evaluation of Teaching in Higher Education** 237
Bob Uttl

Part IV Discussion and Future Directions

- 16 Student Feedback on Teaching in Schools: Current State of Research and Future Perspectives** 259
Wolfram Rollett, Hannah Bijlsma, and Sebastian Röhl

Editors and Contributors

About the Editors

Wolfram Rollett is a Professor of Empirical Educational Research at the University of Education Freiburg and the Freiburg Advanced Center of Education (FACE). Previously he worked as a researcher and lecturer in the field of Educational Science and Psychology at the Universities of Potsdam, Braunschweig, Dortmund, and Wuppertal. His research focuses on school development processes, the quality of extra- and co-curricular activities, educational effectiveness, and classroom composition.

Hannah Bijlsma is a researcher (Ph.D.) at the section of Teacher Professionalization at the University of Twente (The Netherlands) and a primary school teacher (grade 1). Her research focuses on measuring teaching quality and on the use of student perceptions of teaching quality in school contexts. In 2016 she founded a professional association for academic primary school teachers, of which she has been chairman for about five years. She is now a board member of the International Congress for School Effectiveness and Improvement (ICSEI) and a board member of the EARLI SIG on School Effectiveness and Improvement.

Sebastian Röhl has been an Academic Assistant in the Department of Educational Science at the University of Education Freiburg and is a currently Postdoctoral Researcher in the Institute of Education at Tübingen University (Germany). Before that he worked for more than 10 years as a grammar school teacher, school development consultant, and in teacher training. Among other areas, he conducts research in the fields of teaching development and teacher professionalization through feedback, social networks in inclusive school classes, as well as teachers' religiosity and its impact on professionalism. In addition, he is the director of an in-service professional master's study program for teaching and school development.

Contributors

Hannah Bijlsma Section of Teacher Professionalization, University of Twente, Enschede, The Netherlands

Perry den Brok Department of Educational and Learning Sciences, University of Wageningen, Wageningen, The Netherlands

Benjamin Fauth Institute for Educational Analysis (IBBW), Stuttgart, Germany

John W. Fleenor Center for Creative Leadership, Greensboro, NC, USA

Holger Gärtner Freie Universität Berlin, Berlin, Germany

Caterina Gawrilow Department of Psychology, University of Tübingen, Tübingen, Germany

Kerstin Göbel Faculty of Educational Sciences, University of Duisburg Essen, Essen, Germany

Richard Göllner Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

Valerie Hall University of Wolverhampton, Wolverhampton, UK

Mari-Ana Jones Norwegian University of Science and Technology, Trondheim, Norway

Tim Mainhard Department of Education, University of Utrecht, Utrecht, The Netherlands

José Felipe Martínez University of California, Los Angeles, CA, USA

Katharina Neuber Faculty of Educational Sciences, University of Duisburg Essen, Essen, Germany

Meike Raaflaub University of Teacher Education, Bern, Switzerland

Sebastian Röhl Institute for Educational Sciences, University of Education, Freiburg, Germany

Wolfram Rollett Institute for Educational Sciences, University of Education, Freiburg, Germany

Jan-Erik Schmidt Center for School-Quality and Teacher Education, Tübingen, Germany

Jonathan D. Schweig RAND Corporation, Arlington, VA, USA

Bob Uttl Mount Royal University, Calgary, AB, Canada

Rikkert van der Lans LUMC-Curium - Child and Adolescent Psychiatry, Leiden, The Netherlands

Wolfgang Wagner Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

Benedikt Wisniewski Faculty of Philosophy and Social Sciences, University of Augsburg, Augsburg, Germany

Corinne Wyss FHNW School of Education, Brugg, Switzerland

Klaus Zierer Faculty of Philosophy and Social Sciences, University of Augsburg, Augsburg, Germany

Chapter 1

The Process Model of Student Feedback on Teaching (SFT): A Theoretical Framework and Introductory Remarks



Sebastian Röhl, Hannah Bijlsma, and Wolfram Rollett

Abstract Student feedback on teaching in schools, conceptualized as information on student perceptions of teaching, is described by many scholars as an effective instrument for the developmental use of teachers and teaching. Beyond that, various studies show that the productive use of this method is a very complex process in which a variety of aspects must be considered. As an introduction to this volume, this chapter presents a model based on findings from different research areas of feedback and school research, called *Process Model of Student Feedback on Teaching* (SFT). This model follows the steps of the student feedback process, starting with student perceptions of teaching, which must be professionally collected or measured. Subsequently, the teacher perceives and interprets this feedback information, which is linked to cognitive and affective reactions and processes. This can lead to an enhancement of teachers' knowledge about their own teaching and to the initiation of improvement-oriented actions, finally resulting in improved teaching and development of the teachers' professional competence. Thereby, characteristics of the organization, the students, and classes as well as the teachers need to be considered. This model serves as a framework for the subsequent overview of the contributions in this volume.

Keywords Student feedback · Process model · Student perceptions of teaching quality · Teacher development

S. Röhl (✉) · W. Rollett (✉)
University of Education, Freiburg, Germany
e-mail: sebastian.roehl@ph-freiburg.de

W. Rollett
e-mail: wolfram.rollett@ph-freiburg.de

H. Bijlsma
Section of Teacher Professionalization, University of Twente, Enschede, the Netherlands
e-mail: h.j.e.bijlsma@utwente.nl

1 Student Feedback in Schools

Student learning processes are influenced by many different factors, including student, home, school, peer, headteacher, and teacher effects (Hattie, 2009). In schools, teachers are considered to be the most malleable, within-school influence on student learning (Haertel, 2013; Nye et al., 2004), because the teacher determines the events in the classroom to a large extent. In order to be able to work and develop toward their full potential, it is important for teachers to receive information on the quality of their teaching. The teachers can gain some insight into this through the results of monitoring learning success and their own class observations. Such information proves to be particularly helpful when it comes from outside sources. This “information provided by an agent (...) regarding aspects of one’s performance” (Hattie & Timperley, 2007, p. 81) is typically labeled as *feedback*. If feedback on teaching is considered as valuable information about teachers’ performance and used accordingly, it can have positive effects on the professional development of teachers, the quality of teaching and student learning (Garet et al., 2017). Therefore, ideally there would be enough time and available methods to provide teachers with constructive feedback about their teaching in order to improve the quality of their teaching, and, as a follow-up, to positively affect the learning processes of their students.

However, often little energy is invested in education on constructive feedback to teachers about the quality of their teaching (Frase & Streshly, 1994; Voerman et al., 2012). At the same time, classroom observations by an external observer are quite common in many school systems (Darling-Hammond, 2013). Unfortunately, to obtain a truly reliable picture of teaching quality, it is necessary to rate lessons several times, and these observations should be made by several trained observers (Praetorius et al., 2014). This makes the use of classroom observations time-consuming and expensive. On the other hand, using teachers’ self-assessments of their lessons might result in invalid data, because it is questionable whether teachers are able to judge their own lessons—as they see teaching only from their own perspective (Kruger & Dunning, 1999; Visscher, 2017)—and such self-assessments can hardly be looked on as an “information provided by an agent”.

Another way to provide teachers with feedback is to use student perceptions of teaching quality (Muijs, 2006; Peterson et al., 2000). If student perceptions are used, the number both of observed lessons (in cases where students access one teacher’s teaching over several lessons) and of observers (the number of students) is larger than in the case of lesson observations by external persons, which could thus improve the reliability of the feedback scores (Fauth et al., 2014). In addition, student perceptions reflect the perspective of the target group (Kane & Staiger, 2012; Quaglia & Corso, 2014; Staiger, 2012).

Although there are concerns about the validity and reliability of student perceptions of teaching quality (e.g., the extent to which students are able to discriminate between the different facets of teaching: de Jong & Westerhof, 2001; Fauth et al., 2014; Ferguson, 2012; Kunter & Baumert, 2006), recent studies have shown that student perceptions of teaching quality can provide reliable and valid information

both for research purposes and as feedback to teachers for formative evaluation of the quality of their teaching (Burniske & Meibaum, 2012; Ferguson & Danielson, 2014; Kane et al., 2013; Kyriakides, 2005; Peterson et al., 2000).

2 Using Student Perceptions of Teaching for the Development of Teaching and Teachers—The Process Model of Student Feedback on Teaching (SFT)

The basic idea of using student feedback for the development of teaching is to give teachers a comprehensive view of their teaching from the students' perspective, which might result in valuable information or data for teachers about the quality of their teaching. Based on the feedback, they can carry out improvement-oriented actions which might enhance their lessons. This, in turn, could result in more positive perceptions of the teaching by the students and improved learning processes for those students. The first experiments with this form of developing teaching were already being done in the USA in the 1920s (Remmers, 1927). Moreover, the underlying simplistic model of this approach is also commonly used in the data-based decision making research field (Lai et al., 2014; Poortman & Schildkamp, 2016; Schildkamp, 2019; van Geel et al., 2016), where it is stated that the use of teaching-related data, such as the evaluations of students' learning processes, can help to improve teaching and students' learning outcomes. In addition, the process of obtaining feedback from students and the associated student–teacher communication is an educational process, which can promote skills such as giving and receiving feedback, discussability, dealing with criticism, and different points of view (e.g., Bastian, 2010; Zierer & Wisniewski, 2019). Student feedback is still seen as a way of promoting *student voice* (Cook-Sather, 2002, 2007): the voice of students in their own education (Lincoln, 1995). It seems important for such a process to focus on the formative use of student ratings. A summative use of student ratings in schools for accountability purposes could hinder such effects as the teachers would need to justify their teaching. Notably, the use of student feedback for developmental purposes in schools seems to be almost exclusive to Western countries. A systematic literature review on this topic identified studies from Europe, the USA, Australia, and Turkey only, although validated student perception questionnaires for assessing teaching quality also exist for the Asian region, for example (see Chap. 9 of this volume).

Regarding the practical implementation of student feedback, it becomes apparent that the process of gathering and evaluating feedback is not quite that simple. For example, it is necessary both to overcome routines like basing decisions on intuition and instinct (Schildkamp & Kuiper, 2010) and for teachers to be data literate in order to use data systematically for the improvement of their lessons (Kippers et al., 2018; Mandinach & Gummer, 2016). Furthermore, the process of utilizing student perceptions of teaching quality as data to improve teaching is complex, as it is influenced by teacher, student, and class characteristics, and occurs within an organizational context

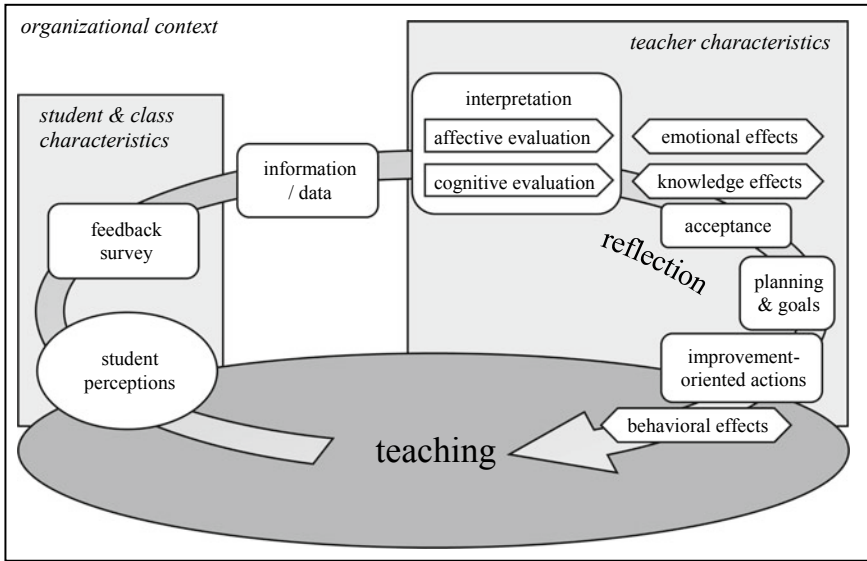


Fig. 1 Process model of student feedback on teaching (SFT, *Source Own*)

(Schildkamp, 2019). Moreover, the success of this process can depend on many situational factors, such as questionnaire characteristics, personal reactions evaluating the feedback, or the choice of improvement-oriented actions based on the feedback. In this introductory chapter, we therefore gradually suggest a more complex model of the use of student feedback for developing teaching and teachers which—among other things—includes these factors. The model is visualized in Fig. 1.

The process starts with the students, who perceive the teaching in class. These perceptions can be captured via a student feedback survey. For this purpose, questionnaires are often used, which include items to be rated or open-ended questions. Student perceptions as well as their teaching quality ratings might be influenced by student and class characteristics (Bijlsma et al., 2019; Fauth et al., 2020; Levy et al., 1992).

Once the feedback is collected (that is, when the information or data are available), it must be understood and interpreted by the teacher. Following this, a cognitive process takes place, which is often described as *reflection* in the context of teacher training and professional development (e.g., Beauchamp, 2006; Korthagen & Wubbels, 1995). Reflection should lead to better understanding of one's own teaching, and subsequently to better teaching practice (Driessen et al., 2008; Ertmer & Newby, 1996).

Research on feedback from organizational psychology could provide important insights with regard to the process of reflection on feedback. According to Ilgen et al. (1979), the processing of received performance feedback follows several steps. First, the feedback message is perceived by the person receiving the feedback, in which the accuracy and intensity of the perception play an important role. Then a decision

is made about the extent to which the perceived feedback message is accepted, i.e., whether the received information is considered to be truthful. As a result, the desire or intention to respond to the feedback can arise, followed by the setting of goals in this regard (intended response) and the implementation of the intended response in practice (see also Kinicki et al., 2004).

A comparable process model is also found in Smither et al. (2005), with the authors proposing the following steps: initial reactions, goal-setting and related actions, taking action and subsequent performance improvement. As an extension of these sequence models, Kahmann and Mulder (2011) included not only cognitive reactions to feedback, but also affective reactions by the person receiving the feedback, which can both eventually result in behavioral effects.

For our theory of action, we combined these sequence models considering the context of teachers as recipients of student feedback. Therefore, we view the perception and interpretation of feedback not only from cognitive perspectives, but also from affective ones.

Regarding these *emotional effects*, student feedback can evoke positive emotions such as satisfaction and joy, or negative ones such as dissatisfaction or defensiveness. These are primarily influenced by the actual as well as the expected positivity or negativity of the feedback, respectively. *Knowledge effects* can occur when feedback provides the teacher with new information about the students' view of his or her teaching or the feedback reinforces the teacher's existing knowledge. Then, a comparison between one's own perceptions and standards for teaching takes place. Discrepancies, which emerge, must be *accepted* in order for the teacher to consider changes in their teaching. Feedback data, which differ strongly from one's own objectives and standards concerning the own teaching together with negative emotional reactions, can lead to *rejection* of the feedback (Kahmann & Mulder, 2011; Kluger & DeNisi, 1996). Furthermore, it should be noted that a discrepancy between the actual state and the target state can also lead to abandonment or modification of the previously set objectives or standards in order to avoid or reduce any further effort (Kluger & DeNisi, 1996).

After the perception and acceptance of a possible area of improvement, *goals* for the elimination of a discrepancy can be set, followed by *planning* for the implementation of the intended response (Smither et al., 2005). Subsequently, as *behavioral effects*, *improvement-oriented actions* can take place, such as adaptive teaching to the different needs of students in the class (Gaertner, 2014), increased attention to specific aspects during teaching (Röhl & Rollett, 2021), discussions with students about the feedback for collaborative improvement (Gaertner, 2014), or participation in special training courses (Balch, 2012). If the actions have the desired effect on teaching practices, this might result in higher ratings from students in subsequent feedback surveys, and/or better learning outcomes.

We consider the presented process model to be a promising tool for structuring research and research questions on student feedback on teaching in schools. The model combines existing research from different research fields and covers what is known about developmental process of teaching and teachers based on student feedback. We acknowledge that the model is an ideal presentation; in real school

settings, influencing factors like the student, class, teacher, and organizational aspects need to be considered. Therefore, the present volume covers a variety of topics linked to these influencing factors. Subsequently, we use this model to arrange and link the contributions and perspectives of this volume in their meaning and connection to this process.

3 Overview of the Volume

In Part One of the volume, student perceptions of teaching quality and their validity and reliability are discussed by considering several theoretical and psychometric issues. These topics address issues which concern theoretical and research questions pertaining the beginning of the cycle of the Process Model of Student Feedback on Teaching (SFT) just introduced. In Chap. 2, Bijlsma et al. introduce the measurement of student perceptions from three psychometric perspectives which dominate contemporary research on teaching quality. They aim to connect psychometric theories and the different perspectives on what (measured) student perceptions are seen to be, as well as the different perspectives regarding how and for what purposes student perceptions should be used. In Chap. 3, Röhl and Rollett—in line with the Process Model of Student Feedback on Teaching (SFT)—discuss theoretically assumed teaching quality dimensions, which can be distinguished in student feedback surveys. Findings on the importance of teachers' communion with students (warmth or cooperation) as a potentially biasing factor in student ratings of instructional quality are also discussed. For Chap. 4, Bijlsma conducted a systematic review on the psychometric quality of student perception questionnaires (SPQ). She presents detailed overviews with general information about the SPQs, the results of the evaluation, and the constructs measured by the SPQs. In Chap. 5, van der Lans focuses on evidence showing that student questionnaires and classroom observation instruments can provide reliable feedback to teachers. He provides empirical evidence indicating that feedback of classroom observations and student questionnaires can be calibrated on the same continuum of instructional effectiveness; he moves on to discuss implications for theory, future research, and practice. In Chap. 6, Schweig and Martínez present an overview of literature from different fields which examines consensus in different measures of teaching quality. They consider these alongside key assumptions and consequences of those measurement models and analytic methods which are commonly used to summarize student survey reports of teaching quality. In Chap. 7, Göllner et al. continue with further findings on the particularities of student ratings of instructional quality, pointing out the importance of considering how exactly the referent and the addressee are noted in survey items and presenting related perspectives for future in-depth research approaches.

Part Two of the book focuses on the use of student feedback for the development of teaching and teachers. Following the SFT model, we arrive here at interpretation, reflection, and the teacher improvement elements. In Chap. 8, Wisniewski and Zierer start with an overview of functions of and success conditions for student feedback

in the development of teaching and teachers. They point out why feedback is important for the professional development of teachers in general, and discuss three basic functions of student feedback in schools. This is followed, in Chap. 9, by Röhl's contribution, in which the first meta-analysis of the effects of student feedback on teaching quality in secondary schools is presented, providing insights on its effectiveness and potential moderating variables. In Chap. 10, Röhl and Gärtner systematize relevant factors influencing the utilization of student feedback by teachers into three domains: personal characteristics of feedback recipients (teachers), characteristics of the organization (school), and characteristics of feedback information (data). The two chapters which follow discuss student feedback from a more practical point of view. Göbel et al. (Chap. 11) focus on the use of student feedback to improve teaching quality during practical phases in teacher education. The authors discuss challenges and opportunities for the use of student feedback as an instrument for reflection on teaching and professional development for pre-service teachers. Schmidt and Gawrillow (Chap. 12) describe the theoretical parameters of reciprocal student–teacher feedback on cooperation between students and teachers, and outline results of an empirical study on the effects of the reciprocal method on the perceived quality of cooperation and on teacher health.

In the next part, three chapters of the volume provide supplementary perspectives on the use of student feedback for developing teaching and teachers, relating to the final part of the feedback cycle of the SFT model. Jones and Hall shed light on the critical pragmatism perspective (Chap. 13), and focus on how student feedback can facilitate dialogue and thus contribute to the development of schools as democratic communities. The multisource feedback perspective in organizations, and the transferability of this perspective to student-to-teacher feedback in schools is discussed by Fleenor (Chap. 14). In Chap. 15, Uttl overviews the lessons to be learned from research on student evaluation of teaching in higher education providing insights to be taken up in research on student feedback on teaching in schools.

Finally, in the concluding chapter of the book, Rollett et al. summarize the findings and conclusions drawn from the chapters in this volume and discuss the directions forward for researchers, policy makers, and schools.

Acknowledgments The editors would like to thank the Dutch Research Council (NWO; project number 36.201.009) and the University of Education Freiburg (Germany) for funding the open access publication of this volume.

References

- Balch, R. T. (2012). *The validation of a student survey on teacher practice*. Doctoral dissertation, Vanderbilt University, Nashville, TN.
- Bastian, J. (2010). Feedbackarbeit in Lehr-Lern-Prozessen [Working with feedback in teaching-learning processes]. *Gruppendynamik und Organisationsberatung*, 41, 21–37. <https://doi.org/10.1007/s11612-010-0097-4>.

- Beauchamp, C. (2006). *Understanding reflection in teaching: A framework for analyzing the literature*. Doctoral dissertation, McGill University, Montreal. <https://escholarship.mcgill.ca/concern/theses/w0892g316>.
- Bijlsma, H. J. E., Glas, C. A. W., & Visscher, A. J. (2019, August 12). *The factors influencing digitally measured student perceptions of teaching quality*. Paper presented at the EARLI conference in Aachen.
- Burniske, J., & Meibaum, D. (2012). *The use of student perception data as a measure of teaching effectiveness*. Texas Comprehensive Center. Retrieved from http://txcc.sedl.org/resources/briefs/number_8/index.php.
- Cook-Sather, A. (2002). Authorizing students' perspectives: Toward trust, dialogue, and change in education. *Educational Researcher*, 31(4), 3–14.
- Cook-Sather, A. (2007). What would happen if we treated students as those with opinions that matter? The benefits to principals and teachers of supporting youth engagement in school. *NASSP Bulletin*, 91(4), 343–362.
- Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College. <https://doi.org/10.1177/0192636507309872>.
- de Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, 4, 51–85. <https://doi.org/10.1023/A:1011402608575>.
- Driessen, E., van Tartwijk, J., & Dornan, T. (2008). The self-critical doctor: Helping students become more reflective. *BMJ*, 336, 827–830. <https://doi.org/10.1136/bmj.39503.608032.AD>.
- Ertmer, P. A., & Newby, T. J. (1996). The expert learner: Strategic, self-regulated and reflective. *Instructional Sciences*, 24, 1–24.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9.
- Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff-Bruchmann, J., Lüdtke, O., et al. (2020). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology*, 112, 1284–1302. <https://doi.org/10.1037/edu0000416>.
- Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, 94(3), 24–28. <https://doi.org/10.1177/003172171209400306>.
- Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 98–144). Jossey-Bass.
- Frase, L. E., & Streshly, W. (1994). Lack of accuracy, feedback, and commitment in teacher evaluation. *Journal of Personnel Evaluation in Education*, 8(1), 47–57. <https://doi.org/10.1007/BF00972709>.
- Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, 42, 91–99. <https://doi.org/10.1016/j.stueduc.2014.04.003>.
- Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., Manzeske, D., et al. (2017). *The impact of providing performance feedback to teachers and principals: Executive summary*. National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Haertel, B. E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores*. William H. Angoff memorial lecture series. Educational Testing Service.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. <https://doi.org/10.1007/s11159-011-9198-8>.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112. <https://doi.org/10.3102/003465430298487>.
- Ilgen, D. R., Fisher, C. D., & Taylor, S. M. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4), 349–371. <https://doi.org/10.1037/0021-9010.64.4.349>.

- Kahmann, K., & Mulder, R. H. (2011). *Feedback in organizations: A review of feedback literature and a framework for future research*. Regensburg. https://www.uni-regensburg.de/psychologie-paedagogik-sport/paedagogik-2/medien/kahmann_mulder_2011.pdf. Accessed 31 October 2019.
- Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers?* Bill & Melinda Gates Foundation. http://www.metproject.org/downloads/MET_Validing_Using_Random_Assignment_Research_Paper.pdf. Accessed 29 August 2020.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bell & Melinda Gates Foundation.
- Kinicki, A. J., Prussia, G. E., Wu, B., & McKee-Ryan, F. M. (2004). A covariance structure analysis of employees' response to performance feedback. *The Journal of Applied Psychology*, 89, 1057–1069. <https://doi.org/10.1037/0021-9010.89.6.1057>.
- Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention? *Studies in Educational Evaluation*, 56, 21–31. <https://doi.org/10.1016/j.stueduc.2017.11.001>.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>.
- Korthagen, F. A. J., & Wubbels, T. (1995). Characteristics of reflective practitioners: Towards an operationalization of the concept of reflection. *Teachers and Teaching*, 1, 51–72. <https://doi.org/10.1080/1354060950010105>.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>.
- Kunter, M., & Baumert, J. (2006). Who's the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environment Research*, 9, 231–251. <https://doi.org/10.1007/s10984-006-9015-7>.
- Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction*, 40(2), 44–66.
- Lai, M. K., Wilson, A., McNaughton, S., & Hsiao, S. (2014). Improving achievement in secondary schools: Impact of a literacy project on reading comprehension and secondary school qualifications. *Reading Research Quarterly*, 49(3), 305–334. <https://doi.org/10.1002/rrq.73>.
- Levy, J., Wubbels, T., & Brekelmans, M. (1992). Student and teacher characteristics and perceptions of teacher communication style. *Journal of Classroom Interaction*, 27, 23–29.
- Lincoln, Y. S. (1995). In search of student voices. *Theory into Practice*, 34, 88–93.
- Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, 60, 366–376. <https://doi.org/10.1016/j.tate.2016.07.011>.
- Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, 12(1), 53–75. <https://doi.org/10.1080/13803610500392236>.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237–257. <https://doi.org/10.3102/01623737026003237>.
- Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education*, 14, 135–153. <https://doi.org/10.1023/A:1008102519702>.
- Poortman, C. L., & Schildkamp, K. (2016). Solving student achievement problems with a data use intervention for teachers. *Teaching and Teacher Education*, 60, 425–433. <https://doi.org/10.1016/j.tate.2016.06.010>.
- Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12. <https://doi.org/10.1016/j.learninstruc.2013.12.002>.
- Quaglia, R., & Corso, M. (2014). *Student voice. The instrument of change*. Corwin Press.

- Remmers, H. H. (1927). The purdue rating scale for instructors. *Educational Administration and Supervision* (6), 399–406.
- Röhl, S., & Rollett, W. (2021). Jenseits von Unterrichtsentwicklung: Intendierte und nicht-intendierte Nutzungsformen von Schülerfeedback durch Lehrpersonen [Beyond teaching development: Intended and non-intended ways of utilization of student feedback by teachers]. In K. Göbel, C. Wyss, K. Neuber, & M. Raafflaub (Eds.), *Quo vadis Forschung zu Schülerrückmeldungen?* [Quo vadis research on student feedback?]. Springer VS. <https://doi.org/10.1007/978-3-658-32694-4>.
- Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, 1–17. <https://doi.org/10.1080/00131881.2019.1625716>.
- Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26(3), 482–496. <https://doi.org/10.1016/j.tate.2009.06.007>.
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multi-source feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58, 33–66. https://doi.org/10.1111/j.1744-6570.2005.514_1.x.
- van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal*, 53(2), 360–394. <https://doi.org/10.3102/0002831216637346>.
- Visscher, A. J. (2017). *Gericht ontwikkelen van leerkrachtkwaliteiten* [Developing teacher qualities in a targeted way]. University of Twente.
- Voerman, L., Meijer, P. C., Korhagen, F. A. J., & Simons, R. J. (2012). Types and frequencies of feedback interventions in classroom interaction in secondary education. *Teaching and Teacher Education*, 28(8), 1107–1115. <https://doi.org/10.1016/j.tate.2012.06.006>.
- Zierer, K., & Wisniewski, B. (2019). *Using student feedback for successful teaching*. Routledge.

Sebastian Röhl has been an Academic Assistant in the Department of Educational Science at the University of Education Freiburg and is currently a Postdoctoral Researcher in the Institute of Education at Tübingen University (Germany). Before that he worked for more than 10 years as a grammar school teacher, school development consultant, and in teacher training. Among other areas, he conducts research in the fields of teaching development and teacher professionalization through feedback, social networks in inclusive school classes, as well as teachers' religiosity and its impact on professionalism. In addition, he is the Director of an in-service professional master's study program for teaching and school development.

Hannah Bijlsma is a researcher (Ph.D.) at the section of Teacher Professionalization at the University of Twente (the Netherlands) and a primary school teacher (grade 1). Her research focuses on measuring teaching quality and on the use of student perceptions of teaching quality in school contexts. In 2016 she founded a professional association for academic primary school teachers, of which she has been chairman for about five years. She is now a board member of the International Congress for School Effectiveness and Improvement (ICSEI) and a board member of the EARLI SIG on School Effectiveness and Improvement.

Wolfram Rollett is a Professor of Empirical Educational Research at the University of Education Freiburg and the Freiburg Advanced Center of Education (FACE). Previously he worked as a researcher and lecturer in the field of Educational Science and Psychology at the Universities of Potsdam, Braunschweig, Dortmund, and Wuppertal. His research focuses on school development processes, the quality of extra- and co-curricular activities, educational effectiveness, and classroom composition.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the Chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the Chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part I
**Measuring Student Perceptions
of Teaching: Reliability, Validity,
and Theoretical Considerations**

Chapter 2

A Reflection on Student Perceptions of Teaching Quality from Three Psychometric Perspectives: CCT, IRT and GT



Hannah Bijlsma, Rikkert van der Lans, Tim Mainhard, and Perry den Brok

Abstract This chapter discusses student perceptions in terms of three psychometric perspectives that dominate contemporary research on teaching quality, namely, Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability Theory (GT). These perspectives function as being exemplars for the connection between psychometric theories and the different perspectives on “what a perception is” as well as on how and for what purposes student perceptions should be used. The main message of the chapter is that the choice of a psychometric theory is not merely a technical matter, but also has implications for how the nature of perceptions is conceptualized. After presenting and linking each psychometric theory, their strengths and weaknesses in the context of student perceptions of teaching quality and issues on practical implementations are discussed.

Keywords Student perceptions · Teaching quality · Classical test theory · Item response theory · Generalizability theory

H. Bijlsma (✉)

Section of Teacher Professionalization, University of Twente,
Enschede, the Netherlands

e-mail: h.j.e.bijlsma@utwente.nl

R. van der Lans

Curium LUMC, Leiden, The Netherlands

e-mail: r.m.van_der_lans@curium.nl

T. Mainhard

Department of Education, University of Utrecht, Utrecht, The Netherlands

e-mail: m.t.mainhard@uu.nl

P. den Brok

Department of Educational and Learning Sciences, University of
Wageningen, Wageningen, The Netherlands

e-mail: perry.denbrok@wur.nl

© The Author(s) 2021

W. Rollett et al. (eds.), *Student Feedback on Teaching in Schools*,
https://doi.org/10.1007/978-3-030-75150-0_2

1 Introduction

Student perceptions of teachers and their behaviours have become an important way to capture what happens in class. Questionnaires that map student perceptions of teaching quality are used, for example, to measure the effectiveness of educational interventions (Burniske & Meibaum, 2012; Kyriakides, 2005). In schools, student perceptions are collected by teachers to obtain feedback for improvement and professional development activities (Bijlsma et al., 2019).

Using student perceptions of teaching quality is a complex process. Typically, perceptions are collected using a standardized questionnaire instrument. When a student selects a response category of an item like “my teacher explains everything clearly to me”, however, many processes may affect the student’s answer. For example, a student may deliberately give a higher rating for the item than their real estimation of their teacher’s skill at explanation because (s)he wants to present him/herself in a socially desirable way, or the student’s perception may be biased by stereotypical impressions. Alternatively, the student might be honest and their perception unbiased, but a misinterpretation of the item content, for example, a different interpretation of what clarity means in this context, may still affect the item response (Maulana & Helms-Lorenz, 2016).

Moreover, items can be formulated according to the level of behaviour at which they are directed (to an individual student or the whole class), and in terms of the level of perception (personal, class). In Chap. 7 by Göllner et al. in this volume, it is referred to as differences in the referent and in the addressee of items. For example, the aforementioned item can be worded as: “This teacher explains things clearly to us/the class” (class perception, behaviour to class), “This teacher explains things clearly to me” (class perception, behaviour to individual), “I find this teacher to explain things clearly” (personal perception, behaviour to class) and “I find this teacher to explain things clearly to me” (personal perception, behaviour to individual). While this may seem trivial, it has consequences for the expected sources of variation in perceptions: items asking about class perceptions or behaviours directed at the whole class are more likely to evoke variation in shared sources of perceptions, while items asking about behaviours directed at individuals or personal perceptions are more likely to evoke variation in idiosyncratic sources of perceptions.

The question of what we actually measure, therefore, has no uniform answer. By completing standardized questionnaires, students give responses to many items and psychometric models are applied to combine the item ratings into an overall student perception of teachers’ teaching (students’ responses are then combined to a numerical value or score). This overall score—not the item ratings—is usually fed back to teachers or is used for research purposes. This approach of combining and integrating ratings into one overall perception score suggests that students cognitively process observations of teaching behaviours similarly and in such a general and integrated way. From this perspective, the psychometric models that connect and integrate the item ratings attempt to reconstruct students’ mental representations of the teachers’ teaching.

This chapter discusses student perceptions in terms of three psychometric perspectives that dominate contemporary research on teaching quality, namely, Classical Test Theory (CTT), Item Response Theory (IRT) and Generalizability Theory (GT). CTT (part 2) is based on the assumption that there is one true score and a variance score (error). The true score is then an average of all students' ratings on certain items that form a dimension or factor. In IRT (part 3), more emphasis is put on how many items relate to each other and what dimensions can be distinguished in the instrument used to collect student perceptions of teaching quality. The potential of GT (part 4) lies in the fact that it tries to disentangle the variability in student ratings beyond a "true score" and error, bringing in aspects such as personal characteristics and dyadic relationships between people. The chapter discusses these psychometric perspectives separately, but there are also integrated approaches that can enable researchers to estimate combinations of the models (Chalmers, 2012; Robitzsch et al., 2020). The connection between the CTT, IRT and GT with latent variable models becomes evident when it is realized that all specify a relationship between the teachers' latent ability level and the responses of students that were stimulated (or elicited) by the items (e.g., Chalmers, 2012; de Boeck et al., 2011; Rizopoulos, 2006; Robitzsch et al., 2020).

The main message of the chapter is that the choice of a psychometric model is not merely a technical matter, but also has implications for how the nature of perceptions is conceptualized. Finally, we acknowledge that the construct of teaching quality is highly contested and consensus about its conceptualization or definition is minimal (Cohen & Goldhaber, 2016). We do not present a definition of teaching quality in this chapter. By leaving the definition completely open, we intend to maximize our flexibility to discuss various possibilities offered by the three psychometric theories. After presenting and linking each psychometric theory, we will discuss their strengths and weaknesses in the context of student perceptions of teaching quality.

2 Classical Test Theory

2.1 *The CTT Model*

According to Classical Test Theory (CTT), student perceptions of teaching quality reflect the teachers' actual teaching quality plus random error variance (e.g., Brennan, 2001; Lord & Novick, 1968; Sijtsma, 2016; Spearman, 1905). The teachers' actual teaching quality is caught by the so-called "true score", which is statistically defined by the mean score over all item responses about that teacher. The error variance consists of all random deviations from the teacher's mean score (Novick, 1966). Furthermore, the CTT model states that all items are equally associated with the broader perceptual representation of the teachers' teaching (i.e., items are supposed to have similar factor loadings).

Table 1 Possible example of feedback form results for one teacher teaching a class of 25 students

| Item My Teacher... | N_{class} | Class mean | Class SD |
|--|--------------------|------------|----------|
| ... | | | |
| ... makes sure that others treat me with respect. | 25 | 3.28 | 0.52 |
| ... makes clear what I need to learn for a test. | 25 | 3.14 | 0.78 |
| ... explains everything clearly to me. | 25 | 2.72 | 0.94 |
| ... uses clear examples. | 25 | 2.82 | 0.93 |
| ... encourages me to cooperate with my classmates. | 25 | 2.08 | 0.80 |
| ... | | | |
| Total | 25 | 2.81 | 0.79 |

Marsh (2007) noted that overall questionnaire outcomes may be uninformative about *specific* teaching behaviours, and therefore recommends structuring questionnaires according to different factors. Factors cluster items that seem to have something in common based on the inter-item correlations. For example, the items, “My teacher explains everything clearly to me” and “My teacher uses clear examples” (see Table 1), are connected to the same factor, which clusters items related to the clarity and structuredness of explanations (Maulana & Helms-Lorenz, 2016). Reporting the class mean for items related to the clarity and structuredness of explanations is considered more informative than just an overall mean for all items.

In educational contexts, the CTT model is usually extended by including multiple nested levels of random error; for example, students are nested within teachers. The key idea of CTT, however, remains, in that only the mean of a factor is informative and variation around the mean is uninformative noise.

Paramount to the logic behind CTT is that item ratings related to the same teacher should show minimal variability *and* that item ratings related to different teachers should show large(r) variation. Hence, item ratings assigned by one student to the same teacher are expected to vary minimally. The mean student questionnaire scores from students within the same class are also expected to show minimal variability. These expectations are routinely examined by estimates of internal consistency (Cronbach, 1951) and intra-class correlations (ICCs; Lüdtke et al., 2009). Internal consistency is sensitive for items showing large variation in ratings compared to the other items’ ratings. The ICC provides an estimate of the variance in mean questionnaire scores from students in different classes as proportionate to the variance of all ratings.

2.2 *An Example of CTT in Practice*

Suppose that 25 students in a class respond to the item “My teacher explains everything clearly to me” by choosing one of the four answer options: 1 = “never”, 2 = “seldom”, 3 = “occasionally” and 4 = “often” (Table 1). If CTT is applied strictly, then the mean class perception (2.72) is the only reliable and, thus, the only informative parameter for the teacher to consider, and individual deviations are *random* noise. This logic can easily be generalized to a broader set of items. For example, the mean of the student questionnaire ratings can be computed and CTT can be applied to these mean scores, which may then be argued to be the most reliable estimate of the teacher’s actual teaching quality. In this example, according to CCT, 2.81 reflects the teacher’s teaching quality based on these five items.

2.3 *Advantages and Limitations of the CTT Approach*

The CTT approach, and Marsh’s (1987, 2007) work in particular, are well-known and studied in the educational sciences. Estimates of internal consistency and ICCs have proven to be stable across different questionnaires (cf. Marsh, 2007; van der Lans & Maulana, 2018). These statistics are also intuitively understandable for many practitioners and the application of CTT requires only a modest level of mathematical and statistical skill, which is not unimportant.

However, the use of CTT reflects high trust in the students as being honest and accurate perceivers. To illustrate this, suppose that students deliberately manipulate their ratings upwards because they like the teacher; then clearly such *systematic* bias or manipulation remains undetected by measures such as internal consistency and ICC, which quantify *random* error variance only (den Brok & Smart, 2007). In general, CTT provides very limited means to empirically investigate systematic biases in perceptions. Second, diagnosing poor item quality by the comparatively large variance in ratings, as is done by internal consistency measures, is only valid if one believes that ratings of all items must be biased by the same amount of (random) error. Suppose again that students deliberately manipulate their ratings upwards because they like the teacher; then their manipulation might well be expressed most in items referring to specific teacher traits that are likable (such as “humour”, or “showing respect”). More in general, CTT fails to make (differentiated) predictions about the response process; for example, when students check a response category, it remains unsolved what latent cognitive representation of the teacher’s teaching students had in mind.

3 Item Response Theory

3.1 Item Response Theory (IRT) Model(s)

According to IRT, student perceptions of teaching are ordered on a latent continuum (Bond & Fox, 2007; Embretson & Reise, 2013). With IRT, researchers estimate the teacher’s position on this latent continuum and this position is then used to predict the most likely teacher behaviour that students will have experienced from this teacher. There are two levels at which IRT can be used to make predictions about what teacher behaviours students likely will have experienced: (1) the level of the item and (2) the level of the construct (Bond & Fox, 2007; Embretson & Reise, 2013). At the level of the item, IRT uses the response categories to make predictions about whether students experienced that particular behaviour seldom, occasionally or often. At the level of the construct, IRT makes predictions about how items jointly represent the teachers’ teaching.

We will explain this by using one of the five items from Table 1 (“My teacher explains everything clearly to me” [explains clearly]). In Fig. 1, the y-axis indicates the probability of checking the higher response category out of two competing response categories and the x-axis indicates the level of teaching quality (θ). Teachers with a level of teaching quality located at the position of the arrow have a high probability of receiving a response “ \geq seldom” on explains everything clearly to me, but a low probability of receiving a response “occasionally”. The probability that students check the higher response category increases only when the teacher—according to the responding student—has achieved the conditions set by the higher response category for the item.

The item response process can be used to predict the most likely frequency with which the behaviour is observed (or the most likely impact, if the item labels are insufficient, sufficient, excellent). This item response process is part of a wider process here referred to as the construct response process. The construct response process predicts how students weigh and position items relative to other items. In IRT, one

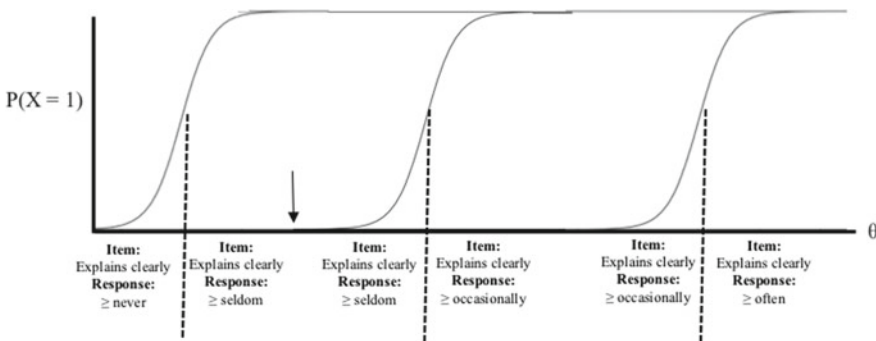


Fig. 1 Visualization of the item and construct response process

Fig. 2 The Guttman scale/simplex construct response (pattern obtained from: Mokken et al., 2001)

| Guttman scale or simplex | | | | | |
|--------------------------|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 | 5 |
| Student A | ✓ | | | | |
| Student B | ✓ | ✓ | | | |
| Student C | ✓ | ✓ | ✓ | | |
| Student D | ✓ | ✓ | ✓ | ✓ | |
| Student E | ✓ | ✓ | ✓ | ✓ | ✓ |
| Student F | ✓ | ✓ | ✓ | ✓ | ✓ |

well-known construct response process is the Guttman scale or simplex¹ (Guttman, 1954; Jöreskog, 1978). In the simplex, item positions depend on their “difficulty”. Some items are much more likely to receive the rating “never” (called “difficult” items), while other items are much more likely to be rated as “often” (the “easy” items). Figure 2 visualizes this pattern using five items. In Fig. 2, the checkmarks indicate a high probability that students perceive the teacher to perform the behaviour described by the item often. Hence, student D is predicted to perceive the teacher as performing the first four behaviours often, but not the fifth. Item 1 would be a “difficult” item, and 5 would be an “easy” item.

To order items, IRT models include a location parameter (sometimes referred to as item difficulty). The location parameter predicts when the item response process changes within the wider construct response process. For example, the response process for item four is predicted to change if the first three items have received high ratings. Other item parameters that can be estimated by IRT models are the discrimination parameter (to predict and correct for systematic deviations from the predicted item response process), and a guessing parameter (to predict and correct for randomness in the item response process). In what follows, we will present an example of research applying IRT to student perceptions to illustrate the above.

3.2 IRT in Research on Student Perceptions

Van de Grift and Kyriakides started independently implementing IRT in the context of teaching quality with student perception data (for details, see Antoniou & Kyriakides, 2013; Kyriakides et al., 2018; Maulana et al., 2015; van de Grift et al., 2011, 2014; van der Lans et al., 2015). Their models hypothesize that teaching effectiveness develops along a latent continuum in which learning to teach starts with learning less complex teaching behaviours (e.g., ensuring a safe classroom climate) and ends

¹ There are two other main classes of construct response processes, namely the Coombs/unfolding and the circumplex (Browne, 1992; de Leeuw & Mair, 2011; Mokken et al., 2001). It goes beyond the scope of this chapter to define and describe these as well. Hence, we focus here on the Guttman scale/simplex construct response.

with learning more complex teaching behaviours (e.g., having students cooperate with classmates). Hence, students' ratings given on questionnaires that list various teaching behaviours should indicate that they perceive some more complex teaching behaviours to be performed successfully less frequently, whereas they perceive other less complex teaching behaviours to be performed successfully more often (by more teachers). These researchers have applied Rasch-family models²—a specific type of IRT model—to test sequences of item complexity and to locate teachers on the latent continuum. After they have located the teacher, they provide the teacher with feedback by indicating the next steps for improvement (i.e., the items located just beyond the teacher's position). In other recent research, IRT has been used to examine issues of validity of student perception data (e.g., Bijlsma et al., submitted; van der Scheer et al., 2018).

3.3 Advantages and Limitations of IRT Models

The comprehensive framework of IRT provides various possibilities for testing hypotheses concerning students' response processes at the level of the item and at the level of the construct. Thereby, IRT is promising as a way to develop and test theories that predict how different formulations of survey items and/or formulations of response categories translate into distinct item response and construct response processes. Substantive theories can also be translated into item and construct response processes, as in the example described in the previous section.

However, the disadvantage of IRT is that it basically assumes that the item response process is unbiased. Take the research we discussed by van de Grift et al. (2014). They predicted that student ratings will follow sequences predicted by theory on teacher development, but this prediction assumes that student ratings are a direct (unbiased) numerical representation of the teacher's actual behaviour. IRT can include a discrimination parameter to correct for systematic biases, but this discrimination parameter corrects the item response process for all biases and generally is uninformative about the potential sources of bias. Various biases will impact the students' item responses, such as social desirability and stereotypical views (Kenny, 1994). As we will detail next, generalizability theory provides a framework for examining such influences on item ratings.

² Rasch-family models are applied to test the theoretical models, because Rasch model fit tests were developed to empirically examine hierarchical orderings in item ratings (Bond & Fox, 2007). Hence, if student perceptions are unbiased, then their responses could be used to locate the teacher on this latent novice–expert continuum.

4 Generalizability Theory

Generalizability Theory (GT) extends Classical Test Theory (CTT) by introducing the possibility of including systematic variance components (or facets) other than error and a teacher's "true score" (Brennan, 2001). The basic idea is that what is called error in CTT can be further sub-divided into systematic facets or sources of variability (Malloy, 2018) that potentially affect student perceptions of teaching quality. When such variance components are considered nuisance parameters, GT conceptually coincides with CTT, as it is viewed by Marsh (2007), for example. Traditionally, in the educational context, GT has been used to determine the number of tasks or raters that yield reliable test results (Shavelson & Webb, 2005). As such, the amount of error that tasks introduce or the degree of consensus between raters is typically GT's main focus. Yet, the strength of GT is that it can also be used to embrace and study 'error' in an attempt to learn more about how these additional sources of variability impact perceptions of social phenomena such as teaching.

4.1 A Practical Example Using GT and Student Ratings

One of the best-known models in social science that applies GT to social perceptions and interactions is Kenny's Social Relations Model (SRM, 1994). The basic assumption of the SRM is that any rating of a social perception has, besides error, three potential sources: an actor or rater effect (i.e., due to the student who responds to an item), a partner or target effect (i.e., due to the teacher who is rated) and a relationship effect (variability introduced due to the specific combination of this student rating that specific teacher). The partner or target effect resembles what is taken to be the teacher's true score or true ability in CTT. The variance in partner effects captures the degree of consensus between students on a certain aspect of teaching quality. Stable response tendencies within students are captured in the actor effect. For example, some students are quick learners and may therefore readily indicate that they understand teacher explanations, irrespective of a specific teacher's quality. There can also be systematic variance in ratings due to the relationship between, or the specific pairing of, students and teachers. Thus, on top of a student's stable tendency to think that teachers can explain things well (rater effect) and the teacher's general ability to explain things (target effect), student A may have experienced instances where teacher B has explained content exceptionally well. This shared interaction history may affect student A's ratings over and above the rater and target effects (Mainhard et al., 2018).

GT and SRM can be applied at the item level, though they are more commonly applied at the construct level (Kenny, 1994, 1996; Kenny et al., 2006). Let us consider an example at the item level. Suppose that students complete the item "my teacher explains everything clearly to me"; then at the item level, the SRM is informative: about the target effect, namely, do students agree that some teachers explain things

well while others are not? about the actor effect, namely, do some students tend to experience all teachers' explanations to be clear while other students tend to perceive all teachers' explanations as hard to understand? and about the dyadic effects, namely, do some students experience the teacher's explanation to be clear over and above their personal actor effect and that teacher's target effect? Note that the actor and relationship variances would be considered as error in CTT. The variability found in these sources can then be explained with predictors, as in regression analysis. For example, students' actor effects may be explained by their general academic ability and teachers' target effects by years of experience. Relationship effects may occur, for example, because some teachers think that certain students require a certain kind of explanation to understand the subject matter.

4.2 Advantages and Limitations of Generalizability Theory

An advantage of dealing with student ratings of teaching quality according to the GT approach is that it is a relatively simple extension of the better-known CTT. Those acquainted with multilevel analyses will find GT quite straightforward (Kenny et al., 2006). Conceptually, GT is more informative about potential variables that impact students' item responses. When items barely show stable variance between students, the responses are only minimally affected by students' personal characteristics and answer tendencies.

However, compared to IRT, GT puts little emphasis on how item ratings can be organized into a broader representation of teaching. Like CTT, GT is applied to sets of items that have a similar association with the latent construct. Further, the GT approach requires complex data sets. It cannot be applied with datasets that pair one class with a teacher. Instead, students need to complete a questionnaire for several teachers, and teachers need to be rated by several classes (see Mainhard et al., 2018 for an example).

5 Discussion

In this chapter, three dominant psychometric theories were discussed within the domain of research on the validity and reliability of student perceptions of teaching quality: Classical Test Theory, Item Response Theory and Generalizability Theory. While each of these models has its specific advantages and disadvantages, together they shed more complete light on what constitutes and determines students' perceptions of teaching quality, disentangling true scores from error, and distinguishing between more systematic and more random sources of variation in perceptions. Together, they present a nuanced and complex picture of what makes a (student) perception, and also how it can be used in research.

The main message of the chapter is that the choice of a psychometric model is not merely a technical matter, but also has implications for how the nature of perceptions is conceptualized. For example, statistical techniques or software are tools that can be of help, but they depend on the specific theory about what teaching (quality) is and what dimensions or constructs and their interrelationships underlie such behaviour. Regardless of the three theories described in this chapter, many instruments measuring student perceptions are based on effectiveness research. It mainly includes variables that have been found to be associated with student outcomes in correlational research, rather than specifying a structure in and between different dimensions of teaching and their likelihood of (co-)occurring (Skourdoumbis & Gale, 2013; Wrigley, 2004). For this purpose, CTT can be applied. Furthermore, many instruments are based on the frequency of occurrence of behaviours, assuming an order or singular dimension in these occurrences that is based on difficulty, routine or other phenomena, which is linked to IRT (Maulana & Helms-Lorenz, 2016; den Brok et al., 2018). However, others have argued that teaching quality is multidimensional in nature, with behaviours being interpretable from various perspectives and adding value to different outcomes at the same time (Doyle, 1986; den Brok, 2001; den Brok et al., 2004; Shuell, 1996). GT can be applied here.

One may argue that basing a theory about teaching quality on the actual presence of behaviour or association with existing student outcomes is conservative, and does not allow exploration of new teaching methods, new organisational forms of education or alternative learning outcomes. However, assumptions behind the occurrence of behaviours may differ depending on the type of perspective taken on teaching, as may their theoretical underpinnings. For example, many interactional theories assume two independent dimensions behind teaching, that order components of behaviour in circumplex structures with specific patterns and interrelations between behaviours (or items) (Fabrigar et al., 1997; Gurtman & Pincus, 2000; Wubbels et al., 2006). The more specified theories are, the easier they can be tested statistically, as many programmes assume or ask for specific relations to be tested when studying perceptions; consider, for example, structural equation modelling, confirmatory factor analyses, IRT analysis or latent variable analysis (den Brok et al., 2018).

6 Putting it all Together

With this chapter, we hope to have provided more insight into the interesting, yet complicated, world of student perceptions of teaching quality. In conclusion, we have a few take-away messages for researchers interested in using student perceptions of teaching.

First, as aforementioned, it is important to be specific about the underlying assumptions one has about the nature of the student perceptions one is interested in. These assumptions should be grounded in prior research conducted on perceptions of the particular teaching behaviours one is interested in. For example, are the

perceptions expected to vary considerably between teachers, classes or schools? Are the perceptions likely to evoke certain psychological processes, such as social desirability or stereotypical responses? Are the behaviours expected to be familiar or unfamiliar to perceivers? Depending on what is known or deemed relevant, researchers can choose between one or several of the theories mentioned in this chapter.

Second, it is important to be specific about the wording of the items capturing the perceptions, as wording may lead to differences in response patterns, and thereby differences in sources of variance that may occur, related to either perceiver, object or the relation between them. Typically, researchers are not that conscious about the choices and assumptions they make about perceptions and the wording they use.

Third, it is important to conceptualize and make explicit the different dimensions or constructs one is interested in and the expected relationships between them, preferably based on theory (and empirical results). As this chapter has shown, constructs may relate to each other in terms of difficulty or chance of occurrence (as with simplex structures), but also in terms of relatedness or independence (as with circumplex structures).

When researchers take all of these reflections into account, interesting insights may be obtained by collecting student perceptions of teaching, and by comparing these with, for example, the perceptions of others, such as teachers themselves. The present chapter provides an overview of techniques and three major theories that may be used to analyse and conceptualize such perceptions.

References

- Antoniou, P., & Kyriakides, L. (2013). A dynamic integrated approach to teacher professional development: Impact and sustainability of the effects on improving teacher behaviour and student outcomes. *Teaching and Teacher Education*, 29(1), 1–12. <https://doi.org/10.1016/j.tate.2012.08.001>.
- Bijlsma, H. J. E., Glas, C. A. W., & Visscher, A. J. (submitted). The reliability and construct validity of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*.
- Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teaching quality? *Technology, Pedagogy & Education*. <https://doi.org/10.1080/1475939X.2019.1572534>.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.
- Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy*. Springer-Verlag. <https://doi.org/10.1007/978-1-4757-3456-0>.
- Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika*, 57(4), 469–497.
- Burniske, J., & Meibaum, D. (2012). *The use of student perception data as a measure of teaching effectiveness*. Texas Comprehensive Center.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6). <https://doi.org/10.18637/jss.v048.i06>.

- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45(6), 378–387. <https://doi.org/10.3102/0013189X16659442>.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. <https://doi.org/10.1007/BF02310555>.
- de Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39(11), 1–28.
- de Leeuw, J., & Mair, P. (2011). *Multidimensional scaling using majorization*. SMACOF in R.
- den Brok, P. (2001). *Teaching and student outcomes: A study on teachers' thoughts and actions from an interpersonal and a learning activities perspective*. W.C.C.
- den Brok, P., & Smart, J. C. (Eds.). (2007). *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Springer. <https://doi.org/10.1007/1-4020-5742-3>.
- den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behavior and student outcomes. *School Effectiveness & School Improvement*, 15, 407–442.
- den Brok, P., Wubbels, T., & Mainhard, T. (2018). Developments in quantitative methods and analyses to study learning environments. In D. Zandvliet & B. Fraser (Eds.), *Thirty years of learning environments: Looking back and looking forward: Advances in learning environments research, volume 11* (pp. 41–58). Brill/Sense.
- Doyle, W. (1986). Classroom organization and management. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 392–431). Macmillan.
- Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
- Fabrigar, L. R., Visser, P. S., & Browne, M. W. (1997). Conceptual and methodological issues in testing the circumplex structure of data in personality and social psychology. *Personality and Social Psychology Review*, 1, 184–203.
- Gurtman, M. B., & Pincus, A. L. (2000). Interpersonal adjective scales: Confirmation of circumplex structure from multiple perspectives. *Personality and Social Psychology Bulletin*, 26, 374–384. <https://doi.org/10.1177/0146167200265009>.
- Guttman, L. L. (1954). A new approach to factor analysis: the radex. In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences*. The Free Press.
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4), 443–477.
- Kenny, D. A. (1994). *Interpersonal perception: A social relation analysis*. Guilford.
- Kenny, D. A. (1996). Models of non-independence in Dyadic research. *Journal of Social and Personal Relationships*, 13(2), 279–294. <https://doi.org/10.1177/0265407596132007>.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. Guilford Press.
- Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction*, 40(2), 44–66.
- Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM Mathematics Education*, 50(3), 381–393.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley Publishing Company.
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, 34(2), 120–131. <https://doi.org/10.1016/j.cedpsych.2008.12.001>.
- Mainhard, M. T., Oudman, S., Hornstra, L., Bosker, R. J., & Goetz, T. (2018). Student emotions in class: The relative importance of teachers and their interpersonal relations with students. *Learning and Instruction*, 53, 109–119. <https://doi.org/10.1016/j.learninstruc.2017.07.011>.
- Malloy, T. E. (2018). *Social relations modeling of behavior in dyads and groups*. Academic Press.

- Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Springer. https://doi.org/10.1007/1-4020-5742-3_9.
- Marsh, H. W. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11(3), 253–388. [https://doi.org/10.1016/0883-0355\(87\)90001-2](https://doi.org/10.1016/0883-0355(87)90001-2).
- Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: Construct representation and predictive quality. *Learning Environments Research*, 19(3), 335–357. <https://doi.org/10.1007/s10984-016-9215-8>.
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. J. C. M. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, 26(2), 169–194. <https://doi.org/10.1080/09243453.2014.939198>.
- Mokken, R. J., van Schuur, W. H., & Leeferink, A. J. (2001). The circles of our minds: A nonparametric IRT model for the circumplex. In *Essays on item response theory* (pp. 339–356). Springer. https://doi.org/10.1007/978-1-4613-0169-1_18.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3(1), 1–18.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25. <https://doi.org/10.18637/jss.v017.i05>.
- Robitzsch, A., Kiefer, T., Wu, M., Robitzsch, M. A., Adams, W., Rupp, L., et al. (2020). *Package 'TAM': Test analysis modules*. Version: 3: 4, 26.
- Shavelson, R. J., & Webb, N. M. (2005). *Generalizability theory*. <https://web.stanford.edu/dept/.../GTheoryAERA.pdf>.
- Shuell, T. J. (1996). Teaching and learning in a classroom context. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 726–764). Macmillan.
- Sijtsma, K. (2016). Classical test theory. In S. J. Henly (Ed.), *Routledge international handbook of advanced quantitative methods in nursing research* (pp. 29–43). Routledge.
- Skourdoumbis, A., & Gale, T. (2013). Classroom effectiveness research: A conceptual critique. *British Educational Research Journal*, 39(5), 892–906. <https://doi.org/10.1002/berj.3008>.
- Spearman, C. (1905). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
- van de Grift, W. J. C. M., van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs [Development of pedagogical didactic skills of primary school teachers]. *Pedagogische Studiën*, 88(6), 416–432.
- van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, 43, 150–159. <https://doi.org/10.1016/j.stueduc.2014.09.003>.
- van der Lans, R. M., & Maulana, R. (2018). The use of secondary school student ratings of their teacher's skillfulness for low-stake assessment and high-stake evaluation. *Studies in Educational Evaluation*, 58, 112–121. <https://doi.org/10.1016/j.stueduc.2018.06.003>.
- van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice*, 34(3), 18–27. <https://doi.org/10.1111/emip.12078>.
- van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. (2018). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*. <https://doi.org/10.1080/09243453.2018.1539015>.
- Wrigley, T. (2004). School effectiveness: The problem of reductionism. *British Educational Research Journal*, 30(2), 227–244. <https://doi.org/10.1080/0141192042000195272>.
- Wubbels, T., Brekelmans, M., Den Brok, P., & Tartwijk, J. (2006). An interpersonal perspective on classroom management in secondary classrooms in the Netherlands. In C. Everson & C. Weinstein (Eds.), *Handbook of classroom management: Research, practice, and contemporary issues* (pp. 1161–1191). Lawrence Erlbaum Associates.

Hannah Bijlsma is a researcher (Ph.D.) at the section of Teacher Professionalization at the University of Twente (the Netherlands) and a primary school teacher (grade 1). Her research focuses on measuring teaching quality and on the use of student perceptions of teaching quality in school contexts. In 2016 she founded a professional association for academic primary school teachers, of which she has been chairman for about five years. She is now a board member of the International Congress for School Effectiveness and Improvement (ICSEI) and a board member of the EARLI SIG on School Effectiveness and Improvement.

Rikkert van der Lans is a postdoctoral researcher currently working at the Department of Child and Adolescent Psychiatry, Curium-Leiden (the Netherlands). Previously he worked as a post-doctoral researcher at the department of teacher education of the University of Groningen, as a lecturer in methods and statistics at the department of educational sciences (GION) of University of Groningen, and as lecturer in the field of psychometrics at the department of methods and statistics of the University of Tilburg (the Netherlands). His research focuses on the evaluation of professional development and the psychometric assessment of quality tests.

Tim Mainhard is Associate Professor at the Department of Education at Utrecht University (the Netherlands). His research focuses on social dynamics in educational settings—specifically teacher-student interactions and relationships—and their impact on student and teacher outcomes, such as emotions, motivation, and academic achievement. Tim teaches classroom management courses at the Utrecht Graduate School for Teaching in both the primary and secondary teacher education programs. Tim has been chair of the Classroom Management Special Interest Group of the American Educational Research Association and is an associate editor for the journal *Learning and Instruction*.

Perry den Brok is full Professor and chair of the Education and Learning Sciences group at Wageningen University and Research (the Netherlands), and chair of the 4TU Centre for Engineering Education. His research focuses on educational innovation in higher education, teacher learning and professional development, teacher-student interpersonal relationships, and educational learning environments—both in-class as well as out-of-school learning environments. He was European editor of the *Learning Environments Research* journal (Springer) for well over 10 years. He has published several review articles and book chapters on teacher effectiveness and teacher collaborative learning.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 3

Student Perceptions of Teaching Quality: Dimensionality and Halo Effects



Sebastian Röhl and Wolfram Rollett

Abstract This chapter deals with the factorial structure of survey instruments for student perception of teaching quality. Often, high intercorrelations occur between different theoretically postulated teaching quality dimensions; other analyses point to a single unified factor in student perceptions of teaching quality, seemingly reflecting a “general impression” instead of a differentiated judgment. At the same time, findings from research on social judgment processes and from classroom research indicate that the teachers’ communion (warmth or cooperation) as well as students’ general subject interest can be important biasing factors in the sense of halo effects in student ratings of teaching quality. After presenting an overview of studies on the dimensionality of various survey instruments, we discuss whether aggregated data is impacted by an overall “general impression”. We confirmed this hypothesis using a sample of $N = 1056$ students from 50 secondary school classes. Moreover, this general impression could be explained at student and class level to a large extent by students’ perception of the teacher’s communion. Student general subject interest showed a medium effect but only at the individual level. These findings indicate that student perceptions of teaching quality dimensions are indeed influenced by a general impression which can be explained largely by teacher’s communion.

Keywords Halo effect · Social judgment · Communion · Student feedback · Teaching quality

S. Röhl (✉) · W. Rollett
Institute for Educational Sciences, University of Education,
Freiburg, Germany
e-mail: sebastian.roehl@ph-freiburg.de

W. Rollett
e-mail: wolfram.rollett@ph-freiburg.de

1 Introduction

An important precondition for using measurements of student perceptions of teaching is the validity of the collected data, both for use as informative feedback to teachers and for the collection of teaching quality within research studies. At the same time, measurements of student perceptions must be consistent with the theoretical assumptions made in the survey instrument with regard to professional competence or quality characteristics. While some other chapters of this volume deal with the question of perspective-specific characteristics of different feedback sources (Chap. 7 by Göllner et al. and Chap. 5 by van der Lans in this volume) or predictive validity (Chap. 6 by Schweig and Martinez in this volume), this chapter focuses on the extent to which students can actually distinguish between different theoretically postulated dimensions in their assessment of individual aspects of teaching. Subsequently, we examine whether the limited ability to differentiate can be explained by overlaying affective attitudes toward teaching or the teacher in the sense of a halo bias.

1.1 Dimensionality of Student Ratings on Teaching Quality

Usually, questionnaires are used to collect student perceptions of teaching, in which a certain number of quality dimensions are differentiated and surveyed separately. However, most of the used instruments show high correlations between the theoretically distinguished quality dimensions. This is also the case when the theoretically postulated structure is confirmed by a confirmatory factor analysis. For example, Krammer et al. (2019) reported intercorrelations ranging from $r = .81$ – $.95$ between the three dimensions “instructional quality”, “teacher-student relationship”, and “performance monitoring” at student and class level. Analyses of the “students’ perceptions of instructional quality” (SPIQ) from Wisniewski et al. (2020) showed correlations from $r = .63$ – $.93$ between the seven dimensions of the instrument. For primary schools, van der Scheer et al. (2019) reported correlations between $r = .74$ and $r = .42$ using an IRT model. One exception seems to be the survey instrument of Fauth et al. (2014), which only shows correlations between the dimensions of $r = .47$, $.50$, and $.70$ at the student and $r = .23$, $.31$, and $.67$ at the class level. However, a closer look reveals fundamental differences between item formulations of different quality dimensions. While the items of two of the dimensions start with “In our science class...”, the third one uses “Our science teacher...”.

Unfortunately, quite a number of the validation studies of student questionnaires on teaching quality did not report the correlations between the included scales (e.g. Bell & Aldridge, 2014; Tripod Education Partners, 2014), or they only tested the unidimensionality of single postulated scales (e.g. van Petegem et al., 2008).

At the same time, there are studies in which the theoretically postulated dimensions could be confirmed factor-analytically, but where they were highly charged

with high standardized loadings to a latent second-order factor (e.g. Nelson et al., 2014 reports $\lambda = .70-1.02$).

A survey instrument which has been intensively analyzed in recent years is the “Tripod” questionnaire (Tripod Education Partners, 2014). Based on explorative factor analyses at the class level, the developers of the instrument postulate seven dimensions. However, in-depth analyses, which simultaneously take into account the nested multi-level structure with student and class level, consistently point to the unidimensionality of this questionnaire (Kuhfeld, 2017; Schweig, 2014; Wallace et al., 2016). A possible further dimension suggested by analyses is only weakly separated and is characterized by items with a certain type of item formulation (Kuhfeld, 2017). When examining other questionnaires, studies found unidimensionality for those with 16 items (Bijlsma et al., 2019) and 64 items (Maulana et al., 2015).

Overall, the question arises how to interpret the high statistical interrelations between theoretically well-distinguished dimensions of instructional quality in student surveys. A possible explanation, which we would like to examine in this chapter, is the impact of an affective overall attitude of students toward the teaching behavior of the evaluated teacher, resulting in biasing effects during the response process to individual items.

In research, different terms are used to describe the phenomenon whereby an overall attitude or impression influences and interferes with the assessment of individual teaching characteristics. For example, Clausen (2002) speaks of the effect of an “affective overall impression”, while other authors use the terms “halo effect” (e.g. Haladyna & Hess, 1994; Wagner, 2008) or “general impression halo” (Lance et al., 1994).

1.2 Possible Explanations for Halo Effects in Student Ratings

One promising path to a better insight into the phenomenon of high intercorrelations is to analyze the subjects’ processing of items. Tourangeau et al. (2000) divide the survey response process into four main cognitive components or steps. In the first step, *comprehension*, the respondent needs to understand the item and to identify its focus. In the subsequent *retrieval* step, the respondent has to generate a retrieval strategy and cues, retrieve specific and generic memories, and fill in missing details. Next, a *judgment* component on the retrieved memories regarding the completeness and relevance of different memories takes place, which ends with an estimation for the subject of the item. In the last step, the person gives a *response* in the requested way, e.g. marking the box with the answering option fitting best.

In case that an overall affective attitude of satisfaction is present throughout the survey answering process, this influences the retrieval and judgment of the information related to the items. Therefore, the rating on a particular aspect is a combination of the overall satisfaction of the person and the actual judgment of the particular aspect (Borg, 2003).

Applied to the situation of students, this would mean that ratings on particular aspects of teaching quality consist of a non-differentiating overall satisfaction with the teacher or class, and a rating component which concerns the particular aspect.

According to the findings from research on social judgments, overall judgments on other persons are based on two fundamental dimensions of perception (Abele et al., 2008; Bakan, 1966). The first dimension, often called “agency”, describes perception in terms of dominance, competence, or individualism. The second dimension, “communion”, refers to perception concerning warmth, cooperation, social and community orientation. In the overall judgment of other people, the perceived communion plays a dominant role and is responsible for much larger parts of variance in character judgments (Abele & Bruckmüller, 2011).

The discussion about the overall impression—which dominates the students’ judgments about teaching and teachers—points in a similar direction. A number of factors were discussed and examined which could well be subsumed under “communion”. Wallace et al. (2016, p. 1859), for example, interpreted the overall factor as a judgment of such forms of teacher interaction which “makes them feel safe, respected, and competent”. Kuhfeld (2017) explained the overall factor as an effect of students’ perception of teachers’ emotional support. Furthermore, findings also indicate that a higher teacher–student communion leads to higher desired learning behaviors of the students (Wubbels et al., 2015), and there is evidence of positive effects on learning achievement for learner-centered teaching approaches (Cornelius-White, 2007).

On the other hand, there are indications that an affective attitude toward the subject being taught could also cause biased ratings, and so the detection of an overall factor. In line with this assumption, findings from research on student ratings on teaching quality point to an influence of students’ general interest in the school subject on the perception of teaching (Ditton, 2002; Eder & Bergmann, 2004; Mayr, 2006; Rahn et al., 2019). Students’ general interest in the school subject is known to show a relatively stable pattern from secondary school onwards (Schurtz & Artelt, 2014), although current teaching characteristics may cause minor changes (Ferdinand, 2014; Lazarides et al., 2015). Findings from Rahn et al. (2019) point out that biasing effects of students’ general interest in the school subject vary considerably between different subjects, particularly with regards to the distinction between compulsory and optional courses. But as Ferdinand (2014) showed, these distortions seem to be largely neutralized in the aggregation of the student ratings of a class. Research findings from higher education also support biasing effects of the perception of the teacher as well as of the subject. For example, in the study by Greimel-Fuhrmann (2014) the interest in the subject and the teachers’ level of student orientation proved to be predictive for the students’ overall rating of teaching quality.

In summary, both the student-perceived communion of a teacher and the general interest in the subject could create an affective overall impression (Clausen, 2002), which as a “general impression halo” (Lance et al., 1994) overlays the student ratings of the individual quality dimensions. This could explain the low statistical separation between the dimensions of teaching quality. Therefore, in the following, we analyze the explainability of a halo bias in student ratings on teaching quality by these two factors in the context of secondary schools.

2 Empirical Part: Explaining Halo Effects in Student Ratings of Teaching Quality Through Students' Perception of Teachers' Communion and Interest in the Subject Being Taught

This study focuses on the following research questions:

RQ1: To what extent can an overlaying second-order factor in the sense of a general impression halo be modeled superordinately to the various dimensions of teaching quality?

RQ2: Can this second-order factor in student ratings on teaching quality be explained by a) teachers' communion perceived by the students and/or b) students' overall subject-specific interest?

RQ3: To what extent can the strength of the correlational structure between the different dimensions of teaching quality be reduced by controlling for one or both of these factors?

These research questions are addressed at the individual as well as at the class level.

2.1 Methods and Sample

2.1.1 Design and Sample

Data used for the following analyses were collected from different secondary schools in the southwestern part of Germany, where teachers obtained student feedback on their teaching and classes. For research purposes, student feedback questionnaires were supplemented by instruments for the survey of teachers' communion and students' general interest in subject taught by the teacher. The sample comprises a total of $N = 1056$ students from 50 classes at lower track schools (*Werkrealschule*, 9.6%), middle track schools (*Realschule*, 35.5%), grammar and high schools (*Gymnasium*, 49.6%), and secondary comprehensive schools (*Gemeinschaftsschule*, 5.3%). The students belong to grades 5–6 (28.0%), 7–8 (20.3%), 9–10 (30.6%), and 10–13 (21.1%), and are aged between 10 and 19 years. Teachers' professional experience and gender were not surveyed for reasons of anonymity, but the sample included both young professionals and very experienced teachers, as well as female and male teachers. The teachers were free to choose the class and course in which they used the questionnaire. Therefore, the sample covers a wide range of taught subjects, including math, German, foreign languages, science, and history, but not physical education.

Table 1 Measurement instruments

| Scale | Number of items | Example | Reliability ω^a |
|--|-----------------|---|------------------------|
| <i>Feedback Questionnaire on Teaching Quality (FQTQ, Röhl, 2015)</i> | | | |
| Clarity of content and explanations | 6 | I understand what I am supposed to learn in each lesson | .87 |
| Activation and use of adaptive methods | 5 | In the lessons, I'm learning to work and learn by myself | .80 |
| Classroom and teaching management | 5 | The noise level in the lessons allows me to work and learn well | .72 |
| Individual care and kindness | 4 | The teacher values my contributions to the lessons | .87 |
| Transparency of assessment | 4 | The gradings of the tests seem to be fair to me | .84 |
| <i>Questionnaire on Teacher Interaction (Wubbels & Levy, 1991)</i> | | | |
| CD: helping/friendly | 6 | We can rely on our teacher | .80 |
| CS: understanding | 6 | If we don't agree our teacher listens to us | .80 |
| Overall subject-related interest | 2 | The subject itself interests me... | .89 |

^aFor the reliability estimator McDonald's Omega see McDonald (1999)

2.1.2 Measures

Feedback Questionnaire on Teaching Quality (FQTQ)

The *Feedback Questionnaire on Teaching Quality (FQTQ, Röhl, 2015)* is based on the characteristics of good teaching according to Meyer (2005), and includes 24 items with a four-level Likert format (“fully agree” to “disagree”). The aim of the instrument is to provide teachers with indications for improving their own teaching and classes. In total, the FQTQ assesses five quality dimensions of teaching: “Clarity of content and explanations”, “Activation and use of adaptive methods”, “Classroom and teaching management”, “Individual care and kindness”, and “Transparency of assessment” (see Table 1). All scales showed satisfactory to good reliability values using the reliability estimator ω (McDonald, 1999), which proved to be particularly reliable for use on short scales and in the context of structural equation modeling (Revelle & Zinbarg, 2009; Teo & Fan, 2013).¹ The formulations in the instrument are kept as low-inferent as possible (Wagner, 2008) and—in order to avoid problems of comprehension (Clausen, 2002)—are formulated positively throughout. In all quality dimensions, both ego- and web-references are used in the item wordings.

¹ McDonald's ω is based on the parameter estimates of the items in a factor model and represents the ratio of the variance due to the common factor to the total variance.

Confirmatory factor analyses indicated a good fit of the theoretically assumed structure with five factors ($\chi^2(242) = 1597.8, p < .001, CFI = .989, TLI = .987, RMSEA = .014$). This was also evident in comparative analyses using a model with one single overall factor, which resulted in less favorable fit statistics ($\chi^2(252) = 3591.7, p < .001, CFI = .923, TLI = .916, RMSEA = .069$).

To survey the students' perception of the "teacher communion", the scales "CD: helping/friendly" and "CS: understanding" of the "Questionnaire on Teacher Interaction" (Wubbels & Levy, 1991) were used, which reflect a high degree of this basic dimension. The response scale for the 12 items comprises five levels (from "1: never" to "5: always").

In addition, we measured students' overall subject-related interest using the two items "The subject itself interests me" and "I like the subject itself", using a five-point scale (from "very" to "not at all").

2.1.3 Data Analysis

The data was analyzed by means of single- and multi-level structural equation analyses using MPlus 8.4 (Muthen & Muthen, 2012–2019). Considering the ordinal level of the four- and five-point rating scales, the "categorical" option was used for the measurement models, which relies on polychoric correlations for the corresponding sub-models. At the same time, the used procedure models response behavior in the sense of a probabilistic latent trait analysis (Uebersax, 2010–15). We chose the robust Weighted Least Square Estimator (WLSMR) as the estimation method, which showed a high reliability for ordinal scaled measurement models in simulation studies (Flora & Curran, 2004). The clustered data structure was considered by the option "type = complex". The high number of parameters of the ordinal measurement models made it necessary to use the less computationally intensive Bayesian estimator for the subsequent multi-level analyses (Asparouhov & Muthen, 2012).

2.2 Findings

2.2.1 Modeling a Latent Second-Order Factor

To examine research question 1 (whether a factor overlaying the dimensions of teaching quality can be modeled reflecting an overall impression) an SEM was specified in which the overall impression is represented as a latent second-order factor. Fit indices pointed to a good fit of the assumed structure ($\chi^2(247) = 414.7, p < .001, CFI = .978, TLI = .975, RMSEA = .021, SRMR = .040$). Analyses indicated medium to large loadings of the five teaching quality dimensions on the second-order factor (clarity: $\beta = .928$, methods: $\beta = .984$, classroom management: $\beta = .739$, care: $\beta = .878$, transparency: $\beta = .772, p < .001$ each).

Table 2 Effects of teacher communion perceived by students and students' general interest in subject on the second-order factor overall impression

| | Model 1 β | Model 2 β | Model 3 β |
|------------------------------------|--------------------|--------------------|--------------------|
| Communion | .84*** | – | .71*** |
| General interest in subject | – | .63*** | .26*** |
| R^2 of overall impression factor | .71 | .40 | .76 |
| χ^2 | 485.5 | 446.1 | 623.9 |
| df | 293 | 293 | 342 |
| p | < .001 | < .001 | < .001 |
| RMSEA | .025 | .029 | .028 |
| CFI | .975 | .970 | .968 |
| TLI | .973 | .967 | .965 |
| SRMR | .040 | .044 | .043 |

*** $p < .001$

In the next step, the effects of the perceived communion of the teacher and students' interest in subject on the overall impression factor were determined to answer research question 2. For this purpose, three regression models were estimated at the student level. First, both possible influencing factors were analyzed individually (models 1 and 2), and then combined in a second step (model 3). The results are summarized in Table 2.

Model 1 examined the perceived teacher communion as an explanatory variable for the second-order factor overall impression. It shows a good model fit and explains more than 70% of the variance of the overall impression. A slightly inferior fit is shown by model 2, which tests students' general interest in subject as the source for the overlaying effect and explains 40%. With the assumption underlying model 3 that both influencing factors jointly explain the overall impression, 76% of the variance can be explained (see Fig. 1). The far greater proportion can therefore be explained by the perceived teacher communion. Both factors correlate with each other on a medium level ($r = .53$).

2.2.2 Correlations Between Teaching Quality Dimensions Controlling for Students' Perception of Teacher Communion

In order to analyze the effects on the intercorrelations between the various quality dimensions of teaching, the amount to which this can be explained by perceived teacher communion and general interest in subject was investigated analogous to the approach of Borg (2003) described above (research question 3). Therefore, a structural equation model was used to determine the direct effects of these factors on the items of teaching quality. This procedure extracted the variance component related to these factors, and only the remaining variance components were loaded onto the quality dimensions.

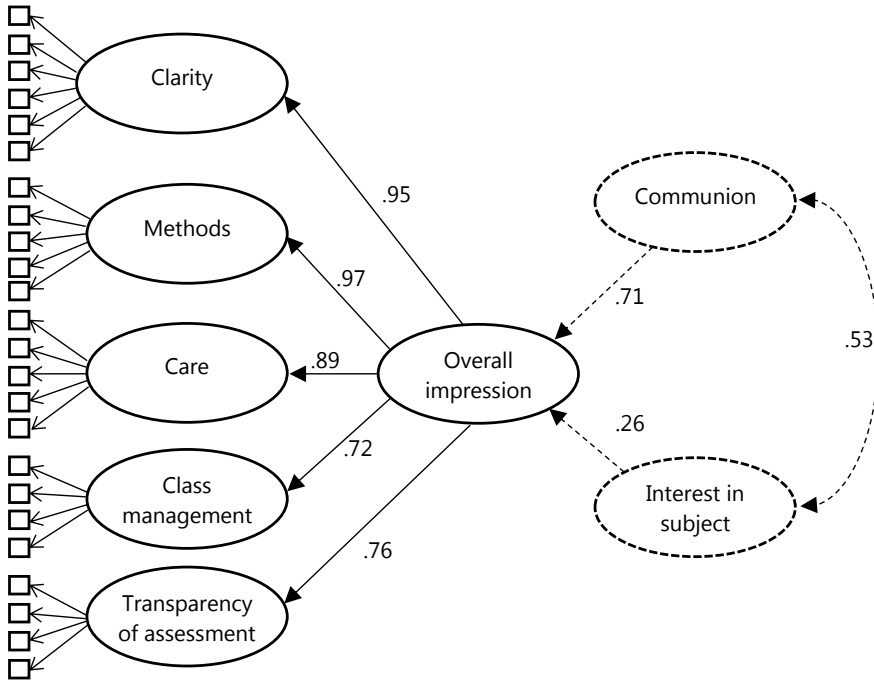


Fig. 1 Structural equation model at student level explaining the overall factor by perceived teacher communion and students’ general interest in subject (model 3). All loadings are standardized and significant at the $p < .001$ -level

At the student level the model showed a good fit ($\chi^2(291) = 477.3$, RMSEA = .025, CFI = .979, TLI = .973, SRMR = .034). The loadings of the items on the quality dimensions remained (with two exceptions) significant ($p < .05$), but decreased substantially (average item loadings: clarity: .29, methods: .24, classroom management: .44, care: .21, transparency: .44). Whereas teacher communion showed highly significant effects on each of the 24 individual items, ranging from $\beta = .21$ –.84 ($p < .001$), analysis of the general interest on subject revealed only eight much less significant effects ($\beta = .10$ –.42, $p < .05$).

At the same time, the intercorrelations between the individual quality dimensions decreased substantially, and in some cases were no longer significant (Table 3). This is especially true for the dimensions “Individual Care and Kindness” and “Transparency of Assessment”, which showed no or only low correlations with the other dimensions. The partially negative correlations of the care dimension can be understood as a suppression effect, since this dimension has the highest content overlap with communion.

Table 3 Intercorrelations between the perceived dimensions of teaching quality at the student level. Below the diagonal = without control of the communion; above the diagonal = with control of the item-related effects of the communion

| | | 1 | 2 | 3 | 4 | 5 |
|----|--|---------|---------|---------|---------|--------|
| 1. | Clarity of content and explanations | X | .606*** | .458*** | -.365 | .203* |
| 2. | Activation and use of adaptive methods | .917*** | X | .635*** | -.280 | .053 |
| 3. | Classroom and teaching management | .715*** | .732*** | X | .381** | .040 |
| 4. | Individual care and kindness | .794*** | .865*** | .665*** | X | -.422* |
| 5. | Transparency of assessment | .728*** | .748*** | .486*** | .703*** | X |

* $p < .05$, ** $p < .01$, *** $p < .001$

2.2.3 Analyses at the Class Level

With regard to the class level, model 3 was extended to a two-level model. The results on the effects at the student level remained almost constant compared to the previous findings. At the class level, the loadings of the individual teaching dimensions on the overall second-order factor showed similarly high values as at the student level (clarity: $\beta = .86$, methods: $\beta = .95$, classroom management: $\beta = .60$, care: $\beta = .98$, transparency: $\beta = .67$, $p < .001$ each). Interestingly, the effect of teacher communion on the overall factor was considerably higher ($\beta = .87$, $p < .001$), whereas interest on subject no longer showed any significant effect ($\beta = .12$, $p = .198$). Replicating the analysis of item-related effects of communion and general subject interest at the class level led to the almost complete elimination of significant item loadings on the dimensions of teaching quality.

3 Discussion

The findings presented here indicate that an overall impression which overlays the perception of teaching quality can be modeled as a latent second-order factor. The modeled overall impression can be explained to a large extent by teacher communion perceived by the students. Students' general interest in the subject taught only shows significant effects at the individual level, and these effects are low. Thus, at the class level the general subject interest does not appear to have any relevant effect on the overall impression, and does not induce a bias for the assessment of teaching quality when the data is aggregated for classes. These results are in line with the findings from Ferdinand (2014). The findings also point to the existence of a "general impression halo" in accordance with Lance et al. (1994), which is based on an affective attitude—to a larger extent toward the teacher and to a lesser extent toward the subject being taught. Furthermore, the modest significant correlation between communion and interest in subject shows that there could be a reciprocal influence in students' perceptions of the subject and the teacher.

Thus, the affective overall impression reported in the literature seems to be predominantly based on students' perception of teachers' communion, which means that the teacher is perceived as being interested in the learning progress of all students and sympathetic to the needs of the learning group. These results show that the theory of social judgments (Abele & Bruckmüller, 2011) provides a valid framework for obtaining a better understanding of the processes of students' assessment of teaching and classes.

The control of direct item-related effects of teachers' communion shows that the high intercorrelations of the dimensions of teaching quality in which the halo effect manifests itself can be drastically reduced—in some cases even to an insignificant level. For the general interest in the subject taught this is only true to a much smaller extent.

However, it can be theoretically argued that a high quality of teaching can indeed go hand in hand with students' perception of a high teacher communion. In this case, students' perception of a high communion of the teacher could be based on an inner attitude of respect and empathy from the teacher, which in turn contributes to an overall higher quality in the different teaching dimensions; conversely, a less empathic attitude from the teacher could lead to a lower quality of teaching (Tausch, 2007). Thus, this inner attitude could lead to teaching being better adapted to the students' learning (from a methodological-didactical point of view), and also to a more comprehensible performance assessment for the students. Conversely, didactics and methodology which are more strongly oriented toward the students could lead to a higher assessment of teacher communion. In this case, the overlaying affective overall impression by the perceived teacher communion would not represent a problematic bias in the context of student feedback or the measurement of teaching quality. As a result of higher teaching quality, it is a central element for its valid measurement.

On the other hand, the perception of a high communion of a teacher could also lead to teaching which has qualitative deficits (with regard to pedagogical action in class) being assessed more positively by the students than might be appropriate. In this case, a weaker quality among teachers with a high communion would be masked by this perception. In other words, in such cases there could actually be a severe bias influencing the measurement of teaching quality. This could explain why many studies showed no or only minor predictive effects of the teaching quality measured by student surveys on learning achievement, and why often large differences in the quality perception of students and external observers are reported (see, for example, Chap. 7 by van der Lans in this volume; Fauth et al., 2014; Kuhfeld, 2017). If this is the case, a way out could be to control the perceived student communion through partial regressions, as was done in the analyses presented here.

However, both phenomena could also exist, which means that on the one hand there are good teachers with high communion and worse teachers with lower communion, for whom the effect described here is not a bias; on the other hand, there are also situations in which good teachers with lower communion are rated worse by the students and worse teachers with high communion are rated much better. In this case, there is a need to clarify whether indicators can be developed to distinguish between

these two situations. These could then be used as a supplement to the classical evaluation procedure for feedback to teachers.

Further research is needed to address the issues raised in this chapter and, if necessary, to develop methods for correcting the measurement of teaching quality through student surveys. This would require longitudinal studies of the perceived quality of teaching over a period of joint work by teachers and classes. In such studies, it would be especially valuable if the dimensions of teaching quality and teacher communion were also assessed by external observers. At the same time, realizing experimental study designs could also be fruitful in which the same teacher's statements with varying communion, e.g. as video vignettes with actors, are rated by students. In addition, studies controlling for the use of ego- and web-references in the item wordings could be helpful in getting a deeper insight into this effect (den Brok et al., 2006).

When using student ratings as classroom feedback, teachers should be aware that there is an overlaying halo effect related to their communion. Teachers perceived more positively by learners in this way should therefore be more critical of the feedback received. Conversely, relatively unfavorable ratings, which can be associated with a lower perception of communion, are an indication to consider and improve related aspects of teaching quality. For a reliable assessment and control of such effects, it would be advantageous to supplement the questionnaires on teaching quality used in practice and research with a scale for measuring the teacher communion perceived by the students. If such information is not available, teachers should bear in mind that the evaluation of the data at the individual level might be less confounded with their communion than the aggregated data at the classroom level. So, it might be advisable to evaluate the data on both levels to gain a better insight into how one's teaching practice is perceived by the students.

References

- Abele, A. E., & Bruckmüller, S. (2011). The bigger one of the “Big Two”? Preferential processing of communal information. *Journal of Experimental Social Psychology*, *47*, 935–948. <https://doi.org/10.1016/j.jesp.2011.03.028>.
- Abele, A. E., Cuddy, A. J. C., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of social judgment. *European Journal of Social Psychology*, *38*, 1063–1065. <https://doi.org/10.1002/ejsp.574>.
- Asparouhov, T., & Muthen, B. (2012). *Comparison of computational methods for high dimensional item factor analysis*. <http://statmodel.com/download/HighDimension.pdf>. Accessed 6 Apr 2016.
- Bakan, D. (1966). *The duality of human existence: Isolation and communion in Western man*. Beacon Press.
- Bell, L. M., & Aldridge, J. M. (2014). Investigating the use of student perception data for teacher reflection and classroom improvement. *Learning Environments Research*, *17*, 371–388. <https://doi.org/10.1007/s10984-014-9164-z>.

- Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, 28, 217–236. <https://doi.org/10.1080/1475939X.2019.1572534>.
- Borg, I. (2003). Affektiver Halo in Mitarbeiterbefragungen [Affective halo in staff surveys]. *Zeitschrift Für Arbeits- Und Organisationspsychologie A&O*, 47, 1–11. <https://doi.org/10.1026//0932-4089.47.1.1>.
- Clausen, M. (2002). *Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität* [Teaching quality: A matter of perspective? Empirical analyses of agreement, construct and criterion validity]. Waxmann.
- Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research*, 77, 113–143. <https://doi.org/10.3102/003465430298563>.
- den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the Questionnaire on Teacher Interaction. *Learning Environments Research*, 9, 199–213. <https://doi.org/10.1007/s10984-006-9013-9>.
- Ditton, H. (2002). Lehrkräfte und Unterricht aus Schülersicht. Ergebnisse einer Untersuchung im Fach Mathematik [Teachers and teaching from a student perspective. Results from a study in Mathematics]. *Zeitschrift für Pädagogik*, 48(2), 262–286.
- Eder, F., & Bergmann, C. (2004). Der Einfluss von Interessen auf die Lehrer-Wahrnehmung von Schülerinnen und Schülern. *Empirische Pädagogik*, 18(4), 410–430.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9.
- Ferdinand, H. D. (2014). *Entwicklung von Fachinteresse: Längsschnittstudie zu Interessenverläufen und Determinanten positiver Entwicklung in der Schule* [Development of interest in subject: Longitudinal study of interest trajectories and determinants of positive development in school]. Waxmann.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491. <https://doi.org/10.1037/1082-989X.9.4.466>.
- Greimel-Fuhrmann, B. (2014). Students' perception of teaching behaviour and its effect on evaluation. *International Journal for Cross-Disciplinary Subjects in Education*, 5(1), 1557–1563.
- Haladyna, T., & Hess, R. K. (1994). The detection and correction of bias in student ratings of instruction. *Research in Higher Education*, 35, 669–687. <https://doi.org/10.1007/BF02497081>.
- Krammer, G., Pflanzl, B., & Mayr, J. (2019). Using students' feedback for teacher education: Measurement invariance across pre-service teacher-rated and student-rated aspects of quality of teaching. *Assessment and Evaluation in Higher Education*, 44, 596–609. <https://doi.org/10.1080/02602938.2018.1525338>.
- Kuhfeld, M. R. (2017). When students grade their teachers: A validity analysis of the Tripod Student Survey. *Educational Assessment*, 22, 253–274. <https://doi.org/10.1080/10627197.2017.1381555>.
- Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, 79, 332–340. <https://doi.org/10.1037/0021-9010.79.3.332>.
- Lazarides, R., Ittel, A., & Juang, L. (2015). Wahrgenommene Unterrichtsgestaltung und Interesse im Fach Mathematik von Schülerinnen und Schülern [Students' perceived teaching style and interest in the subject of mathematics]. *Unterrichtswissenschaft*, 43(1), 67–82.
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, 26, 169–194. <https://doi.org/10.1080/09243453.2014.939198>.
- Mayr, J. (2006). Klassenführung in der Sekundarstufe II: Strategien und Muster erfolgreichen Lehrerhandelns [Classroom management in upper secondary schools: Strategies and patterns of

- successful teacher action]. *Schweizerische Zeitschrift Für Bildungswissenschaften*, 28(2), 227–241.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates.
- Meyer, H. (2005). *Was ist guter Unterricht?* [What is good teaching?] (2nd ed.). Cornelsen Scriptor.
- Muthen, L. K., & Muthen, B. (2012–2019). *MPlus* (7th ed.). Muthén & Muthén.
- Nelson, P. M., Demers, J. A., & Christ, T. J. (2014). The Responsive Environmental Assessment for Classroom Teaching (REACT): The dimensionality of student perceptions of the instructional environment. *School Psychology Quarterly*, 29, 182–197. <https://doi.org/10.1037/spq0000049>.
- Rahn, S., Gruehn, S., Fuhrmann, C., & Keune, M. S. (2019). Schülerfeedback – fächerübergreifend vergleichbar? [Is it adequate to compare students' feedback beyond subjects?]. *Unterrichtswissenschaft*, 47, 383–404. <https://doi.org/10.1007/s42010-019-00043-w>.
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74, 145–154. <https://doi.org/10.1007/S11336-008-9102-Z>.
- Röhl, S. (2015). *Feedbackfragebogen zur Unterrichtsqualität (FFU)* [Feedback Questionnaire on Teaching Quality (FQTQ)]. University of Education Freiburg.
- Schurtz, I. M., & Artelt, C. (2014). Die Entwicklung des Fachinteresses Deutsch, Mathematik und Englisch in der Adoleszenz: Ein personenzentrierter Ansatz [The development of subject interest in German, mathematics, and English in adolescence: A person-centered approach]. *Diskurs Kindheits- Und Jugendforschung*, 9(3), 285–301.
- Schweig, J. D. (2014). Cross-level measurement invariance in school and classroom environment surveys. *Educational Evaluation and Policy Analysis*, 36, 259–280. <https://doi.org/10.3102/0162373713509880>.
- Tausch, R. (2007). Lernförderliches Lehrerverhalten: Zwischenmenschliche Haltungen beeinflussen das fachliche und persönliche Lernen der Schüler [Teacher behaviors that promote learning: Interpersonal attitudes influence students' subject and personal learning]. In W. Mutzeck (Ed.), *Professionalisierung von Sonderpädagogen: Standards, Kompetenzen und Methoden* (pp. 14–29, Beltz-Bibliothek). Beltz.
- Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher*, 22, 209–213. <https://doi.org/10.1007/s40299-013-0075-z>.
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press.
- Tripod Education Partners. (2014). *Tripod's 7Cs: Technical manual*.
- Uebersax, J. S. (2010–15). *The tetrachoric and polychoric correlation coefficients: Statistical methods for rater agreement*. <http://john-uebersax.com/stat/tetra.htm>. Accessed 6 Apr 2016.
- van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*, 30, 30–50. <https://doi.org/10.1080/09243453.2018.1539015>.
- van Petegem, P., Deneire, A., & de Maeyer, S. (2008). Evaluation and participation in secondary education: Designing and validating a self-evaluation instrument for teachers to solicit feedback from pupils. *Studies in Educational Evaluation*, 34, 136–144. <https://doi.org/10.1016/j.stueduc.2008.07.002>.
- Wagner, W. (2008). Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI (Deutsch Englisch Schülerleistungen International) der Kultusministerkonferenz [Methodological problems in the analysis of classroom perception from the students' point of view - using the example of the study DESI (Deutsch Englisch Schülerleistungen International) of the Standing Conference of the Ministers of Education and Cultural Affairs of the Federal Republic of Germany]. Dissertation, University of Koblenz-Landau.
- Wallace, T. L., Kelcey, B., & Ruzek, E. A. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, 53, 1834–1868. <https://doi.org/10.3102/0002831216671864>.

- Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. H. (2020). Obtaining students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction*. <https://doi.org/10.1016/j.learninstruc.2020.101303>.
- Wubbels, T., Brekelmans, M., den Brok, P., Wijsman, L., Mainhard, T., & van Tartwijk, J. (2015). Teacher–student relationships and classroom management. In E. T. Emmer & E. J. Sabornie (Eds.), *Handbook of classroom management* (2nd ed., pp. 363–386). Routledge.
- Wubbels, T., & Levy, J. (1991). A comparison of interpersonal behavior of Dutch and American teachers. *International Journal of Intercultural Relations*, *15*, 1–18. [https://doi.org/10.1016/0147-1767\(91\)90070-W](https://doi.org/10.1016/0147-1767(91)90070-W).

Sebastian Röhl has been an Academic Assistant in the Department of Educational Science at the University of Education Freiburg and is currently a Postdoctoral Researcher in the Institute of Education at Tübingen University (Germany). Before that he worked for more than 10 years as a grammar school teacher, school development consultant, and in teacher training. Among other areas, he conducts research in the fields of teaching development and teacher professionalization through feedback, social networks in inclusive school classes, as well as teachers' religiosity and its impact on professionalism. In addition, he is the Director of an in-service professional master's study program for teaching and school development.

Wolfram Rollett is a Professor of Empirical Educational Research at the University of Education Freiburg and the Freiburg Advanced Center of Education (FACE). Previously he worked as a researcher and lecturer in the field of Educational Science and Psychology at the Universities of Potsdam, Braunschweig, Dortmund, and Wuppertal. His research focuses on school development processes, the quality of extra- and co-curricular activities, educational effectiveness, and classroom composition.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4

The Quality of Student Perception Questionnaires: A Systematic Review



Hannah Bijlsma

Abstract Student perceptions of teaching are promising for measuring the quality of teaching in primary and secondary education. However, generating valid and reliable measurements when using a student perception questionnaire (SPQ) is not self-evident. Many authors have pointed to issues that need to be taken into account when developing, selecting, and using an SPQ in order to generate valid and reliable scores. In this study, 22 SPQs that met the inclusion criteria used in the literature search were systematically evaluated by two reviewers. The reviewers were most positive about the theoretical basis of the SPQs and about the quality of the SPQ materials. According to their evaluation, most SPQs also had acceptable reliability and construct validity. However, norm information about the quality rating measures was often lacking and few sampling specifications were provided. Information about the features of the SPQs, if available, was also often not presented in an accessible way by the instrument developers (e.g., in a user manual), making it difficult for potential SPQ users to obtain an overview of the qualities of available SPQs in order to decide which SPQs best fit their own context and intended use. It is suggested to create an international database of SPQs and to develop a standardized evaluation framework to evaluate the SPQ qualities in order to provide potential users with the information they need to make a well-informed choice of an SPQ.

Keywords Student perception questionnaires · Teaching quality · Systematic review

1 Introduction

Student perceptions of teaching are promising as a way to measure the quality of teaching in primary and secondary education (Ferguson, 2012; Ferguson & Danielson, 2014; Schulz et al., 2014; van der Scheer et al., 2018). The scores provided by students with regard to their teachers' teaching can be used for research

H. Bijlsma (✉)
Section of Teacher Professionalization, University of Twente,
Enschede, the Netherlands
e-mail: h.j.e.bijlsma@utwente.nl

© The Author(s) 2021
W. Rollett et al. (eds.), *Student Feedback on Teaching in Schools*,
https://doi.org/10.1007/978-3-030-75150-0_4

and accountability purposes, for example, or for teacher professional development (Timperley et al., 2007). A student perception questionnaire (SPQ) is often used for collecting student perceptions of teaching quality, and usually consists of a set of items about the quality of teaching that students have to respond to using a numeric scale. An SPQ also involves the information and activities required to use the instrument as intended. Therefore, ideally, an SPQ also includes a user manual, scoring rules, and sampling specifications (for example, specifications regarding the subject of the lesson). Although the use of SPQs is not new, several studies (Bijlsma et al., under review; Lüdtke et al., 2006; van der Scheer et al., 2018; Wallace et al., 2016) have recently generated renewed interest in reliability and validity issues surrounding SPQs and other teacher evaluation approaches, such as classroom observation systems (Bell et al., 2018; Dobbelaer, 2019).

Generating valid and reliable scores using an SPQ is not guaranteed. Many authors (e.g., Bijlsma et al., under review; van der Lans & Maulana, 2018; and Chap. 7 by Göllner et al., in this volume) have pointed to issues that need to be taken into account when developing and/or using an SPQ in order to generate valid and reliable scores. These include issues regarding the theoretical basis of the items and the constructs that the SPQ aims to measure. However, these issues are often not (fully) addressed by SPQ users or developers, bringing the reliability and validity of the student scores into question. Moreover, there is no overview of the student perception questionnaires available for use in primary and secondary education that identifies what their psychometric characteristics are and what teaching quality constructs they aim to measure. Therefore, in this study, a systematic review was conducted of SPQs in primary and secondary education. The SPQs found were reviewed, based on an evaluation framework developed for this study. An overview with general information about the SPQs and the results of the evaluation and an overview with the constructs measured by the SPQs are presented. The overviews contribute to an increased awareness of the complexity of SPQs by developers and users, and to the deliberate design and use of SPQs. Note that in this chapter, a clear or standard definition of teaching quality is not presented, because consensus about its conceptualization or definition across SPQs is minimal (Cohen & Goldhaber, 2016).

2 The Evaluation Framework

A selection of SPQs was reviewed using an evaluation framework (available in Dutch) consisting of seven standards: the theoretical basis of the questionnaire, the quality of the questionnaire, the quality of the manual, norms, reliability, construct validity, and criterion validity. The standards in the framework were drawn from two strands of literature: the literature on SPQs and the literature on testing and performance

assessment (the COTAN evaluation standards for test quality, Evers et al., 2010¹). The evaluation framework and the underlying theory are outlined in the following paragraphs. Additionally, SPQ assessment purposes are distinguished and discussed.

2.1 Evaluation Standard 1—The Theoretical Basis of the Questionnaire

Each SPQ in this review includes a scoring format or tool consisting of items that are scored on a rating scale. In most SPQs, several items can form a construct² (Marsh, 2007), an aspect of teaching quality as perceived by students (Maulana & Helms-Lorenz, 2016). All SPQs measure the quality of teaching; however, they can focus on different constructs that are perceived by students. Based on several meta-analyses related to effective teaching (Praetorius et al., 2018; Bell et al., 2018; Creemers, 1994; Pianta & Hamre, 2006; Sammons et al., 1995), nine constructs that are known to be effective for student learning are distinguished in this study: a safe and stimulating classroom climate, classroom management, the involvement and motivation of students, explanation of subject matter, the quality of subject-matter representation, cognitive activation, assessment for learning, differentiated instruction, and teaching learning strategies and student self-regulation. The constructs in SPQs can be derived from different theories, research, or standards, but all should have a solid scientific basis (American Educational Research Association [AERA] et al., 1999) and the items should cover the theoretical constructs (Evers et al., 2010). Although the theoretical basis of the questionnaires was evaluated in this review, it was not feasible to evaluate the quality of the research underlying the questionnaire, as well.

2.2 Evaluation Standard 2—Quality of the Questionnaires

To evaluate the quality of the questionnaires (this corresponds to “material” in general psychological tests), the item design of the SPQ is considered and it is determined whether the scoring system and procedure are standardized. The items on the SPQ can be subject-specific (e.g., designed to capture the quality of mathematics teaching) or generic (items that can be used across subjects), and can focus on teachers’ actions, students’ actions, or both (Bell et al., 2018). The number of items included in the

¹ The COTAN evaluation standards are used by the Dutch Committee on Tests and Testing (COTAN) to evaluate the quality of psychological tests available in the Netherlands. COTAN has audited over 750 tests published for professional use.

² Others have used terms such as dimension, scale, or pattern to refer to what I am calling “construct,” but in my opinion such terms do not capture well the conceptual link with the aspects of teaching quality perceived by the student.

scoring tools can differ as well as the response categories in the rating scale. Strong (2011) pointed out that a large number of items can be problematic for students because “there is an upper limit of a rater’s ability to match his or her responses to a given set of stimuli” (the channel capacity; Strong, 2011, p. 88). Although utilizing a small number of items may reduce students’ cognitive load and be adequate for evaluating teaching quality, more items enable providing richer feedback to teachers on their strengths and weaknesses, which is needed for improvement (Marzano, 2012).

2.3 Evaluation Standard 3—The Quality of the Manual

A description of the scoring tools should enable potential users to judge whether an SPQ is suitable for their purposes and should therefore include a description of the constructs the SPQ aims to measure, the type of use for which the SPQ has been developed and who/what can be observed by using the SPQ (Evers et al., 2010).

2.4 Evaluation Standard 4—Norms

Numeric ratings usually result in a raw score, which is partly determined by the characteristics of the SPQ. The norms evaluation standard evaluates whether the SPQ provides a meaningful interpretation of its results (Evers et al., 2010). Two ways of “scaling” or categorizing can be used to interpret the raw scores (American Educational Research Association (AERA), (APA), and (NCME), 1999). First, a set of scaled norms may be derived from the distribution of the raw scores of a reference group. This is called norm-referenced interpretation (Drenth & Sijtsma, 2006; Bechger et al., 2009). Second, standards may be derived from a domain or subject matter to be mastered or from the results of empirical validity research: the domain-referenced interpretation and the criterion-referenced interpretation, respectively (Berk, 1986; Vos & Knuver, 2000). The raw scores are given meaning by providing norms or standards and it makes the SPQ more user-friendly.

2.5 Evaluation Standard 5—Reliability

Providing reliability evidence is primarily the responsibility of SPQ developers since prospective users need this information to make an informed choice among alternative instruments or other measurement approaches and prospective users will generally be unable to conduct reliability studies prior to the operational use of an SPQ (AERA et al., 1999). In this evaluation, the assessment purpose was taken into account. A higher reliability coefficient (or a similar measure) is more critical for

high-stakes decisions (e.g., tenure decisions) than for low-stakes decisions such as teacher professional development activities. The quality of the research was also taken into account (as suggested by Evers et al., 2010), namely, whether the (analysis) procedures followed were correct, whether the research had been conducted in the target group of the SPQ (e.g., an SPQ designed for primary education should be investigated in primary education) and whether developers provided enough information to thoroughly judge the reliability of the SPQ scores. Various types of reliability measurements can be used, such as parallel-form and split-half reliability, reliability on the basis of inter-item covariance, test-retest reliability, interrater reliability, generalization theory, and structural equation models. Other methods for reliability testing are Guttman's lambda2 (Guttman, 1945) and the greatest lower bound (glb; ten Berge & Socan, 2004). For a more detailed description of Generalizability Theory (as well as Classical Test Theory and Item Response Theory), see Chap. 2 by Bijlsma et al., in this volume.

2.6 Evaluation Standard 6—Construct Validity

Validity reflects the extent to which an SPQ fulfills its purpose (the measurement of a specific construct; Drenth & Sijtsma, 2006). This shows whether the instrument is useful or not. Validity is “a matter of degree rather than an all-or-none property and validation is an unending process” (Nunnally, 1967, p. 75). Several methods can be used to support construct validity such as research on the (uni)dimensionality of the items (explanatory or confirmatory factor analyses), the quality of the items and the fit of the items to a model (e.g., IRT model, see Chap. 2 by Bijlsma et al., in this volume), and the correlations between relevant scales.

2.7 Evaluation Standard 7—Criterion Validity

To demonstrate the relationship of variables (e.g., do the SPQ scores for teachers' teaching quality relate to student achievement), criterion validity (also called predictive validity) can be investigated. Some SPQs are developed specifically for the purpose of investigating criterion validity. In addition, aspects of validity could also be captured through interviews with the students, which could relate to the content and factual accuracy of their understanding of the items. But this type of validity research is not always conducted.

2.8 Possible SPQ Assessment Purposes

The evaluation standards described above are based on the circumstances of psychological testing procedures and not, for example, on their use as a pure feedback instrument via reporting student perceptions of teaching. The necessity of meeting each standard is related to the assessment purpose of the SPQ. For example, norms are not necessarily needed when you use the student perception scores as feedback to teachers, nor are construct and criterion validity (Kane, 2006). In the context of this chapter, I distinguish four SPQ assessment purposes (Mislevy, 2013; Bell, 2019): assessment for research practice (e.g., for measuring intervention effects), assessment as a feedback loop (e.g., for improvement at the teacher level [teaching quality] or school level), assessment as an evidentiary argument (e.g., to develop claims that are supported by measurements for personnel decisions), and assessment as a measurement (e.g., to specify and test assessment models as a way to work toward models that enable representation of real-world resources). In the results section, I overview which SPQs seem suitable for which assessment purpose.

3 Method

Before searching for SPQs, a review protocol was developed in collaboration with an information specialist. This protocol included the aim of the review, the research questions, the inclusion and exclusion criteria, and the search strategy.

3.1 Inclusion and Exclusion Criteria for Questionnaires

It specifies the characteristics of populations, interventions, contexts, and outcomes:

Population: The SPQs were designed for use in primary and secondary education.

Intervention: The SPQs measure teaching quality. Aspects of learning climate and classroom climate were included as well as the extent to which students are involved in their lessons.

Context: The SPQs were designed for teacher-oriented lessons, such as mathematics, language, or reading.

Outcomes of interest: SPQs were only included if research had been done on the psychometric quality of the instrument. No exclusions were made based on the methodology of the study.

Other criteria for inclusion were forms of publication, language, and time period (Littell et al., 2008).

Forms of publications: To deal with publication bias, only peer-reviewed publications, dissertations, and unpublished articles were included.

Language: No criteria were set based on the country where the instrument was developed. However, for practical reasons, only questionnaires published in Dutch and English were included.

Time period: Only questionnaires developed between 1970 and 2016 were included. SPQs developed earlier than 1970 would be very outdated and inappropriate for the current educational context, while 2016 was the year in which the research project took place.

3.2 Search Strategy

Littell et al. (2008) described a procedure to efficiently plan a systematic literature review. The steps of this procedure are: searching in bibliographic and scientific databases, using terms and strings; searching for sources of unpublished articles and dissertations, and asking for personal contacts. Based on the experiences of fellow researchers, the step of “hand searching” was not conducted, because it turned out that nothing relevant was found that way (Dobbelaer et al., 2015).

Following this procedure, first, six databases were searched (ERIC, PsycInfo, Web of Science, ProQuest, Scopus, and Narcis³) with the last search conducted in April 2016. The search terms included: evaluation, perception, student, teacher, education, school, and psychometric. Synonyms for each individual concept were combined with the Boolean operator *OR* and all the different concept lists were combined using the Boolean operator *AND*. For Narcis, however, the same terms translated to Dutch were used for the search. The search terms are presented in the Appendix. All searches were run in all the databases in everything (title, abstracts, keywords) except full text. Backward-snowballing (reviewing the reference list of identified articles) resulted in more relevant SPQ publications and grey literature (unpublished articles and dissertations). The contacted field experts were all researchers conducting research into teaching evaluation, or had developed or used an SPQ in their research. A total of 92 researchers from 13 different countries were contacted for the purpose of both a systematic review of classroom observation systems (Dobbelaer, 2019) and for the current study. They were asked to name SPQ that met the inclusion criteria.

For the systematic review, the PRISMA reporting guidelines were followed (Liberati et al., 2009; Moher et al., 2009). A total of 1,544 publications were identified. After screening abstracts and titles, to check whether the articles were specifically about newly developed SPQs and not (for example about research done with an already available SPQ), 290 articles remained. Duplicates (56) were then removed and an additional 160 articles were removed after scanning the texts and reading the abstracts and titles more closely. 74 publications were then selected for full-text review. As a result of the full-text review, 49 additional publications were omitted based on the inclusion/exclusion criteria. Other publications, in which research had been done *with* the instrument, were read as well, to address practicality and usability

³ Dutch database for scientific research.

issues or—if available—interrater reliability. These were in total 25 questionnaires. The whole search was documented, using Refworks and Excel.

During the evaluation phase of this study, three questionnaires were determined to be not possible to evaluate. The Science-Technology Learning Environment Questionnaire (STLEQ) could not be evaluated because it did not meet the inclusion criterion of context. Comenius was the second questionnaire that could not be evaluated. This was due to language difficulties, which were not noticed during the search. The questionnaire was developed in Serbian and English, but the user guide and all other available sources were only available in Serbian. After contact with the developer, the decision was made not to evaluate the questionnaire. Lastly, the Learning Environment Inventory (LEI) was too out-of-date. The research done with the questionnaire was carried out in 1982, but it was originally developed in 1967. Both evaluators concluded that the questionnaire was too outdated, so it was excluded from the list. The elimination of these three questionnaires left a total of 22 questionnaires that were evaluated. Figure 1 shows the full flowchart of the PRISMA guidelines.

3.3 Description of the SPQ

A range of sources provided information about the SPQs: user manuals, manuals for SPQ use in research projects, peer-reviewed publications, dissertations, grey literature, websites, and personal contact with the authors. All relevant information was described in an overview for each SPQ, which included information on the general characteristics of the SPQ and the results of the evaluation. Additionally, the SPQ assessment purposes were analyzed and presented together with the descriptives for the SPQs.

3.4 Evaluation Procedure

The SPQs were reviewed based on the evaluation framework described above. After instruction in the use of the evaluation framework, two reviewers independently evaluated two SPQs. Based on the results, the two reviewers clarified all uncertainties and, where necessary, refined the evaluation framework. Next, two strongly differing questionnaires (in terms of development date, content, and intended use) were evaluated by both reviewers. The interrater reliability was then 0.87 (Kappa). Because this is considered high enough interrater reliability, the two reviewers independently evaluated all the remaining SPQs. After evaluating all SPQs, the scores were discussed by the two reviewers to come to an agreement about the final judgment.

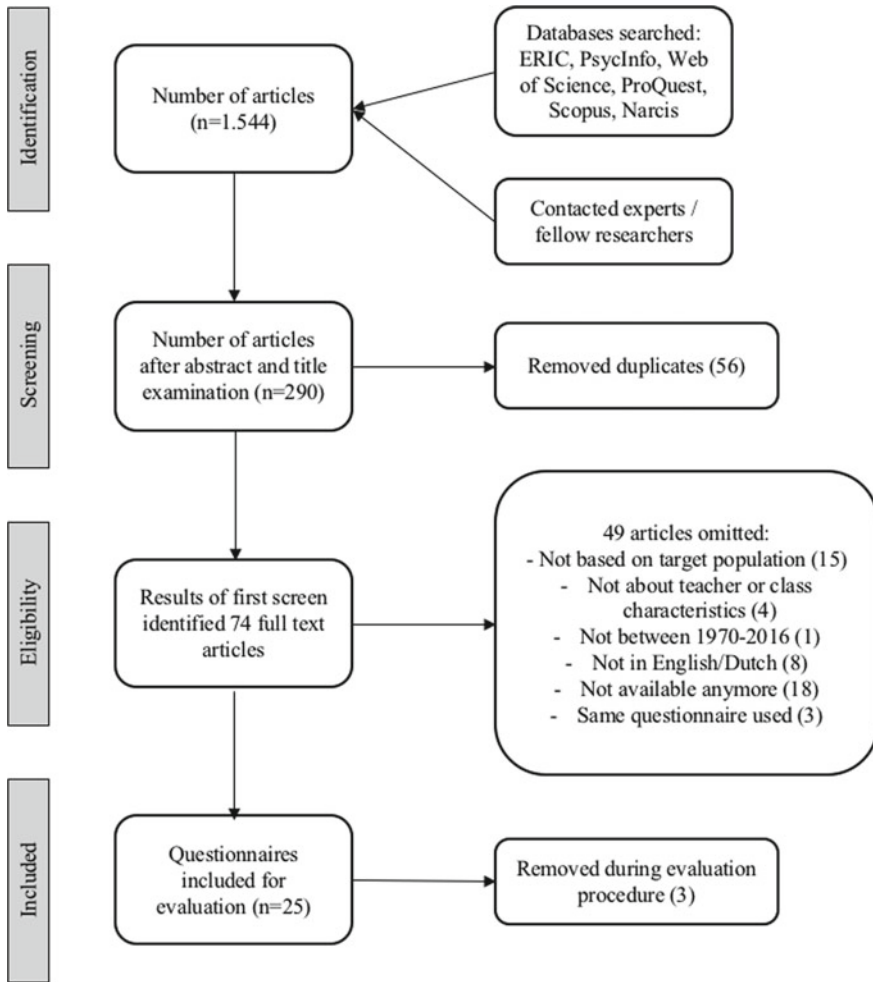


Fig. 1 Flowchart of the PRISMA guidelines followed in this study

3.5 Evaluation Framework

The evaluation framework consists of seven evaluation standards (as described earlier in this chapter) with 26 questions about the quality of the SPQ (ranging from 3 to 5 questions per standards). The reviewers scored all questions on a dichotomous scale (met or not met). The reviewers were instructed to assign the score “not met” if there was not enough evidence to evaluate a standard. The reviewers were required to give a reason for every response.

3.6 Analysis

The results of this review are descriptively consisting of description of the SPQs based on information about the SPQs and the evaluation framework, descriptive statistics from the results of the review based on the evaluation standards in the evaluation framework, and, if relevant, description of the reasons for a score. These reasons were analyzed qualitatively by open and axial coding.

4 Results

4.1 General Information

An overview of the 22 questionnaires can be found in Table 1. The number of items differed, ranging from 16 to 96. Eleven questionnaires used a 5-point rating scale, seven questionnaires a 4-point rating scale; one questionnaire a 3-point rating scale, two questionnaires a 2-point rating scale, and one questionnaire used both a 3 and a 5-point rating scale. Fifteen questionnaires were developed for secondary education, four for primary education, and three could be used for both (according to the authors of the questionnaire). Questionnaires differed in their date of development. Eleven questionnaires were developed between 2005 and 2014, five between 1995 and 2004, three questionnaires between 1985 and 1994, and the oldest three between 1970 and 1994. SPQ assessment purposes differed across the questionnaires. All SPQs were intended for research practices, while some were additionally used as a feedback tool (5) or for measurement purposes (4). Only two SPQs were developed to make an evidentiary argument (2). Remarkable, criterion validity could be evaluated in SPQs that were intended for measurement assessment (e.g., intended to specify and test assessment models as a way to work toward models that enable representation of real-world resources). All SPQs measured at least one of the teaching quality constructs described earlier (a safe and stimulating classroom climate, classroom management, the involvement and motivation of students, explanation of subject matter, the quality of subject-matter representation, cognitive activation, assessment for learning, differentiated instruction, and teaching learning strategies and student self-regulation). See Table 2 for an overview of the teaching quality constructs measured by the SPQs.

4.2 Evaluation Results

The evaluation results can be found in Table 1. Regarding the theoretical basis of the questionnaire as specified in the evaluation framework (standard one), all SPQs met the evaluation standard. This means that the constructs could be derived from different theories, researches, or standards, but all had a solid scientific basis and the

Table 1 Overview of the SPQs' general information and evaluation results

| Instrument | General information | | | | Evaluation results | | | | | | | | |
|------------|---|--------------------------|--------------------------|--------|----------------------------|--------------|---------------------|-----------------------|---------------------|-------|-------------|--------------------|--------------------|
| | Full name | Main reference | Target pop. ^a | #Items | Assess. purp. ^b | Scale points | Theory ^c | Material ^d | Manual ^e | Norms | Reliability | Construct validity | Criterion validity |
| 1 | SAES Students' perceptions of the classroom Assessment Environmental Scale | Alkharusi (2011) | PE SE ✓ | 16 | RP | 5 | ✓ | ✓ | | | ✓ | ✓ | |
| 2 | COLES Constructivist-Oriented Learning Environment Survey | Bell and Aldridge (2014) | ✓ | 88 | RP | 5 | ✓ | ✓ | | | ✓ | ✓ | |
| 3 | CES Short form of the Classroom Environment Scale | Fraser and Fisher (1983) | ✓ | 24 | RP | 2 | ✓ | ✓ | | | ✓ | ✓ | |
| 4 | MCI Short form of the My Class Inventory | Fraser and Fisher (1983) | ✓ | 25 | RP/EA | 2 | | | | | | | |
| 5 | ICEQ Short form of the Individualized Classroom Environment Questionnaire | Fraser and Fisher (1983) | ✓ | 25 | RP | 5 | | | | | | | |
| 6 | SPOCQ Student Perception Of Classroom Quality | Gentry and Owen (2004) | ✓ | 38 | RP/FL | 5 | ✓ | ✓ | | | ✓ | ✓ | |
| 7 | ICALT ICALT student perception questionnaire | Maulana et al. (2015) | ✓ | 59 | RP/M/FL | 4 | ✓ | ✓ | | | ✓ | ✓ | ✓ |

(continued)

Table 1 (continued)

| Instrument | General information | | | | Evaluation results | | | | | | | | |
|------------|---|---------------------------|--------------------------|--------|----------------------------|--------------|---------------------|-----------------------|---------------------|-------|-------------|--------------------|--------------------|
| | Full name | Main reference | Target pop. ^a | #Items | Assess. purp. ^b | Scale points | Theory ^c | Material ^d | Manual ^e | Norms | Reliability | Construct validity | Criterion validity |
| 8 | REACT Responsive Environmental Assessment for Classroom Teaching | Nelson et al. (2014) | PE SE | 27 | RP | 4 | ✓ | ✓ | | | ✓ | ✓ | |
| 9 | SAFL-Q Assessment for Learning Questionnaire for students | Pat-El et al. (2013) | | 28 | RP | 4 | ✓ | ✓ | | | ✓ | ✓ | |
| 10 | TBQ-S Teaching Behavior Questionnaire | Possel et al. (2013) | | 37 | RP | 4 | ✓ | ✓ | | | ✓ | ✓ | |
| 11 | TLP-SE Teachers Learn from Pupils-Secondary Education | van Petegem et al. (2008) | | 53 | RP/FL | 5 | ✓ | ✓ | ✓ | | ✓ | ✓ | |
| 12 | WHIC What is Happening In this Class? | Dorman (2003) | | 56 | RP/EA | 5 | ✓ | ✓ | | | ✓ | ✓ | |
| 13 | SLEI Science Laboratory Environmental Inventory | Fraser et al. (1993) | | 34 | RP/M | 4 | ✓ | ✓ | | | ✓ | ✓ | ✓ |

(continued)

Table 1 (continued)

| Instrument | General information | | | | Evaluation results | | | | | | | | |
|------------|--|------------------------------------|--------------------------|--------|----------------------------|--------------|---------------------|-----------------------|---------------------|-------|-------------|--------------------|--------------------|
| | Full name | Main reference | Target pop. ^a | #Items | Assess. purp. ^b | Scale points | Theory ^c | Material ^d | Manual ^e | Norms | Reliability | Construct validity | Criterion validity |
| SPQ | | | PE | SE | | | | | | | | | |
| 14 | Learners' perception questionnaire from the SIBO project | Vandenbergh et al. (2011a, 2011b) | ✓ | | 77 | RP | 5 | ✓ | ✓ | | ✓ | ✓ | |
| 15 | Patterns of Adaptive Learning Scales | Midgley et al. (1998) | ✓ | ✓ | 94 | RP | 5 | ✓ | ✓ | ✓ | ✓ | | |
| 16 | Teacher As Social Context (short form & long form) | Belmont et al. (1988) | | ✓ | 52 | RP | 4 | ✓ | ✓ | | ✓ | ✓ | |
| 17 | PIRLS learners' perception questionnaire | Mullis et al. (2012) | ✓ | | 36 | RP | 4 | ✓ | ✓ | ✓ | | | |
| 18 | Tripod education partners | Tripod Educational Partners (2014) | ✓ | ✓ | 35/27 | RP/M | 5/3 | ✓ | ✓ | | ✓ | ✓ | ✓ |
| 19 | Questionnaire on instructional behavior ("vragenlijst Instructiegedrag") | Lamberts et al. (1999) | | ✓ | 33 | RP | 5 | ✓ | ✓ | | ✓ | ✓ | |

(continued)

Table 1 (continued)

| Instrument | General information | | | Evaluation results | | | | | | | | | |
|------------|--|-------------------------------|--------------------------|--------------------|----------------------------|--------------|---------------------|-----------------------|---------------------|-------|-------------|--------------------|--------------------|
| | Full name | Main reference | Target pop. ^a | #Items | Assess. purp. ^b | Scale points | Theory ^c | Material ^d | Manual ^e | Norms | Reliability | Construct validity | Criterion validity |
| 20 | QTI Questionnaire on Teacher Interaction ("Vragenlijst Interactief Leraargedrag") | Wubbels and Levy (1883) | PE ✓ SE ✓ | 50 | RP/M/FL | 5 | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| 21 | QLA Questionnaire on Lesson Activities ("Vragenlijst Lesactiviteiten") | den Brok et al. (1997) | ✓ | 96 | RP | 5 | ✓ | ✓ | | ✓ | ✓ | ✓ | |
| 22 | ZEBO ZEBO self-evaluation instrument for schools | Hendriks and Bosker (2003) | ✓ | 70 | RP/FL | 3 | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ |

Note ^aPE = Primary education; SE = Secondary education, ^bAssess. purp. = SPQ Assessment purpose, RP = Research practice; FL = Feedback loop; EA = Evidentiary argument; M = Measurement. ^cTheory = The theoretical basis of the instrument. ^dMaterial = Quality of the material. ^eManual = Quality of the manual

Table 2 Nine teaching quality constructs in the SPQs

| | SPQ | Climate ^a | Management ^b | Motivation ^c | Explanation ^d | Quality of representation ^e | Cognitive activation ^f | AfL ^g | Differentiation ^h | Learning strategies ⁱ | Total |
|----|--------|----------------------|-------------------------|-------------------------|--------------------------|--|-----------------------------------|------------------|------------------------------|----------------------------------|-------|
| 1 | SAES | ✓ | | | | | | | | | 1 |
| 2 | COLES | ✓ | | ✓ | ✓ | | ✓ | ✓ | ✓ | | 6 |
| 3 | CES | ✓ | ✓ | ✓ | ✓ | | ✓ | | | | 5 |
| 4 | MCI | ✓ | | ✓ | | | ✓ | ✓ | ✓ | | 5 |
| 5 | ICEQ | ✓ | | ✓ | | | ✓ | | ✓ | | 4 |
| 6 | SPOCQ | | | ✓ | ✓ | | ✓ | | ✓ | ✓ | 5 |
| 7 | ICALT | ✓ | ✓ | | ✓ | | ✓ | | ✓ | ✓ | 6 |
| 8 | REACT | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | 6 |
| 9 | SAFL-Q | | | | | | | ✓ | | | 1 |
| 10 | TBQ-S | ✓ | ✓ | ✓ | ✓ | | | | | | 4 |
| 11 | TLP-SE | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | | ✓ | 7 |
| 12 | WHC | ✓ | | ✓ | ✓ | | ✓ | | ✓ | | 5 |
| 13 | SLEI | ✓ | ✓ | ✓ | | ✓ | ✓ | | | | 5 |
| 14 | SIBO | ✓ | | ✓ | | ✓ | | ✓ | | | 4 |
| 15 | PALS | ✓ | ✓ | ✓ | | | | | | | 3 |
| 16 | TASC | ✓ | ✓ | ✓ | ✓ | | | | ✓ | | 5 |

(continued)

Table 2 (continued)

| SPQ | Climate ^a | Management ^b | Motivation ^c | Explanation ^d | Quality of representation ^e | Cognitive activation ^f | AfL ^g | Differentiation ^h | Learning strategies ⁱ | Total |
|--------------|----------------------|-------------------------|-------------------------|--------------------------|--|-----------------------------------|------------------|------------------------------|----------------------------------|-------|
| 17 | PIRLS | | ✓ | | | | | | | 1 |
| 18 | Tripod | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | | 7 |
| 19 | QIB | ✓ | ✓ | ✓ | | | | | | 4 |
| 20 | QTI | ✓ | ✓ | | | | | | | 2 |
| 21 | QLA | ✓ | ✓ | ✓ | | ✓ | ✓ | | | 6 |
| 22 | ZEBO | ✓ | ✓ | ✓ | | ✓ | | ✓ | | 5 |
| Total | 18 | 12 | 18 | 13 | 3 | 12 | 8 | 10 | 3 | 97 |

Note ^aClimate = Safe and stimulating learning climate. ^bManagement = Classroom management. ^cMotivation = Involvement/motivation of students. ^dExplanation = Explanation of subject matter. ^eQuality of representation = Quality of subject representation. ^fCognitive activation = Cognitive activation of the students during the lesson. ^gAfL = Assessment for Learning. ^hDifferentiation = Differentiated/adaptive instruction. ⁱLearning Strategies = Teaching and learning strategies and student self-regulation

items covered the theoretical constructs. Although the theoretical basis of the SPQs was evaluated in this review, it was not feasible to evaluate the quality of the research underlying the questionnaire, as well. Constructs also differed across SPQs in the way they were operationalized in items.

Standard two of the evaluation framework (the quality of the material) was met by 86% of the SPQs. A list of items (material) was useful for evaluation and two SPQs lacked such a list (except for items that were included in publications). Evaluation standard three considered the quality of the user guide. A user guide was available in five questionnaires, from which one out of the five was considered inadequate. Norms for the SPQ scores (standard four) were only given in two of the 22 SPQs. In one of the two cases in which the norms were given, norms were not satisfactory (ZEBO). Sample sizes were small and the research conducted on the norms was weak.

Regarding the psychometric quality of the SPQs (standard 5), for all but one SPQ, information about reliability was available and considered acceptable. A wide range of measurements was reported, for example, percentages of exact agreement between students, Cohen's kappa coefficient, or Cronbach's alpha correlation coefficient. The different measures of reliability are hard to compare because some imply absolute student score agreement while others imply student score consistency. For example, Fraser et al. (1993) reported the internal consistency using Cronbach's alpha (.70–.95), while van Petegem et al. (2008) tested the split-half reliability of the results using the Spearman-Brown coefficient on the student level (r_s .77–.90) and class level (r_s .84–.95) and they presented the r-output of the Mokken scaling analyses as yet another measure of reliability. However, for most SPQs, an acceptable reliability coefficient was reported in at least one of the publications.

For all but two SPQs, information about construct validity (standard 6) was available provided by an exploratory or confirmatory factor analysis (EFA or CFA). Factor loadings on the teaching quality construct an SPQ aimed to measure were acceptable in all SPQs. Five SPQs reported measures of criterion validity (standard 7). For example, Tripod Educational Partners (2014) demonstrated the criterion validity of the 7Cs framework by examining the correlation between the 7Cs and other commonly used measures of teaching effectiveness. Another example is Maulana et al. (2015), who examined predictive validity by evaluating the link between the scores on the questionnaire and student self-report of academic engagement. The criterion validity was acceptable in these studies.

5 Conclusion, Discussion, and Next Steps

Generating valid and reliable measurements by means of a student perception questionnaire (SPQ) is not self-evident. However, as an overview of the SPQs that have been developed and of their quality has been lacking, users have been hampered in making deliberate choices with respect to which SPQ to use in their own context (for examples, see Chap. 8 by Wisniewski and Zierer, in this volume). From a scientific

perspective, it is also valuable to evaluate the extent to which SPQs meet the general standards for measurement instruments. Therefore, this review gives an overview of SPQs that are available for measuring teaching quality in primary and secondary education. It provides information about the (psychometric) quality of the SPQs and about what constructs they measure. After conducting a systematic review, 22 SPQs were evaluated based on seven evaluation standards. The evaluation was conducted by two reviewers. After reviewing all SPQs separately, the reviewers came to a final judgment.

5.1 *What Was Learned*

The results of this study show that the reviewers were most positive about the theoretical basis of the SPQs and about the quality of the SPQ materials. For most SPQs, the reviewers were positive about the availability of empirical evidence regarding the reliability and validity of scores (except criterion validity). Moreover, the scoring tools in most SPQs were based on theory, research, and/or national standards. According to the reviewers, most SPQs also had acceptable reliability and construct validity. This could be partly explained by the well-known publication bias, in which only acceptable and confirmatory results get written about in order to increase the probability of publication (Falagas & Vangelis, 2008).

However, norm information about the quality rating measures was often lacking and few sampling specifications were provided. Information about the features of the SPQs, if available, was often not presented in an accessible way by instrument developers (e.g., in a user manual) making it difficult for potential SPQ users to obtain an overview of the qualities of available SPQs and to decide which SPQs best fits their own context and intended use. This raises questions about the factors that may have caused this. For example:

- There might have been no perceived need to make the SPQ materials accessible to other users, as not all SPQs in this review were developed for use by third parties (for example, some were developed by researchers for use in their own research project) and/or instrument developers were not aware that their SPQ could be of use to others.
- The required resources, such as time and financing, might not have been available to develop materials for external users, to make SPQ materials accessible, and to keep the information on the SPQs updated.
- Questionnaire developers might not be aware of the information SPQ users need for proper SPQ use and/or to decide whether the SPQ is useful in their own context and for their own assessment purpose.
- Conventional standards for research into the qualities of SPQs might not be available.

5.2 *Limitations of the Study*

It is important to note that the evaluation framework used in this study was based on the COTAN evaluation standards of Evers et al. (2010). These standards were developed for psychological tests in the social sciences and it is possible that the COTAN evaluation standards are not the best way to evaluate SPQs, although the conclusions drawn from the results would be similar in both cases.

Although a systematic search was conducted for all available SPQs, questionnaires could be missing from the final overview. Search terms were chosen, inclusion and exclusion criteria were formulated, and different search strings were used. Twenty-two questionnaires were found and evaluated in this review. With slightly different search strings, more or less extensive criteria, or with searching in different databases, other SPQs (e.g., in other languages) could have been found. This might also have to do with the timeframe during which the review was conducted. All information was gathered in 2016. Since multiple SPQs could still have been under development at that time, this review might not include all the information that is now available. However, there is no reason to believe that the conclusions of this review no longer apply.

5.3 *Next Steps*

In line with Dobbelaer (2019), who reviewed classroom observation systems, the creation of an international database with all SPQs, that have been developed, their materials and the research into them, could make SPQ developers more aware of the value of their SPQ for others. If SPQ materials are accessible to others, SPQs can also be researched and developed further (even if the instrument developers themselves did not have the resources to do so), which could also help to reduce the need to constantly develop new SPQs.

Furthermore, a standardized evaluation framework with quality standards and research standards for SPQs, designed to evaluate the quality of SPQs and the empirical evidence for the reliability and validity of SPQ scores, could make instrument developers more aware of the complexity of SPQ development and the research that is needed (for potential users). It could also make it easier for potential SPQ users to compare (the empirical evidence regarding) SPQs. Moreover, independent evaluations of the SPQs using a standardized evaluation framework can provide potential users with the information they need to make a well-informed choice of a SPQ as well as to examine the aims of the SPQs, the operationalization and definitions of the constructs being measured, the different ways to use them and to make sure the desired information will be obtained.

Appendix: Search Terms

| S1 | S2 | S3 | | S4 | | S5 | |
|---------------|-----------------------|---------|-------|------------|---------|------------------------|--------------------------|
| Evaluat* | Perception | Student | | Primary | | Valid | |
| Assess* | Feedback | Pupil | | Secondary | | Reliab* | |
| Questionnaire | “Teacher observation” | Learner | | Elementary | | Unreliab* | |
| Instrument | “student* view” | | | | | Repeated | |
| Survey | “student* voice” | | | | Educat* | “internal consistency” | |
| | | | + N10 | | School* | “item correlation” | (continue S5) |
| | | | | | | “item selection” | Test-retest |
| | | | | | | “internal correlation” | (test AND retest) |
| | | | | | | “factor analysis” | Stability |
| | | | | | | Cronbach* | “factor analyses” |
| | | | | | | Psychometr* | Dimension* |
| | | | | | | “internal consistency” | Subscale* |
| | | | | | | Error | “interscale correlation” |
| | | | | | | Rat* scale | Alpha |
| | | | | | | Kappa* | Correlation* |
| | | | | | | “perception variation” | Interrater |

(continued)

References

- Alkharusi, H. (2011). Development and datametric properties of a scale measuring students' perceptions of the classroom assessment environment. *International Journal of Instruction*, 4, 105–120.
- American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. American Educational Research Association.
- Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering* [About the use of continuous standardization]. CITO.
- Bell, C., & Aldridge, J. M. (2014). Investigating the use of student perception data for teacher reflection and classroom improvement. *Learning Environments Research*, 17, 371–388.
- Bell, C., Dobbelaer, M. J., Klette, K., & Visscher, A. J. (2018). *Qualities of classroom observation systems: School effectiveness and school improvement*. <https://www.tandfonline.com/doi/10.1080/09243453.2018.1539014?scroll=top&needAccess=true>. Accessed 6 Aug 2020.
- Bell, C. A. (2019). *What we see depends on how we look*. QUINT Conference.
- Belmont, M., Skinner, E., Wellborn, J., & Connell, J. (1988). *Teacher as social context: A measure of student perceptions of teacher provision of involvement, structure, and autonomy support (tech. rep. no. 102)*. University of Rochester.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research in Higher Education*, 137–172.
- Bijlsma, H. J. E., Glas, C., & Visscher, A. J. (Under review). Are student perceptions reliable measures to evaluate teaching quality? *Assessment*.
- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45, 378–387. <https://doi.org/10.3102/0013189X16659442>.
- Creemers, B. M. P. (1994). *The effective classroom*. Cassall.
- den Brok, P., Brekelmans, M., & Wubbels, T. (1997). *Vragenlijst lesactiviteiten*. IVLOS.
- Dobbelaer, M. J. (2019). *The quality and qualities of classroom observation systems*. Ipskamp Printing.
- Dobbelaer, M. J., Visscher, A. J., & Janssens, F. J. G. (2015, June). Review van lesobservatie-instrumenten [Review of classroom observation systems]. *Onderwijs Research Dagen*.
- Dorman, J. P. (2003). Cross-national validation of the what is happening in this class? (WIHIC) questionnaire using confirmatory factor analysis. *Learning Environment Research*, 6, 231–245. <https://doi.org/10.1023/A:1027355123577>.
- Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie* [Test theory]. Bohn Stafleu van Loghum.
- Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). *COTAN beoordelingsstelsel voor de kwaliteit van tests* [COTAN evaluation system for the evaluation of test quality]. Heijnis & Schipper.
- Falagas, M. E., & Vangelis, G. A. (2008). The top-ten in journal impact factor manipulation. *Archivum Immunologiae et Therapiae Experimentalis*, 56, 223–226. <https://doi.org/10.1007/s00005-008-0024-5>.
- Ferguson, R. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, 94, 24–28. <https://doi.org/10.1177/003172171209400306>.
- Ferguson, R., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems*. Jossey-Bass.
- Fraser, B. J., & Fisher, D. L. (1983). Development and validation of short forms of some instruments, measuring student perceptions of actual and preferred classroom learning environment. *Science Education*, 67, 115–131.
- Fraser, B. J., McRobbie, C. J., & Giddings, G. J. (1993). Development and cross-national validation of a laboratory classroom environment instrument for senior high school science. *Science Education Assessment Instruments*, 77, 1–24.

- Gentry, M., & Owen, S. V. (2004). Secondary student perceptions of classroom quality: Instrumentation and differences between advanced/honors and non-honors classes. *Journal of Secondary Gifted Education*, 16, 20–29.
- Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10, 255–282.
- Hendriks, M., & Bosker, R. J. (2003). *ZEBO: Instrument voor zelfevaluatie in het basisonderwijs* [ZEBO: instrument for self-evaluation in primary education]. Twente University Press.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*. American Council on Education and Praeger Publishers.
- Lamberts, R., Den Brok, P., Derksen, K., & Bergen, T. (1999). Het concept activerende instructie gemeten via perceptie van leerlingen [The concept of activating instruction measured through student perceptions]. *Pedagogische studiën*, 76, 36–50.
- Liberati, A., Altman, D. G., Tetzlaff, J., Murrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med*, 6, e1000100. <https://doi.org/10.1371/journal.pmed.1000100>.
- Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford University Press. <https://doi.org/10.1177/1049731508318552>.
- Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, 9, 215–230. <https://doi.org/10.1007/s10984-006-9014-8>.
- Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*. Springer. https://doi.org/10.1007/1-4020-5742-3_9.
- Marzano, R. J. (2012). The two purposes of teacher evaluation. *Educational Leadership*, 70, 14–19.
- Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: Construct representation and predictive quality. *Learning Environments Research*, 19, 335–357.
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, 26, 169–194.
- Midgley, C., Kaplan, A., Middleton, M., Urdan, T., Maehr, M. L., Hicks, L., & Roeser, R. W. (1998). Development and validation of scales assessing students' achievement goal orientation. *Contemporary Educational Psychology*, 23, 113–131.
- Mislevy, R. J. (2013). *Four metaphors we need to understand assessment: Report of The Gordon Commission on the future of assessment in education*. Educational Testing Service.
- Moher, D., Liberati, A., Etzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, 6.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international reading results*. TIMSS & PIRLS International Study Center.
- Nelson, P. M., Demers, J. A., & Christ, T. J. (2014). The responsive environmental assessment for classroom teaching (REACT): The dimensionality of student perceptions of the instructional environment. *School Psychology Quarterly*, 29, 182–197. <https://doi.org/10.1037/spq0000049>.
- Nunnally, J. C. (1967). *Psychometric theory*. McGrawhill.
- Pat-El, R., Jonathan, R. J., Tillema, H., Segers, M., & Vedder, P. (2013). Validation of assessment for learning questionnaires for teachers and students. *British Journal of Educational Psychology*, 83, 98–113. <https://doi.org/10.1111/j.2044-8279.2011.02057.x>.
- Pianta, R. C., & Hamre, B. K. (2006). Conceptualization, measurement, and improvement of classroom processes: Standardized observation van leverage capacity. *Educational Researcher*, 38, 109–119. <https://doi.org/10.3102/0013189X09332374>.
- Possel, P., Moritz-Rudasill, K., Adelson, J. L., Bjerg, A. C., Wooldridge, D. T., & Black, S. W. (2013). Teaching behavior and well-being in students: Development and concurrent validity of

- an instrument to measure student-reported teaching behavior. *International Journal of Emotional Education*, 5, 5–30.
- Praetorius, A. K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of three basic dimensions. *ZDM Mathematics Education*, 50, 407–426.
- Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. Institute of Education.
- Schulz, J., Sud, G., & Crowe, B. (2014). *Lessons from the field: The role of student surveys in teacher evaluation and development*. Bellwether Education Partners.
- Strong, M. (2011). *The highly qualified teacher: What is teacher quality and how do we measure it?* Teachers College Press.
- ten Berge, J. M. F., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69, 613–625. <https://doi.org/10.1007/BF02289858>.
- Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. Ministry of Education.
- Tripod Educational Partners U.K. (2014). *Tripod's 7 cs. technical manual*. Tripod Educational Partners.
- van der Lans, R. M., & Maulana, R. (2018). The use of secondary school student ratings of their teacher's skillfulness for low-stake assessment and high-stake evaluation. *Studies in Educational Evaluation*, 58, 112–121.
- van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. (2018). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement*.
- van Petegem, P., Deneire, A., & de Maeyer, S. (2008). Evaluation and participation in secondary education: Designing and validating a self-evaluation instrument for teachers to solicit feedback from pupils. *Studies in Educational Evaluation*, 34, 136–144. <https://doi.org/10.1016/j.stueduc.2008.07.002>.
- Vandenbergh, N., Cortois, L., De Bilde, V., Verschueren, K., & van Damme, J. (2011a). *Longitudinaal onderzoek in het basisonderwijs. Leerlingperceptie vragenlijst in het zesde leerjaar* [Longitudinal research in primary education: Student perception questionnaire in the fourth grade]. Steunpunt SSL.
- Vandenbergh, N., Cortois, L., De Bilde, V., Verschueren, K., & van Damme, J. (2011b). *Longitudinaal onderzoek in het basisonderwijs. Leerlingvragenlijst einde basisonderwijs* [Longitudinal research in primary education: Student perception questionnaire for sixth grade]. Steunpunt SSL.
- Vos, H. J., & Knuver, J. W. M. (2000). Standaarden in onderwijsevaluatie [Standards in educational evaluation]. In R. J. Bosker (Ed.), *Onderwijskundig lexicon (Editie III), Evalueren in het onderwijs* [Educational lexicon (Edition III), Evaluation in education]. Samsom.
- Wallace, T., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching. *American Educational Research Journal*, 53(6), 1834–1868. <https://doi.org/10.3102/0002831216671864>.

Wubbels, T., & Levy, J. (1883). *Do you know what you look like? Interpersonal relationships in education*. The Falmer Press.

Hannah Bijlsma is a researcher (Ph.D.) at the section of Teacher Professionalization at the University of Twente (The Netherlands) and as a primary school teacher (grade 1). Her research focuses on measuring teaching quality and on the use of student perceptions of teaching quality in school contexts. In 2016 she founded a professional association for academic primary school teachers, of which she has been chairman for about five years. She is now a board member of the International Congress for School Effectiveness and Improvement (ICSEI) and a board member of the EARLI SIG on School Effectiveness and Improvement.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5

A Probabilistic Model for Feedback on Teachers' Instructional Effectiveness: Its Potential and the Challenge of Combining Multiple Perspectives



Rikkert van der Lans

Abstract This chapter describes research into the validity of a teacher evaluation framework that was applied between 2012 and 2016 to provide feedback to Dutch secondary school teachers concerning their instructional effectiveness. In this research project, the acquisition of instructional effectiveness was conceptualized as unfolding along a continuum ranging from ineffective novice to effective expert instructor. Using advanced statistical models, teachers' current position on the continuum was estimated. This information was used to tailor feedback for professional development. Two instruments were applied to find teachers' current position on the continuum, namely the *International Comparative Assessment of Learning and Teaching (ICALT)* observation instrument and the *My Teacher–student questionnaire (MTQ)*. This chapter highlights background theory and central concepts behind the project and it introduces the logic behind the statistical methods that were used to operationalize the continuum of instructional effectiveness. Specific attention is given to differences between students and observers in how they experience teachers' instructional effectiveness and the resulting disagreement in how they position teachers on the continuum. It is explained how this disagreement made feedback reports less actionable. The chapter then discusses evidence of two empirical studies that examined the disagreement from two methodological perspectives. Finally, it makes some tentative conclusions concerning the practical implications of the evidence.

Keywords Teacher evaluation · Teaching quality · Measures · Teacher development · Feedback

1 Introduction

Tess is a school leader who must decide how to spend the school's resources on professional development. Today she has an assessment interview with John of whom

R. van der Lans (✉)
LUMC-Curium - Child and Adolescent Psychiatry, Leiden, The Netherlands
e-mail: r.m.van_der_lans@lumc.nl

she has just visited a lesson yesterday. Her rubric scorings signal some clear directions for improvement in the clarity of John's front class explanations, yet the student questionnaire administered one month ago gives few signs of poor explanation skills. Instead, the student results signal that John could improve on the interactivity of his instructions. Tess wants to use the interview to plan and guide John's further professionalization. However, how can she use the available information to provide John with actionable feedback that likely adds to John's teaching? Also, Tess knows that John wants to discuss opportunities to participate in further training. Yet, the evidence of John's instructional skill is inconclusive. Hence, on what grounds should she accept or refuse the request?

Imaginary situation based on conversations with school leaders and teachers.

The last decade teacher evaluation has had a central position in policies aiming to improve educational quality in many countries (e.g., Doherty & Jacobs, 2013; Isoré, 2009; Nusche et al., 2014). In the Dutch context, the Ministry of Education published the "teacher agenda" which documented several challenges, objectives, and policy measures meant to increase the quality of the Dutch teacher workforce. One objective was to increase the frequency of performance evaluations in schools (Nusche et al., 2014; OECD, 2016). In the eyes of the policy-makers, performance evaluations were a means to turn schools into "learning organizations" by functioning as a yearly update and a reminder of teachers' and school leaders' commitment to increase educational quality. To realize this objective, the councils for primary (PO-raad) and secondary (VO-raad) education and the teacher labor unions together agreed to install a new differentiated payment system and to assign every teacher a personal professionalization budget (Nusche et al., 2014; OECD, 2016). This created an incentive for teachers to request for a performance evaluation interview to discuss evidence of instructional effectiveness. If evidence of effectiveness was insufficient to qualify for a salary raise, the teacher should be informed about steps and/or skills required to qualify. Teachers could use their personal professionalization budget to train these skills. In practice these policies implied that school leaders, like Tess, were confronted with the task to distinguish between "average", "good" and "excellent" teachers, to give teachers feedback about what they needed to learn, and to organize the conditions under which teachers can start to learn.

The research project described in this chapter took place within this context and examined statistical methods and models that could assist school leaders to distinguish between teachers in terms of their level of instructional effectiveness. Furthermore, the new methods needed to result in feedback that clearly indicated a specific direction for improvement. Instruments used to collect data were student questionnaires and classroom observations. These two instruments were chosen because they share the strength that they collect direct observations of teachers' classroom behavior (Darling-Hammond, 2013; Goe et al., 2008; Peterson, 2000). However, as the situation of Tess and John shows, the feedback resulting from the student questionnaire and the feedback resulting from the classroom observation instrument do not always agree.

The chapter starts with an introduction of central concepts and how these were operationalized. The applied methods are introduced at a conceptual level and it is discussed how the models relate to other statistical methods that are commonly applied. After the background section, the chapter focuses on the problem of agreement between feedback sampled with student and classroom observation instruments.

2 Background Theory and Definitions of Central Concepts

In the sketch at the beginning of this chapter, Tess is wondering how she may use student questionnaires and classroom observation instruments as two complementary instruments to provide John with actionable feedback. This highlights some central concepts of this chapter, namely instructional effectiveness, improvement, and actionable feedback.

2.1 *Instructional Effectiveness*

In this chapter, instructional effectiveness is viewed as an estimation of the degree to which teachers' classroom behavior is expected to give students the opportunity to maximize their learning potential. By stating that instructional effectiveness provides students with opportunities to learn it is clarified that instructional effectiveness is associated with, but not identical to, student achievement and school success which are realizations of these opportunities. Furthermore, the definition clarifies that instructional effectiveness is estimated meaning that any claim about it is surrounded by some level of uncertainty. The research described in this chapter has operationalized instructional effectiveness using two instruments, namely the International Comparative Assessment of Learning and Teaching (ICALT)—which is a classroom observation instrument—and the My Teacher questionnaire—which is a student questionnaire. The ICALT and My Teacher questionnaire instruments conceptualize an effective instructor as a teacher that scores high on six domains of instruction. The six domains are labeled “safe and stimulating learning climate”, “efficiency of classroom management”, “clear and structured explanations”, “intensive and interactive instructions”, “teaching students learning strategies”, and the “adaptation of instructions to individual student needs”. Table 1 details the conceptualizations of each domain. Several studies from different countries provide evidence suggesting that the items included in the ICALT and My Teacher questionnaire cluster according to these six domains (André et al., 2020; Maulana & Helms-Lorenz, 2016; van de Grift et al., 2011).

Table 1 An overview of the six domain and their conceptualization

| Domain | Conceptualization |
|--|---|
| • Safe and stimulating learning climate | A safe and stimulating learning climate is established when the teacher and students trust and respect each other. |
| • Efficient classroom management | An efficient classroom management is structured by clear procedures, routines, and rules about where and how learning takes place. |
| • Clear and structured explanations | Clear and structured explanations prompt students' prior knowledge, emphasize critical knowledge, and regularly checks on students' comprehension of content. |
| • Intensive and interactive instructions | Intensive and interactive instructions stimulate teacher–student and student–student interaction by questioning, collaborative group work, having students explain topics to one another, or asking students to think aloud. |
| • Teaching students learning strategies | Teaching students learning strategies enhance students' metacognitive skills and self-regulated learning. |
| • Adaptation of instructions to individual student needs | Adaptation of instruction means that teachers adjust their instructional practice to specific students' learning needs by, for example, allowing flexible time to complete assignments or providing additional explanation to small groups. |

2.2 *Improvement*

Another central term in this chapter is improvement, which suggests that teachers can learn or be trained to become more effective instructors. Analogous to Berliner's (2004) novice-expert continuum, the research project and evidence discussed here conceptualizes the improvement of instructional effectiveness as unfolding along a continuum ranging from completely ineffective instruction to completely effective instruction. To illustrate how we operationalized this continuum we start with a one-item example, "This teacher uses time efficiently". Research suggests that more effective instructors in general use time more efficiently than less effective instructors (Muijs et al., 2014). The *x*-axis in Fig. 1 visualizes the continuum of effective instruction. The *y*-axis represents the probability on a positive item score. In line with the above statement, Fig. 1 predicts that highly effective instructors have near 100% probability on positive scores and that low effective instructors have near 0% probability. Furthermore, only when teachers have acquired a certain level of instructional effectiveness, are they predicted to start to learn how to use time efficiently. This can be inferred from Fig. 1 by observing that the probability on positive scores starts to rise at a certain location on the continuum. Key to the perspective taken in this

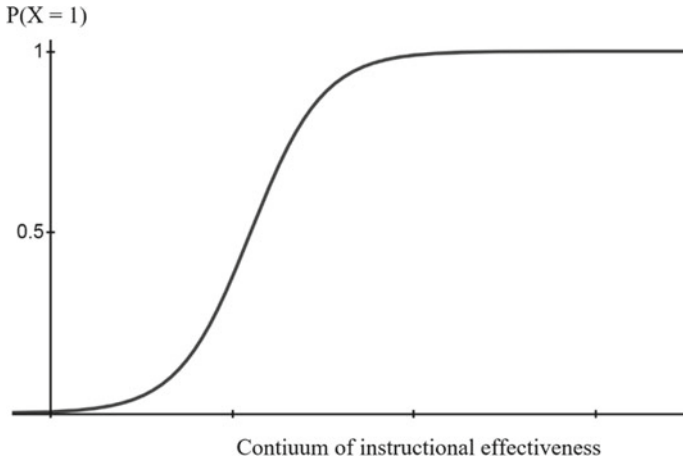


Fig. 1 Increase in probability on a positive response to the classroom observation item “This teacher uses time efficiently”

chapter is that training of teachers’ instructional effectiveness is optimal when it is focused on items that match the teacher’s location on the continuum.

2.2.1 Relation of the Applied Relatively Novel Statistical Model to Other Statistical Models

The proposition that all items measure instructional effectiveness are associated with a single continuum seems conflicting with research that groups items of instructional effectiveness according to dimensions of teaching quality which load on separate (statistical) factors (e.g., studies applying factor analysis). Figure 2 is used to discuss and visualize the relationship between the continuum discussed in this chapter and factor analysis results. Figure 2 again visualizes the continuum of instructional effectiveness, but now includes multiple items. Solid, dashed, and dotted lines indicate clusters of items that have high(er) inter-item correlations (i.e., load on separate factors). The reader can move the icon directly below the Figure to logically derive this. For example, teachers positioned at the icon have an approximately 50% probability on positive scores on the dotted items, but near 0% probability on positive scores on the dashed and solid ones. When the teacher moves up the continuum, the probability on positive scores on the dashed items increases first, while the probability on the dotted items remains high and the solid items remains low. Thus, some teachers likely have high scores on dotted items and low on all other items, other teachers likely have high scores on the dotted and high on the dashed items but low on the solid items, and yet others likely have high scores on all items. However, it is unlikely that teachers score high on the solid but low on the other items. This scoring pattern, which in factor analytic literature is referred to as the simplex pattern, is

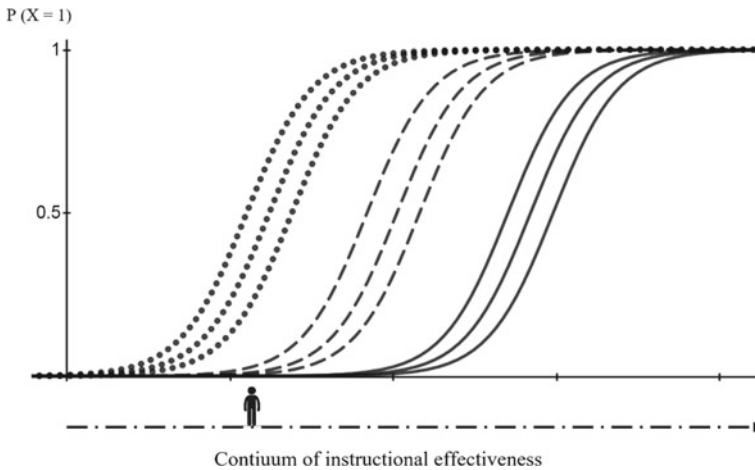


Fig. 2 The continuum of instructional effectiveness in which indicators are grouped into three factors

detected by factor analysis as a sign that the dotted items and dashed items are in distinct clusters (interested readers may consult Browne [1992] or Jöreskog [1978] for further details). It is acknowledged that the Figure presents an oversimplification of the relationship of factor analysis with the continuum described in the chapter. For example, item slopes, i.e., the steepness with which the s-curved lines increase, are rarely exactly parallel. Such differences in item slope also impact on the assignment of items to factors. However, Fig. 2 illustrates the basic rationale of how different factors on a single continuum.

2.2.2 The Sequence of Clusters Along the Continuum

The research group at the University of Groningen has given much empirical attention to the ordering of these factors along the continuum of instructional effectiveness (e.g., Maulana et al., 2015a, 2015b; van de Grift et al., 2011, 2014; van der Lans et al., 2015, 2018, 2021). The results indicated an ordering of the factors in the sequence in which the factors are presented in Table 1, thus: (1) safe and stimulating learning climate, (2) efficient classroom management, (3) clear and structured explanations, (4) intensive and interactive instructions, (5) teaching students learning strategies, and (6) adaptations of instructions to the individual students' learning needs. The validity of this ordering was further corroborated by other research in Cyprus that applied the same statistical models to their own questionnaire and observation instruments and which reported broadly similar results (e.g., Kyriakides et al., 2009, 2018).

Based on these results, the author developed a feedback report to provide teachers with information of our best estimate of their current position on the continuum of instructional effectiveness and our best estimate of what the teacher could improve

on next. Figure 3 presents two reports that were applied by the author to give teacher feedback. The left report concerns the classroom observation instrument and the right concerns the student questionnaire instrument. The reports show a table with three columns. In the column “level” are the six identified levels (or domains) of instructional effectiveness. The column “item” lists the items included in the instruments. Finally, the column, “teacher score” indicates what probably went well (darkest grey top area), what probably can be learnt next (lightest grey middle area), and what probably is beyond the teachers’ competency to learn yet (grey lowest area). The Asterisk indicates the exact teacher position on the continuum of instructional effectiveness.

2.3 Actionable Feedback

The third central concept in this chapter is actionable feedback. Cannon and Witherspoon (2005) describe actionable feedback as feedback that leads to learning and increased performance. Evidence indicates that feedback is actionable when, (1) it is directed at the task—not the person, (2) it is unambiguous and specific, and (3) it has clear implications for action (Cannon & Witherspoon, 2005; Kluger & DeNisi, 1996). The aim was to design feedback reports such that they would assist the feedback giver, which presumably is a school leader or coach, to be actionable. Therefore, the reports emphasize on what the teacher does (e.g., the teacher involves students, explains clearly), it attempts to communicate as specific as possible about what went well and what can be improved. In addition, reports accompanied this information with specific implications that would now require action. Participating teachers found this approach informative and often recognized themselves. Nonetheless, the actionability of the feedback was hindered by various organizational and psychometric factors. Organizationally, schools lacked an infrastructure to support training activities at this level of precision. Though, this was not part of the research project discussed here, it is nonetheless important to mention. Psychometrically, the disagreement between students and observers created uncertainty about the reliability of the estimates. Take, for example, the two feedback reports in Fig. 3. The students on average positioned the teacher on the item “my teacher involves me in the lesson” and signal that the teacher should focus to improve on the clarity and structuredness of explanation. The classroom observer, however, positioned the teacher on the item “encourages students to apply what they have learnt” and signals that the teacher should improve on teaching students learning strategies. Moreover, the observer’s report suggests no problems with the clarity and structuredness of the teacher’s explanations. Hence, in case of disagreement the feedback reports no longer unambiguously communicate what went well and what domains of instruction needed improvement. Also, the implications for action were no longer clear.

| Teacher: Class: School: | #Name #Name #Name | Level | Item: | My teacher... | Teaching skill score | |
|-------------------------------|-------------------------|-------|---|---|--|--|
| | | | safe learning climate safe learning climate efficient classroom management safe learning climate clear and structured explanation efficient classroom management clear and structured explanation clear and structured explanation safe learning climate efficient classroom management clear and structured explanation clear and structured explanation clear and structured explanation clear and structured explanation intensive and interactive instructions intensive and interactive instructions intensive and interactive instructions intensive and interactive instructions intensive and interactive instructions intensive and interactive instructions intensive and interactive instructions teaching learning strategies teaching learning strategies teaching learning strategies teaching learning strategies adaptation of instructions adaptation of instructions teaching learning strategies teaching learning strategies teaching learning strategies adaptation of instructions adaptation of instructions | shows respect for students in behavior and language creates a relaxed atmosphere ensures effective class management supports student self-confidence explains the subject matter clearly ensures that the lesson runs smoothly gives well-structured lessons uses learning time efficiently clearly explains teaching tools and tasks ensures mutual respect checks during processing whether students are carrying out tasks properly gives feedback to students involves all students in the lesson encourages students to do their best checks during instruction whether students have understood the subject matter uses teaching methods that activate students provides interactive instruction asks questions that encourage students to think encourages students to reflect on solutions explains the lesson objectives at the start of the lesson has students think out loud boosts the self-confidence of weak students encourages students to think critically teaches students how to simplify complex problems encourages students to apply what they have learned checks whether the lesson objectives have been achieved adapts processing of subject matter to student differences encourages the use of checking activities asks students to reflect on approach strategies teaches students to check solutions offers weak students additional learning and instruction time adapts instruction to relevant student differences | approaches me with respect. helps me if I do not understand answers my questions. prepares his/her lessons well. makes sure that I know what to do. explains how I need to do things. makes clear what I need to learn for a test. makes sure that I treat others with respect. states clearly when assignments/tasks are due. makes sure that I pay attention. makes clear to me why my answers are good or not. uses clear examples. pays attention to me explains everything clearly to me. involves me in the lesson. makes sure that I do my best. explains the purpose of the lesson clearly. makes sure that I use my time effectively. encourages me to think. states the lesson objectives. repeats what we have learnt in the previous lesson. makes sure that I keep on working. motivates me to think. makes connections to what I already know. asks me questions that I need to think about. takes into account what I already know. motivates me. stimulates me to cooperate with my classmates. checks whether I have understood the content of the lesson. talks interestingly. makes me feel self-confident with difficult tasks. teaches me to check my solutions. tells how I should learn something. knows what I have difficulty with. lets me explain to him/her how I tackled the task/assignment. lets me explain the content of the lesson to other students. lets me summarize the content of the lesson. asks me how I am going to learn the content of the lesson. | |
| | | | Safe Learning Climate Safe Learning Climate Efficient Classroom Management Efficient Classroom Management Clear and Structured Explanation Safe Learning Climate Efficient Classroom Management Clear and Structured Explanation Clear and Structured Explanation Efficient Classroom Management Safe Learning Climate Clear and Structured Explanation Clear and Structured Explanation Efficient Classroom Management Clear and Structured Explanation Intensive and Interactive Instruction Clear and Structured Explanation Efficient Classroom Management Intensive and Interactive Instruction Adaptation of Instructions Intensive and Interactive Instruction Intensive and Interactive Instruction Adaptation of Instructions Intensive and Interactive Instruction Adaptation of Instructions Intensive and Interactive Instruction Intensive and Interactive Instruction Intensive and Interactive Instruction Adaptation of Instructions Intensive and Interactive Instruction Adaptation of Instructions Teaching Learning Strategies Teaching Learning Strategies Adaptation of Instructions Teaching Learning Strategies Teaching Learning Strategies Teaching Learning Strategies | | * | |

Fig. 3 ICALT (left) and My Teacher (right) feedback reports. Darkest-grey = probably skilled (top area), light grey (middle area) = try to learn now, grey (lowest area) = probably unskilled in (Note These two feedbacks are slightly adapted versions of the ones that were originally reported to teachers)

3 Prior Research on the Disagreement Between Classroom Observation and Student Questionnaires

Prior research suggests that the disagreement between students and observers may be frequent and/or substantial. Studies documenting correlations between observation and survey measures mostly report modest correlations in the range of 0.15–0.30 (e.g., De Jong & Westerhof, 2001; Ferguson & Danielson, 2014; Howard et al., 1985; Martínez et al., 2016; Maulana & Helms-Lorenz, 2016). Designs varied considerably between studies, however. For example, De Jong and Westerhof (2001) report on the correlation between a classroom observation instrument and a student questionnaire that had considerably different factor structure, whereas Maulana and Helms-Lorenz report on the correlation of the ICALT and My Teacher questionnaire that have an overlapping factor structure. Because the study by Maulana and Helms-Lorenz (2016) applied the same instruments, their results are most relevant to the discussion in this chapter. They report a correlation of 0.26. This correlation was replicated in the data that was used in the research project that is reported on in this chapter. This modest correlation suggests that feedback reports of students and observers will more often disagree than agree. An exception to the above list of studies reporting modest correlations is the study by Murray (1983), who reports a correlation of 0.76. We will return on Murray's study somewhat later in this chapter.

4 Studying Evidence of Agreement and Disagreement Between Questionnaires and Classroom Observation Instruments

Two perspectives can be taken to compare the feedback reports presented in Fig. 3. The first perspective focuses on the teacher's position and as we have seen this leads to the conclusion that the students and observers disagree. The alternative perspective focuses on the ordering of the items and domains on the continuum of instructional effectiveness. From this perspective the students and observers mostly agree. Both the classroom observation and student questionnaire feedback report start with items related to safe and stimulating learning climate and end with items related to teaching students learning strategies and adaption of instruction to individual students' learning needs. The only two domains that are ordered differently by the two methods are the just mentioned final two domains.

Van der Lans et al. (2019) went one step further and showed that the My Teacher–student questionnaire and the ICALT classroom observation instrument items can be concurrently calibrated on the same continuum of instructional effectiveness. Table 2 lists the joint item ordering mixing observation and questionnaire items. Items denoted with an “s” are student questionnaire items and items denoted with an “o” are classroom observation items.

Table 2 Item ordering that resulted from the concurrent calibration of ICALT observation and My Teacher questionnaire items. This table was originally published in van der Lans et al. (2019) (O = ICALT observation item; S = My teacher questionnaire item)

| Level | Item | Description: This/My teacher... |
|-------------------------|------|---|
| Climate | O1 | Shows respect for students in behavior and language |
| Climate | S21 | Treats me with respect |
| Climate | O2 | Creates a relaxed atmosphere |
| Management | S20 | Prepares his/her lesson well |
| Management | O7 | Ensures effective class management |
| Climate | O3 | Supports student self-confidence |
| Climate | S40 | Helps me if I do not understand |
| Clear explanation | O9 | Explains the subject matter clearly |
| Climate | S6 | Answers my questions |
| Management | O5 | Ensures that the lesson runs smoothly |
| Climate | O4 | Ensures mutual respect |
| Clear explanation | O14 | Gives well-structured lessons |
| Management | S3 | Makes clear what I need to study for a test |
| Management | O8 | Uses learning time efficiently |
| Management | S19 | Makes clear when I should have finished an assignment |
| Climate | S8 | Ensures that I treat others with respect |
| Climate | S1 | Ensures that others treat me with respect |
| Clear explanation | S13 | Explains the purpose of the lesson |
| Clear explanation | S24 | Uses clear examples |
| Management | S23 | Ensures that I pay attention |
| Management | S26 | Applies clear rules |
| Management | O6 | Checks during processing whether students are carrying out tasks properly |
| Management | S2 | Ensures that I use my time effectively |
| Clear explanation | O15 | Clearly explains teaching tools and tasks |
| Clear explanation | O10 | Gives feedback to students |
| Clear explanation | O11 | Involves all students in the lesson |
| Clear explanation | S39 | Involves me in the lesson |
| Clear explanation | O13 | Encourages students to do their best |
| Clear explanation | S33 | Ensures that I know the lesson goals |
| Interactive instruction | S17 | Encourages me to think for myself |
| Interactive instruction | O19 | Asks questions that encourage students to think |
| Interactive instruction | S12 | Ensures that I keep working |
| Interactive instruction | O16 | Uses teaching methods that activate students |
| Interactive instruction | S30 | Stimulates my thinking |

(continued)

Table 2 (continued)

| Level | Item | Description: This/My teacher... |
|-------------------------|------|---|
| Interactive instruction | O21 | Provides interactive instruction |
| Clear explanation | O12 | Checks during instruction whether students have understood the subject matter |
| Interactive instruction | O20 | Has students think out loud |
| Differentiation | S25 | Connects to what I am capable of |
| Differentiation | S34 | Checks whether I understood the subject matter |
| Learning strategies | O30 | Encourages students to apply what they have learned |
| Learning strategies | S16 | Teaches me to check my own solutions |
| Learning strategies | O31 | Encourages students to think critically |
| Differentiation | S36 | Knows what I find difficult |
| Differentiation | O23 | Checks whether the lesson objectives have been achieved |
| Learning strategies | O28 | Encourages the use of checking activities |
| Learning strategies | O29 | Teaches students to check solutions |
| Differentiation | O25 | Adapts processing of subject matter to student differences |
| Differentiation | O26 | Adapts instruction to relevant student differences |

Studying Table 2 teaches us that similarly phrased questionnaire and observation items occasionally have similar positions on the continuum. Examples are S39 “my teacher involves me in the lesson” (student) and O11 “this teacher involves all students in the lesson” (classroom observation) and: S17 “my teacher encourages me to think for myself” (student) and O19 “this teacher asks questions that encourage students to think” (classroom observation). However, the questionnaire and classroom observation instrument also contained several items that were instrument unique, but which nevertheless could be calibrated on the same continuum. The considerable overlap between the questionnaire and observation instrument has clear practical implications. Suppose that two reports in Fig. 3 would locate a teacher on items related to the same domain—e.g., both on the domain clear and structured explanation—then the observation and student feedback reports give identical suggestions for improvement and, thus, would be more actionable. That is, there is no thinkable scenario in which feedback reports are actionable and in which the ordering of teaching behaviors (items) on the continuum varies between the methods.

The combination of agreement in item ordering and disagreement in teacher location also has theoretical implications, because these findings do not fit well with most prior beliefs, theory, and hypotheses concerning the disagreement between questionnaire and observation instruments. For example, a long-standing tradition in educational psychology studies biases in instrument scores. Bias is generally examined by regressing the teacher scores on variables other than instructional effectiveness. Studies in the MET project, like Martínez et al. (2016), regressed the “teacher scores” on variables that were hypothesized to bias measurement. This resulted in a set of ‘bias-corrected’ teacher scores related to the student questionnaire and a set

of ‘bias-corrected’ teacher scores related to the classroom observation instrument. These corrected scores were then correlated. However, the resulting correlations were similar to the correlations reported in studies that do not correct for bias (cf. Maulana & Helms-Lorenz, 2016; Martínez et al., 2016). More in general, hypotheses reflecting the belief that inferences based on scores need to be corrected for bias do not fit well with the evidence discussed so far. When scores obtained with the instruments are biased and not indicative of instructional effectiveness, then how can we explain the high similarity in the item ordering along the continuum.

The evidence also is difficult to align with another prominent hypothesis, namely the perspective-specific validity hypothesis stated by Kunter and Baumert (2006). Kunter and Baumert proposed that, despite disagreement, scores obtained with distinct instruments can be used to make valid inferences about teachers’ instructional effectiveness, given that the instruments are well-designed and administered. Kunter and Baumert do not clearly define what they mean with “perspectives” and this makes it complex to empirically assess their hypothesis (see also Fauth et al., 2020). However, many seem to understand different perspectives as meaning that some instruments might be more sensitive to tap certain aspects of instructional effectiveness. The difference in sensitivity explains the modest correlation. Also, using multiple instruments could help to offset blind spots thereby allowing for a fuller and richer picture of instructional effectiveness. The current evidence is insufficient to completely verify this idea, but an analysis of the unique items in Table 2 provides surprisingly limited support for it. Take, for example, the item S3 “my teacher makes clear what I need to study for a test”. Despite that this item mentions unique content (“what students need to study for a test” is not part of the ICALT observation list because observers usually cannot know this), the item S3 does not have a unique position on the continuum. We could leave out item S3 without losing much information about teachers’ instructional effectiveness. As another example, take item S34 “my teacher checks whether I understood the subject matter”. The phrasing “whether I understood” focuses on the individual student and such focus is not included in any of the classroom observation instrument items. Nonetheless, item S34 is very closely located to the item O12 “this teacher checks during instruction whether students have understood the subject matter” which measures the same content, but has observers focus on the “average” student in the class. In sum, the evidence provides no strong indications that differences between instruments in terms of item content and item focus can explain disagreement in how students and observers position teachers on the continuum of teaching effectiveness.

The disagreement in teachers’ position on the continuum was examined in another study. Central in that study was the hypothesis that disagreement in teachers’ position on the continuum changes as a function of measurement reliability (van der Lans, 2018). Central in the study were two claims of which the correctness was empirically assessed. First, it was claimed that the scores assigned by observers reflect the average student in the class. Therefore, agreement was expected to increase when classroom observation scores were correlated with class average questionnaire score, instead of scores assigned by a single student. Secondly, it was claimed that student responses to questionnaire items reflect the teachers’ typical teaching across

many lessons. Therefore, the agreement was expected to increase when classroom observation scores sampled from different lessons were averaged. The study applied generalizability theory to test these predictions and found support for both of them. The predicted correlation is lowest when scores of a single student's questionnaire are correlated with one classroom observation score of one single lesson and the more student questionnaires are sampled the higher the predicted correlation with the observation score of a single lesson becomes. Also, the predicted correlation increases when the observation scores are aggregated over multiple lessons, and, again, the more lessons are sampled the higher the predicted correlation. The study results suggest that the correlation between questionnaire and classroom observation instruments increases to 0.76 when the classroom observation scores concern an aggregate of seven different lesson visits performed by seven different observers and when the student questionnaire is administered in the same class and spans scores of 25 different students. This correlation of 0.76 was interesting because of its correspondence to the correlation reported by Murray (1983), which was also 0.76. Murray estimated that correlation based on the aggregate classroom observation score of six to eight lesson visits by three different observers and a student questionnaire administered in the same class and which score was aggregated over all students in the class. The increase in the expected correlation between the questionnaire and classroom observation instrument is graphically presented in Fig. 4. The y-axis in Fig. 4 gives the predicted correlation. The x-axis indicates the number of students in the class. The separate lines indicate how predictions differ when the number of classroom

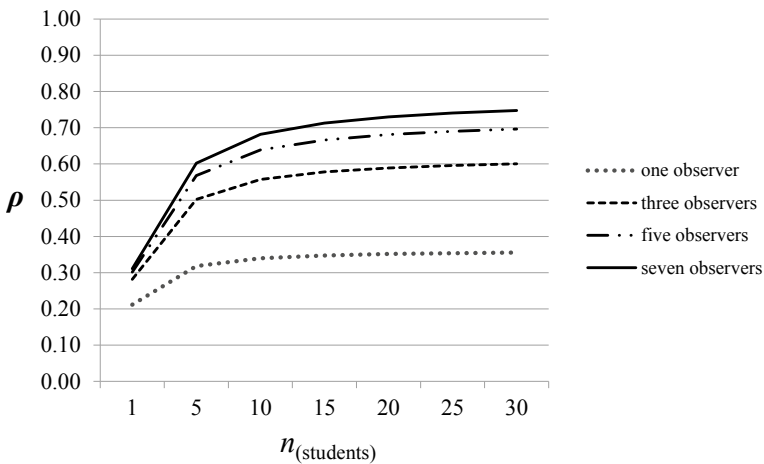


Fig. 4 Predicted increase in correlations (ρ) between the MTQ student questionnaire and ICALT classroom observation instrument for an increasing number of administered classroom observations. The correlations apply when questionnaires and classroom observations are performed within the same class and span no more than one school year. Van der Lans (2018) reports predictions related to other situations (e.g., questionnaires and observations spanning different classes)

observers and lesson moments sampled with the classroom observation instrument varies.

The implications of the results in Fig. 4 are not yet well understood. There are varying possible interpretations. One interpretation is that more valid inferences are made with the questionnaire and classroom observation instruments when scores are aggregated over many students and over many lessons and observers, respectively. This interpretation aligns well with studies suggesting that the reliability of single student questionnaire scores is unreliable (Marsh, 2007) and that classroom observations of one single lesson are unreliable snap-shots (Hill et al., 2012; Praetorius et al., 2014; van der Lans et al., 2016). This interpretation has considerable implications for the number of classroom observations and student questionnaires that need to be administered at schools. However, this interpretation does not align well with the rationale behind the hypotheses of van der Lans (2018). That is, the only reason why it is predicted that the correlation between the classroom observation instrument and the student questionnaire increases as a function of the number of lesson visits included in the aggregate is that the students are also expected to aggregate their experiences across many lessons when scoring the questionnaire. If questionnaires would be able to tap students' experiences about one particular lesson, then it would be predicted that the correlation is highest when the questionnaire results are correlated with classroom observations concerning that same particular lesson. This study was unable to empirically examine this, however. Similarly, the finding that questionnaire scores obtained with single students have low correlation with the classroom observation scores is, in the study by van der Lans (2018), explained by the assumption that classroom observers score the instructional effectiveness towards the "average" student. If observers would have been instructed to score the classroom observation items in relation to one particular student, the correlation is predicted to be highest for the particular student-observer dyad (compared to other dyads). Again, the study was unable to empirically examine this claim.

5 Discussion and Conclusion

5.1 *Potential Implication Teacher Evaluation in Schools*

Based on the above discussions, what can we advise Tess and John? What we can say is that the evidence so far suggests that single classroom visits rarely will show agreement with a single administration of the student questionnaire. Also, the evidence generally indicates that this has little to do with the interpretation of item content by students and observers. The item ordering estimated across all students is very similar to the item ordering across all observers. This does not imply that the quality of the item phrasing is unimportant, however. The evidence indicates that when items are well-formulated the students can score items related to the same domains of instructional effectiveness very similar. We might advise Tess to postpone the

performance evaluation interview and administer some more classroom observations. The evidence presented in this chapter indicates that this increases the chance on agreement. However, it is not always possible to schedule additional classroom observations. Alternatively, we might advise Tess and John to focus on one result. Perhaps John wants to improve on his instructional effectiveness when teaching certain subject matter and because the classroom observation visit took place when John was teaching this particular subject matter, the classroom observation results are favored over the student questionnaire results. However, while this last advice might be intuitive to some, we must acknowledge that it is full of untested claims and hypotheses.

5.2 *What to Do Next?*

One direction for future research concerns the construction of student questionnaires that can help us to make valid inferences about the instructional effectiveness of single lessons. The Impact! tool might be a potential example of such an instrument (Bijlsma et al., 2019). The alternative item phrasing of the Impact! questionnaire provides opportunities to assess the hypothesis that correlations between student questionnaires and classroom observation instruments are attenuated because students aggregate their experiences over many lessons when scoring regular questionnaire items.

Another direction for future research concerns the commonly shared understanding among researchers that some instruments have higher sensitivity to measure certain aspects/behaviors of instructional effectiveness and that using multiple instruments could help to offset blind spots. When instruments have blind spots concerning the measurement of instructional effectiveness, then we would expect “gaps” in the continuum of instructional effectiveness. Such “gaps” can only become visible when multiple instruments are concurrently calibrated to continuum. The resulting ordering in item positions may reveal that items of one instrument have unique locations on the continuum. The current evidence assessing this idea is very limited. Hopefully, the counterintuitive result—few evidence supporting the idea—motivates future researchers to improve on the study designs, content of the instruments and psychometric methods applied by van der Lans et al. (2019) to more thoroughly study this idea empirically.

References

- André, S., Maulana, R., Helms-Lorenz, M., Telli, S., Chun, S., Fernández-García, C. M., et al. (2020). Student perceptions in measuring teaching behavior across six countries: A multi-group confirmatory factor analysis approach to measurement invariance. *Frontiers in Psychology*, *11*, 273.

- Berliner, D. C. (2004). Describing the behavior and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society*, 24(3), 200–212.
- Bijlsma, H. J., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, 28(2), 217–236.
- Browne, M. W. (1992). Circumplex models for correlation matrices. *Psychometrika*, 57, 469–497.
- Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Perspectives*, 19(2), 120–134.
- Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press.
- De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research*, 4(1), 51–85.
- Doherty, K. M., & Jacobs, S. (2013). *Connect the dots—using evaluations of teacher effectiveness to inform policy and practice*. National Council on Teacher Quality.
- Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift für Pädagogik. Beiheft*, 66(1), 138–155.
- Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems*. Wiley.
- Goe, L., Bell, C., & Little, O. (2008). Approaches to evaluating teacher effectiveness: A research synthesis. *National Comprehensive Center for Teacher Quality*.
- Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When interrater-reliability is not enough: Teacher observation systems and a case for the generalizability theory. *Educational Researcher*, 41, 561cat1. <https://doi.org/10.3102/0013189X12437203>.
- Howard, G. S., Conway, C. G., & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. *Journal of Educational Psychology*, 77(2), 187–196.
- Isoré, M. (2009). *Teacher evaluation: Current practices in OECD countries and a literature review*. OECD Education Working Papers, No. 23. OECD Publishing (NJ1).
- Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika*, 43(4), 443–477.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254.
- Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9(3), 231–251.
- Kyriakides, L., Creemers, B. P. M., & Antaniou, P. (2009). Teacher behavior and student outcomes: Suggestions for research on teacher training and professional development. *Teaching and Teacher Education*, 25, 12–23.
- Kyriakides, L., Creemers, B. P., & Panayiotou, A. (2018). Using educational effectiveness research to promote quality of teaching: The contribution of the dynamic model. *ZDM Mathematics Education*, 50(3), 381–393.
- Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective*, (pp. 319–383). The Netherlands, Dordrecht: Springer.
- Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, 38(4), 738–756.
- Maulana, M., & Helms-Lorenz, R. (2016). Observations and student perceptions of pre-service teachers' teaching behavior quality: Construct representation and predictive quality. *Learning Environments Research*, 19(3), 335–357. <https://doi.org/10.1007/s10984-016-9215-8>.

- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015a). Development and evaluation of a survey measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, 26(2), 169.
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015b). Pupils' perceptions of teaching behaviour: Evaluation of an instrument and importance for academic motivation in Indonesian secondary education. *International Journal of Educational Research*, 69, 98–112.
- Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, 25(2), 231–256.
- Murray, H. G. (1983). Low-inference classroom teaching and student ratings of college teaching effectiveness. *Journal of Educational Psychology*, 75(1), 138–149.
- Nusche, D., Braun, H., Halász, G., & Santiago, P. (2014). *OECD reviews of evaluation and assessment in education: Netherlands 2014*. OECD Reviews of Evaluation and Assessment in Education, OECD Publishing. <https://doi.org/10.1787/9789264211940-en>.
- OECD. (2016). *Netherlands 2016: Foundations for the future*. OECD Publishing. <https://doi.org/10.1787/9789264257658-en>.
- Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practice*. Corwin Press.
- Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, 31, 2–12.
- van de Grift, W. J. C. M., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, 43, 150–159. <https://doi.org/10.1016/j.stueduc.2014.09.003>.
- van de Grift, W. J. C. M., van der Wal, M., & Torenbeek, M. (2011). Ontwikkeling in de pedagogische didactische vaardigheid van leraren in het basisonderwijs [Primary teachers' development of pedagogical didactical skill]. *Pedagogische Studiën*, 88, 416–432.
- van der Lans, R. M. (2018). On the “association between two things”: The case of student surveys and classroom observations of teaching quality. *Educational Assessment, Evaluation and Accountability*, 30(4), 347–366.
- van der Lans, R. M., Maulana, R., Helms-Lorenz, M., Fernández-García, C.-M., Chun, S., Jager, T., Iridayanti, Y., Inda-Caro, M., Lee, O., Coetzee, T., Fadhilah, N., Jeon, M., & Moorer, P. (2021). Student perceptions of teaching quality in five countries: A Partial Credit Model approach to assess measurement invariance. Manuscript submitted for publication.
- van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice*, 34(3), 18–27.
- van der Lans, R. M., van de Grift, W. J., & Van Veen, K. (2018). Developing an instrument for teacher feedback: Using the rasch model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education*, 86(2), 247–264.
- van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2019). Same, similar, or something completely different? calibrating student surveys and classroom observations of teaching quality onto a common metric. *Educational Measurement: Issues and Practice*, 38(3), 55–64.
- van der Lans, R. M., van de Grift, W. J., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, 50, 88–95.

Rikkert van der Lans is a Postdoctoral Researcher currently working at the Department of Child and Adolescent Psychiatry, LUMC-Curium Leiden (The Netherlands). Previously he worked as a Postdoctoral Researcher at the Department of Teacher Education of the University of Groningen, as a lecturer in methods and statistics at the Department of Educational Sciences (GION) of the University of Groningen, and as a lecturer in the field of psychometrics at the Department of

Methods and Statistics of the University of Tilburg (The Netherlands). His research focuses on the evaluation of professional development and the use of questionnaire and observation data to inform decisions.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6

Understanding (Dis)Agreement in Student Ratings of Teaching and the Quality of the Learning Environment



Jonathan D. Schweig and José Felipe Martínez

Abstract Student surveys are increasingly being used to collect information about important aspects of learning environments. Research shows that aggregate indicators from these surveys (e.g., school or classroom averages) are reliable and correlate with important climate indicators and with student outcomes. However, we know less about whether within-classroom or within-school variation in student survey responses may contain additional information about the learning environment beyond that conveyed by average indicators. This question is important in light of mounting evidence that the educational experiences of different students and student groups can vary, even within the same school or classroom, in terms of opportunities for participation, teacher expectations, or the quantity and quality of teacher–student interactions, among others. In this chapter, we offer an overview of literature from different fields examining consensus for constructing average indicators, and consider it alongside the key assumptions and consequences of measurement models and analytic methods commonly used to summarize student survey reports of instruction and learning environments. We also consider recent empirical evidence that variation in student survey responses within classrooms can reflect systematically different experiences related to features of the school or classroom, instructional practices, student background, or a combination of these, and that these differences can predict variation in important academic and social-emotional outcomes. In the final section, we discuss the implications for evaluation, policy, equity, and instructional improvement.

Keywords Student ratings · Learning environment · Teaching quality

J. D. Schweig (✉)
RAND Corporation, Arlington, VA, USA
e-mail: Jonathan_Schweig@rand.org

J. F. Martínez
University of California, Los Angeles, CA, USA
e-mail: jfmtz@ucla.edu

1 Introduction

Educators are increasingly turning to student surveys as a valuable source of information about important features of school and classroom learning environments, ranging from time on task and content coverage to more qualitative aspects of teaching—e.g., the extent to which classes are well-managed, teachers foster student cognitive engagement, or students feel emotionally, physically, and intellectually safe (Baumert et al., 2010; Klieme et al., 2009; Pianta & Hamre, 2009). Considerable research shows that student survey reports can be aggregated into reliable indicators of constructs that have been variously identified in the literature with terms like *learning environment*, *classroom climate*, *instructional practice*, or *teaching quality*. These constructs may or may not be exchangeable across areas of study, but irrespective of terminology, the literature shows that student survey aggregates tend to correlate significantly with each other, with indicators derived through other methods (e.g., classroom observation), and with a range of desirable student outcomes. However, there is a gap in research investigating whether within-classroom or within-school variability in such student survey responses may offer additional information beyond that conveyed by average indicators. This question is important in light of emerging evidence that the educational experiences of individual students can vary considerably within schools, and even within the same classroom, including opportunities for student participation (Reinholz & Shah, 2018; Schweig et al., 2020), and the quantity and quality of teacher–student interactions (e.g., Connor et al., 2009), among others. In this chapter we review literature that examines aggregate survey indicators in different fields, and consider the key assumptions and consequences of various measurement models and analytic methods commonly used to summarize student survey reports of teaching. We then examine the growing literature that investigates the variability in student survey responses within classrooms and schools, and whether this variation may relate to educational experiences and outcomes. We illustrate the potential implications of this kind of variation using a hypothetical example case. In the final section, we discuss the implications of this research for evaluation policy and instructional improvement.

2 Student Surveys, Teaching, and the Learning Environment

There are many reasons why educators are increasingly interested in student surveys as a source of information about learning environments. Perhaps most importantly, students can spend over 1,000 hours in their schools every year, and thus have unmatched depth and breadth of experience interacting with teachers and peers (Ferguson, 2012; Follman, 1992; Fraser, 2002). Students also provide a unique perspective compared to other reporters (Downer et al., 2015; Feldlaufer et al., 1988). Probing students about their perceptions of teaching and the learning environment

acknowledges their voice (Bijlsma et al., 2019; Lincoln, 1995), and the significance of their school-based experiences (Fraser, 2002; Mitra, 2007). Second, a growing body of research suggests that students can provide trustworthy information about important aspects of the learning environment (Marsh, 2007). For example, survey-based aggregate indicators can reliably distinguish among instructional practices (Fauth et al., 2014; Kyriakides, 2005; Wagner et al., 2013), and aspects of teaching quality (e.g., Benton & Cashin, 2012). These aggregates are furthermore significantly and positively associated with other measures of teaching quality (e.g., Burniske & Meibaum, 2012; Kane & Staiger, 2012).

Like other measures, student survey responses can be susceptible to error (e.g., recall, inconsistency in interpretation; see e.g., Popham, 2013; van der Lans et al., 2015), bias (e.g., acquiescence), and halo effects (perceptions of one aspect of teaching influencing those of other aspects; see e.g., Fauth et al., 2014; Chap. 3 by Röhl and Rollett of this volume) that may influence their psychometric properties (see for example, Follman, 1992; Schweig, 2014; Wallace et al., 2016). Nevertheless, most existing studies suggest that these biases are generally small in magnitude and do not greatly influence comparisons across teachers or student groups, or how aggregates relate with one another and with external variables (Kane & Staiger, 2012; Vriesema & Gehlbach, 2019). Research also demonstrates that aggregated student survey responses are associated with important student outcomes including academic achievement (Durlak et al., 2011; Shindler et al., 2016), engagement (Christle et al., 2007), and self-efficacy and confidence (e.g., Fraser & McRobbie, 1995).

Student surveys also have the benefit of being cost-effective, relatively easy to administer, and feasible to use at scale (e.g., Balch, 2012; West et al., 2018). This is a particular advantage when contrasted with other commonly used methods for measuring teaching and the quality of the learning environment, including direct classroom observation. In large school districts, an observation system closely tied to professional development can require dozens of full-time positions, with yearly costs in the millions of dollars (Balch, 2012; Rothstein & Mathis, 2013). As a result, the use of student surveys has seen remarkable growth over the last two decades for evaluating educational interventions (Augustine et al., 2016; Gottfredson et al., 2005; Teh & Fraser, 1994), and monitoring and assessing educational programs and practices (Hamilton et al., 2019). In particular, student surveys are commonly used to inform teacher evaluation and accountability systems—summatively as input for setting actionable targets (Burniske & Meibaum, 2012; Little et al., 2009), or formatively to provide feedback and promote teacher reflection and instructional improvement (Bijlsma et al., 2019; Gehlbach et al., 2016; Wubbels & Brekelmans, 2005).

3 Psychological Climate, Organizational Climate, and Student Surveys

In most contexts, schooling is an inherently social activity, and students typically experience schooling in organizational clusters (Bardach et al., 2019). The common pattern of student clustering within classrooms and schools presents challenges and choices in using surveys to understand teaching and the quality of the learning environment. One of the first choices is whether to focus the survey on understanding the personal perceptions and experiences of individual class members, or more broadly on shared elements of teaching quality relevant to the class or school as a whole (Bliese & Halverson, 1998; Den Brok et al., 2006; Echterhoff et al., 2009).

Surveys that aim to capture individual student interpretations of teaching quality or of the learning environment are described as reflecting *psychological climate*, and include items that ask for individual self-perceptions and personal beliefs (Glick, 1985; Maehr & Midgley, 1991). A long history of educational research suggests that psychological climate is a key proximal determinant of academic beliefs, behaviors, and emotions (Maehr & Midgley, 1991; Ryan & Grolnick, 1986). Because psychological climate variables treat individual perceptions as interpretable, it is appropriate to analyze them at the individual level (Stapleton et al., 2016), and differences among individual respondents are considered as substantively meaningful. Individuals can react in different ways to the same practices, procedures that seem fair to one individual might seem unfair to another individual, and so forth. Psychological climate variables can be aggregated to describe the composition of an organization (Sirotnik, 1980).

On the other hand, surveys that focus on the classroom or the school as a whole are described as reflecting *organizational climate* (see e.g., Lüdtke et al., 2009; Marsh et al., 2012), a concept that has a rich history in industrial and social psychology (Bliese & Halverson, 1998; Chan, 1998). Unlike psychological climate, organizational climate emerges from the collective perceptions of individuals as they experience policies, practices, and procedures (e.g., Hoy, 1990; Ostroff et al., 2003). Aggregating individual perceptions produces measures of organizational level phenomena (Sirotnik, 1980). These new variables can be interpreted to reflect an *overall* or shared perception of the environment (Lüdtke et al., 2009). The concept of organizational climate informs the design and use of many student surveys, which are typically directed toward students as a group, often asking for observations of the behavior of others (e.g., classmates, teachers; see Den Brok et al., 2006).

When conceived as measures of organizational climate, aggregating survey responses essentially positions students as informants or judges of a classroom or school level trait, similar to observers who would provide ratings using a standardized protocol. To illustrate this assumption, consider the following claims in Table 1 regarding three widely used student surveys.

Thus, while psychological climate variables treat interindividual differences as substantively interpretable, organizational climate variables emerge based on shared student experiences, and assume that students have similar mental images of their

Table 1 Measurement claims for three widely used surveys

| Survey | Claim |
|---------------------------------------|---|
| Tripod Survey | The variance between teachers provides the “signal” we are interested in... while the variability among students within a classroom may be regarded as “rater variance.” In effect, Tripod casts each student within a class as an informant, or rater, of the quality of the classroom; inconsistencies among student responses within a class are therefore regarded as “rater error” and are thus part of the measurement error <i>Source: Raudenbush and Jean (2014), p. 179</i> |
| National Education Longitudinal Study | In the measurement sense, students are considered <i>judges</i> or raters of the disciplinary climate of the school. If the variation of ratings within schools is small, we consider inter-rater agreement to be strong <i>Source: Ma and Willms (2004), p. 174</i> |
| Learning Environment Scale | Ratings obtained by multiple informants within a structural class (e.g., multiple students within a school or multiple teachers within a school) can be considered interchangeable because they share a more common role and a presumed more similar perspective than informants from different structural groups <i>Source: Konold and Cornell (2015)</i> |

classroom or school (Fraser, 1998). Students in a particular classroom or school are treated as *exchangeable* (Lüdtke et al., 2009), and interindividual differences are treated as idiosyncratic measurement error. Lüdtke et al. (2006, p. 207) noted that in the ideal scenario, “each student would assign the same rating, such that the responses of students in the same class would be interchangeable.” Because organizational climate variables treat individual perceptions as error, it is appropriate to analyze them at the classroom or school level (Stapleton et al., 2016). However, while the distinction between psychological and organizational variables is frequently drawn in the theoretical and methodological literature, much-applied literature does not explicitly or consistently consider student survey-based ratings of teaching quality and the learning environment as either psychological or organizational level measures (Lam et al., 2015; Schweig, 2014; Sirotnik, 1980). This in part reflects the fact that most student surveys occupy a gray area between these two classifications. On one hand classrooms and schools are shared spaces, students interact socially and build social relationships with their peers and with their teachers, and some aspects of teaching quality are more or less equally applicable to all students in the classroom (Lam et al., 2015; Urdan & Schoenfelder, 2006). At the same time, students’ school-based experiences can and often do differ, making their responses not exchangeable; students are not objective, external observers, but active participants involved in complex interactions with other students, teachers, and features of the classroom and school environment. Teachers often interact with students through multiple modes

and formats, both individually, and as a group (whole-class instruction, group work); and of course students interact directly with one another individually and as a group (Den Brok et al., 2006; Glick, 1985; Sirotnik, 1980).

4 Reporting Survey Results: Common Practices and Opportunities for Improvement

In the previous section, we argued that research often does not explicitly state the measurement assumptions that underlie their use of student surveys. In particular, researchers are not always explicit about the unit of interest (e.g., the individual or the group), and what this implies for the interpretability of individual student responses. These issues also arise in how survey developers choose to summarize and report survey results. In practice, nearly all survey platforms report measures of teaching and the quality of the learning environment by aggregating individual student responses to create classroom-level or school-level scores. It is these aggregates that are subsequently communicated to stakeholders or practitioners through data dashboards or survey reports (Bradshaw, 2017; Panorama Education, 2015). These aggregates can reflect simple averages (Balch, 2012; Bijlsma et al., 2019), percentages of respondents that report a certain experience or behavior (Panorama Education, 2015), or more sophisticated statistical models (e.g., IRT, or other latent variable models, see e.g., Maulana et al., 2014).

Irrespective of whether the survey developers are interested in individual or school- or classroom-level variables, this approach to score reporting often does not include information about the variability of student responses within classrooms or schools (Chan, 1998; Lüdtke et al., 2006). Thus, whether by accident or design, survey reports are ultimately firmly rooted in the notion of organizational climate in industrial or organizational psychology described previously: the shared learning environment is the central substantive focus, students are assumed to react similarly to similar external stimuli, and individual variation is assumed idiosyncratic or reflective of random measurement error (Chan, 1998; Lüdtke et al., 2009; Marsh et al., 2012).

However, while aggregated scores are useful for characterizing the overall learning experiences of a typical student, a growing body of research shows that these experiences can in fact vary greatly within schools and classrooms. Croninger and Valli (2009), for example, found that the vast majority (more than 80 percent) of the variance in the quality of spoken teacher–student exchanges occurred among lessons delivered by the same teachers. Den Brok et al. (2006) found that the majority of variance in student survey reports reflects differences among students within the same classroom (between 60 and 80 percent of the total variance). Crucially, emerging research also suggests that disagreement among students in their reports of the learning environment does not reflect only error, and indeed can provide important additional insights into teaching and learning, not captured by classroom or

school aggregates. In a study of elementary school students, Griffith (2000) found that schools with higher levels of agreement in student and parent survey reports of order and discipline tended to have higher levels of student achievement and parent engagement. Recent work by Bardach and colleagues (2019) found that within-classroom consensus on student reports of classroom goal structures was positively associated with socio-emotional and academic outcomes.

4.1 An Example Case of Within-Classroom Variability

Examining the distribution of student reports can open up possibilities for using information about the nature and extent of student disagreements for diagnostic and formative uses, and focused professional development opportunities for teachers, among others. The three hypothetical Classrooms in Fig. 1 illustrate how different within-classroom distributions can produce the same aggregate classroom climate rating (e.g., Lindell & Brandt, 2000; Lüdtke et al., 2006).

For the purposes of this example, students in each of these classrooms are asked about their perceptions of cognitive activation in the classroom, and the extent to which they are presented with questions that encourage them to think thoroughly and explain their thinking (Lipowski et al., 2009). Figure 1 displays the ratings provided by twenty students in each of the three classrooms. All three classrooms have the same average score of 3.42 on a 5-point scale.

In Classroom 1 there is noticeable disagreement in student survey responses, and students provide responses all across the allowable score range. In Classroom 2, there is also a lot of variability in student responses, but student perceptions seem polarized: there is a large group of students that feel very positively about the level

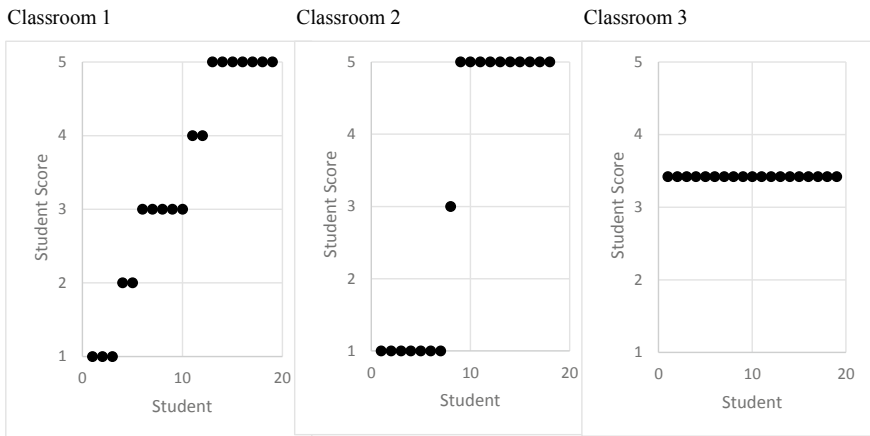


Fig. 1 Three hypothetical distributions of student climate ratings yielding the same average of 3.42

of cognitive activation, while a large group of students feel very negatively. Finally, in Classroom 3, there is perfect agreement among all students—this is the hypothetical ideal classroom described in Lüdtke and colleagues (2006) where all students experience classroom climate the same way. These scenarios raise important questions for practice. In principle it does not seem justifiable to give the three schools in Fig. 1 the same feedback and professional development recommendations for teachers—thus omitting the fact that the patterns of within-classroom variation are dramatically different. A more sensible approach would likely entail considering whether the within-classroom variability in student reports can potentially be informative for purposes of diagnosing and improving teaching quality. It is not possible to determine from this raw quantitative display *why* students in these three classrooms perceived cognitive activation in different ways. However, examining the distribution of student reports can open up possibilities for using this information for diagnostic and formative uses, and focused professional development opportunities for teachers. In the remainder of this chapter, we summarize and discuss relevant literature for understanding these interindividual differences.

5 School and Classroom Factors Associated with Variation in Student Perceptions of Teaching Quality

Within classrooms or schools, interindividual differences in the perception of teaching or the learning environment can arise for many reasons. We begin this section discussing the standard assumption invoked by common approaches to survey score reporting (that within-classroom or school variation reflects measurement error) and subsequently present four alternative interpretations that have support in the literature in other areas: (1) differential expectations and teacher treatment, (2) diversity of student needs and expectations, (3) diversity of student backgrounds, experiences, cultural values, and norms, and (4) teacher characteristics.

5.1 *Measurement Error*

Interindividual variability in student perceptions can be assumed to involve some idiosyncratic component of measurement error—i.e., random fluctuations around the “true” score of a school or classroom, related to memory, inconsistency, and unpredictable interactions among time, location, and personal factors. Individual students may also vary in terms of their standards of comparison (Heine et al., 2002), or the internal scales they use to calibrate their perceptions (Guion, 1973). This can create differences in student scores analogous to rater effects in studies of observational protocols: some students may be more lenient or severe than others. Thus, some differences among students are not substantively interpretable (Marsh et al., 2012;

Stapleton et al., 2016), Moreover, to the extent students are not systematically sorted into classrooms based on *stringency*, these differences are not expected to induce bias and are best treated as measurement error (West et al., 2018). If interindividual variability were idiosyncratic and random, however, we would generally not expect within-classroom student ratings to be associated with other measures of teaching quality or student outcomes. However, a number of prior studies have demonstrated that individual perceptions of school or classroom climate can be positively associated with student achievement. Griffith (2000) and Schweig (2016), found that learning environments with more intraindividual disagreements about order, discipline, and the quality of classroom management had lower academic performance, even holding average ratings constant. Schenke et al. (2018) found that lower levels of heterogeneity among students' perceptions of emotional support, autonomy support, and performance focus are negatively associated with mathematics achievement. Martínez (2012) found that individual perceptions of opportunity to learn (OTL) were predictive of reading achievement, even after controlling for class and school level OTL. Such findings strongly suggest that within-classroom variability in student reports is not entirely reflective of measurement error.

5.2 *Differential Expectations and Teacher Treatment*

Teacher expectations are a critical determinant of student learning (Muijs et al., 2014). Teachers may consciously or unconsciously have differential expectations for subgroups of students, which may translate into different sets of rules, classroom environments, and pedagogical strategies (Babad, 1993; Brophy & Good, 1974), potentially leading to opportunity gaps (Flores, 2007). Research has shown some teachers can have lower achievement expectations for students of color (Banks & Banks, 1995; Oakes, 1990). Teachers may also have lower achievement expectations for female students (Lazarides & Watt, 2015), and offer them less reinforcement and feedback (e.g., Simpson & Erickson, 1983). Teacher expectations may also differ based on perceptions of student ability. At higher grades, research has shown that prior academic achievement is the most significant influence on teacher expectations (Lockheed, 1976). More recent research suggests that learning tasks are often differentially assigned to students based on teacher beliefs about student ability. For example, “mathematically rich” instruction (tasks requiring reasoning and creativity, multiple concepts and methods, and application to novel contexts) is often reserved for students perceived to be *high-achieving*, while those perceived as lower achieving spend more time developing and practicing basic skills (Schweig et al., 2020; Stipek et al., 2001). Thus, within-classroom variability in student survey reports could point to suboptimal or inequitable participation opportunities and instructional experiences for students of different groups (Gamoran & Weinstein, 1998; Seidel, 2006), which may, in turn, result in achievement gaps (Voight et al., 2015).

5.3 *Diversity of Student Needs and Expectations*

Student perceptions of teaching and the learning environment may reflect different student needs and expectations—learning experiences and instructional practices that are successful with some students may not be effective with others, and student socio-emotional needs and expectations may also differ substantially within classrooms. Levy et al. (2003) provide an example that students with lower self-esteem may have greater needs with respect to the establishment of a supportive climate. Lüdtke et al. (2006) suggest that higher and lower ability students may differ in their perceptions of certain aspects of instructional practice, including pacing or task difficulty. English learners (ELs) and students with disabilities tend to report their schools to be less safe and supportive than their peers (Crosnoe, 2005; De Boer et al., 2013; Watkins & Melde, 2009). ELs face challenges with language comprehension, particularly with academic or mathematical language (Freeman & Crawford, 2008), and this may create differential perceptions of the clarity of classroom procedures. On the other hand, Hough and colleagues (2017) found that ELs had systematically more favorable perceptions of their teachers and classrooms than their peers on several aspects of climate. ELs students could be more engaged, more challenged, and better behaved, which influences their overall perception of the classroom (LeClair et al., 2009). In this way, ELs could also be more proactive at seeking out additional support from teachers, or that teachers are particularly sensitive to the needs of ELs (LeClair et al., 2009).

Alternatively, teachers may use instructional strategies that are responsive to and supportive of students' diverse needs and expectations, potentially causing student perceptions of the quality of their learning experiences to be *more similar*. For example, teachers may use complex instruction structured to promote student engagement, support critical thinking, and to connect content in meaningful ways to students' lives (Averill et al., 2009; Freeman & Crawford, 2008). Thus, to the extent that within-classroom agreement is associated with the use of instructional strategies responsive to students' diverse needs and expectations, there may be more equitable opportunities for all students. In a recent mixed-methods study of science classrooms, we found that classrooms with higher levels of student agreement tended to provide more collaborative learning opportunities for students, including more group work, and to have more structured systems for eliciting student participation (Schweig et al., n.d.).

5.4 *Diversity of Student Backgrounds, Experiences, Cultural Values, and Norms*

Reports of teaching and the quality of the learning environment may reflect cultural or contextual factors that cause students to perceive the learning environment differently (Bankston & Zhou, 2002; West et al., 2018). There is also research suggesting that

student perceptions of the learning environment may also differ by grade level (West et al., 2018). In the United States, research has shown that Black and Hispanic/Latino students often report feeling less connected to their schools, feel less positively about their relationships with teachers and administrators, and feel less safe in some areas of the school (Lacoe, 2015; Voight et al., 2015). However, recent literature suggests that this may not always be the case. Hough and colleagues (2017) found that while Black students had systematically lower ratings of school connectedness, discipline, and safety than their peers, Hispanic/Latino students tended to report systematically higher perceptions. These findings are not inherently at odds, and other literature suggests that perceptions of the learning environment can differ even from one area of the school to the other. Using data from New York City, Lacoe (2015) found that, for example, Black students have systematically lower perceptions of safety than their white peers in classrooms, but have systematically higher perceptions of safety in hallways, bathrooms, and locker rooms. In our own work, we found that classes with higher proportions of ELs and low-achieving students tended to have more intraindividual disagreements about teaching and the quality of the learning environment (Schweig, 2016; Schweig et al., 2017) in mathematics and science classrooms, and we also found significant within-classroom gaps between Black and white students on several aspects of teaching and the quality of the learning environment, with Black students typically having more positive perceptions relative to their white peers (Perera & Schweig, 2019).

The perception of some teacher behaviors, including the extent to which teachers make students feel cared for, may depend strongly on cultural conceptions of caring (Garza, 2009). Calarco (2011) highlighted several ways in which economically disadvantaged students help-seeking behaviors differed from their classmates in ways that could impact perceptions of teaching quality. Specifically, Calarco found that economically disadvantaged students sought less teacher assistance, and as a result, received less guidance from their teachers. Atlay and colleagues (2019) found that students from higher socioeconomic backgrounds were more critical about teacher assistance, perhaps reflecting a sense of entitlement (Lareau, 2002). Students' perceptions of teaching quality can also be influenced by out-of-school experiences. For example, there may be differential exposure to external stressors that influence feelings of school safety (Bankston & Zhou, 2002; Lareau & Horvat, 1999).

5.5 *Teacher Characteristics*

A number of teacher characteristics can influence survey-based reports. Past work, for example, has shown that student perceptions of teachers are associated with teacher experience, and in particular, that more experienced teachers are perceived as more dominant and strict (Levy & Wubbels, 1992). More experienced teachers, however, are not generally perceived as more caring or supportive by their students (Den Brok et al., 2006; Levy et al., 2003). Teacher race and ethnicity can also play a role in survey-based ratings of teaching quality. Newly emerging research suggests

that race-based disparities in perceptions of teaching quality can be ameliorated by the presence of teachers of color. Specifically, teacher–student race congruence may positively influence students’ perceptions of teaching quality (Dee, 2005; Gershenson et al., 2016). In our own research, however, we did not find evidence that observable teacher characteristics, including teacher race, gender, years of experience, and level of education explain variation in race-based perceptual gaps (Perera & Schweig, 2019; Schweig, 2016).

6 Conclusion

A growing body of evidence suggests that in considering instructional climate, researchers and school leaders may want to look beyond aggregate indicators, and consider also the extent of variation (or consensus) in student survey reports, as a potential indicator of important aspects of the school or classroom environment. In fact, the ability to capture within-school or within-classroom variability in student experiences is one of the defining strengths of student survey-based measures. Other commonly used measurement modes (including teacher self-report and structured classroom observations) are structurally not well-equipped to capture differential student experiences. Classroom observation protocols, for example, are typically not designed to measure whether or how teachers engage with individual students (Cohen & Goldhaber, 2016; Douglas, 2009). Student surveys, on the other hand, offer information that goes beyond typical experiences and can allow teachers and instructional leaders better understand how instruction, socio-emotional support, and other aspects of the learning environment are experienced by different students or groups of students.

Collectively, the research presented in this chapter suggests that variation in student survey reports of their learning environment may reflect a variety of factors and influences, ranging from strategic instructional choices, responsive pedagogy, and classroom structures implemented by teachers, varying needs and perceptions of particular students or groups of students, contextual factors, and the interactions among these. Importantly, variation can also reflect more pernicious influences like differential teacher expectations, and other structural disadvantages for some group of students. Our example case also raises important questions about whether within-school or within-classroom variability should be considered as ignorable measurement error when examining student survey reports of teaching quality and learning environments. Should we give the three classrooms in Fig. 1 the same feedback and professional development recommendations for teachers? Or is there evidence in the within-classroom variability in student reports that can potentially be informative for these purposes? Recent policy guidelines in the United States either explicitly require or implicitly move in the latter direction, advising education agencies to provide schools not only aggregated survey-based indicators, but also indicators disaggregated by student subgroup (Holahan & Batey, 2019; Voight et al., 2015). A growing consensus also sees attending to these subgroup differences as a key for

school-wide adoption of instructional improvement strategies that meet the learning needs of the most vulnerable students (Kostyo et al., 2018).

Considering the diversity of student perspectives and experiences can be particularly useful for informing efforts to promote equitable learning and outcomes. Ultimately, whether the climate is conceived as a psychological or organizational climate, or both, if subgroups of students experience school life in meaningfully different ways, reliance on aggregated survey indicators as measures of teaching quality can potentially obscure diagnostic information (Roberts et al., 1978), and compromise the validity and utility of these measures to inform teacher reflection or feedback, and other improvement processes within schools (Gehlbach, 2015; Lüdtke et al., 2006).

References

- Atlay, C., Tieben, N., Fauth, B., & Hillmert, S. (2019). The role of socioeconomic background and prior achievement for students' perception of teacher support. *British Journal of Sociology of Education*, 40(7), 970–991. <https://doi.org/10.1080/01425692.2019.1642737>.
- Augustine, C. H., McCombs, J. S., Pane, J. F., Schwartz, H. L., Schweig, J., McEachin, A., & Siler-Evans, K. (2016). *Learning from summer: Effects of voluntary summer learning programs on low-income urban youth*. RAND Corporation.
- Averill, R., Anderson, D., Easton, H., Maro, P. T., Smith, D., & Hynds, A. (2009). Culturally responsive teaching of mathematics: Three models from linked studies. *Journal for Research in Mathematics Education*, 157–186.
- Babad, E. (1993). Teachers' differential behavior. *Educational Psychology Review*, 5(4), 347–376. <https://doi.org/10.1007/BF01320223>.
- Balch, R. T. (2012). *The validation of a student survey on teacher practice*. Doctoral dissertation, Vanderbilt University, Nashville, TN.
- Banks, C. A., & Banks, J. A. (1995). Equity pedagogy: An essential component of multicultural education. *Theory Into Practice*, 34(3), 152–158. <https://doi.org/10.1080/00405849509543674>.
- Bankston III, C. L., & Zhou, M. (2002). Being well vs. doing well: Self esteem and school performance among immigrant and nonimmigrant racial and ethnic groups. *International Migration Review*, 36(2), 389–415. <https://doi.org/10.1111/j.1747-7379.2002.tb00086.x>.
- Bardach, L., Yanagida, T., Schober, B., & Lüftenegger, M. (2019). Students' and teachers' perceptions of goal structures—will they ever converge? Exploring changes in student-teacher agreement and reciprocal relations to self-concept and achievement. *Contemporary Educational Psychology*, 59. <https://doi.org/10.1016/j.cedpsych.2019.101799>.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., Klusmann, U., Krauss, S., Neubrand, M., & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180. <https://doi.org/10.3102/0002831209345157>.
- Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature*. IDEA Center. https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_50.pdf. Accessed 5 August 2020.
- Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Bernard, P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, 28(2), 217–236. <https://doi.org/10.1080/1475939X.2019.1572534>.
- Bliese, P. D., & Halverson, R. R. (1998). Group consensus and psychological well-being: A large field study. *Journal of Applied Social Psychology*, 28(7), 563–580. <https://doi.org/10.1111/j.1559-1816.1998.tb01720.x>.

- Bradshaw, R. (2017). *Improvement in Tripod student survey ratings of secondary school instruction over three years*. Doctoral dissertation, Boston University, Boston, MA.
- Brophy, J. E., & Good, T. L. (1974). *Teacher-student relationships: Causes and consequences*. Holt, Rinehart & Winston.
- Burniske, J., & Meibaum, D. (2012). *The use of student perceptual data as a measure of teaching effectiveness*. Texas Comprehensive Center. https://sedl.org/txcc/resources/briefs/number_8/find_ex.php. Accessed 5 August 2020.
- Calarco, J. M. (2011). "I need help!" Social class and children's help-seeking in elementary school. *American Sociological Review*, 76(6), 862–882. <https://doi.org/10.1177/0003122411427177>.
- Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, 83(2), 234. <https://doi.org/10.1037/0021-9010.83.2.234>.
- Christle, C. A., Jolivet, K., & Nelson, C. M. (2007). School characteristics related to high school dropout rates. *Remedial And Special Education*, 28(6), 325–339. <https://doi.org/10.1177/07419325070280060201>.
- Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45(6), 378–387. <https://doi.org/10.3102/0013189X16659442>.
- Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S., Piasta, S. B., Crowe, E. C., & Schatschneider, C. (2009). The ISI classroom observation system: Examining the literacy instruction provided to individual students. *Educational Researcher*, 38(2), 85–99.
- Croninger, R. G., & Valli, L. (2009). "Where is the action?" Challenges to studying the teaching of reading in elementary classrooms. *Educational Researcher*, 38(2), 100–108. <https://doi.org/10.3102/0013189X09333206>.
- Crosnoe, R. (2005). Double disadvantage or signs of resilience? The elementary school contexts of children from Mexican immigrant families. *American Educational Research Journal*, 42(2), 269–303. <https://doi.org/10.3102/00028312042002269>.
- de Boer, A., Pijl, S. J., Post, W., & Minnaert, A. (2013). Peer acceptance and friendships of students with disabilities in general education: The role of child, peer, and classroom variables. *Social Development*, 22(4), 831–844. <https://doi.org/10.1111/j.1467-9507.2012.00670.x>.
- Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter? *American Economic Review*, 95(2), 158–165. <https://doi.org/10.1257/000282805774670446>.
- Den Brok, P., Brekelmans, M., & Wubbels, T. (2006). Multilevel issues in research using students' perceptions of learning environments: The case of the questionnaire on teacher interaction. *Learning Environments Research*, 9(3), 199. <https://doi.org/10.1007/s10984-006-9013-9>.
- Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher*, 38(7), 518–521. <https://doi.org/10.3102/0013189x09350881>.
- Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2015). Measuring effective teacher-student interactions from a student perspective: A multi-level analysis. *The Journal of Early Adolescence*, 35(5–6), 722–758. <https://doi.org/10.1177/0272431614564059>.
- Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, 82, 405–432. <https://doi.org/10.1111/j.1467-8624.2010.01564.x>.
- Echterhoff, G., Higgins, E. T., & Levine, J. M. (2009). Shared reality: Experiencing commonality with others' inner states about the world. *Perspectives on Psychological Science*, 4(5), 496–521. <https://doi.org/10.1111/j.1745-6924.2009.01161.x>.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>.
- Feldlaufer, H., Midgley, C., & Eccles, J. S. (1988). Student, teacher, and observer perceptions of the classroom environment before and after the transition to junior high school. *The Journal of Early Adolescence*, 8(2), 133–156. <https://doi.org/10.1177/0272431688082003>.

- Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, 94(3), 24–28. <https://doi.org/10.1177/003172171209400306>.
- Flores, A. (2007). Examining disparities in mathematics education: Achievement gap or opportunity gap? *The High School Journal*, 91(1), 29–42. <https://doi.org/10.1353/hsj.2007.0022>.
- Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal*, 75(3), 168–178.
- Fraser, B. J. (1998). Classroom environment instruments: Development, validity and applications. *Learning Environments Research*, 1(1), 7–34.
- Fraser, B. J. (2002). Learning environments research: Yesterday, today and tomorrow. *Studies in educational learning environments: An international perspective* (pp. 1–25). World Scientific.
- Fraser, B. J., & McRobbie, C. J. (1995). Science laboratory classroom environments at schools and universities: A cross-national study. *Educational Research and Evaluation*, 1, 289–317. <https://doi.org/10.1080/1380361950010401>.
- Freeman, B., & Crawford, L. (2008). Creating a middle school mathematics curriculum for English-language learners. *Remedial and Special Education*, 29(1), 9–19. <https://doi.org/10.1177/0741932507309717>.
- Gamoran, A., & Weinstein, M. (1998). Differentiation and opportunity in restructured schools. *American Journal of Education*, 106(3), 385–415. <https://doi.org/10.1086/444189>.
- Garza, R. (2009). Latino and white high school students' perceptions of caring behaviors: Are we culturally responsive to our students? *Urban Education*, 44(3), 297–321. <https://doi.org/10.1177/0042085908318714>.
- Gehlbach, H. (2015). Seven survey sins. *The Journal of Early Adolescence*, 35(5–6), 883–897.
- Gehlbach, H., Brinkworth, M. E., King, A. M., Hsu, L. M., McIntyre, J., & Rogers, T. (2016). Creating birds of similar feathers: Leveraging similarity to improve teacher–student relationships and academic achievement. *Journal of Educational Psychology*, 108(3), 342. <https://doi.org/10.1037/edu0000042>.
- Gershenson, S., Holt, S. B., & Papageorge, N. W. (2016). Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, 52, 209–224. <https://doi.org/10.1016/j.econedurev.2016.03.002>.
- Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review*, 10(3), 601–616. <https://doi.org/10.2307/258140>.
- Gottfredson, G. D., Gottfredson, D. C., Payne, A. A., & Gottfredson, N. C. (2005). School climate predictors of school disorder: Results from a national study of delinquency prevention in schools. *Journal of Research in Crime and Delinquency*, 42(4), 412–444. <https://doi.org/10.1177/0022427804271931>.
- Griffith, J. (2000). School climate as group evaluation and group consensus: Student and parent perceptions of the elementary school environment. *The Elementary School Journal*, 101(1), 35–61. <https://doi.org/10.1086/499658>.
- Guion, R. M. (1973). A note on organizational climate. *Organizational Behavior and Human Performance*, 9(1), 120–125.
- Hamilton, L. S., Doss, C. J., & Steiner, E. D. (2019). *Teacher and principal perspectives on social and emotional learning in America's schools: Findings from the American educator panels*. RAND Corporation.
- Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference-group effect. *Journal of Personality and Social Psychology*, 82(6), 903. <https://doi.org/10.1037/0022-3514.82.6.903>.
- Holahan, C., & Batey, B. (2019). *Measuring school climate and social and emotional learning and development: A navigation guide for states and districts*. Council of Chief State School Officers. <https://ccsso.org/sites/default/files/2019-03/CCSSO-EdCounsel%20SE%20and%20School%20Climate%20measurement.pdf>. Accessed 5 August 2020.
- Hough, H., Kalogrides, D., & Loeb, S. (2017). *Using surveys of students' social-emotional learning and school climate for accountability and continuous improvement*. Policy Analysis for California

- Education. https://edpolicyinca.org/sites/default/files/SEL-CC_report.pdf. Accessed 5 August 2020.
- Hoy, W. K. (1990). Organizational climate and culture: A conceptual analysis of the school workplace. *Journal of Educational and Psychological Consultation*, 1(2), 149–168. https://doi.org/10.1207/s1532768xjepc0102_4.
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation. <http://k12education.gatesfoundation.org/resource/gathering-feedback-on-teaching-combining-high-quality-observations-with-student-surveys-and-achievement-gains-3/>. Accessed 5 August 2020.
- Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Waxmann.
- Konold, T., & Cornell, D. (2015). Multilevel multitrait–multimethod latent analysis of structurally different and interchangeable raters of school climate. *Psychological Assessment*, 27(3), 1097. <https://doi.org/10.1037/pas0000098>.
- Kostyo, S., Cardichon, J., & Darling-Hammond, L. (2018). *Making ESSA's equity promise real: State strategies to close the opportunity gap*. Learning Policy Institute. <https://learningpolicyinstitute.org/product/essa-equity-promise-report>. Accessed 5 August 2020.
- Kyriakides, L. (2005). Drawing from teacher effectiveness research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *The Journal of Classroom Interaction*, 44–66.
- Lacoe, J. R. (2015). Unequally safe: The race gap in school safety. *Youth Violence and Juvenile Justice*, 13(2), 143–168. <https://doi.org/10.1177/1541204014532659>.
- Lam, A. C., Ruzek, E. A., Schenke, K., Conley, A. M., & Karabenick, S. A. (2015). Student perceptions of classroom achievement goal structure: Is it appropriate to aggregate? *Journal of Educational Psychology*, 107(4), 1102.
- Lareau, A. (2002). Invisible inequality: Social class and childrearing in black families and white families. *American Sociological Review*, 747–776.
- Lareau, A., & Horvat, E. M. (1999). Moments of social inclusion and exclusion race, class, and cultural capital in family-school relationships. *Sociology of Education*, 37–53. <https://doi.org/10.2307/3088916>.
- Lazarides, R., & Watt, H. M. (2015). Girls' and boys' perceived mathematics teacher beliefs, classroom learning environments and mathematical career intentions. *Contemporary Educational Psychology*, 41, 51–61. <https://doi.org/10.1016/j.cedpsych.2014.11.005>.
- LeClair, C., Doll, B., Osborn, A., & Jones, K. (2009). English language learners' and non-English language learners' perceptions of the classroom environment. *Psychology in the Schools*, 46(6), 568–577. <https://doi.org/10.1002/pits.20398>.
- Levy, J., & Wubbels, T. (1992). Student and teacher characteristics and perceptions of teacher communication style. *The Journal of Classroom Interaction*, 23–29.
- Levy, J., Wubbels, T., den Brok, P., & Brekelmans, M. (2003). Students' perceptions of interpersonal aspects of the learning environment. *Learning Environments Research*, 6(1), 5–36.
- Lincoln, Y. S. (1995). In search of students' voices. *Theory Into Practice*, 34(2), 88–93. <https://doi.org/10.1080/00405849509543664>.
- Lindell, M. K., & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *Journal of Applied Psychology*, 85(3), 331.
- Lipowski, F., Rakoczy, K., Drollinger-Vetter, B., Klieme, E., Reusser, K., & Pauli, C. (2009). Quality of geometry instructions and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction*, 19(1), 527–537.
- Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. National Comprehensive Center for Teacher Quality. <https://files.eric.ed.gov/fulltext/ED543776.pdf>. Accessed 5 August 2020.

- Lockheed, M. (1976). Some determinants and consequences of teacher expectations concerning pupil performance. In *Beginning teacher evaluation study: Phase II*. ETS.
- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology*, 34(2), 120–131. <https://doi.org/10.1016/j.cedpsych.2008.12.001>.
- Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, 9(3), 215–230. <https://doi.org/10.1007/s10984-006-9014-8>.
- Ma, X., & Willms, J. D. (2004). School disciplinary climate: Characteristics and effects on eighth grade achievement. *Alberta Journal of Educational Research*, 50(2), 169–188.
- Maehr, M. L., & Midgley, C. (1991). Enhancing student motivation: A schoolwide approach. *Educational Psychologist*, 26(3–4), 399–427. https://doi.org/10.1207/s15326985ep2603&4_9.
- Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Springer.
- Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47(2), 106–124. <https://doi.org/10.1080/00461520.2012.670488>.
- Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: An illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement*, 23(3), 305–326. <https://doi.org/10.1080/09243453.2012.678864>.
- Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2014). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement*, 26(2), 169–194. <https://doi.org/10.1080/09243453.2014.939198>.
- Mitra, D. (2007). Student voice in school reform: From listening to leadership. In D. Thiessen & A. Cook-Sather (Eds.), *International handbook of student experience in elementary and secondary school*. Springer.
- Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art—Teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, 25(2), 231–256. <https://doi.org/10.1080/09243453.2014.885451>.
- Oakes, J. (1990). *Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science*. RAND Corporation.
- Ostroff, C., Kinicki, A. J., & Tamkins, M. M. (2003). Organizational climate and culture. *Comprehensive Handbook of Psychology*, 12, 365–402. <https://doi.org/10.1002/0471264385.wei1222>.
- Panorama Education. (2015). Validity brief: Panorama student survey. *Panorama Education*. https://go.panoramaed.com/hubfs/Panorama_January2019%20Docs/validity-brief.pdf. Accessed 5 August 2020.
- Perera, R. M., & Schweig, J. D. (2019). *The role of student- and classroom-based factors associated with classroom racial climate gaps*. Presented at the annual meeting of the American Educational Research Association, Toronto, Canada.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <https://doi.org/10.3102/0013189X09332374>.
- Popham, W. J. (2013). *Evaluating America's teachers: Mission possible?* Corwin Press.
- Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems* (pp. 170–201). Jossey Bass.

- Reinholz, D. L., & Shah, N. (2018). Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. *Journal for Research in Mathematics Education*, 49(2), 140–177.
- Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). *Developing an interdisciplinary science of organizations*. Jossey-Bass.
- Rothstein, J., & Mathis, W. J. (2013). *Review of “Have we identified effective teachers?” and “A composite estimator of effective teaching: Culminating findings from the measures of effective teaching project”*. National Education Policy Center. <https://nepc.colorado.edu/sites/default/files/ttr-final-met-rothstein.pdf>. 5 August 2020.
- Ryan, R. M., & Grolnick, W. S. (1986). Origins and pawns in the classroom: Self-report and projective assessments of individual differences in children’s perceptions. *Journal of Personality and Social Psychology*, 50(3), 550. <https://doi.org/10.1037/0022-3514.50.3.550>.
- Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2018). To the means and beyond: Understanding variation in students’ perceptions of teacher emotional support. *Learning and Instruction*, 55, 13–21. <https://doi.org/10.1016/j.learninstruc.2018.02.003>.
- Schweig, J. D. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, 36(3), 259–280. <https://doi.org/10.3102/0162373713509880>.
- Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research*, 19(3), 441–462. <https://doi.org/10.1007/s10984-016-9216-7>.
- Schweig, J. D., Kaufman, J. H., & Opfer, V. D. (2020). Day by day: Investigating variation in elementary mathematics instruction that supports the common core. *Educational Researcher*, 49(3), 176–187. <https://doi.org/10.3102/0013189X20909812>.
- Schweig, J. D., Martínez, J. F., & Langi, M. (2017). *Beyond means: Investigating classroom learning environments through consensus in student surveys*. Presented at the biennial EARLI Conference for Research on Learning and Instruction, Tampere, Finland.
- Schweig, J. D., Martínez, J. F., & Schnittka, J. (2020). Making sense of consensus: Exploring how classroom climate surveys can support instructional improvement efforts in science. Manuscript under review.
- Seidel, T. (2006). The role of student characteristics in studying micro teaching–learning environments. *Learning Environments Research*, 9(3), 253–271.
- Shindler, J., Jones, A., Williams, A. D., Taylor, C., & Cardenas, H. (2016). The school climate–student achievement connection: If we want achievement gains, we need to begin by improving the climate. *Journal of School Administration Research and Development*, 1(1), 9–16.
- Simpson, A. W., & Erickson, M. T. (1983). Teachers’ verbal and nonverbal communication patterns as a function of teacher race, student gender, and student race. *American Educational Research Journal*, 20(2), 183–198. <https://doi.org/10.2307/1162593>.
- Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement*, 17(4), 245–282. <https://doi.org/10.1111/j.1745-3984.1980.tb00831.x>.
- Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, 41(5), 481–520. <https://doi.org/10.3102/1076998616646200>.
- Stipek, D. J., Givvin, K. B., Salmon, J. M., & MacGyvers, V. L. (2001). Teachers’ beliefs and practices related to mathematics instruction. *Teaching and Teacher Education*, 17(2), 213–226. [https://doi.org/10.1016/S0742-051X\(00\)00052-4](https://doi.org/10.1016/S0742-051X(00)00052-4).
- Teh, G. P., & Fraser, B. J. (1994). An evaluation of computer-assisted learning in terms of achievement, attitudes and classroom environment. *Evaluation & Research in Education*, 8(3), 147–159. <https://doi.org/10.1080/09500799409533363>.
- Urduan, T., & Schoenfelder, E. (2006). Classroom effects on student motivation: Goal structures, social relationships, and competence beliefs. *Journal of School Psychology*, 44(5), 331–349. <https://doi.org/10.1016/j.jsp.2006.04.003>.

- van der Lans, R. M., van de Grift, W. J., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice*, 34(3), 18–27. <https://doi.org/10.1111/emip.12078>.
- Voight, A., Hanson, T., O'Malley, M., & Adekanye, L. (2015). The racial school climate gap: Within-school disparities in students' experiences of safety, support, and connectedness. *American Journal of Community Psychology*, 56(3–4), 252–267. <https://doi.org/10.1007/s10464-015-9751-x>.
- Vriesema, C. C., & Gehlbach, H. (2019). *Assessing survey satisficing: The impact of unmotivated questionnaire respondents on data quality*. Policy Analysis for California Education. <https://files.eric.ed.gov/fulltext/ED600463.pdf>. Accessed 5 August 2020.
- Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, 28, 1–11. <https://doi.org/10.1016/j.learninstruc.2013.03.003>.
- Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal*, 53(6), 1834–1868. <https://doi.org/10.3102/0002831216671864>.
- Watkins, A. M., & Melde, C. (2009). Immigrants, assimilation, and perceived school disorder: An examination of the “other” ethnicities. *Journal of Criminal Justice*, 37(6), 627–635. <https://doi.org/10.1016/j.jcrimjus.2009.09.011>.
- West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE districts. *Journal of Applied Developmental Psychology*, 55, 119–129. <https://doi.org/10.1016/j.appdev.2017.06.001>.
- Wubbels, T., & Brekelmans, M. (2005). Two decades of research on teacher–student relationships in class. *International Journal of Educational Research*, 43(1–2), 6–24. <https://doi.org/10.1016/j.ijer.2006.03.003>.

Jonathan D. Schweig is a Social Scientist at the RAND Corporation (USA). His research focuses on the measurement of teaching quality, and features the learning environment, the development and implementation of teacher evaluation policies, and student social and emotional learning. His recent publications focus on the implications which methodological choices can have on inferences about instructional practice, classroom climate, and program effectiveness. Jonathan earned an M.A. in Curriculum and Teacher Education from Stanford University, an M.S. in Statistics from the University of California, Los Angeles, and a Ph.D. in Education from the University of California, Los Angeles (USA).

José Felipe Martínez is an Associate Professor of Social Research Methodology at UCLA's School of Education and Information Studies (USA). His research examines the application of measurement theory and methods to issues in education policy, specifically in relation to teacher, school, and program evaluation. A particular focus comprises instruments and tools for measuring classroom climate and instructional practice in mathematics and science, including teacher portfolios, student surveys, and observation protocols. His work has been supported by the National Science Foundation, and the Spencer, WT Grant, and McDonnell foundations. Dr. Martinez teaches courses on measurement, research design, and survey methodology.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 7

Student Ratings of Teaching Quality

Dimensions: Empirical Findings and Future Directions



Richard Göllner, Benjamin Fauth, and Wolfgang Wagner

Abstract This chapter discusses current issues in research on the validity of student ratings of teaching quality. We first discuss the advantages and limitations of student ratings of teaching quality based on theoretical considerations regarding the teaching quality concept. Research reveals that the validity of student ratings differs depending on the aspect of teaching quality being rated (i.e., classroom management, cognitive activation, or student support). Extending this research, we propose that future studies on the validity of student ratings should take into account students' cognitive processing while responding to survey items. We discuss three areas that seem promising for future research: the complexity and comprehensibility of survey items, the referent and addressee of items, and finally, the idiosyncratic nature of student ratings.

Keywords Student ratings · Teaching quality · Dimensions · Validity · Theoretical considerations

1 Introduction

Assuring reliable and valid measures is a key issue in assessing teaching quality in schools or classrooms for evaluative purposes. In general, student ratings represent a promising way to evaluate teaching because they provide firsthand impressions and are more efficient in assessing teaching quality than alternatives such as classroom observations. On the other hand, however, scholars have expressed concerns about

R. Göllner (✉) · W. Wagner
Hector Research Institute of Education Sciences and Psychology, University of Tübingen,
Tübingen, Germany
e-mail: richard.goellner@uni-tuebingen.de

W. Wagner
e-mail: wolfgang.wagner@uni-tuebingen.de

B. Fauth
Institute for Educational Analysis (IBBW), Stuttgart, Germany
e-mail: Benjamin.Fauth@ibbw.kv.bwl.de

students' ability to provide reliable and valid information about teaching quality. In the following chapter, we first describe a common framework of teaching quality and then present recent findings on the differential validity of student ratings for conceptually different aspects of teaching quality. Finally, we show that the way in which students are asked about teaching quality in surveys raises awareness of the potential and limitations of student ratings and can help us identify existing gaps in the field of teaching quality research.

2 The Concept of Teaching Quality

Teaching quality is widely understood as rooted in a teacher's actual behavior, but it is also influenced by student–teacher interactions (Doyle, 2013; Fauth et al., 2020b; Göllner et al., 2020; Hamre & Pianta, 2010; Kunter et al., 2013). Thus, conceptually, teaching quality refers to teacher behavior in the classroom as well as students' reactions to this behavior and vice versa. One implication of this is that the context and conditions in which teaching takes place always need to be considered. Teaching quality has been described and assessed in a number of different frameworks, many of which show a great deal of overlap (e.g., Creemers & Kyriakides, 2008; Danielson, 2007; Pianta et al., 2008). A very common conception of teaching quality subdivides it into three superordinate quality domains, namely classroom management, teachers' learning support, and cognitive activation (see Hamre & Pianta, 2010; Praetorius et al., 2018). Classroom management has traditionally been seen as a central element of good teaching and has an important place in many conceptualizations of teaching quality. Relevant characteristics include a lack of student misbehavior and effective management of time and classroom routines (Evertson & Weinstein, 2006). Student support is based on a positive student–teacher relationship and a learning environment in which, for example, students are given constructive feedback on how to improve their performance or see the subject matter as more relevant (Brophy, 2000). Finally, cognitive activation encompasses, for example, providing challenging tasks that clarify the connection between different concepts or link new learning content to prior knowledge (e.g., Kunter et al., 2013). These aspects of quality have received substantial empirical attention in recent years. Most importantly for the present chapter, they serve as the foundation for survey instruments and observation protocols that can then be used to examine the empirical relevance of teaching quality for students' achievement and learning-related outcomes (e.g., students' interest, motivation, self-efficacy; e.g., Kunter et al., 2013).

3 Why Should Student Ratings Be Used to Assess Teaching Quality?

Teaching quality—in terms of teachers’ classroom management, the support teachers provide to students, or the extent to which learning is cognitively demanding—can be assessed in different ways, each of which entails a number of advantages and disadvantages (Derry et al., 2010; Desimone et al., 2010; Fraser & Walberg, 1991; Wubbels et al., 1992). For instance, classroom observations are viewed as the gold standard in teaching quality research. They are considered the most objective method of measuring teaching practices and represent a central element in teacher training (Pianta et al., 2008). On the other hand, it is widely recognized that classroom observation is not without problems. Observers need to be specially trained, their observations provide only snapshots, and it is unclear whether the presence of observers systematically changes the behavior of teachers and students (e.g., Derry et al., 2010).

In contrast to classroom observations, student ratings of teaching quality are much easier to obtain. They are considered to be more cost effective, and they are directly tied to students’ day-to-day classroom experiences. Moreover, they are not merely the result of a single or quite limited number of observations, and they ensure a reliable assessment of teaching quality (Lüdtke et al., 2009). Research has shown that the psychometric properties of a class’ average teaching quality perceptions are not systematically inferior to those from observational measures (e.g., Clausen, 2002; de Jong & Westerhof, 2001; Maulana & Helms-Lorenz, 2016). In addition, there is empirical evidence that students are able to provide valid ratings of teaching quality, although differences between quality dimensions need to be taken into account (Fauth et al., 2014; Kuhfeld, 2017; Nelson et al., 2014; Schweig, 2014; Wagner et al., 2013; Wallace et al., 2016; see also Chap. 5 by van der Lans in this volume). Specifically, previous research has shown that student ratings of classroom management typically emerge as a clearly identifiable teaching quality aspect, which exhibits significant associations with observational as well as teacher self-report data and predicts students’ learning in terms of their achievement, interest, and motivation (e.g., Kunter et al., 2007; Lipowsky et al., 2009). Furthermore, student ratings of classroom management are comparable across different learning contexts (e.g., different school subjects; Wagner et al., 2013) and even reveal time-specificity. That is, student ratings have proven to be sensitive enough to capture differences in teachers’ classroom management over the course of several weeks or months (Wagner et al., 2016). In contrast, the psychometric properties of student ratings of learning support and cognitive activation are less clear. In the case of cognitive activation, this is because measures cannot be generally applied to all subjects but need to reflect the specificity and requirements of each individual subject (e.g., mathematics, languages, the arts, etc.). Consequently, the majority of existing student surveys of teaching quality do not include cognitive activation measures, making it much harder to evaluate the validity of student ratings with respect to this dimension. Nevertheless, the few studies that do exist show that even ratings by primary school students reveal substantial differences in cognitive activation between classrooms. In addition, cognitive activation

ratings have been shown to be separable from classroom management ratings and to a lesser extent from learning support ratings, and to be statistically significant associations with student learning outcomes (e.g., subject-related interest; Fauth et al., 2014). The situation for student ratings of learning support is even more complex. Previous research has shown that student ratings of learning support exhibit relatively low agreement with classroom observations and even low agreement across students in the same classroom. One potential explanation for this is that students' perceptions of teachers' learning support do not exclusively function as a quality characteristic that differs across classrooms but are also affected by students' individual experiences within classrooms (Aldrup et al., 2018; Atlay et al., 2019; den Brok et al., 2006a; Göllner et al., 2018). For a long time, these within-classroom differences were considered the result of factors external to teaching quality, such as students' rating tendencies (e.g., harshness or leniency) or perceptual mindsets (e.g., halo error; e.g., Lance et al., 1994). However, recent research has shown that these differences can also reflect effects stemming from the dyadic relationships between each individual student and his or her teacher. Specifically, a recent study by Göllner and colleagues (2018) used national longitudinal data from the Program for International Student Assessment (PISA) database and showed that rating differences in student perceptions of learning support partially result from teacher-independent rater tendencies, but also reflect the dyadic relationship between an individual student and one specific teacher. Therefore, students' ratings of teaching quality provide important information about their individual experiences in their classroom learning environments.

4 Future Directions for the Use of Students' Ratings

Although student ratings of teaching quality have become a prominent way to obtain student feedback on teaching quality in schools and classrooms, scholars and practitioners have also criticized their use in both summative and formative assessments (Abrami et al., 2007; Benton & Cashin, 2012). They emphasize the specific nature of student ratings, as students are not trained to provide valid assessments of teaching quality in the same way as adult observers. Thus, it is important to acknowledge potential limitations of student ratings, which raises the question of how student ratings for evaluative purposes can be improved. We believe that a more detailed examination of existing survey instruments can be a fruitful approach to finding out how student ratings work and what we can do to achieve reliable and valid ratings. From a very general perspective, a student survey can be seen as ordinary text material (i.e., textual information presented in the form of separate items), requiring students to read and interpret a question to understand what is meant, retrieve the requested information from memory, and form a judgment based on their knowledge and expertise (Tourangeau et al., 2000). Building upon this foundation, this chapter presents three areas of recent research that might help provide a deeper understanding of students' teaching quality rating and exploit future research directions.

4.1 *Complexity and Comprehensibility*

At first glance, existing student surveys fundamentally differ in their linguistic complexity, which shapes student responses (e.g., Krosnick & Presser, 2010; Tourangeau et al., 2000). It is surprising to see that even frequently used surveys are linguistically challenging, particularly for younger respondents (e.g., Fauth et al., 2014; Wagner et al., 2013). Consequently, it can be argued that many reporting problems (i.e., low interrater agreement) arise because students encounter difficulties in comprehending the survey. Survey items include many linguistic features, including surface aspects (e.g., the length of words and sentences) and characteristics that require more linguistic analysis (e.g., the number of complex noun phrases). For example, the following items might be used to assess teachers' sensitivity to and awareness of students' level of academic functioning: "In math, the individual students often do different tasks" and "In math lessons, the teacher asks different questions, depending on how able the student is." However, the items differ in their linguistic characteristics: number of words (9 vs. 15), structure of sentences (1 vs. 2 clauses), average word length (5.33 characters vs. 5.00 characters), and number of complex noun phrases per clause (2 vs. 0.5). In addition, students may be less familiar with certain words used in the items (e.g., "individual," "depending") or have to make many interpretations because single words do not refer to specific, denotable, and relatively objective behavior (i.e., high-inference ratings; e.g., Roch et al., 2009; Rosenshine, 1970). Despite the large body of literature on traditional best practices in the construction of survey questions (see Krosnick & Presser, 2010), only a few studies have examined the impact of these and other linguistic characteristics on student surveys' ability to reliably and validly assess teaching quality. One of these studies showed that the use of measures with a lower specificity and higher level of abstraction (high-inference ratings) leads to higher interrater reliability in student ratings, but lower agreement with expert assessments. Contrary to common expectations, rater agreement increased as the behavioral observability of the measures decreased (Roch et al., 2009). The authors argue that raters might compensate for uncertainty in high-inference ratings by more strongly adjusting their ratings to match their general impression, which might in turn be unrelated or only partially related to the teaching quality dimension in question. Such findings impressively demonstrate that the association between linguistic features and psychometric properties of student ratings is anything but trivial, and a more rigorous consideration of linguistic forms in existing surveys is needed.

4.2 *Framing*

Student surveys also differ in characteristics apart from linguistic complexity. Specifically, the referent and addressee of survey items are two salient characteristics that might affect the information obtained from student ratings of teaching quality but

received less attention in research on student perceptions of teaching quality (den Brok et al., 2004, 2006b; McRobbie et al., 1998). The referent can be defined as the subject to which an item refers. At first glance, student rating items that refer more to the classroom (e.g., “In math class, the lesson is often disrupted”) than to the teacher (e.g., “Our math teacher always knows exactly what is happening in class”) tend to exhibit more favorable psychometric properties in terms of interrater agreement or distinctiveness from other theoretically relevant aspects of teaching quality (see Fauth et al., 2020a; Göllner et al., 2020). However, the use of surveys that refer more to the classroom than to the teacher might result in serious constraints. First, items referring more to the classroom than to the teacher are frequently used to assess classroom management, but much rarer for items assessing learning support or cognitive activation. This raises the question of whether the well-established distinctiveness of classroom management compared to other quality aspects is also due to systematic differences in the referent used. Second, previous findings have shown that when classroom management items refer to the classroom, measures are more prone to classroom composition effects (e.g., proportion of male students or performance composition). Even though existing analytical procedures can be used to account for such differences in classroom composition, it is unclear whether such analytical adjustments result in fair comparisons or relatively favor or penalize certain individual teachers. Irrespective of this, classroom management measures referring more to students than to the teacher need to be seen from an interactionist perspective that includes both teachers and students they teach (Fauth et al., 2020a). In addition, the target of the teacher’s behavior that is addressed in a survey is important. In the simplest case, this can be either the responding student him/herself (e.g., “The teacher motivates me”) or all students in the classroom (e.g., “The teacher motivates us”). An examination of existing surveys shows that the “me-addressee” is predominantly used when assessing the support teachers provide to students, whereas the “we-addressee” is more frequently used for classroom management and cognitive activation (e.g., BIJU, Baumert et al., 1996; Tripod survey; e.g., Prenzel et al., 2013; Wallace et al., 2016). At the same time, previous studies have shown that student support dimensions usually fail to predict student learning outcomes on the classroom level but are more consistent predictors at the individual student level (e.g., Aldrup et al., 2018). These results raise the question of whether support can be better conceptualized as a dyadic phenomenon between a teacher and an individual student or whether they merely reflect how teacher support is assessed. Experimentally varying the addressee for items assessing multiple teaching quality dimensions will enable us to examine whether the addressee affects the information obtained from student ratings at the student and classroom level. The findings might also be interesting for analytical modeling procedures used in teaching quality research. First, findings from multilevel models applied to separate students’ shared (student level) and non-shared (classroom level) perceptions of teaching quality might be directly affected by the used addressee. Whereas the “me-addressee” assumed to provide valid information about students individual learning experiences at the student level, the “we-addressee” might be more adequate to give insights in students’ learning at the classroom level; or in other words, one cannot simply assume that different

item wordings can interchangeably be used at different levels of analysis (den Brok, 2001). Second, there is increased interest in more recent analytical procedures that model classroom heterogeneity in student ratings as an additional indicator of good teaching (e.g., Schenke et al., 2018). Applying these modeling procedures to surveys with a “me-addressee” might be a better way to assess student–teacher fit in classrooms and teacher adaptivity than surveys with a “we-addressee.” If surveys with a “we-addressee” are considered, different levels of heterogeneity between classes might be more a reflection of class-specific measurement precision (i.e., more or less agreement in classes). In other words, the choice of addressee in surveys can be assumed to have very serious consequences for teaching quality assessment and the view we take on students’ learning in classrooms.

4.3 The Idiosyncratic Nature of Student Ratings

Finally, it is important to ask what we can fundamentally expect from student ratings of teaching quality and to what extent student ratings of teaching quality reveal idiosyncrasies, i.e., are systematically different from alternative methods. Even when we take special care to use comprehensible and age-appropriate surveys and make more intentional decisions about the referent and addressee in survey items, the specific nature of student ratings needs to be considered. One main objective of previous research has been to determine the degree of idiosyncrasy in student ratings by comparing them to alternative assessment methods (e.g., Clausen, 2002; Kunter & Baumert, 2006). This research has shown that student ratings, particularly those assessing learning support and cognitive activation, exhibit substantial differences to classroom observations or teacher self-report data, which might lead to the conclusion that students are less able to provide valid information on teaching quality and its theoretically proposed dimensions (e.g., Abrami et al., 2007). However, this focus on limitations and biases of student ratings bears the risk of neglecting the expertise students naturally acquire through their everyday experiences in classrooms. Thus, future research needs to better appreciate the unique information we obtain from student ratings (e.g., Leighton, 2019). In order to do so, however, we need to learn much more about the mental models that underlie students’ ratings and the extent to which these models differ from those of adult observers evaluating teaching quality. A recent study by Jaekel et al., (2021) found that student ratings of teaching quality in one school subject (mathematics or German language) did not only result from students’ daily experiences in the subject at hand, but were also affected by their experiences in the respective other subjects. Students seem to make use of comparative information when objective criteria for good teaching is not available. In addition, there is a need to understand how a developmental perspective can help us understand idiosyncrasies in student ratings of teaching quality. That is, it is reasonable to assume that student ratings of teaching quality are affected by the age-related developmental stages in which ratings take place. For instance, students’ need to define their own identity and stronger need for autonomy during adolescence (e.g., Eccles et al., 1993)

might function as a guiding perspective when students have to rate teaching quality. A recent study by Wallace and colleagues (2016) based on the Tripod survey identified two dimensions of students' ratings of teaching quality: one specific classroom management factor and one broad general factor. Interestingly, the quality indicators with the highest loadings on the general factor were indicators that clearly capture students' perceptions of teachers' learning support and student–teacher relationship (Schweig, 2014; Wallace et al., 2016). The same is true for student ratings of cognitive activation. It is interesting to note that even though cognitive activation is considered a central aspect explaining students' achievement, cognitive activation measures are much less common in existing surveys than classroom management or learning support measures. One major reason for this is that assessing teachers' ability to use stimulating learning materials, the quality of questions teachers ask during lessons, or the quality of classroom discussion from students perspective is seen as a particularly challenging task because it requires special knowledge and skills which is beyond students' firsthand experiences of participation in the classroom. Whether and to what extent students are really able to provide information on these and other aspects of cognitive activation in line with an adult view remains an open question that needs to be addressed in future research. As part of this process, we have to think about further refining existing measures that capture central aspects of cognitive activation in a wide variety of learning situations and by making more explicit use of other principles getting learners to learn long, complex, and difficult things. Alternative ways of conceptualizing and measuring effective learning contexts from related disciplines (e.g., discourse analysis in linguistic research; Turner & Meyer, 2000) or entirely different research fields (e.g., game-based learning; Gee, 2007) can provide a good foundation for improving existing cognitive activation measures.

5 Closing Remarks

As the work we reviewed in this chapter makes clear, student ratings have become a vibrant part of teaching quality research. We are particularly excited about two aspects of this research. The first is the usefulness of student ratings in research and practice. Even though differences across teaching quality dimensions need to be considered, students can provide a valid perspective on teaching quality and are thus in no way generally inferior to alternative assessments such as classroom observations or teacher self-reports. Second, students provide a plethora of information on teaching quality at both the classroom and the student level, with the latter referring to students' individual learning experiences within a classroom in a way that is beyond the scope of alternative assessments. As research on student ratings progresses, it will be critical to take a deeper and more consequential look at the characteristics of existing surveys to determine what we can learn about teaching quality from the students' perspective. We look forward to participating in work on these topics in the future.

References

- Abrami, P. C., D'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: What we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 385–456). Springer.
- Aldrup, K., Klusmann, U., Lüdtke, O., Göllner, R., & Trautwein, U. (2018). Social support and classroom management are related to secondary students' general school adjustment: A multilevel structural equation model using student and teacher ratings. *Journal of Educational Psychology, 110*, 1066–1083. <https://doi.org/10.1037/edu0000256>.
- Atlay, C., Tieben, N., Fauth, B., & Hillmert, S. (2019). The role of socioeconomic background and prior achievement for students' perception of teacher support. *British Journal of Sociology of Education, 40*, 970–991. <https://doi.org/10.1080/01425692.2019.1642737>.
- Baumert, J., Roeder, P. M., Gruehn, S., Heyn, S., Köller, O., Rimmel, R., et al. (1996). Bildungsverläufe und psychosoziale Entwicklung im Jugendalter (BIJU) [Educational pathways and psychosocial development in adolescence]. In K.-P. Treumann, G. Neubauer, R. Moeller, & J. Abel (Eds.), *Methoden und Anwendungen empirischer pädagogischer Forschung* [Methods and applications of empirical educational research] (pp. 170–180). Waxmann.
- Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature* (IDEA Paper No. 50). Manhattan: Kansas State University, Center for Faculty Evaluation and Development.
- Brophy, J. (2000). *Teaching. Educational practices series, 1*. Brüssel: International Academy of Education (IAE).
- Clausen, M. (2002). *Qualität von Unterricht: Eine Frage der Perspektive?* [Quality of instruction as a question of perspective?]. Waxmann.
- Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. Routledge.
- Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). ASCD.
- de Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*, 51–85. <https://doi.org/10.1023/A:1011402608575>.
- den Brok, P. (2001). *Teaching and student outcomes*. W. C. C.
- den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal teacher behaviour and student outcomes. *School Effectiveness and School Improvement, 15*, 407–442. <https://doi.org/10.1080/09243450512331383262>.
- den Brok, P., Brekelmans, M., & Wubbels, T. (2006a). Multilevel issues in research using students' perceptions of learning environments: The case of the Questionnaire on Teacher Interaction. *Learning Environments Research, 9*, 199–213. <https://doi.org/10.1007/s10984-006-9013-9>.
- den Brok, P., Fisher, D., Rickards, T., & Bull, E. (2006b). Californian science students' perceptions of their classroom learning environments. *Educational Research and Evaluation, 12*, 3–25. <https://doi.org/10.1080/13803610500392053>.
- Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences, 19*(1), 3–53. <https://doi.org/10.1080/10508400903452884>.
- Desimone, L. M., Smith, T. M., & Frisvold, D. E. (2010). Survey measures of classroom instruction: Comparing student and teacher reports. *Educational Policy, 24*(2), 267–329. <https://doi.org/10.1177/0895904808330173>.
- Doyle, W. (2013). Ecological approaches to classroom management. In C. M. Evertson & C. S. Weinstein (Eds.), *Handbook of classroom management* (pp. 107–136). Routledge.
- Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage–environment fit on adolescents' experiences in schools and families. *American Psychologist, 48*, 90–101.

- Evertson, C. M., & Weinstein, C. S. (2006). *Handbook of classroom management: Research, practice, and contemporary issues*. Lawrence Erlbaum Associates Publishers.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>.
- Fauth, B., Göllner, R., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020a). Who sees what? Conceptual considerations on the measurement of teaching quality from different perspectives. *Zeitschrift Für Pädagogik, 66*, 138–155.
- Fauth, B., Wagner, W., Bertram, C., Göllner, R., Roloff-Bruchmann, J., Lüdtke, O., Polikoff, M. S., Klusmann, U., & Trautwein, U. (2020b). Don't blame the teacher? The need to account for classroom characteristics in evaluations of teaching quality. *Journal of Educational Psychology, 112*, 1284–1302. <https://doi.org/10.1037/edu0000416>.
- Fraser, B. J., & Walberg, H. J. (1991). *Educational environments: Evaluation, antecedents and consequences*. Pergamon Press.
- Gee, J. P. (2007). *What video games have to teach us about learning and literacy* (2nd ed.). Palgrave Macmillan.
- Göllner, R., Fauth, B., Lenske, G., Praetorius, A.-K., & Wagner, W. (2020). Do student ratings of classroom management tell us more about teachers or classrooms composition? *Zeitschrift Für Pädagogik, 66*, 156–172.
- Göllner, R., Wagner, W., Eccles, J. S., & Trautwein, U. (2018). Students' idiosyncratic perceptions of teaching quality in mathematics: A result of rater tendency alone or an expression of dyadic effects between students and teachers? *Journal of Educational Psychology, 110*, 709–725. <https://doi.org/10.1037/edu0000236>.
- Hamre, B. K., & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. In J. Meece & J. Eccles (Eds.), *Handbook of research on schools, schooling, and human development* (pp. 25–41). Routledge.
- Jaekel, A.-K., Göllner, R., & Trautwein, U. (2021). How students' perceptions of teaching quality in one subject are impacted by the grades they receive in another subject—Dimensional comparisons in student evaluations of teaching quality. *Journal of Educational Psychology, 113*, 1037/edu0000488. <https://doi.org/10.1037/edu0000488>.
- Krosnick, J. A., & Presser, S. (2010). Questionnaire design. In J. D. Wright & P. V. Marsden (Eds.), *Handbook of survey research* (2nd ed.). Emerald Group.
- Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod Student Survey. *Educational Assessment, 22*(4), 253–274. <https://doi.org/10.1080/10627197.2017.1381555>.
- Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*, 231–251. <https://doi.org/10.1007/s10984-006-9015-7>.
- Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S., & Neubrand, M. (Eds.). (2013). *Cognitive activation in the mathematics classroom and professional competence of teachers: Results from the COACTIV project*. Springer.
- Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17*, 494–509. <https://doi.org/10.1016/j.learninstruc.2007.09.002>.
- Lance, C. E., LaPointe, J. A., & Fiscicar, S. A. (1994). Tests of three causal models of halo rater error. *Organizational Behavior and Human Decision Processes, 57*, 83–96. <https://doi.org/10.1006/obhd.1994.1005>.
- Leighton, J. P. (2019). Students' interpretation of formative assessment feedback: Three claims for why we know so little about something so important. *Journal of Educational Measurement, 56*, 793–814. <https://doi.org/10.1111/jedm.12237>.
- Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean theorem. *Learning and Instruction, 19*, 527–537. <https://doi.org/10.1016/j.learninstruc.2008.11.001>.

- Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings in multilevel modelling. *Contemporary Educational Psychology*, 34, 123–131. <https://doi.org/10.1016/j.cedpsych.2008.12.001>.
- Maulana, R., & Helms-Lorenz, M. (2016). Observations and student perceptions of the quality of preservice teachers' teaching behaviour: Construct representation and predictive quality. *Learning Environments Research*, 19, 335–357. <https://doi.org/10.1007/s10984-016-9215-8>.
- McRobbie, C. J., Fisher, D. L., & Wong, A. F. L. (1998). Personal and class forms of classroom environment instruments. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education* (pp. 581–594). Kluwer.
- Nelson, P. M., Demers, J. A., & Christ, T. J. (2014). The responsive environmental assessment for classroom teaching (REACT): The dimensionality of student perceptions of the instructional environment. *School Psychology Quarterly*, 29, 182–197. <https://doi.org/10.1037/spq0000049>.
- Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS: PreK-3)*. Brookes.
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM Mathematics Education*, 50, 407–426. <https://doi.org/10.1007/s11858-018-0918-4>.
- Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., Pekrun, R., Rolff, H.-G., Rost, J., & Schiefele, U. (2013). *PISA-I-Plus 2003, 2004*. IQB—Institute for Educational Quality Improvement.
- Roch, S. G., Paqin, A. R., & Littlejohn, T. W. (2009). Do raters agree more on observable items? *Human Performance*, 22, 391–409. <https://doi.org/10.1080/08959280903248344>.
- Rosenshine, B. (1970). Evaluation of classroom instruction. *Review of Educational Research*, 40, 279–300. <https://doi.org/10.3102/00346543040002279>.
- Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2018). To the means and beyond: Understanding variation in students' perceptions of teacher emotional support. *Learning and Instruction*, 55, 13–21. <https://doi.org/10.1016/j.learninstruc.2018.02.003>.
- Schweig, J. (2014). Multilevel factor analysis by model segregation: New applications for robust test statistics. *Journal of Educational and Behavioral Statistics*, 39(5), 394–422. <https://doi.org/10.3102/1076998614544784>.
- Tourangeau, R., & Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
- Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist*, 35, 69–85. https://doi.org/10.1207/S15326985EP3502_2.
- Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and domain-generalizability of domain-independent assessments. *Learning and Instruction*, 104, 148–163. <https://doi.org/10.1016/j.learninstruc.2013.03.003>.
- Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, 108, 705–721. <https://doi.org/10.1037/edu0000075>.
- Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, 53, 1834–1868. <https://doi.org/10.3102/0002831216671864>.
- Wubbels, T., Brekelmans, M., & Hooymayers, H. P. (1992). Do teacher ideals distort the self-reports of their interpersonal behavior? *Teaching and Teacher Education*, 8(1), 47–58.

Richard Göllner is a Professor of Educational Effectiveness and Trajectories at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen (Germany). His work focuses on teaching quality, specifically on the measurement of instructional

practice from an interdisciplinary perspective, and its impact on students' achievement. Furthermore, he is interested in students' personality development within schools and the use of simulated learning contexts in experimental research in education.

Benjamin Fauth is Head of the Department for Empirical Educational Research at the Institute for Educational Analysis (IBBW) in Stuttgart (Germany) and Associate Professor at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen (Germany). His research focuses on the quality of teaching, in particular questions of the theoretical conceptualization, assessment, and the impact of teaching quality. Furthermore, his research focuses on the professional competence of teachers and on questions of applied evaluation research.

Wolfgang Wagner studied psychology at the University of Koblenz-Landau (Germany) and now works as a research assistant at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen (Germany). His main research interests include the assessment of characteristics of learning environments and their effects on the development of targeted outcomes (in particular, academic achievement), as well as methodological issues in the field of (multilevel) latent variable models.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part II
Using Student Feedback
for the Development
of Teaching and Teachers

Chapter 8

Functions and Success Conditions of Student Feedback in the Development of Teaching and Teachers



Benedikt Wisniewski and Klaus Zierer

Abstract The term “student feedback” is often used synonymously with evaluation, assessment, or ratings of teaching, but can be conceptually delimited from these concepts, distinguishing formative and summative aspects. Obtaining feedback is a core component of teachers’ professional development. It is the basis for critical self-reflection, a prerequisite of reducing discrepancies between one’s performance and set goals, a tool to identify blind spots, and a means of correcting false self-assessments. Student feedback opens up opportunities for teachers to improve on their teaching by comparing students’ perspectives on instructional quality to their own perspectives. Feedback can also help teachers to implement democratic principles, and experience self-efficacy. Conditions are discussed that need to be fulfilled for student feedback to be successful.

Keywords Student feedback · Professional development · Democratization · Teacher satisfaction

1 Introduction

Student feedback is a fundamental part of professional teaching practice. In contrast to forms of organizational assessment such as teacher evaluations, which always serve an allocation or selection purpose (e.g. promotion, access to functional positions), feedback has the aim of personal professional development. This development requires a critical reflection that compares one’s own experiences with external information, and students can provide this information in a reliable and valid way.

Among wide media interest, two attempts (in 2017 and 2019) were made in Germany and Austria to create online platforms that allowed students to rate their teachers publicly. These platforms (spickmich.de and lernsieg.de) both claimed to

B. Wisniewski (✉) · K. Zierer
Faculty of Philosophy and Social Sciences, University of Augsburg, Augsburg, Germany
e-mail: benedikt.wisniewski@phil.uni-augsburg.de

K. Zierer
e-mail: klaus.zierer@phil.uni-augsburg.de

provide feedback for teachers in order to improve teaching. By means of categories such as “professional competence”, “motivation”, “popularity”, “clothing”, “fair examinations”, or “physical appearance”, teachers could be evaluated anonymously with grades. After a certain number of ratings per teacher, the results were then made publicly accessible. Due to several complaints by rated teachers and by teachers’ unions, both platforms were turned off.

What both platforms had in common were partly irrelevant evaluation criteria (e.g., clothing), evaluation criteria that included areas which could not be (sufficiently) assessed by students (e.g., professional competence), and—as the most critical aspect—a publication of the results.

Starting with these two negative examples, we will show how the functions of student feedback can be defined in a professional context: After a conceptual delimitation, we will point out why feedback is important for the professional development of teachers in general. After that, we will discuss three basic functions of student feedback: the development of teaching, the democratization of schools, and the improvement of teachers’ satisfaction and health. In the last step, we will—in brief—propose success conditions of student feedback.

2 Feedback, Evaluation, Assessment, and Rating—A Conceptual Delimitation

Because grading plays a central role in most school systems around the world and teachers usually provide feedback in the form of grades, student feedback is often equated with grading teachers (Elstad et al., 2017). The terms “student feedback”, “student assessments”, “student ratings”, and “student evaluations” are used many times in a more or less synonymous way. It is assumed that students grade their teachers—similar to how teachers grade their students. Feedback is considered primarily a summative form of evaluation, rather than a formative form of providing information for professional development. Consequently, parallels are drawn between student feedback in school and student evaluations of teaching at university, the latter of which are widely used for selecting and promoting academic staff. The problems with evaluations of teaching in higher education have been discussed by Sproule (2000, 2002), who argues that the adoption of the “consumer” model of education does not capture the pedagogical process in its entirety, overlooking the students’ influence on this process, and that false consequences are drawn from SETs. Research in the higher education context also shows no or only minimal correlation between SET and learning outcomes (Uttl et al., 2017, see Chap. 15 of this volume). Of course, findings like these could be used as arguments against student feedback, but the conceptual blur resulting from different concepts requires a delimitation of what feedback means, what distinguishes it from evaluation, assessment, and ratings (Table 1), and then to define what student feedback really means when we talk about its functions.

Table 1 Conceptual delimitation (Zierer & Wisniewski, 2018)

| | |
|------------|---|
| Feedback | Data-based exchange of information between people aimed at development and serving to adapt one's own behavior in response to feedback from others. |
| Evaluation | Investigation of whether and to what extent a behavior is suitable for achieving a desired target state or fulfilling a purpose. |
| Assessment | Verification of the extent to which a person's behavior or qualities are consistent with the evaluators' standards, usually expressed in terms of statements such as "good" or "bad". |
| Rating | Measures of personal characteristics, performance, and social behavior, usually expressed in terms of predicates, e.g., in the form of grades. |

A delimitation is of great importance for further discourse on this subject. Student feedback in schools is not synonymous to student evaluations of teaching or student ratings (concepts primarily used in higher education). Basically, and primarily, it provides information for the teachers who obtain it in order to get an impression of how their students experience their teaching. However, studies show that—just like in the higher education context (Marsh & Dunkin, 1992)—student feedback in schools is very often used for evaluation and assessment purposes rather than as an opportunity for personal change (Elstad et al., 2017) and that instruments are used which do not do justice to the actual purpose, for example by being inappropriate for innovative forms of teaching (Kember et al., 2002). When feedback is used at the end of a term, students believe that their feedback to teachers does not change anything in the classroom (Chen & Hoshower, 2003; Spencer & Schmelkin, 2002). When evaluation rather than professional development is emphasized, teachers see student feedback as a controlling tool (Harvey, 2002; Newton, 2000). The formative and summative components of feedback are not categorically incompatible, but an over-emphasis of the summative components can undermine the use of feedback and negatively affect school climate (Ford et al., 2018).

In the following, we will focus on functions of student feedback in schools obtained by teachers in order to acquire information on how students perceive teaching in a formative sense and neglect a more detailed discussion of summative functions used by school administrations to select or promote teachers.

3 Why Student Feedback Is Important

The explanation and prediction of the feeling of professional success and professional satisfaction of teachers are often attributed to largely unchangeable and unlearnable personality traits. This attribution is evident in both beginners and experienced teachers (Bromme & Haag, 2004). If one holds the view that stable personality traits are largely responsible for one's professional success, feedback is mostly irrelevant. However, empirical research shows that the concept of "the born teacher" is

outdated. It is not the unchangeable characteristics that primarily influence the quality of teaching but rather professional skills and knowledge, motivation, self-regulation, and attitudes (Zierer, 2015). All these are qualities to work on that require constant reflection based on data.

Feedback contains an oral or written external perception after a data collection, whereby these data can be in the micro range as sensory impressions or perceptions of a counterpart (for example the perception of facial expressions and gestures), and in the macro range of an observer in the form of multi-perspective data collection with differentiated methods and instruments, for example, feedback questionnaires (Buhren, 2015). Increasingly, teachers are confronted with the expectation of being reflective practitioners (Schön, 1987) who can develop their professional skills throughout their professional lives (Staub, 2001). There are numerous, partly very different, definitions of professional development (Reh, 2004), but, despite differently substantiated theoretical concepts, a large consensus can be established that reflexivity is a core area of professionalism (ibid.). A (self-)critical reflection that uses both one's own experience and external information forms the core of pedagogical professionalism (Paseka et al., 2011). For this reason, obtaining feedback is a core component of teachers' professional development. As active directors of instruction, they have a very high impact on their students' achievement (Hattie, 2009). However, not all teachers have the same influence. It is particularly high when they try to see teaching through the eyes of their students, when they try to understand how their teaching impacts the learners (ibid.).

According to control theory (Carver & Scheier, 1982), people constantly compare their performance to a behavioral goal and, when they detect a discrepancy, attempt to reduce this discrepancy. Feedback is a necessary prerequisite of professional reflection, increasing the awareness of behaviors and the impact of these behaviors. It helps to question automatic processes, habits, and routines, providing opportunities for behavioral change. Additionally, feedback influences motivational processes by reducing negative emotions caused by an observed discrepancy between goals and performance and fostering positive emotions by decreasing such a discrepancy (Deci et al., 1999). Furthermore, performers do better on tasks for which higher quality feedback is available (Northcraft et al., 2011).

When teachers state that they do not need feedback because they know best how effective their teaching is, it must be noted that the self-assessment of one's own competences is often wrong. This can generally be proven for different tasks and requirements (Kruger & Dunning, 1999). In the worst case, the consequence is that students become bored in class, learn less than they could, and the teacher still assumes that he or she is offering the best possible instruction. Feedback serves to prevent such misjudgments by providing information that is only accessible through an external perspective (Wisniewski & Zierer, 2019).

Feedback is an essential prerequisite for goal-oriented and self-reflective processes because teachers, like any other professional group, have so-called "blind spots" in their professional practice, as described in the model of the Johari window

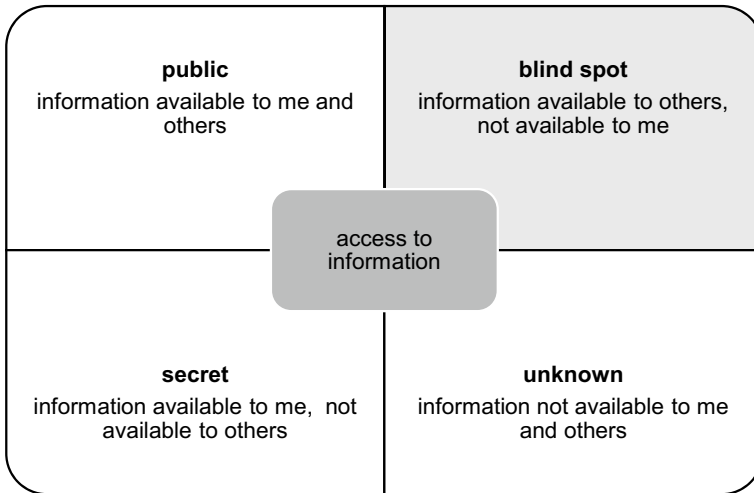


Fig. 1 Johari window for corporate settings (Luft & Ingham, 1955)

(Luft & Ingham, 1955, see Fig. 1), a model developed for corporate settings. Like in any other professional context, there is certain relevant information for teachers that is not accessible to a person him or herself but only accessible to others. The relevance of blind spots can range from minor to major—from the frequent repetition of a certain filler word and unfavorable non-verbal signals to the fact that a teacher explains content too quickly or too incomprehensibly (Wisniewski & Zierer, 2019). The only way to gain access to such blind spots is feedback.

A classic blind spot of teachers is, for example, their estimation of their own speaking time in class. Thus, Helmke and colleagues (2008) were able to show that teachers' estimation of their speaking time during a lesson differs considerably from the time objectively measured. In short: Teachers talk way more than they think they do (Fig. 2). The example shows that there are highly relevant characteristics of teaching that are not accessible through pure self-reflection but need to be communicated from an external perspective.

In this sense, feedback offers the opportunity to reveal blind spots by comparing perspectives. Blind spots can refer to critical aspects of behavior (like in the presented example), but also to strengths and resources that a teacher does not perceive from his or her own perspective. Student feedback can provide teachers with information on both, unknown strengths and unknown weaknesses.

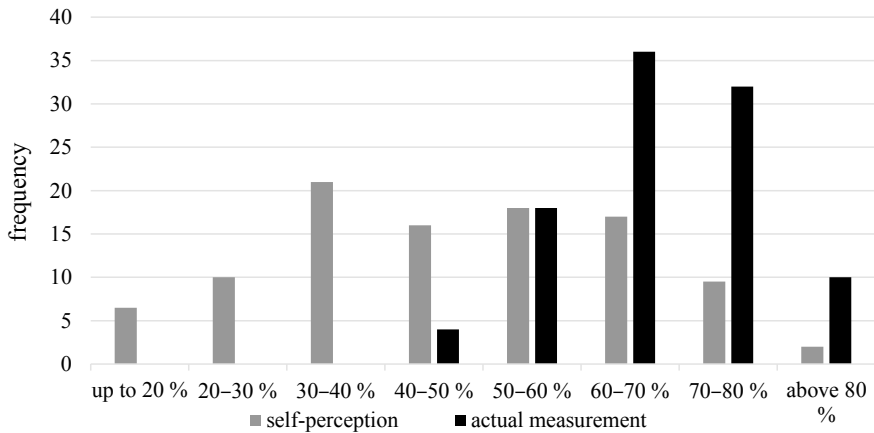


Fig. 2 Speaking time by teachers in the classroom—estimation and actual measurement (Helmke et al., 2008, p. 139)

4 Developing Teaching and Teachers with Student Feedback

4.1 Development of Teaching

What teachers actually do in their classrooms is one of the strongest predictors of students' learning outcomes (Hattie, 2009; Helmke, 2017; Seidel & Shavelson, 2007). Consequently, it is crucial to pinpoint what works in the classroom. Student feedback is supposed to help teachers improve the quality of teaching (Ditton & Arnoldt, 2004; Gärtner, 2007, 2013; Helmke, 2017) by providing diagnostic information on teaching characteristics that determine if students feel sufficiently challenged, engaged, and comfortable asking for help, telling teachers where they need to focus so that their current students benefit, suggest students' misunderstandings, and diagnose teachers' specific attempts at clarification (Gates Foundation, 2012).

Theoretically, improvements of teaching quality by student feedback can be explained in three ways: Firstly, feedback helps teachers to gain information about relevant lesson characteristics that are not accessible through pure self-reflection. Procedures that question learners consciously and directly about the core components of teaching provide opportunities for developing instructional quality. Kunter and Voss (2013) distinguish surface structures (characteristics that are directly observable, e.g., social forms, forms of teaching, methods, media use) from deep structures of teaching (characteristics that become visible through the interpretation of the teaching-learning process, and classroom interaction). Effectiveness of teaching depends largely on the latter (Hattie, 2009) and student feedback is a relatively reliable and valid information source on deep structures (Wisniewski et al., 2020a; see Chap. 7 of Göllner et al. in this volume). They can reveal if one teaching method

produces better learning results than another, if assignments were clear and produced the intended effect, if students felt comfortable and challenged, if content was properly consolidated, if learning time was used efficiently, if students were able to work without disturbance, if the feedback which students got from the teacher was helpful, and so on. Teachers get an impression which of these aspects were seen critical, but also which of them were perceived favorably by students. It is becoming apparent that positive feedback leads to a further strengthening of the methods that have been successfully used in class, and effects can be seen in the tendency to make teaching more transparent and to regularly reflect on the lessons with the pupils (Gärtner, 2013). Ideally, comparing the student perspective to the teacher perspective and subsequent discussion leads to conclusions on how to optimize teaching (Desimone, 2009). This allows a shift of focus from surface structures of teaching and formal specifications (Have all curriculum goals been achieved?) to actual learning processes that have (or have not) happened in the classroom and helps to answer the question why this was the case.

There can be an action-guiding function of feedback: People, in general, act differently, when they expect feedback (Carver & Scheier, 1990). Feedback increases general self-awareness, and—consecutively—increases an individual’s capability to inhibit behaviors that are undesired or dysfunctional (Alberts et al., 2011). When teachers know that they will get feedback based on certain criteria, they are likely to pay particular attention to these criteria (a reason why valid criteria for student feedback are crucial, see Chap. 4 of this volume). Assessing expectancies *before* getting feedback can already cause behavioral change. For example, a teacher who expects to get feedback about clarity will most probably be more aware of this aspect and put more emphasis on clarifying than without the expectancy. Similarly, a teacher who expects feedback on classroom management will monitor student behavior more carefully than without the expectancy.

Thirdly, feedback can help to implement innovations in teaching. Professional development aims to achieve change with regard to teachers’ attitudes, beliefs, and perceptions that will result in improved student achievement or other desired outcomes. It has been shown by research that changing teachers’ attitudes, beliefs, or perceptions requires the experience of successful implementation (Guskey, 2002). Thus, student feedback is one key element in the implementation process, being able to demonstrate that an innovation works (or doesn’t work). Student feedback was found to increase implementation of innovations (Mortenson & Witt, 1998; Noell et al., 2002), providing information on whether innovations have a positive effect.

4.2 *Democratization of Schools*

The development of teaching is often focused on effectiveness, aiming at an increase in student achievement. Student feedback can contribute to the development of teaching in an additional way by promoting democratic attitudes. Feedback between the participants of school life is a basic condition for participation and therefore the

experience of democratic structures. While professional feedback from teachers to students is the basis for an appreciative climate and successful teacher–student relationships, student feedback is the basic form of freedom of expression with regard to successful learning conditions and the prerequisite for a dialogue about teaching and learning (Wisniewski et al., 2019).

School has an indirect or latent influence on the political socialization of pupils in the sense of the social-cognitive learning theory. A prerequisite for successful interaction is the granting of mutual recognition and appreciation, not as a sufficient, but at least as a necessary condition for a democratic form of school life. “When developing a political standpoint, young people apparently pay less attention to bold confessions and teachings than to the nuances both in interpersonal relationships and in the context of educational institutions” (Kleeberg-Niepage, 2012, p. 13, translated). The participation of young people in discussions and co-determination processes in educational institutions plays an important role (see Chap. 13 of this volume).

Student feedback contains several components of a basic understanding of democracy: Students are given the opportunity to express their opinions in a differentiated way. They have to think about how different criteria for the quality of teaching are to be assessed in each individual case instead of assessing teaching in general as “great” or “bad”. They realize that their own opinion is not to be seen as absolute, but that there are different perspectives on a subject. They learn to engage in a dialogue with their teachers on how changes can lead to better conditions for all those involved and thus influence an area relevant to them—nothing other than social participation in the school system. And finally, feedback offers the opportunity for mutual appreciation between teachers and learners.

4.3 Improving Teachers’ Satisfaction and Health

Student feedback can help to improve teaching, but an additional—and often overlooked—potential benefit it provides is teachers’ development of a professional experience that is more satisfactory, and—as a consequence—healthier. Although, this may seem contra-intuitive because feedback can (and often does) include criticism, research suggests a cautious assumption of such positive effects.

Teachers’ satisfaction is a key affective reaction to working conditions and an important predictor of teacher attrition (Ford et al., 2018). It is related to their expectations of self-efficacy, in other words the belief that they can produce desirable changes in student achievement (Ford et al., 2018; Skaalvik & Skaalvik, 2007; Wang et al., 2015), which in turn is enhanced by feedback. Teachers with low self-efficacy expectations do not believe that they can successfully provide instruction that will increase student performance (Finnegan, 2013), whereas teachers who experience that their use of feedback leads to positive changes in their practice have higher satisfaction than those who don’t (Ford et al., 2018). When teachers are given areas to improve or reflect on, their perception of the effectiveness is higher than when only praise is given (Milanowski & Heneman, 2001).

Enns and colleagues (2002) have been able to demonstrate that teachers who seek regular feedback in their professional practice

- have the feeling of being encouraged as teachers,
- gain in perceived safety,
- put their own weaknesses into perspective,
- establish working partnerships,
- establish a research-oriented attitude in the classroom,
- develop openness and sensitivity,
- increase their job satisfaction,
- reduce stress factors,
- experience self-efficacy, and
- benefit from recognition.

In this sense, feedback does not—as one might expect—demotivate teachers by criticism, but, contrary to this, support and encourage them. Feedback even has this motivating effect, regardless of whether it is positive or negative (Pritchard et al., 2002). Further, it leads to a more realistic self-assessment (Mayo et al., 2012), promotes a solution-oriented approach to problems (Enns et al., 2002), and increases the experience of self-efficacy. Considering this, the reflection on lessons with the help of external data can be one of the most important resources for satisfactory professional practice.

Finally, job satisfaction has an effect on teachers' health. Symptoms of burnout (emotional exhaustion and depersonalization) are negatively related to teacher self-efficacy (Skaalvik & Skaalvik, 2007) and teachers with a high sense of efficacy seem to employ a pattern of strategies that minimized negative emotions (Finnegan, 2013). It is at least plausible that an increase in the above-mentioned areas will in turn have a positive effect on the quality of teaching. Reciprocally, students give more positive feedback to teachers who—in the sense of a low psychosocial risk for stress symptoms—show a favorable combination of work commitment, resilience and emotions, a high degree of resistance to professional problems, and a higher level of positive emotions (Klusmann et al., 2006). Consequently, student feedback can make a significant contribution not only to job satisfaction, but to the health of teachers.

5 Success Conditions of Student Feedback

We have tried to show in this chapter that student feedback has a number of important functions for the development of teaching and teachers. However, there are several success conditions that are a prerequisite for student for feedback to be able to really fulfill these functions. Therefore, we propose the following four criteria:

1. The aim of student feedback needs to be transparent to all participants.

Formative student feedback with the purpose of personal development must be clearly separated from any forms of summative evaluations, assessments, or ratings that are used for administrative decisions. Transparency is also needed with regard to the availability of feedback results: the obtaining teacher should be able to decide who has access to these results.

2. Student feedback needs to be informative.

Feedback is most useful when it contains a high amount of information (Hattie & Timperley, 2007; Wisniewski et al., 2020b). Consequently, student feedback should provide information that allows the teacher to gain detailed insight into strengths and weaknesses of her or his teaching, pointing at opportunities to make suitable changes and reinforcing functional behavior.

3. Student feedback needs to be based on sound criteria.

In many schools, ad hoc instruments that are mainly based on everyday assumptions and not on sound theory are used to obtain student feedback (Ory & Ryan, 2001). This brings the disadvantage that criteria are highly subjective and arbitrary. Useful student feedback is based on criteria whose importance is supported by empirical evidence and which cover deep structures of teaching (with positive effects on student learning).

4. Teachers need support when dealing with student feedback.

The most crucial step in the process of using student feedback is not obtaining information but dealing with the information. Penny and Coe (2004) have shown the importance of supporting teachers when dealing with feedback information. High impact was found when teachers had various support systems at hand, including counseling and coaching.

6 Conclusion

The various functions of student feedback suggest that it is a self-evident part of teachers' professional development, providing valuable information with no or low cost. It is therefore rather astonishing, that it is still not a matter of course in schools. Student feedback helps to get into conversation about teaching and learning. Sometimes this is the beginning of a real feedback culture.

References

- Alberts, H. J., Martijn, C., & de Vries, N. K. (2011). Fighting self-control failure: Overcoming ego depletion by increasing self-awareness. *Journal of Experimental Social Psychology*, 47(1), 58–62. <https://doi.org/10.1016/j.jesp.2010.08.004>.
- Bromme, R., & Haag, L. (2004). Forschung zur Lehrerpersönlichkeit [Research on teacher personality]. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* [Manual of school research] (pp. 777–794). VS Verlag. https://doi.org/10.1007/978-3-663-10249-6_31.
- Buhren, C. G. (2015). Feedback–Definitionen und Differenzierungen [Feedback definitions and differentiations]. In *ibid.* (Ed.), *Handbuch Feedback in der Schule* [Manual on feedback in schools] (pp. 11–30). Beltz. https://doi.org/10.1007/978-3-658-10223-4_2.
- Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological Bulletin*, 92(1), 111–135. <https://doi.org/10.1037/0033-2909.92.1.111>.
- Carver, C. S., & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychological Review*, 97(1), 19–35. <https://doi.org/10.1037/0033-295X.97.1.19>.
- Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education*, 28(1), 71–88. <https://doi.org/10.1080/02602930301683>.
- Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–668. <https://doi.org/10.1037/0033-2909.125.6.627>.
- Desimone, L. M. (2009). Improving impact studies on teachers’ professional development: Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199. <https://doi.org/10.3102/0013189X08331140>.
- Ditton, H., & Arnoldt, B. (2004). Wirksamkeit von Schülerfeedback zum Fachunterricht [Effectiveness of student feedback on subject teaching]. In J. Doll (Ed.), *Bildungsqualität von Schule [Quality of education in schools]* (pp. 152–170). Waxmann.
- Elstad, E., Lejonberg, E., & Christophersen, K. A. (2017). Student evaluation of high-school teaching: Which factors are associated with teachers’ perception of the usefulness of being evaluated? *Journal for Educational Research Online*, 9(1), 99–117.
- Enns, E., Rüegg, R., Schindler, B., & Strahm, P. (2002). *Lehren und Lernen im Tandem. Porträt eines partnerschaftlichen Fortbildungssystems* [Teaching and learning in tandem: Portrait of a continuing education system based on partnership]. Zentralstelle für Lehrerinnen- und Lehrerfortbildung Kanton.
- Finnegan, R. S. (2013). Linking teacher self-efficacy to teacher evaluations. *Journal of Cross-Disciplinary Perspectives in Education*, 6(1), 18–25.
- Ford, T. G., Urlick, A., & Wilson, A. S. (2018). Exploring the effect of supportive teacher evaluation experiences on US teachers’ job satisfaction. *Education Policy Analysis Archives*, 26(59), 83–93. <https://doi.org/10.14507/epaa.26.3559>.
- Gärtner, H. (2007). *Unterrichtsmonitoring [Classroom monitoring]*. Münster, Germany: Waxman.
- Gärtner, H. (2013). Wirksamkeit von Schülerfeedback als Instrument der Selbstevaluation von Unterricht [Effectiveness of student feedback as a tool for self-evaluation of teaching]. In J. Hense, S. Rädiker, W. Böttcher, & T. Widmer (Eds.), *Forschung über Evaluation. Bedingungen, Prozesse und Wirkungen* [Research on evaluation: Conditions, processes and effects] (pp. 107–124). Waxmann.
- Gates Foundation. (2012). *Asking students about teaching: Student perception surveys and their implementation* (Policy & Practice Brief). Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf.
- Guskey, T. R. (2002). Professional development and teacher change. *Teachers and Teaching: Theory and Practice*, 8(3/4), 381–391. <https://doi.org/10.1080/135406002100000512>.

- Harvey, L. (2002). The end of quality? *Quality in higher education*, 8(1), 5–22. <https://doi.org/10.1080/13538320220127416>.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. <https://doi.org/10.4324/9780203887332>.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>.
- Helmke, A. (2017). Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts [Quality of teaching and teacher professionalism: Diagnosis, evaluation and improvement of teaching]. Klett-Kallmeyer.
- Helmke, T., Helmke, A., Schrader, F.-W., Wagner, W., Nold, G., & Schröder, K. (2008). Die Videostudie des Englischunterrichts [The video study of English teaching]. In DESI-Konsortium (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* [Teaching and competence acquisition in German and English: Results of the DESI study]. (pp. 345–363). Beltz.
- Kember, D., Leung, D. Y., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment & Evaluation in Higher Education*, 27(5), 411–425. <https://doi.org/10.1080/0260293022000009294>.
- Kleeberg-Niepage, A. (2012). Zur Entstehung von Rechtsextremismus im Jugendalter—oder: Lässt sich richtiges politisches Denken lernen? [On the emergence of right-wing extremism in adolescence—Or: Can proper political thinking be learned]. *Journal für Psychologie [Journal for Psychology]*, 20(2), 1–30.
- Klusmann, U., Kunter, M., Trautwein, U., & Baumert, J. (2006). Lehrerbelastung und Unterrichtsqualität aus der Perspektive von Lehrenden und Lernenden [Teacher workload and teaching quality from the perspective of teachers and students]. *Zeitschrift für pädagogische Psychologie [Journal for Educational Psychology]*, 20(3), 161–173. <https://doi.org/10.1024/1010-0652.20.3.161>.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>.
- Kunter, M., & Voss, T. (2013). The model of instructional quality in COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss, & M. Neubrand (Eds.), *Cognitive activation in the mathematics classroom and professional competence of teachers* (pp. 85–113). Springer. <https://doi.org/10.1007/978-1-4614-5149-5>.
- Luft, J., & Ingham, H. (1955). *The Johari window, a graphic model for interpersonal relations*. University of California.
- Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research*. Agathon Press.
- Mayo, M., Kakarika, M., Pastor, J. C., & Brutus, S. (2012). Aligning or inflating your leadership self-image? A longitudinal study of responses to peer feedback in MBA teams. *Academy of Management Learning & Education*, 11(4), 631–652.
- Milanowski, A. T., & Heneman, H. G. (2001). Assessment of teacher reactions to a standards-based teacher evaluation system: A pilot study. *Journal of Personnel Evaluation in Education*, 15(3), 193–212. <https://doi.org/10.1023/A:1012752725765>.
- Mortenson, B. P., & Witt, J. C. (1998). The use of weekly performance feedback to increase teacher implementation of a prereferral academic intervention. *School Psychology Review*, 27(4), 613–627.
- Newton, J. (2000). Feeding the Beast or Improving Quality? academics' perceptions of quality assurance and quality monitoring. *Quality in Higher Education*, 6(2), 153–163. <https://doi.org/10.1080/713692740>.
- Noell, G. H., Duhon, G. J., Gatti, S. L., & Connell, J. E. (2002). Consultation, follow-up, and implementation of behavior management interventions in general education. *School Psychology Review*, 31, 217–234.

- Northcraft, G. B., Schmidt, A. M., & Ashford, S. J. (2011). Feedback and the rationing of time and effort among competing tasks. *Journal of Applied Psychology*, 96(5), 1076–1086. <https://doi.org/10.1037/a0023221>.
- Ory, J., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, 109, 27–44. <https://doi.org/10.1002/ir.2>.
- Paseka, I., Schraz, M., & Schrittmesser, I. (2011). Professionstheoretische Grundlagen und thematische Annäherung [Professional theoretical foundations and thematic approach]. In *ibid.* (Eds.), *Pädagogische Professionalität quer denken – undenken – neu denken* [Thinking pedagogical professionalism from a different angle—Rethinking—Thinking new] (pp. 187–21). Facultas.
- Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, 74(2), 215–253. <https://doi.org/10.3102/00346543074002215>.
- Pritchard, R. D., Holling, H., Lammers, F., & Clark, B. D. (2002). *Improving organizational performance with the productivity measurement and enhancement system: An international collaboration*. Nova Science.
- Reh, S. (2004). Abschied von der Profession, von Professionalität oder vom Professionellen? [Farewell to the profession, to professionalism or the professional]. *Zeitschrift für Pädagogik*, 50(3), 358–372.
- Schön, D. A. (1987). *Educating the Reflective Practitioner*. Jossey-Bass.
- Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. <https://doi.org/10.3102/0034654307310317>.
- Skaalvik, E. M., & Skaalvik, S. (2007). Dimensions of teacher self-efficacy and relations with strain factors, perceived collective teacher efficacy, and teacher burnout. *Journal of Educational Psychology*, 99(3), 611–625. <https://doi.org/10.1037/0022-0663.99.3.611>.
- Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education*, 27(5), 397–409. <https://doi.org/10.1080/0260293022000009285>.
- Sproule, R. (2000). Student evaluation of teaching: Methodological critique. *Education Policy Analysis Archives*, 8(50), 125–142. <https://doi.org/10.14507/epaa.v8n50.2000>.
- Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review*, 21(3), 287–294. [https://doi.org/10.1016/S0272-7757\(01\)00025-5](https://doi.org/10.1016/S0272-7757(01)00025-5).
- Staub, F. (2001). Fachspezifisch-pädagogisches Coaching: Theoriebezogene Unterrichtsentwicklung zur Förderung von Unterrichtsexpertise [Specialised pedagogical coaching: Theoretical teaching development to promote teaching expertise]. *Beiträge zur Lehrerinnen-und Lehrerbildung*, 19(2), 175–198.
- Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty’s teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22–42. <https://doi.org/10.1016/j.stueduc.2016.08.007>.
- Wang, H., Hall, N. C., & Rahimi, S. (2015). Self-efficacy and causal attributions in teachers: Effects on burnout, job satisfaction, illness, and quitting intentions. *Teaching and Teacher Education*, 47, 120–130. <https://doi.org/10.1016/j.tate.2014.12.005>.
- Wisniewski, B., Engl, M., & Zierer, K. (2019). Mehr Demokratie wagen: Warum Schülerfeedback Schule demokratischer macht [Dare more democracy: Why student feedback makes schools more democratic]. *Schulverwaltung BY*, 3, 68–72.
- Wisniewski, B., & Zierer, K. (2019). Visible Feedback—From Research to Reality. *Kappa Delta Pi Record*, 55, 66–71. <https://doi.org/10.1080/00228958.2019.1580984>.
- Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. (2020a). Obtaining students’ perceptions of instructional quality—Two-level structure and measurement invariance. *Learning and Instruction*, 66(2). <https://doi.org/10.1016/j.learninstruc.2020.101303>.

- Wisniewski, B., Zierer, K., & Hattie, J. (2020b). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>.
- Zierer, K. (2015). Nicht nur Wissen und Können, sondern auch und vor allem Wollen und Werten. Das K3W-Modell im Zentrum pädagogischer Expertise [Not only knowledge and ability, but also and above all willingness and values. The K3W model at the centre of pedagogical expertise]. *Pädagogische Rundschau [Pedagogical Review]*, *69*(1), 91–98. <https://doi.org/10.1163/25890581-091-01-90000008>.
- Zierer, K., & Wisniewski, B. (2018). *Using Student Feedback for Successful Teaching*. Routledge. <https://doi.org/10.4324/9781351001960>.

Benedikt Wisniewski is a school psychologist, former teacher, and teacher trainer. As a researcher and lecturer, he works at the University Augsburg (Germany). His research focuses on students' and teachers' perceptions of instructional quality and the validity of student feedback. In 2014, he co-founded *FeedbackSchule*, a German online platform for obtaining professional student feedback.

Klaus Zierer is a Professor of School Education at the University of Augsburg (Germany). He taught for five years in primary and secondary schools. In 2009 he was a Visiting Research Fellow at the Department of Education, University of Oxford (UK). He is currently an Associate Research Fellow of the ESRC Centre on Skills, Knowledge and Organisational Performance (SKOPE). His research interests include international aspects of school education, learning and teaching, teacher education, and professionalization.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 9

Effects of Student Feedback on Teaching and Classes: An Overview and Meta-Analysis of Intervention Studies



Sebastian Röhl

Abstract Based on a comprehensive literature review of student feedback intervention studies in schools, this chapter provides an overview of found effects on teachers and teaching. The first part summarizes the self-reported cognitive, affective, and motivational effects of student feedback on teachers, which can subsequently lead to behavioral changes in the classroom. In the second part, the focus is on the extent to which these behavioral changes are perceived by students. For the first time, a meta-analysis of changes in students' perceptions of teaching was carried out for the 18 existing longitudinal studies for this purpose. A small but significant positive weighted mean effect size of $d=0.21$ for students' perceived improvement of teaching quality was found, while more in-depth analyses pointed to a beneficial effect of individual support measures for teachers regarding reflection and subsequent development of teaching. Implications for further research and practical implementation of student feedback in schools are discussed.

Keywords Student feedback · Meta-analysis · Effects · Intervention studies · Teacher · Teaching development

1 Introduction

Feedback can be understood as a communicative process “in which some sender [...] conveys a message to a recipient. In the case of feedback, the message comprises information about the recipient” (Ilgen et al., 1979, p. 350). This information can be used by the recipient to improve task performance (Kluger & DeNisi, 1996) or to enable and develop learning processes (Hattie & Timperley, 2007). In the case of student feedback, the feedback recipients are teachers, who receive information on teaching from their students in class as senders. As described in the Introduction of this volume and Chap. 8 by Wisniewski and Zierer, the received feedback

S. Röhl (✉)
Institute for Educational Sciences, University of Education,
Freiburg, Germany
e-mail: sebastian.roehl@ph-freiburg.de

should contain useful and meaningful information for the given teacher. As a first step, the feedback could therefore have positive cognitive and possibly also affective and motivational effects on the teacher. Subsequently, this could lead to changes in teacher behavior, thus promoting development and improvement of teaching and professionalism. This in turn could lead to a more positive perception of teaching by students.

This overview chapter follows this process and is based on a comprehensive literature review of studies dealing with student feedback as an intervention for the improvement of the teaching quality of fully trained teachers. In the first part, findings on teacher-reported effects from student feedback are summarized. The second part contains a meta-analysis of findings of longitudinal student feedback intervention studies, which almost exclusively examined changes in teaching and classes from the perspective of students in secondary schools. Remarkably, no studies could be found which were conducted in grades one to four.

This chapter complements Chap. 11 by Göbel et al., which describes the use of student feedback in the context of the first and second phases of teacher training. In Chap. 12, Schmidt and Gawrilow describe how student feedback can be used to improve the cooperation between teachers and students. Furthermore, teachers' productive use of student feedback depends on various individual and situational characteristics, and this is described by Röhl and Gärtner in Chap. 10 and in the Introduction of this volume.

2 Self-Reported Effects of Student Feedback on Teachers

Whether a feedback message leads to visible changes in the recipient's behavior depends on the effects of the feedback message on the recipient—in this case the teacher. Therefore, this part offers an overview of literature on self-reported effects of student feedback on teachers. For the teacher obtaining feedback, student feedback can have effects at different levels (see *processes and effects of student feedback model (PESF)* in the Introduction of this book). Here, a distinction can be made between affective, cognitive, and behavioral effects, which in turn are related to motivational processes.

Regarding *cognitive* effects of obtaining student feedback, several studies reported an increasing amount of reflection by teachers on their actual practice due to aspects of teaching quality included in the used feedback questionnaires (Gärtner & Vogt, 2013; Göbel & Neuber, 2019; Mandouit, 2018). As a result of the feedback received, teachers express an improvement regarding their understanding of how students perceive their teaching and classes (Gage, 1963; Thorp et al., 1994; Wyss et al., 2019). Furthermore, student feedback can help teachers to find students' misconceptions about learning (Mandouit, 2018). Subsequently, teachers identified possible areas for improvement (Barker, 2018; Gaertner, 2014). As a side effect, the first-time use of student feedback can lead to a more positive attitude towards this instrument

(Brown, 2004; Campanale, 1997; Gaertner, 2014), although opposite effects such as a higher skepticism have also been observed (Dretzke et al., 2015).

On the *affective* level, many teachers experience emotions of happiness and curiosity during the feedback reception and reflection, especially if the feedback is perceived as positive (Villa, 2017). Other teachers reported emotions of anger due to feedback perceived as negative, or sadness due to helplessness regarding a possible improvement of their own teaching (Brown, 2004; Gärtner & Vogt, 2013; Villa, 2017).

Both cognitive and affective effects can impact motivational processes and lead to changes on the *behavioral* level. Teachers expressed that they paid more attention to identified improvement areas during preparation and teaching, sometimes resulting in a self-perceived improvement (Balch, 2012; Gaertner, 2014; Rösch, 2017). In addition, some teachers planned to participate in relevant professional training programs (Balch, 2012). Another behavioral outcome is the discussion about feedback received and teaching with the corresponding class, which was seen by many teachers as an important further source of information about their own teaching and a common ground for changing teaching practices (Gaertner, 2014; Thorp et al., 1994). In addition, teachers mentioned changes in their behavior before obtaining student feedback. While reflecting on the feedback questionnaire, they prepared the lessons in which the instrument was to be used more carefully, in line with the questionnaire's quality criteria (Balch, 2012; Rösch, 2017).

3 A Meta-Analysis of Longitudinal Studies on the Teaching-Related Effects of Student Feedback Interventions

Without a doubt, it is desirable that positive effects of student feedback are not only reported by teachers, but that they also become evident in student perceptions and learning achievement. Based on the process model of Student Feedback on Teaching (SFT, see the Introduction to this book), this process can only be achieved if several conditions are met. First of all, the students have to report back that there is a need for improvement. This must be perceived and accepted by the teacher in the feedback reports. Furthermore, it is necessary that the teacher creates a desire for change or sets goals and then pursues them. Subsequently, a teacher's behavioral change should improve students' learning processes—and students have to perceive this behavioral change—before a positive effect of student feedback on teaching and classes becomes visible.

While intervention studies on the use of students' achievement data for the instructional development in schools also focus on student achievement (e.g. Keuning et al., 2019; van der Scheer & Visscher, 2018) or the improvement of teachers' instructional skills (van der Scheer et al., 2017), the overwhelming focus of investigations into student feedback has been on the effects of student perception of teaching behavior. In

the literature review performed here, one single study (Novak, 1972) was found which additionally analyzed several audio-recorded lessons before and after the student feedback intervention on changes in teacher behavior. The findings of this study pointed to significantly lower proportions of teacher talk and lectures during the lessons following repeated reception of student feedback. Regarding possible effects of student feedback interventions on students' motivation, findings of a single study (Tozoglu, 2006) indicated a small positive effect ($d = 0.289$), but only for teachers who received enhanced support for interpreting feedback and teaching development. No effect was found for teachers who received only student feedback mean scores without any support. In a dissertation study, Kime (2017) measured students' achievement scores in the context of teaching evaluations based on student ratings, comparing a group of teachers receiving student feedback only with another group which carried out additional peer coaching on the feedback received. Contrary to Kime's expectations, analysis could not prove a significant effect on achievement scores for the peer coaching condition. However, a comparison with teachers who did not receive student feedback was not possible due to the lack of an appropriate control group. With regard to the question of the extent to which primary school pupils perceive an improvement in the quality of teaching, a study by van der Scheer (2016) resulted in no changes in pupils' rating of teaching quality during a data-based decision-making intervention, whereas pupils' learning achievement significantly improved. Research concerning effects on students' learning achievement, comparing teachers receiving student feedback with non-receivers, is still absent.

While in the field of university and college teaching some meta-analyses of effects of students' mid-term feedback on classes already exist (e.g. Cohen, 1980; L'Hommedieu et al., 1990; Penny & Coe, 2004), a meta-analysis regarding effects in schools is still pending. The meta-synthesis regarding feedback by Hattie (2009), which resulted in $d = 0.73$, and also a recent and thorough meta-analysis of the underlying primary studies with a lower effect size of $d = 0.48$ (Wisniewski et al., 2020), mainly include feedback from teachers to students, with the exception of three meta-analyses of effects of student feedback in higher education. For the context of higher education, Cohen's (1980) meta-analysis of 17 intervention studies resulted in an effect size of $d = 0.20$ on students' end-of-semester ratings of classes for providing mid-term feedback to university teachers. If the feedback is accompanied by further measures such as individual consultation, this effect increases to an average of $d = 0.64$. Penny and Coe (2004) found an average effect size of $d = 0.69$ for student feedback augmented with peer and expert consultation in their analysis of 11 intervention studies. The analysis of 28 studies by L'Hommedieu et al., (1990) resulted in $\Delta = 0.34$. Uttl et al. (2017) conducted a meta-analysis of 51 studies on the relation between student evaluation of teaching ratings and student learning achievement. The results indicated no significant overall correlation. In order to close this research gap in the field of primary and secondary schools, a meta-analysis is now presented here, which includes student feedback intervention studies while surveying changes in students' perception of teaching quality.

3.1 Measures and Methods

3.1.1 Literature Search

For this overview, a comprehensive literature search using the terms “student feedback”, “pupil feedback”, and “self-evaluation” was conducted in the databases ERIC, PsycInfo, Scopus, Web of Science, ProQuest, and OpenDissertations. As most of the studies found focus on student feedback in higher education, the search was limited to publications which did not contain this keyword. In a second step, articles with a theoretical or practical focus were excluded. Next, only intervention studies which reported pre- and post-measures were selected. In addition, some non-catalogued studies mentioned in scientific articles on student feedback were found. More details about the studies included can be seen in Part III.

3.1.2 Study Coding

Regarding possible moderators, most of the different study characteristics are explicitly reported, such as the existence of a control group, the number of feedback reports, the duration of the treatment, and the publication type. For the level of provided support for the participating teachers (see below), a coding was conducted by two trained raters. The inter-rater agreement was high ($\rho = 0.85$, $p < 0.001$), and in the subsequent discussion a consensus was reached on the different opinions.

3.1.3 Effect Size Calculation and Analysis

The dependent variable in this meta-analysis is the student-perceived change in the quality of teaching. As the included studies use different questionnaires for student feedback, single scales or constructs are not comparable across the studies. Therefore, in order to achieve comparability of the effects, it was decided to calculate the arithmetic mean of all reported effect sizes included in each study for the students' perception of teaching as an overall effect.

Effect sizes are calculated using Cohen's d with groups-size-adjusted standard deviation (σ_{pooled} , Morris & DeShon, 2002). Effect size variances were estimated following Lipsey and Wilson (2001, pp. 44–49). If available, d was estimated using the reported means and standard deviations of pre- and post-measurements on teacher level. Otherwise, available t , F , and χ^2 statistics were used.

In this meta-analysis, longitudinal studies with and without a control group design are included. This led to some problems in the estimation of comparable effect sizes and variances:

- (a) Several studies without control groups didn't include the standard deviations of the measurements and the correlation between the pre- and post-test scores. While comparable effect sizes can be estimated without this information using

reported t - or F -values (Lipsey & Wilson, 2001), the variances of the effect sizes can only be estimated if the standard deviations or correlations are available. For this meta-analysis, several solutions were considered. The most conservative approach would be to assume no correlation between the two measurement time points, which would lead to a strong overestimation of variances. However, many studies report quite high consistency of student ratings on teaching quality over time (e.g. Polikoff, 2015; Rowley et al., 2019). In addition, the calculation of the correlation between teachers' pre- and post-measures using available data from two studies (Bartel, 1970; Ditton & Arnold, 2004) results in values of $r > 0.73$. Therefore, following the suggestions of Borenstein et al. (2009), we assumed a lower limit of $r = 0.70$ for the estimation of effect sizes variances.

- (b) Many studies with a control group design showed a moderate decrease of control groups' student ratings on teaching quality between the measurement time points (Buurman et al., 2018; Gage, 1963; Nelson et al., 2015; Tacke & Hofer, 1979; Tuckman & Oliver, 1968). For the studies using a control group design, this tendency is already considered in the estimation of effect sizes. However, assuming that this effect is also evident in the treatment group, this could lead to an underestimation of the strength of the effect in designs without the control group. Therefore, possible moderator effects regarding the design of the study are included in our analyses.

Because of the heterogeneity of treatment and design characteristics of the included studies, random-effect models appeared to be more suitable than fixed-effect models for this meta-analysis (Borenstein et al., 2009). For the estimation of assumed moderator effects of study and treatment characteristics, separate mean weighted effect sizes and confidence intervals for every subgroup were estimated (Borenstein et al., 2009). Regarding continuous study characteristics such as the number of feedback reports and the intervention duration, the studies were split at the median. Estimation of the overall and moderator effect sizes and confidence intervals was done using the package *metafor* (Viechtbauer, 2010) in *R* (R Core Team, 2019). In addition, as three studies included several effect sizes by different intervention groups, a sensitivity analysis was conducted with regard to bias due to possible dependencies (Hedges et al., 2010). This revealed that the resulting biases are about $d = 0.0001$, and therefore negligible.

Analysis on possible outliers or influential studies was conducted. We chose to use Cook's distance (Cook & Weisberg, 1982) test statistics for residual heterogeneity when each study is removed in turn (Viechtbauer, 2010), and the distribution of weights of the included studies as indicators.

3.2 Characteristics of Included Studies

In the literature review, 18 longitudinal studies with student feedback treatments published between 1960 and 2019 were identified (see Table 1). The design of these studies is experimental or quasi-experimental. Thus, all studies include at least one pre- and one post-measurement of students' perception of teaching quality, but not all of them provide a control group comparison. Seven of the studies were conducted in the USA, three more took place each in Australia and Germany, two in the Netherlands, and one each in Great Britain, Turkey, and Austria.

All studies utilized questionnaires which were mainly based on closed questions or rating scales. The research teams carried out the counting and provided a feedback report to the teachers. One study used a digital smartphone-based feedback system for this purpose (Bijlsma et al., 2019). All included studies were conducted in grade 5–13. While five interventions were limited to exactly one grade level, the other studies involved teachers from different levels. Three interventions also continued to restrict the subject matter for a better comparability of the classes. Novak (1972) focused on biology teachers, Rösch (2017) on physics, and Bijlsma et al. (2019) on mathematics.

The findings on the effects of a student feedback intervention on changes in teaching behavior perceived by students are heterogeneous in the studies. While two studies show clearly negative treatment effects (Bennett, 1978, $d = -0.30$; Knox, 1973, $d = -0.24$),¹ most studies report effects ranging from $d = 0.1$ to $d = 0.5$.

Furthermore, some studies instruct teachers to focus on only one to three areas for improvement in subsequent classroom development (Fraser & Fisher, 1986; Fraser et al., 1982; Nelson et al., 2015; Thorp et al., 1994). However, information on which aspects were selected by teachers for improvement is only available for the three case studies. As expected, results show the highest improvements in the targeted areas (up to $d = 0.8$), whereas the other scales do not change. Another study (Mayr, 1993, 2008) examined only individual areas of teaching which had been agreed with the teachers. However, as there is a complete lack of such information for all other studies, the individual prioritization of certain areas by individual teachers cannot be considered in this meta-analysis, and so we used the average effect sizes of all scales in each study. This also means that the average overall effects of all included scales are smaller than the reported bigger improvements in some selected scales.

The *sample size* differs greatly between the studies. Whereas some have reported case studies with single teachers (Fraser & Fisher, 1986; Fraser et al., 1982; Thorp et al., 1994) or one team of five teachers (Mandouit, 2018), the other studies used sample sizes ranging from $N = 10$ to $N = 508$ teachers. Also, the duration of the intervention varied between the studies from one month to one year, with an average of $M = 3.06$ months. During these periods, a different *number of feedbacks* were reported to the teachers. In most of the studies, the last feedback report was used

¹ Noteworthy, both studies were conducted by persons from the school administration. To what extent the negative effect can be explained by possible refusal attitudes of subordinate teachers cannot be clarified here due to the small number of studies.

Table 1 Studies included in the meta-analysis

| No. | Study | Country, types of school and grades | N teachers | | Duration in months | Number of feedbacks | Level of support | Effect size <i>d</i> | Publication type |
|--|---|-------------------------------------|------------|---------|--------------------|---------------------|------------------|----------------------|------------------|
| | | | Treatment | Control | | | | | |
| <i>Intervention studies with control group</i> | | | | | | | | | |
| 1 | Gage et al. (1960, 1963) | USA, secondary, 6 | 86 | 90 | 1.5 | 1 | Low | 0.17 | Journal |
| 2 | Tuckman and Oliver (1968) | USA, high and vocational, 10–13 | 186 | 100 | 3 | 1 | Low | 0.30 | Journal |
| 3 | Bartel (1970) • Feedback Group • Counselling Group | USA, secondary, 7–12 | 14 | 14 | 3 | 1 | Low | 0.08 | Dissertation |
| | | | 14 | 14 | 3 | 1 | High | 0.47 | |
| 4 | Bennett (1978) | USA, high, 9–12 | 19 | 18 | 2 | 3 | Low | −0.30 | Dissertation |
| 5 | Tacke and Hofer (1979) | Germany, secondary, 7–10 | 22 | 22 | 1 | 1 | Low | 0.24 | Journal |
| 6 | Tozoglu (2006) • Feedback only Group • Enhanced Feedback Group ^a | Turkey, high, 9–12 | 20 | 20 | 5 | 1 | Low | 0.61 | Dissertation |
| | | | 20 | 20 | 5 | 1 | High | 1.61 | |
| 7 | Nelson et al. (2015) | USA, middle, 6–8 | 16 | 15 | 1 | 1 | Medium | 0.17 | Journal |
| 8 | Buurman et al. (2018) | Netherlands, vocational, 11–13 | 116 | 126 | 12 | 1 | Low | 0.11 | Report |
| 9 | Bijlsma et al. (2019) | Netherlands, secondary, 9 | 28 | 25 | 4 | 4–17 | Low | 0.13 | Journal |

Intervention studies without control group

| No | Study | Country, types of school and grades | N Teachers | Duration in months | Number of feedbacks | Level of support | Effect size <i>d</i> | Publication type |
|----|--|--|------------|--------------------|---------------------|------------------|----------------------|------------------|
| 10 | Novak (1972) | USA, high, 10 | 18 | 1 | 5 | Low | 0.19 | Dissertation |
| 11 | Knox (1973) | USA, high, 10-12 | 20 | 3 | 2 | Medium | -0.24 | Dissertation |
| 12 | Fraser (1980, 1982) | Australia, elementary & secondary, 6&7 | 2 | 1.5 | 1 | High | 0.21 | Journal |
| 13 | Mayr (1993, 2008) | Austria, lower secondary track | 27 | 2 | 1 | High | 0.83 | Report |
| 14 | Thorp et al. (1994) | Great Britain, comprehensive, 8 | 1 | 1 | 1 | High | 0.65 | Journal |
| 15 | Ditton and Arnold (2004) | Germany, secondary and high, 5-13 | 46 | 8 | 1 | Low | -0.09 | Journal |
| 16 | Bell and Aldridge (2014) • Reflection Group ^a • Action Research Group | Australia, secondary, "upper grades" | 508 52 | 2 2 | 1 1 | Low High | 0.17 0.42 | Journal |
| 17 | Rösch (2017) | Germany, middle and grammar, 9 | 10 | 3 | 4 | Low | 0.12 | Dissertation |
| 18 | Mandouit (2018) | Australia, secondary, n/a | 5 | 3 | 3 | High | 0.47 | Journal |

^aExcluded after outlier analysis

as post-measure of changes in the student perceived teaching quality or teacher behavior. Therefore, for comparability reasons, we counted the number of student feedback reports before the last measurement. Whereas 11 studies reported only one student feedback measurement to the teachers, the other studies obtained and reported feedback up to five times. A special case in point is the study by Bijlsma et al. (2019), where teachers could use the smartphone app to obtain feedback as often as they wanted. The frequency varied between 4 and 17 feedback measurements, with an average of 6.7 for these teachers.

The studies reported here differ also in the manner and amount of *support* provided for the feedback interpretation and subsequent developmental processes. In line with the meta-analysis results from higher education described above (Cohen, 1980; Penny & Coe, 2004), findings on teachers' use of students' achievement data pointed out that solely providing data rarely leads to subsequent changes in teaching (Schildkamp et al., 2015). Thus, it seems to be important to consider this characteristic of the interventions. Furthermore, three of the included studies analyzed different treatment conditions (Bartel, 1970; Bell & Aldridge, 2014; Tozoglu, 2006). One part of the teachers received written feedback without further instructions, while the other part received additional reflection impulses and counseling. All three studies showed significantly more positive effects for the latter condition. For this reason, the effects of these different treatments are reported as two separate effect sizes for each of these studies in the meta-analysis. During the coding process of the support by the raters it became apparent that the following three levels of support can be distinguished:

- Low level of support: General training of student feedback use. This support level includes introductory explanations and training on the use of student feedback before the start of the intervention. These were partly given in written form but also in face-to-face sessions. Also, studies which do not contain explicit descriptions of this topic were assigned to this level. If the information is missing, we assume that the participating teachers were appropriately instructed in the use of the feedback questionnaires and reports.
- Medium level of support: Individual reflection support for the feedback received. This more intense kind of support includes an individualized feedback report with the special marking of possible developmental areas. This occurs in written form and also in face-to-face meetings.
- High level of support: Individual support for subsequent teaching development. Furthermore, some interventions also included ongoing advice on the subsequent development processes through individual or group consultations, counseling, or professional learning communities.

A further distinguishing feature of the studies is the *type of publication*. While the findings of some studies were published in peer-reviewed journals, others were only available as reports or university theses and required a high search effort to find them. If only studies from scientific journals are included in meta-analyses, this easily leads to a so-called "publication bias", since these usually contain higher effects and more significant findings than those not included in such journals (Lipsey & Wilson,

2001). An analysis on differences of effects between publication types could provide indications on whether a publication bias also exists for this research field (Borenstein et al., 2009). Of course, this leaves the question unanswered to what extent further studies exist which could not or cannot be found.

3.3 Results of the Meta-Analysis

A first estimation of the mean weighted effect size using all 21 effect sizes found in a random-effects model resulted in $d = 0.23$ ($p < 0.001$, 95%-C.I.: 0.13–0.33). Analyses of influential studies pointed to an overweight of the reflection group in the study of Bell and Aldridge (2014) because of the exceptional sample size. In addition, analysis of the residual heterogeneity led to the exclusion of the enhanced feedback group from Tozoglu (2006) due to outlier characteristics of this subsample.

For the remaining 19 effect sizes, the estimation of the overall mean weighted effect size led to $d = 0.21$ ($p < 0.001$) with a 95% confidence interval of $0.11 < d < 0.32$. The effect sizes with confidence intervals of all included studies are plotted in Fig. 1.

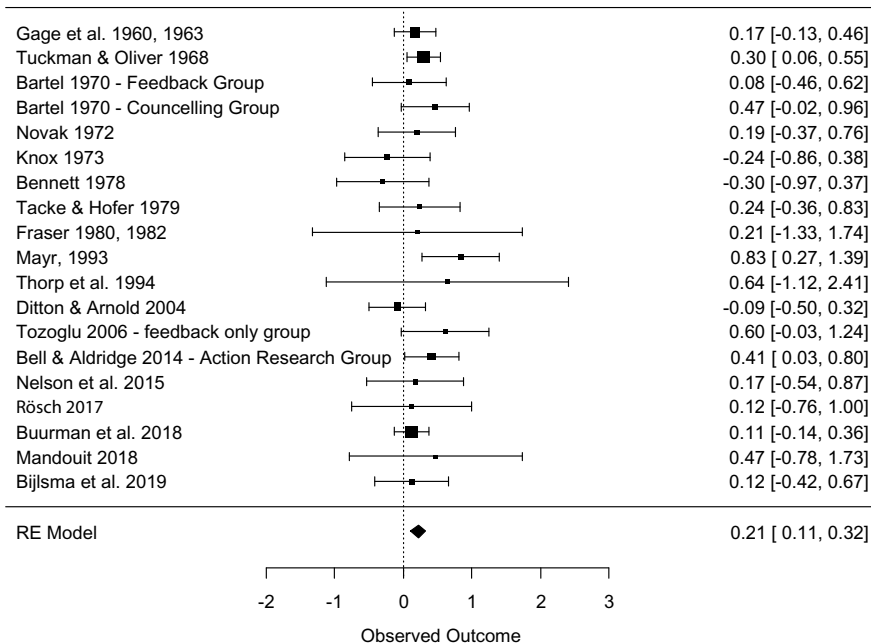


Fig. 1 Forest plot of effect sizes and 95% confidence intervals of included studies and the mean weighted effect size

The inspection of the heterogeneity test statistics ($Q(18) = 16.62, p = 0.549$) reveals that the homogeneity of the effect size is statistically sufficient (Lipsey & Wilson, 2001).

3.3.1 Moderator Analysis

The resulting mean effect sizes and 95% confidence intervals of the subgroups split along the moderator variables are presented in Table 2. In line with the relatively small numbers of studies found, confidence intervals overlap mostly between the different subgroups.

The only study characteristic which turned out to be a significant moderator is the level of support. Treatments with a high level of individual support for reflecting on feedback and teaching development (level 3) showed a significantly higher effect size ($d = 0.52, p = 0.010$) than studies with a medium or low supportive level. Contrary to the assumptions, no significant differences were found between the effect sizes of studies including control groups and studies without ($d = 0.21$ vs. $d = 0.24$). The differences (presumed as considerable) between studies with only one or with more feedback reports ($d = 0.25$ vs. $d = 0.01$) were not statistically relevant ($p = 0.123$). The same applies to the differences regarding the treatment duration of the

Table 2 Analysis of moderator effects regarding study and treatment characteristics

| | <i>n</i> | <i>d</i> | 95%-CI | <i>p</i> ^a |
|-------------------------------|----------|----------|---------------|-----------------------|
| Design of study | | | | |
| with control group | 10 | 0.21 | [0.01; 0.81] | 0.817 |
| without control group | 9 | 0.24 | [-0.04; 0.52] | |
| Level of support ^b | | | | |
| low | 11 | 0.16 | [0.04; 0.28] | 0.089 |
| medium | 2 | -0.06 | [-0.53; 0.40] | 0.235 |
| high | 6 | 0.52 | [0.26; 0.77] | 0.010 |
| Number of feedback reports | | | | |
| 1 | 13 | 0.25 | [0.13; 0.36] | 0.123 |
| 2 or more | 6 | 0.01 | [-0.26; 0.29] | |
| Duration of treatment | | | | |
| 1–2 months | 9 | 0.27 | [0.08; 0.46] | 0.461 |
| 3–12 months | 10 | 0.18 | [0.05; 0.32] | |
| Publication type | | | | |
| Peer reviewed journal | 10 | 0.22 | [0.08; 0.37] | 0.824 |
| Thesis or report | 9 | 0.22 | [0.00; 0.44] | |

^aSignificance of moderator, ^bDummy-coded

intervention and to whether the studies are published in scientific journals or only accessible as theses or reports.

4 Conclusion and Discussion

In this overview chapter, findings of a comprehensive literature review on effects of student feedback interventions in schools were presented. In the first step, effects on teachers were summarized from the literature found. Regarding cognitive effects, studies reported reflective thinking processes on teachers' own perceptions and goals of teaching—initiated by feedback reports and also by questionnaire topics—which could lead to an identification of areas for improvement. In addition, a fostering effect on teachers' understanding of students' perception of teaching and learning processes was observed. Both positive (happiness, joy) and negative (sadness, feelings of helplessness) affective reactions are found with regard to the feedback received. Cognitive and affective processes can result in motivational effects, which could change teachers' behavior in classes. According to teachers' self-reports, these behavioral changes are apparent in a more intense preparation of lessons and a stronger perception and control of one's own actions in class, if they consider the feedback points as critical. Furthermore, teachers initiated discussions with students about the received feedback and the improvement of teaching and collaboration within the school class.

In a second step, this chapter examined whether and to what extent behavioral changes by teachers were perceived by the students. To answer this question, the first meta-analysis of effects of student feedback interventions on student-perceived teaching quality in schools was conducted, including 18 studies with 19 effect sizes. Using a random-effects model, a weighted mean effect size of $d = 0.21$ was found. Although this effect seems to be relatively small, it is significant and lies in a similar range to meta-analyses from student feedback use in higher education (Cohen, 1980; L'Hommedieu et al., 1990). Furthermore, it should be noted that these analyses were based on all the teaching characteristics assessed by the students, but teachers often focused only on specific areas for improvement. For the target areas, the case studies in particular showed considerably greater effects. In addition, the effect sizes varied to a considerable extent between the different scales of teaching dimensions used in the larger studies.

Additional moderator analysis showed an increase in the effect size to $d = 0.52$ for additional individual support, which is also in line with findings for college and university teachers (Penny & Coe, 2004). Other moderator analyses showed no significant effects. This emphasizes the important impact of providing appropriate teacher support for the feedback-related teaching development process, whereas other structural treatment characteristics play no or only a minor role. However, there were indications that further studies should pay particular attention to the number of feedback reports provided in longer-term studies.

Considering the findings of the first part of this chapter on the teacher-reported effects of feedback, the teacher's perception processes and reactions are the "needle's

eye” for improving teaching. Therefore, support for teachers using student feedback should aim at facilitating a constructive cognitive processing of feedback and accompanying affective reactions, so that teachers can develop action alternatives and thus the motivation for change is fostered.

As a limiting factor for the meta-analysis presented, it should be noted that only relatively few studies were found. This reduces the power of the analyses of possible moderators. However, the similarity of the findings presented here to meta-analyses from higher education points toward validity of these results, together with the fact that there is no indication of a publication bias or design effect of the included studies. This chapter thus provides evidence for the effectiveness of student feedback as a tool for improving the quality of teaching perceived by students. It provides a comprehensive overview of the effects on teachers which have so far only been considered in isolation in studies. Furthermore, an extensive literature review and meta-analysis of intervention studies on student feedback in schools was presented for the first time.

Simultaneously, there are various implications for further research on the effects of student feedback in schools:

- With one exception, only intervention studies which measure changes in teaching based on student perceptions or teacher self-reports have been conducted to date. Hence, there is an urgent need for studies which measure changes in teaching using other methods such as video analysis or student achievement.
- The findings of this study point to the importance of additional support to teachers for productive use of student feedback. However, it has not yet been controlled to what extent the supporting measures would have the same positive effect if, for example, self-assessments of teachers were used instead of student feedback.
- Studies should include which areas of improvement have been identified by teachers and analyze these effects separately.
- In addition, there is also a lack of studies which focus both on teachers’ reflection processes on feedback together with the subsequent changes in teaching, perceived by students or external observers.

For the practical use of student feedback for teaching development in schools, this meta-analysis also results in several implications. Most importantly, the findings emphasize the need for support for teachers on using student feedback. This does not only concern the subsequent lesson development, but also support for the interpretation of feedback reports, dealing with accompanying emotions, identification of improvement areas, and how to work on them. This can for example take place through coaching and supervision, but also in collegial settings such as professional learning communities.

Additionally, when planning the implementation of student feedback in schools, there is a need to consider organizational characteristics which are beneficial for constructively dealing with feedback, as presented in Chap. 10 by Röhl and Gärtner in this volume.

Acknowledgements The author wants to thank Prof. Dr. Martin Schwichow and Prof. Dr. Wolfram Rollett from the University of Education Freiburg for their valuable suggestions and support for this study.

References

*Studies included in the meta-analysis are marked with an asterisk**

- Balch, R. T. (2012). *The validation of a student survey on teacher practice*. Dissertation. Vanderbilt University, Nashville, Tennessee.
- Barker, S. (2018). *Student voice to improve instruction: Leading transformation of a school system*. Digital Commons @ ACU.
- *Bartel, B. W. (1970). *The effectiveness of student feedback in changing teacher classroom image*. University of Minnesota.
- *Bell, L. M., & Aldridge, J. M. (2014). Investigating the use of student perception data for teacher reflection and classroom improvement. *Learning Environments Research*, 17, 371–388. <https://doi.org/10.1007/s10984-014-9164-z>.
- *Bennett, C. R. (1978). *A developed and field-tested experiment to test the effect of increased student feedback on specific teacher performance behaviors*. Dissertation. Iowa State University, Ames, IA.
- *Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education*, 28, 217–236. <https://doi.org/10.1080/1475939X.2019.1572534>.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.
- Brown, S. L. (2004). *High school students' perceptions of teacher effectiveness: Student ratings and teacher reflections*. Dissertation. University of New Mexico, Albuquerque, NM.
- *Buurman, M., Delfgaauw, J. J., Dur, R. A. J., & Zoutenbier, R. (2018). *The effects of student feedback to teachers: Evidence from a field experiment* (Tinbergen Institute Discussion Paper). Amsterdam & Rotterdam.
- Campanale, F. (1997). Autoévaluation et transformations de pratiques pédagogiques. *Mesure Et Évaluation En Éducation*, 20(1), 1–24.
- Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education*, 13, 321–341. <https://doi.org/10.1007/BF00976252>.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall.
- *Ditton, H., & Arnold, B. (2004). Wirksamkeit von Schülerfeedback zum Fachunterricht [Effectiveness of student feedback on teaching]. In J. Doll & M. Prenzel (Eds.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsentwicklung* (pp. 152–172). Waxmann.
- Dretzke, B. J., Sheldon, T. D., & Lim, A. (2015). What do K-12 teachers think about including student surveys in their performance ratings? *Mid-Western Educational Researcher*, 27(3), 185–206.
- *Fraser, B. J., & Fisher, D. L. (1986). Using short forms of classroom climate instruments to assess and improve classroom psychosocial environment. *Journal of Research in Science Teaching*, 23, 387–413. <https://doi.org/10.1002/tea.3660230503>.
- *Fraser, B. J., Seddon, T., & Eagleson, J. (1982). Use of student perceptions in facilitating improvement in classroom environment. *Australian Journal of Teacher Education*. <https://doi.org/10.14221/ajte.1982v7n1.3>.
- Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, 42, 91–99. <https://doi.org/10.1016/j.stueduc.2014.04.003>.

- *Gage, N. L. (1963). A method for “improving” teacher behavior. *Journal of Teacher Education*, 14, 261–266. <https://doi.org/10.1177/002248716301400306>.
- Gärtner, H., & Vogt, A. (2013). Wie Lehrkräfte Ergebnisse eines Schülerfeedbacks verarbeiten und nutzen [How teachers process and use results of student feedback]. *Unterrichtswissenschaften*, 41(3), 252–267.
- Göbel, K., & Neuber, K. (2019). Lernende geben Rückmeldungen zum Unterricht: Potenziale der Nutzung von Schülerfeedback und deren Bedingungen [Learners give feedback on teaching: potentials of using student feedback and their conditions]. *Friedrich Jahresheft*, 37, 48–49.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (1st ed., Educational research). Routledge.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77, 81–112. <https://doi.org/10.3102/003465430298487>.
- Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, 1, 39–65. <https://doi.org/10.1002/jrsm.5>.
- Ilgen, D. R., Fisher, C. D., & Taylor, S. M. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4), 349–371.
- Keuning, T., Geel, M., Visscher, A., & Fox, J.-P. (2019). Assessing and validating effects of a data-based decision-making intervention on student growth for mathematics and spelling. *Journal of Educational Measurement*, 56, 757–792. <https://doi.org/10.1111/jedm.12236>.
- Kime, S. J. M. (2017). *Student evaluation of teaching: can it raise attainment in secondary schools? A cluster randomised controlled trial*. Dissertation. Durham University, Durham. <http://etheses.dur.ac.uk/12267/>.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284.
- *Knox, J. T. (1973). *Student feedback as a means of improving teacher effectiveness*. Dissertation. Wayne State University, Detroit, MI.
- L’Hommedieu, R., Menges, R. J., & Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback from student ratings. *Journal of Educational Psychology*, 82, 232–241. <https://doi.org/10.1037/0022-0663.82.2.232>.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Applied Social Research Methods Series, Vol. 49). SAGE.
- *Mandouit, L. (2018). Using student feedback to improve teaching. *Educational Action Research*, 13, 1–15. <https://doi.org/10.1080/09650792.2018.1426470>.
- *Mayr, J. (1993). *Mitarbeit und Störung im Unterricht: Beschreibung und Evaluierung eines Konzepts zur Verbesserung pädagogischen Handelns* [Collaboration and classroom disturbance: description and evaluation of a concept to improve pedagogical action]. Linz.
- *Mayr, J. (2008). Forschungen zum Führungshandeln von Lehrkräften: Wie qualitative und quantitative Zugänge einander ergänzen können [Research on teacher leadership: How qualitative and quantitative approaches can complement each other]. In F. Hofmann, C. Schreiner, & J. Thonhauser (Eds.), *Qualitative und quantitative Aspekte: Zu ihrer Komplementarität in der erziehungswissenschaftlichen Forschung* (pp. 321–341). Waxmann.
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7, 105–125. <https://doi.org/10.1037/1082-989X.7.1.105>.
- *Nelson, P. M., Ysseldyke, J. E., & Christ, T. J. (2015). Student perceptions of the classroom environment: Actionable feedback to guide core instruction. *Assessment for Effective Intervention*, 41, 16–27. <https://doi.org/10.1177/1534508415581366>.
- *Novak, J. H. (1972). *A study of the effects of the use of a pupil response instrument on the behaviors of biological science teachers*. Final Report. Pittsburgh, PA.

- Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research*, 74, 215–253. <https://doi.org/10.3102/00346543074002215>.
- Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education*, 121(2), 183–212.
- R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- *Rösch, S. (2017). *Wirkung und Wirkmechanismen von regelmäßigem Schülerfeedback in der Sekundarstufe: Eine explorative Untersuchung im Physikunterricht* [Effect and impact principles of regular student feedback in secondary education: An exploratory study in physics classrooms]. Dissertation. Universität Basel, Basel.
- Rowley, J. F. S., Phillips, S. F., & Ferguson, R. F. (2019). The stability of student ratings of teacher instructional practice: Examining the one-year stability of the 7Cs composite. *School Effectiveness and School Improvement*, 30, 549–562. <https://doi.org/10.1080/09243453.2019.1620293>.
- Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2015). Data teams for school improvement. *School Effectiveness and School Improvement*, 27, 228–254. <https://doi.org/10.1080/09243453.2015.1056192>.
- *Tacke, G., & Hofer, M. (1979). Behavioral changes in teachers as a function of student feedback: A case for the achievement motivation theory? *Journal of School Psychology*, 17, 172–180. [https://doi.org/10.1016/0022-4405\(79\)90025-6](https://doi.org/10.1016/0022-4405(79)90025-6).
- *Thorp, H. S., Burden, R. L., & Fraser, B. J. (1994). Assessing and improving classroom environment. *School Science Review*, 75, 107–113.
- *Tozoglu, D. (2006). *Effects of student ratings feedback on instructional practices, Teaching Effectiveness, and Student Motivation*. Florida State University.
- *Tuckman, B. W., & Oliver, W. F. (1968). Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology*, 59, 297–301. <https://doi.org/10.1037/h0026022>.
- Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22–42. <https://doi.org/10.1016/j.stueduc.2016.08.007>.
- van der Scheer, E. A. (2016). *Data-based decision making put to the test*. Dissertation. University of Twente, Twente.
- van der Scheer, E. A., Glas, C. A. W., & Visscher, A. J. (2017). Changes in teachers' instructional skills during an intensive data-based decision making intervention. *Teaching and Teacher Education*, 65, 171–182. <https://doi.org/10.1016/j.tate.2017.02.018>.
- van der Scheer, E. A., & Visscher, A. J. (2018). Effects of a data-based decision-making intervention for teachers on students' mathematical achievement. *Journal of Teacher Education*, 69(3), 307–320.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48.
- Villa, L. A. L. (2017). *Teachers taking action with student perception survey data*. Arizona State University.
- Wyss, C., Raaflaub, M., & Hüsler, N. (2019). Schülerrückmeldungen zur Förderung der Partizipation in der Schule [Student feedback to promote participation in school]. In S. Hauser & N. Nell-Tuor (Eds.), *Mündlichkeit* (1st ed., pp. 181–209, Sprache und Partizipation im Schulfeld, Vol. 6). hep.
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>.

Sebastian Röhl has been an Academic Assistant in the Department of Educational Science at the University of Education Freiburg and is currently a Postdoctoral Researcher in the Institute of Education at Tübingen University (Germany). Before that he worked for more than 10 years as a grammar school teacher, school development consultant, and in teacher training. Among other

areas, he conducts research in the fields of teaching development and teacher professionalization through feedback, social networks in inclusive school classes, as well as teachers' religiosity and its impact on professionalism. In addition, he is the director of an inservice professional master's study program for teaching and school development.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 10

Relevant Conditions for Teachers' Use of Student Feedback



Sebastian Röhl and Holger Gärtner

Abstract Based on the findings from research on organizational feedback and data use in schools, this chapter systematizes relevant factors influencing the use of student feedback by teachers in three domains: (1) personal characteristics of feedback recipients (teachers), (2) characteristics of the organization (school), and (3) characteristics of feedback information (data). We identified teachers' self-efficacy, attribution styles, goal orientations, and age or professional experience as relevant individual characteristics. In addition, teachers' attitude toward students' trustworthiness or competence as a feedback provider appeared to be relevant for the use of student feedback. Beyond that, findings on organizational characteristics for teachers' successful dealing with feedback pointed to the importance of a feedback culture and organizational safety, leadership, supportive measures, and perceived function of feedback as control vs. development. Furthermore, relevant characteristics of feedback information were identified as comprehensibility, valence, and specificity. Although such findings from other fields of research have been known for some time, studies on student feedback concerning these aspects are rare. Finally, practical measures are derived for each of the three domains in order to increase the use of student feedbacks by teachers.

Keywords Student feedback · Teacher characteristics · Organizational characteristics · Feedback information · Data use

S. Röhl (✉)
University of Education, Freiburg, Germany
e-mail: sebastian.roehl@ph-freiburg.de

H. Gärtner
Freie Universität Berlin, Berlin, Germany
e-mail: Holger.Gaertner@isq-bb.de

1 Introduction

Since the publication of Hattie's meta-analysis *Visible Learning* (2009), the use of student feedback as an effective method for improving the quality of teaching has moved strongly into the focus of educational practice (see Chap. 8 by Wisniewski and Zierer in this volume). Besides that, student feedback is also discussed in models of data-based decision-making, and is seen as an informative addition to the analysis of student performance data in order to provide teachers with well-founded information about areas in their teaching which they can improve (Lai & Schildkamp, 2013). The use of student feedback consists of two different phases: First, teachers ask their students for feedback on teaching and their perception of the learning environment. In a second step, this information can be used to develop teaching and professional competencies. This chapter focuses on the questions of (a) which factors have an influence on whether teachers collect feedback on their teaching and (b) whether they use this feedback to improve their teaching. To answer these questions, we examine two different research approaches. In one approach, the results of student surveys can be understood as feedback on teaching, so that theories and findings from the field of organizational feedback research can provide important information (e.g., London & Smither, 2002; Smither et al., 2005a). In the other approach, the results of student surveys can be seen as relevant data about teaching quality. In this case, theories and findings from the field of data-based decision-making in schools (e.g., Schildkamp, 2019; Schildkamp et al., 2015) are helpful in providing relevant information on this topic.

The research on feedback in organizational psychology focuses more on cognitive, emotional, and motivational processes of the feedback receiver, which can subsequently lead to behavioral changes, such as improved performance and increased commitment. From this point of view, it is necessary that the receiver accepts the feedback, desires to respond, develops and aims for alternative actions, and finally implements these (Ilgen et al., 1979; Smither et al., 2005a; Kahmann & Mulder, 2011; Introduction to this volume by Röhl et al.). For this process, the characteristics of (1) the feedback recipient, (2) the sender and his or her relation to the recipient, (3) the feedback message, and (4) the organization have proven to be important factors.

Models of data-based decision-making describe the process of using data to support school quality development (e.g., Brunner & Light, 2008; Helmke & Hosenfeld, 2005). Lai and Schildkamp (2013) define the process as consisting of five steps: (1) First of all, it is necessary to clarify within the school which question the data should answer, i.e. the purpose of the data collection. (2) Subsequently, the data considered as relevant are collected (e.g., performance data, classroom observations, survey data, school administration data, etc.). (3) The data collected are analyzed with regard to the initial question. This is followed by (4) an interpretation in terms of meaning for the initial question and the consequences they should have for school and teaching development. (5) The last step is an implementation of the planned measures in everyday school life. Based on well-known models of data-based school and teaching development, relevant influencing factors on the process of data use

can be distinguished at different levels (Coburn & Turner, 2012; Visscher & Coe, 2003). Schildkamp et al. (2017), for example, differentiate between three relevant influencing factors on the process of data use: (1) characteristics of the school organization, such as the existence of support structures or the significance of data use for school management, (2) characteristics of the existing data, such as user-friendliness or timely provision, and (3) characteristics of data users, such as how qualified teachers are in analyzing data or what attitudes they have towards the use of data.

In addition, however, a differentiation should be made as to whether the collection of student feedback is voluntary or not, and for what purpose the feedback should be used. The following situations can be distinguished: (a) teachers voluntarily searching for feedback on their own initiative, (b) student feedback is delivered to teachers as established practice or given by the organization, but without official accountability purpose, and (c) student feedback with accountability purposes. Most of the literature included in this overview refers to situations (a) and (b).

2 What Are Relevant Conditions for Teachers' Use of Student Feedback?

This chapter summarizes the empirical findings from literature on relevant influencing factors according to the three areas mentioned in both research fields (teachers as feedback recipients and data users, school organization, and feedback message or data). As the feedback senders in this context are uniformly students, teachers' perception of the students as competent in this point is particularly relevant.

2.1 *Teachers as Feedback Recipients And Data Users: Relevant Individual Characteristics*

Studies on feedback use from organizational research usually focus on one or more of the following aspects: feedback-seeking behavior, acceptance, perceived usefulness, and performance improvement due to feedback.

Older employees or those with a longer professional experience perceive feedback as less useful (Ilgen et al., 1979). This tendency is also evident for teachers: older teachers seek less feedback from colleagues or peers (Kunst et al., 2018; Runhaar et al., 2010). Regarding student feedback, teachers with longer professional experience are more skeptical of the usefulness (Dretzke et al., 2015) and older teachers use student feedback less often (Ditton & Arnold, 2004b). Some findings on *gender* effects regarding feedback show that female teachers seek more feedback from colleagues (Runhaar et al., 2010) and tend to improve their teaching more after a student feedback intervention (Buurman et al., 2018).

Many findings from social cognitive psychology showed *self-efficacy* to be a particularly important personal factor in the use of feedback (e.g., Bandura, 1997; Heslin & Latham, 2004; Lyden et al., 2002; Sedikides & Strube, 1995; Stajkovic & Sommer, 2000). This seems obvious, as receivers of critical or negative feedback are more likely to respond with additional effort if the person is convinced of achieving an improvement. In addition, feedback is seen as less threatening to self-esteem if persons are convinced that they could respond productively to criticism. In this way, teachers with a higher self-efficacy seek more feedback and are more willing to reflect upon it (Runhaar et al., 2010). A higher self-efficacy correlates with a positive attitude toward school evaluation results (Schneewind, 2007). Ditton and Arnold (2004b) find a differential effect of teachers' self-efficacy on the improvement of teaching after a student feedback intervention.

Closely related to self-efficacy is the concept of *attribution styles* (Weiner, 1985). People differ in whether they attribute the causes of a performance result or feedback more to themselves (internally) or to other people and circumstances (externally). Additionally, the causes can be seen as stable or variable. Persons further differentiate whether these causes are controllable or not. Since attribution styles are particularly associated with motivational and emotional effects as well as convictions regarding individual freedom for action and options, they are considered relevant for dealing with feedback (Strijbos & Müller, 2014). For example, a change in effort or action which is based on negative or corrective feedback can only take place if the recipient assesses the feedback cause as changeable by him- or herself—this corresponds to an attribution as internal, variable, and controllable. With an internal attribution, the receiver can assume responsibility for his or her own results. If, in addition, the cause of a negative result is regarded as controllable and changeable, adjustment processes can be initiated to achieve the performance target. If the cause of negative feedback is considered to be internal, but stable and not controllable, e.g., attributed to one's own lack of ability, this will not lead to a motivation for change. Furthermore, an attribution to one's own personality can lead to negative, performance reducing affects (Kluger & DeNisi, 1996) and to a weakening of the self-concept (Ilgen & Davis, 2000). In the same way, positive feedback can only increase self-efficacy if the cause is assessed as controllable and internal (Bandura, 1997; Lyden et al., 2002; Tolli & Schmidt, 2008). In order to maintain a positive self-concept, recipients of negative feedback tend toward an external attribution (Korn et al., 2016; Sedikides & Strube, 1995). Studies in the school context identified teachers' attribution styles as crucial for the sensemaking process in the context of student achievement and self-evaluation data use (Bertrand & Marsh, 2015; Schildkamp & Visscher, 2009). For example, teachers who attributed students' achievement to their own instruction as internal, variable, and controllable improved their teaching successfully, while the causal attribution to student or test characteristics inhibited instructional improvement (Bertrand & Marsh, 2015). Although these effects have been much elaborated in psychological research, as far as we know only Tacke and Hofer (1979) analyzed effects of teachers' ($n = 20$) internal and external attributions of received positive and negative student feedback. Their results could not show any associations regarding teachers' improvement of teaching.

As a further relevant personal factor, various studies show significant effects of persons' *goal orientations* on the processing and use of feedback (Elliott & Dweck, 1988). Most research approaches distinguish between: mastery goal orientation, which focuses on the development of competence or task mastery; performance-approach goal orientation, with a focus on presenting competence relative to others; and performance-avoidance goal orientation, which concentrates on the avoidance of demonstrating incompetence (Elliot, 1999). Performance goal orientations are often linked with the belief that abilities (e.g., intelligence) or competencies are internal, stable, and not controllable. By contrast, people with a high mastery goal orientation tend to assume that their own abilities can be changed or improved, thus tending toward internal, variable, and controllable attribution. Furthermore, studies from organizational psychology point out significant correlations between persons' self-efficacy and mastery goal orientation (Runhaar et al., 2010; VandeWalle, 2001). For the processing and constructive reaction to negative feedback, a high mastery goal orientation proves to be favorable (Elliott & Dweck, 1988; He et al., 2016; Heslin & Latham, 2004), whereas a high performance-avoidance goal orientation often leads to lower performance in this situation (VandeWalle et al., 2001). When feedback is positive, performance-approach goal-oriented persons tend to react with increased effort, whereas mastery goal-oriented individuals retain their performance or show a weaker increase of effort (Cianci et al., 2010; Donovan & Hafsteinnsson, 2006).

Studies on teachers' handling of collegial feedback from colleagues or peers also confirm these effects (Funk, 2016; Kunst et al., 2018), whereas teachers with a high mastery and a low performance-avoidance goal orientation showed the highest level of feedback-seeking behavior (Runhaar et al., 2010). Therefore, we assume that motivational goal orientations show similar effects for the teachers' use of student feedback, although the only study we found in this regard (Tacke & Hofer, 1979) did not prove any such effects, albeit using an older conceptualization of general achievement motivation.

Even though goal orientations are largely stable personality traits (Praetorius et al., 2014), they can partly be controlled by prompting effects. If an evaluation or feedback is presented as a learning opportunity, this tends to lead to a higher mastery orientation, whereas the description as a control instrument is associated with a stronger performance-achievement or performance-avoidance goal orientation, depending on whether the person considers him/herself to be competent and capable (Cianci et al., 2010). This could also explain the strong relation of teachers' perceived control purpose of student feedback with their resistance against this instrument and so lower acknowledgment of the feedback (Elstad et al., 2015). Conversely, a perceived developmental purpose of student feedback is linked to a higher appreciation of usefulness (Elstad et al., 2017).

In the context of data-based decision-making, teachers' *data literacy* is mentioned as an important factor for the use of student achievement data (Schildkamp et al., 2017). This means that the teacher must be able to understand the results (data),

which are often numerical, draw the right conclusions, and translate them into action-leading steps for their subsequent teaching (Mandinach & Gummer, 2016). Obviously, teachers must also have a kind of data literacy when using student feedback. However, as far as we know, no study results are available in this regard.

Studies investigating the relationships between the Big 5 *personality traits* and the use of feedback show inconsistent and sometimes contradictory findings (Strijbos & Müller, 2014). Some analyses suggest a positive effect of a higher agreeableness and conscientiousness on the feedback use and acceptance (e.g., Bell & Arthur, 2008; Guo et al., 2017). Other findings show a negative effect of extraversion and neuroticism (Smither et al., 2005b), or indeed failed to find any significant correlation (Walker et al., 2010).

Findings on effects of teachers' *stress* experience indicate a negative association with student feedback use (Ditton & Arnold, 2004b). Simultaneously, another analysis indicates that self-evaluations can lead to additional stress experiences among teachers, which in turn lead to a stronger rejection of the procedure and a lower acceptance of student feedback (Elstad et al., 2015).

As many studies from organizational psychology point out, the perceived trustworthiness and competence or expertise of the feedback provider has an important effect on the acceptance and usage of feedback (Cherasaro et al., 2016; Ilgen et al., 1979; Lechermeier & Fassnacht, 2018; Steelman et al., 2004; Raemdonck & Strijbos, 2013). In the context of student feedback to teachers, this means that teachers' attitudes to student judgment accuracy and trustworthiness have an important impact. Findings of several studies confirm this point (Balch, 2012; Ditton & Arnold, 2004b; Elstad et al., 2017), whereas the analyses of Gärtner (2014) show a positive correlation between skepticism regarding student responses and reported usage of feedback. Especially skepticism about feedback from young students is reflected in studies which focus on the quality of feedback from primary school students (De Jong & Westerhof, 2001). This skepticism may explain the low use of student surveys in primary schools. Gärtner (2010), for example, shows (on the basis of the usage statistics of an online portal for student surveys in the German federal states of Berlin and Brandenburg) that only a few feedback-surveys take place in primary schools (8.3% in grade levels 3 & 4 and 18.7% in grade levels 5 & 6), although about half of all students are taught in primary schools. On the other hand, there is also evidence of the validity of primary school students' perception of teaching (Fauth et al., 2014; Gärtner & Brunner, 2018; van der Scheer et al., 2019). These results prompt the questions: (a) under which circumstances do teachers trust younger students to be able to assess teaching competently? (Iglar et al., 2019); or rather (b) which student characteristics influence whether teachers assess their students as competent feedback providers (age, achievement level, socio-economic status, language skills, etc.)? In addition, studies are still lacking on the adjacent question of the extent to which teachers' student orientation is linked with the acceptance and perceived usefulness of student feedback.

2.2 School: Relevant Organizational Characteristics

Several studies summarize different organizational characteristics, such as support for giving and interpreting feedback, a non-threatening atmosphere, or the value of feedback to improve, as *feedback culture* of an organization. The results provide evidence of the importance of this overall concept as moderator for the use of feedback in organizations (Kahmann, 2009; London & Smither, 2002; Mulder, 2013). Also, for the systematic use of student feedback, feedback culture within teaching staff has been found relevant (Gaertner, 2014). In addition, compliance-oriented cultures in schools appeared to hinder developmental use of students' achievement data (Farrell & Marsh, 2016).

However, studies have also shown that specific organizational characteristics have effects on the productive use of feedback. The perceived *team psychological safety*, which means that team members share a belief that the team is safe for interpersonal risk taking, has proved to be relevant for feedback use (Edmondson, 1999; Harvey et al., 2019; Semmer & Jacobshagen, 2010).

In addition, a beneficial approach to increase the productive use of feedback in organizations is to offer special support in understanding feedback, setting goals, and implementing them in practice; this includes coaching, group reflections, and counseling (Luthans & Peterson, 2003; Smither et al., 2003; Walker et al., 2010). In the context of data use at schools, training and support for teachers with regard to data analysis and interpretation has been found instrumental for instructional data use (Farrell & Marsh, 2016; Kerr et al., 2006; Schildkamp & Visscher, 2009). In the context of student feedback, in particular, those intervention studies which provided supportive measures for reflection and teaching development show significantly higher positive effects (see Chap. 9 by Röhl in this volume).

In all of this, *leadership* plays an important role in feedback usage processes. In organizational research, transformational leadership (Bass, 1985) proved to be advantageous for team learning, feedback processes, and reflection in working groups and school teams (Lam, 2002; Runhaar et al., 2010; Tuytens et al., 2019). According to this concept, school principals should provide a clear vision for the future, inspire teachers, give the work a greater sense of meaning, and stimulate the questioning of old assumptions. Findings from research on data-based decision-making processes in schools pointed to the importance of encouragement from principals (Schildkamp & Visscher, 2009) and teachers' feeling of autonomy to make decisions about their instruction in data use processes in schools (Kerr et al., 2006; Prenger & Schildkamp, 2018). Whether teachers interpret the obtaining of student feedback more as a control or as a development opportunity depends on the communication from the school leaders (Elstad et al., 2017; Lejonberg et al., 2017). Active encouragement by school leaders of teachers to seek student feedback is also supportive, as extrinsically motivated feedback use is just as beneficial to reported improvements in teaching as is intrinsically motivated feedback use (Gaertner, 2014).

2.3 *Feedback Message as Data: Relevant Feedback Characteristics*

As several studies reveal, also the characteristics of the feedback message show relevant effects on the processing and use of feedback (Coe, 1998; Ilgen et al., 1979; Kluger & DeNisi, 1996). In this way, the comprehensibility of feedback results proves to be an important predictor for feedback use both in school performance studies (Groß Ophoff, 2013) and in the context of student feedback (Ditton & Arnold, 2004a; Rösch, 2017). The findings of a study by Merk et al. (2019) on online-based student feedback indicate that teachers feel more confident with the presentation of scale averages than with the display of single values or box plots. However, since the information on the variance of student perceptions as well as the individual item scores contain relevant information about one's own teaching (see Chap. 6 by Schweig and Martínez and Chap. 3 by Röhl and Rollett in this volume), a promotion of teachers' data literacy appears to be an important prerequisite for productive use (Schildkamp, 2019).

Findings from organizational psychology point out that the perception of the valence or positivity of feedback with regard to one's own actions is accompanied by a more precise reception, easier remembering of the feedback contents, and better acceptance of the feedback (Ilgen et al., 1979; Lyden et al., 2002). If feedback is perceived as negative, there is a tendency to adopt a defensive attitude, which serves to protect one's own self-concept, and so reduces the intensity of perception, the acceptance, and the willingness to change (Lechermeier & Fassnacht, 2018; Sedikides & Strube, 1995). In most cases, this defensive attitude is also expressed in an external attribution of the reasons for the negative feedback (see above). These reactions on the feedback valence are also found in the context of student feedback (Ditton & Arnold, 2004a; Rösch, 2017). However, whether the student feedback is more positive or more negative than teachers' self-perception has not been found to be significant for student feedback use (Buurman et al., 2018; Gaertner, 2014).

In organizational research, the specificity of feedback, i.e., the accuracy and extent of the exemplary reference to the task and its improvement, shows differential effects depending on the expertise of the recipient. Highly specific feedback leads to positive effects if the recipient is in an early exercise phase with regard to the task. However, long-term learning performance is negatively influenced by this highly specific feedback (Lechermeier & Fassnacht, 2018). On the other hand, low-specificity or summarized feedback, which refers to several tasks or a longer period, has the opposite effect: Short-term exercise performance is worse, whereas long-term learning performance is better. A tentative explanation is that low-specificity feedback could lead to an active search for possible improvements and a deeper processing of the necessary information, which in turn leads to deepening learning effects (Schmidt & Bjork, 1992). Since most teachers can be classified as professionals with many years of teaching experience, a less specific student feedback for this group could generate greater usage effects. More specific forms of feedback, which include concrete suggestions for improvement, could be beneficial for novice

teachers or in teacher training. In addition, in a discussion following the feedback, students can give important concrete advice on how to improve teaching (see also Chap. 12 by Schmidt and Gawrilow in this volume).

Studies on the effects of *timing* of feedback mainly refer to the accuracy of experimental learning tasks and distinguish immediate feedback from feedback given between 10 min and 24 h after completion of the learning task (Lechermeier & Fassnacht, 2018). The delay retention effect shown here, according to which late feedback leads to higher long-term learning success (Kulhavy & Anderson, 1972), seems to be based on the recapitulation of the associated learning content, which leads to a more in-depth memorization (Smith & Kimball, 2010). Since feedback from students to the teacher is usually only given after a certain time interval from the teaching activities, e.g., at the end of a lesson, week, or learning unit, these findings are only of limited significance for this context. However, Coe (1998) argues for the school context that feedback to teachers on students' learning achievement or teaching should be given as soon as possible in order to have the maximum effect on the further development of teaching.

Furthermore, a survey instrument which is valid and reliable should be selected for a successful use of student feedback (see also Chap. 4 by Bijlsma in this volume). However, it should also suit the age of the students and the type of teaching in which it is to be used. For example, a questionnaire designed for use in the context of self-regulated learning may provide little helpful information if it is used in the context of strongly teacher-directed learning.

3 Conclusion and Outlook on Future Practice and Research

This chapter summarizes existing evidence on the use of student feedback according to three relevant influencing factors: characteristics of the feedback recipient, characteristics of the organization, and characteristics of the feedback information.

The following personal characteristics of teachers which influence the use of feedback were identified: *self-efficacy, attribution styles, goal orientations, perceived trustworthiness and competence of students as feedback providers, data literacy, and age and professional experience*. The reported findings indicated relevant characteristics of schools as organizations: *feedback culture, leadership, safety, support measures, and perceived function of feedback as control vs. development*. Relevant characteristics of feedback information were identified as: *timeliness, comprehensibility, valence, and specificity*.

For many of the reported teacher and feedback characteristics only evidence from organizational research exists. Although some findings on teachers' use of feedback from colleagues or school leaders point to a transferability of results found to the school context, there are no or only rare studies which would confirm the results for the context of student feedback. Regarding the characteristics of schools as organizations, a little more is known from data use studies, but also only a few findings concerning student feedback exist. With regard to future research, we believe

that there is a particular need for complex intervention studies which examine the effectiveness of student feedback in teaching development while controlling factors identified as relevant.

The findings presented in this chapter reveal a number of indications for the beneficial use of student feedback in schools. Firstly, with regard to organizational conditions, it seems helpful to communicate student feedback as a learning opportunity for teachers and not as a control instrument. The school management should ensure a safe environment in order to realize the use of student feedback, especially in a collegial setting, and thus build up a feedback culture in the long term. Transformational and feedback-encouraging leadership can help to enable reflective and developmental processes in schools overall and so foster productive feedback use. Finally, and in the best case, student feedback can be implemented in such a way that, at the same time, support measures are in place for the joint development of teaching.

Positive experiences in dealing with student feedback can thus possibly also change teachers' attitudes toward student feedback, such as the perceived trustworthiness of students as feedback providers (Gärtner & Vogt, 2013). In addition, this could also have a positive influence on relevant personality traits such as self-efficacy, attribution styles, and goal orientations.

With regard to the preparation of reports from student feedback, it seems helpful to make it as comprehensible as possible, especially with regard to statistical parameters (Merk et al., 2019), but also to include information about heterogeneous views among students on individual aspects of teaching. Furthermore, it appears to be beneficial for the developmental use of feedback to enrich reports with concrete suggestions for improving teaching activities (specificity), especially for less experienced teachers. Moreover, positive results should be particularly emphasized, so that negative results can also be better accepted.

References

- Balch, R. T. (2012). *The validation of a student survey on teacher practice*. Dissertation. Vanderbilt University, Nashville, TN.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman. <https://doi.org/10.5860/choice.35-1826>.
- Bass, B. M. (1985). *Leadership and performance beyond expectations*. Free Press.
- Bell, S. T., & Arthur, W. (2008). Feedback acceptance in developmental assessment centers: The role of feedback message, participant personality, and affective response to the feedback session. *Journal of Organizational Behavior*, 29, 681–703. <https://doi.org/10.1002/job.525>.
- Bertrand, M., & Marsh, J. A. (2015). Teachers' sensemaking of data and implications for equity. *American Educational Research Journal*, 52, 861–893. <https://doi.org/10.3102/0002831215599251>.
- Brunner, C., & Light, D. (2008). From knowledge management to data-driven instructional decision-making in schools: The missing link. In A. Breiter, A. Lange, & E. Stauke (Eds.), *School information systems and data-based decision-making* (pp. 37–48). Peter Lang.

- Buurman, M., Delfgaauw, J. J., Dur, R. A. J., & Zoutenbier, R. (2018). *The effects of student feedback to teachers: Evidence from a field experiment* (Tinbergen Institute Discussion Paper). Amsterdam & Rotterdam.
- Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter?* U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central.
- Cianci, A. M., Schaubroeck, J. M., & McGill, G. A. (2010). Achievement goals, feedback, and task performance. *Human Performance, 23*, 131–154. <https://doi.org/10.1080/08959281003621687>.
- Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. *American Journal of Education, 118*(2), 99–111. <https://doi.org/10.1086/663272>.
- Coe, R. (1998). Can feedback improve teaching? A review of the social science literature with a view to identifying the conditions under which giving feedback to teachers will result in improved performance. *Research Papers in Education, 13*(1), 43–66. <https://doi.org/10.1080/0267152980130104>.
- De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research, 4*, 51–85. <https://doi.org/10.1023/A:1011402608575>.
- Ditton, H., & Arnold, B. (2004a). Schülerbefragungen zum Fachunterricht: Feedback an Lehrkräfte [Student surveys on subject teaching: feedback to teachers]. *Empirische Pädagogik, 18*(1), 115–139.
- Ditton, H., & Arnold, B. (2004b). Wirksamkeit von Schülerfeedback zum Fachunterricht [Effectiveness of student feedback on subject teaching]. In J. Doll & M. Prenzel (Eds.), *Bildungsqualität von Schule: Lehrerprofessionalisierung, Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsentwicklung* (pp. 152–172). Waxmann.
- Ditton, H., & Müller, A. (Eds.). (2014). *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* [Feedback: Theoretical foundations, empirical findings, practical application fields]. Waxmann.
- Donovan, J. J., & Hafsteinsson, L. G. (2006). The impact of goal-performance discrepancies, self-efficacy, and goal orientation on upward goal revision. *Journal of Applied Social Psychology, 36*, 1046–1069. <https://doi.org/10.1111/j.0021-9029.2006.00054.x>.
- Dretzke, B. J., Sheldon, T. D., & Lim, A. (2015). What do K-12 teachers think about including student surveys in their performance ratings? *Mid-Western Educational Researcher, 27*(3), 185–206.
- Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly, 44*, 350. <https://doi.org/10.2307/2666999>.
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*, 169–189. https://doi.org/10.1207/s15326985ep3403_3.
- Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*, 5–12. <https://doi.org/10.1037/0022-3514.54.1.5>.
- Elstad, E., Lejonberg, E., & Christophersen, K.-A. (2015). Teaching evaluation as a contested practice: Teacher resistance to teaching evaluation schemes in Norway. *Education Inquiry, 6*, 375–399. <https://doi.org/10.3402/edui.v6.27850>.
- Elstad, E., Lejonberg, E., & Christophersen, K.-A. (2017). Student evaluation of high-school teaching: Which factors are associated with teachers' perception of the usefulness of being evaluated? *Journal for Educational Research Online, 9*(1), 99–117.
- Farrell, C. C., & Marsh, J. A. (2016). Contributing conditions: A qualitative comparative analysis of teachers' instructional responses to data. *Teaching and Teacher Education, 60*, 398–412. <https://doi.org/10.1016/j.tate.2016.07.010>.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>.

- Funk, C. M. (2016). *Kollegiales Feedback aus der Perspektive von Lehrpersonen* [Peer feedback from the perspective of teachers]. Springer Fachmedien. <https://doi.org/10.1007/978-3-658-13062-6>.
- Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, 42, 91–99. <https://doi.org/10.1016/j.stueduc.2014.04.003>.
- Gärtner, H. (2010). Das ISQ-Selbstevaluationsportal. Konzeption eines Online-Angebots, um die Selbstevaluation in Schule und Unterricht zu unterstützen. *Die Deutsche Schule*, 102(2), 163–175.
- Gärtner, H., & Brunner, M. (2018). Once good teaching, always good teaching? The differential stability of student perceptions of teaching quality. *Educational Assessment, Evaluation and Accountability*, 30(2), 159–182. <https://doi.org/10.1007/s11092-018-9277-5>.
- Gärtner, H., & Vogt, A. (2013). Selbstevaluation des Unterrichts: Wie Lehrkräfte Ergebnisse eines Schülerfeedbacks rezipieren [Self-evaluation of teaching: how teachers receive results of student feedback]. *Unterrichtswissenschaft*, 41(3), 255–270.
- Groß Ophoff, J. (2013). *Lernstandserhebungen: Reflexion und Nutzung* [Learning assessments: Reflection and use]. Waxmann.
- Guo, Y., Zhang, Y., Liao, J., Guo, X., Liu, J., Xue, X., et al. (2017). Negative feedback and employee job performance: Moderating role of the big five. *Social Behavior and Personality: An International Journal*, 45, 1735–1744. <https://doi.org/10.2224/sbp.6478>.
- Harvey, J.-F., Johnson, K. J., Roloff, K. S., & Edmondson, A. C. (2019). From orientation to behavior: The interplay between learning orientation, open-mindedness, and psychological safety in team learning. *Human Relations*, 72, 1726–1751. <https://doi.org/10.1177/0018726718817812>.
- Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- He, Y., Yao, X., Wang, S., & Caughron, J. (2016). Linking failure feedback to individual creativity: The moderation role of goal orientation. *Creativity Research Journal*, 28, 52–59. <https://doi.org/10.1080/10400419.2016.1125248>.
- Helmke, A., & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation [Standard related teaching evaluation]. In G. Brägger, B. Bucher, & N. Landwehr (Eds.), *Schlüsselfragen zur externen Schulevaluation* (pp. 127–151). hep.
- Heslin, P. A., & Latham, G. P. (2004). The effect of upward feedback on managerial behavior. *Applied Psychology*, 53(1), 23–37. <https://doi.org/10.1111/j.1464-0597.2004.00159.x>.
- Igler, J., Ohle-Peters, A., & McElvany, N. (2019). Mit den Augen eines Grundschulkindes [Through the eyes of a primary school child]. *Zeitschrift Für Pädagogische Psychologie*, 33, 191–205. <https://doi.org/10.1024/1010-0652/a000243>.
- Ilgen, D. R., & Davis, C. A. (2000). Bearing bad news: Reactions to negative performance feedback. *Applied Psychology*, 49, 550–565. <https://doi.org/10.1111/1464-0597.00031>.
- Ilgen, D. R., Fisher, C. D., & Taylor, S. M. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4), 349–371. <https://doi.org/10.1037/0021-9010.64.4.349>.
- Kahmann, K. (2009). *Die Erfassung der Feedbackkultur in Organisationen: Konstruktion und psychometrische Überprüfung eines Messinstrumentes* [Measuring feedback culture in organizations: Construction and psychometric testing of a measurement instrument]. Dr. Kovac.
- Kahmann, K., & Mulder, R. H. (2011). *Feedback in organizations: A review of feedback literature and a framework for future research* (Research Report 6). Regensburg. https://www.uni-regensburg.de/psychologie-paedagogik-sport/paedagogik-2/medien/kahmann_mulder_2011.pdf. Accessed 31 October 2019.
- Kerr, K. A., Marsh, J. A., Ikemoto, G. S., Darilek, H., & Barney, H. (2006). Strategies to promote data use for instructional improvement: Actions, outcomes, and lessons from three urban districts. *American Journal of Education*, 112, 496–520. <https://doi.org/10.1086/505057>.

- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>.
- Korn, C. W., Rosenblau, G., Rodriguez Buritica, J. M., & Heekeren, H. R. (2016). Performance feedback processing is positively biased as predicted by attribution theory. *PLoS one*, 11. <https://doi.org/10.1371/journal.pone.0148581>.
- Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, 63, 505–512. <https://doi.org/10.1037/h0033243>.
- Kunst, E. M., van Woerkom, M., & Poell, R. F. (2018). Teachers' goal orientation profiles and participation in professional development activities. *Vocations and Learning*, 11, 91–111. <https://doi.org/10.1007/s12186-017-9182-y>.
- Lai, M. K., & Schildkamp, K. (2013). Data-based decision making: An overview. In K. Schildkamp, M. K. Lai, & L. Earl (Eds.), *Data-based decision making in education* (pp. 9–22). Springer. https://doi.org/10.1007/978-94-007-4816-3_2.
- Lam, Y. L. J. (2002). Defining the effects of transformational leadership on organisational learning: A cross-cultural comparison. *School Leadership & Management*, 22, 439–452. <https://doi.org/10.1080/1363243022000053448>.
- Lechermeier, J., & Fassnacht, M. (2018). How do performance feedback characteristics influence recipients' reactions? A state-of-the-art review on feedback source, timing, and valence effects. *Management Review Quarterly*, 68, 145–193. <https://doi.org/10.1007/s11301-018-0136-8>.
- Lejonberg, E., Elstad, E., & Christophersen, K. A. (2017). Teaching evaluation: Antecedents of teachers' perceived usefulness of follow-up sessions and perceived stress related to the evaluation process. *Teachers and Teaching*, 24, 281–296. <https://doi.org/10.1080/13540602.2017.1399873>.
- London, M., & Smither, J. W. (2002). Feedback orientation, feedback culture, and the longitudinal performance management process. *Human Resource Management Review*, 12, 81–100. [https://doi.org/10.1016/S1053-4822\(01\)00043-2](https://doi.org/10.1016/S1053-4822(01)00043-2).
- Luthans, F., & Peterson, S. J. (2003). 360-degree feedback with systematic coaching: Empirical analysis suggests a winning combination. *Human Resource Management*, 42, 243–256. <https://doi.org/10.1002/hrm.10083>.
- Lyden, J. A., Chaney, L. H., Danehower, V. C., & Houston, D. A. (2002). Anchoring, attributions, and self-efficacy: An examination of interactions. *Contemporary Educational Psychology*, 27, 99–117. <https://doi.org/10.1006/ceps.2001.1080>.
- Mandinach, E. B., & Gummer, E. S. (2016). *Data literacy for educators: Making it count in teacher preparation and practice*. Teachers College Press.
- Merk, S., Poindl, S., & Bohl, T. (2019). Wie sollten Rückmeldungen von quantitativ erfasstem Schülerfeedback (nicht) gestaltet werden? Wahrgenommene Informativität und Interpretationssicherheit von quantitativen Rückmeldungen zur Unterrichtsqualität [Which statistical information of feedback data from student questionnaires should (not) be reported to teachers? Perceived informativity and validity of interpretation of feedback about instructional quality]. *Unterrichtswissenschaft*, 47, 457–494. <https://doi.org/10.1007/s42010-019-00048-5>.
- Mulder, R. H. (2013). Exploring feedback incidents, their characteristics and the informal learning activities that emanate from them. *European Journal of Training and Development*, 37, 49–71. <https://doi.org/10.1108/03090591311293284>.
- Praetorius, A.-K., Nitsche, S., Janke, S., Dickhäuser, O., Drexler, K., Fasching, M., et al. (2014). Here today, gone tomorrow? Revisiting the stability of teachers' achievement goals. *Contemporary Educational Psychology*, 39, 379–387. <https://doi.org/10.1016/j.cedpsych.2014.10.002>.
- Prenger, R., & Schildkamp, K. (2018). Data-based decision making for teacher and student learning: A psychological perspective on the role of the teacher. *Educational Psychology*, 38, 734–752. <https://doi.org/10.1080/01443410.2018.1426834>.
- Raemdonck, I., & Strijbos, J.-W. (2013). Feedback perceptions and attribution by secretarial employees: Effects of feedback-content and sender characteristics. *European Journal of Training and Development*, 37, 24–48. <https://doi.org/10.1108/03090591311293275>.

- Rösch, S. (2017). *Wirkung und Wirkmechanismen von regelmäßigem Schülerfeedback in der Sekundarstufe: Eine explorative Untersuchung im Physikunterricht* [Effect and impact mechanisms of frequent student feedback in secondary education: an exploratory study in physics classrooms]. Dissertation, Universität Basel, Basel.
- Runhaar, P., Sanders, K., & Yang, H. (2010). Stimulating teachers' reflection and feedback asking: An interplay of self-efficacy, learning goal orientation, and transformational leadership. *Teaching and Teacher Education*, 26, 1154–1161. <https://doi.org/10.1016/j.tate.2010.02.011>.
- Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, 61, 257–273. <https://doi.org/10.1080/00131881.2019.1625716>.
- Schildkamp, K., Poortman, C. L., & Handelzalts, A. (2015). Data teams for school improvement. *School Effectiveness and School Improvement*, 27, 228–254. <https://doi.org/10.1080/09243453.2015.1056192>.
- Schildkamp, K., Poortman, C., Luyten, H., & Ebbeler, J. (2017). Factors promoting and hindering data-based decision making in schools. *School Effectiveness and School Improvement*, 28(2), 242–258. <https://doi.org/10.1080/09243453.2016.1256901>.
- Schildkamp, K., & Visscher, A. (2009). Factors influencing the utilisation of a school self-evaluation instrument. *Studies in Educational Evaluation*, 35, 150–159. <https://doi.org/10.1016/j.stueduc.2009.12.001>.
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–217. <https://doi.org/10.1111/j.1467-9280.1992.tb00029.x>.
- Schneewind, J. (2007). *Wie Lehrkräfte mit Ergebnisrückmeldungen aus Schulleistungsstudien umgehen* [How teachers deal with results feedback from school performance assessments]. <https://doi.org/10.17169/refubium-15430>. Accessed 2 January 2020.
- Sedikides, C., & Strube, M. J. (1995). The multiply motivated self. *Personality and Social Psychology Bulletin*, 21(12), 1330–1335. <https://doi.org/10.1177/01461672952112010>.
- Semmer, N. K., & Jacobshagen, N. (2010). Feedback im Arbeitsleben – eine Selbstwert-Perspektive [Feedback at work - a self-esteem perspective]. *Gruppendynamik Und Organisationsberatung*, 41, 39–55. <https://doi.org/10.1007/s11612-010-0104-9>.
- Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of experimental psychology. Learning, Memory, and Cognition*, 36, 80–95. <https://doi.org/10.1037/a0017407>.
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multi-source feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, 58, 33–66. https://doi.org/10.1111/j.1744-6570.2005.514_1.x.
- Smither, J. W., London, M., & Richmond, K. R. (2005). The relationship between leaders' personality and their reactions to and use of multisource feedback. *Group & Organization Management*, 30, 181–210. <https://doi.org/10.1177/1059601103254912>.
- Smither, J. W., London, M., Flautt, R., Vargas, Y., & Kucine, I. (2003). Can working with an executive coach improve multisource feedback ratings over time? *A Quasi-Experimental Field Study. Personnel Psychology*, 56(1), 23–44. <https://doi.org/10.1111/j.1744-6570.2003.tb00142.x>.
- Stajkovic, A. D., & Sommer, S. M. (2000). Self-efficacy and causal attributions: Direct and reciprocal links. *Journal of Applied Social Psychology*, 30, 707–737. <https://doi.org/10.1111/j.1559-1816.2000.tb02820.x>.
- Steelman, L. A., Levy, P. E., & Snell, A. F. (2004). The feedback environment scale: Construct definition, measurement, and validation. *Educational and Psychological Measurement*, 64, 165–184. <https://doi.org/10.1177/0013164403258440>.
- Srijbos, J.-W., & Müller, A. (2014). Personale Faktoren im Feedbackprozess [Individual factors in the feedback process]. In H. Dittton & A. Müller (Eds.), *Feedback und Rückmeldungen: Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (pp. 83–134). Waxmann.

- Tacke, G., & Hofer, M. (1979). Behavioral changes in teachers as a function of student feedback: A case for the achievement motivation theory? *Journal of School Psychology, 17*, 172–180. [https://doi.org/10.1016/0022-4405\(79\)90025-6](https://doi.org/10.1016/0022-4405(79)90025-6).
- Tolli, A. P., & Schmidt, A. M. (2008). The role of feedback, causal attributions, and self-efficacy in goal revision. *The Journal of Applied Psychology, 93*, 692–701. <https://doi.org/10.1037/0021-9010.93.3.692>.
- Tuytens, M., Moolenaar, N., Daly, A., & Devos, G. (2019). Teachers' informal feedback seeking towards the school leadership team. A social network analysis in secondary schools. *Research Papers in Education, 34*, 405–424. <https://doi.org/10.1080/02671522.2018.1452961>.
- van der Scheer, E. A., Bijlsma, H. J. E., & Glas, C. A. W. (2019). Validity and reliability of student perceptions of teaching quality in primary education. *School Effectiveness and School Improvement, 30*(1), 30–50. <https://doi.org/10.1080/09243453.2018.1539015>.
- VandeWalle, D. (2001). Goal orientation: Why wanting to look successful doesn't always lead to success. *Organizational Dynamics, 30*, 162–171. [https://doi.org/10.1016/S0090-2616\(01\)00050-X](https://doi.org/10.1016/S0090-2616(01)00050-X).
- VandeWalle, D., Cron, W. L., Slocum, J. W., & J. R. (2001). The role of goal orientation following performance feedback. *The Journal of Applied Psychology, 86*, 629–640. <https://doi.org/10.1037/0021-9010.86.4.629>.
- Visscher, A. J., & Coe, R. (2003). School performance feedback systems: Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement, 14*(3), 321–349. <https://doi.org/10.1076/sesi.14.3.321.15842>.
- Walker, A. G., Smither, J. W., Atwater, L. E., Dominick, P. G., Brett, J. F., & Reilly, R. R. (2010). Personality and multisource feedback improvement: A longitudinal investigation. *Journal of Behavioral and Applied Management, 11*(2), 175–204.
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548–573. https://doi.org/10.1007/978-1-4612-4948-1_6.

Sebastian Röhl has been an Academic Assistant in the Department of Educational Science at the University of Education Freiburg and is currently a Postdoctoral Researcher in the Institute of Education at Tübingen University (Germany). Before that he worked for more than 10 years as a grammar school teacher, school development consultant, and in teacher training. Among other areas, he conducts research in the fields of teaching development and teacher professionalization through feedback, social networks in inclusive school classes, as well as teachers' religiosity and its impact on professionalism. In addition, he is the Director of an in-service professional master's study program for teaching and school development.

Holger Gärtner is Scientific Director of the Institute for School Quality (ISQ) as well as Professor at the department for the Evaluation of School and Teaching Quality at the Freie Universität Berlin (Germany). After obtaining his doctorate in psychology, he initially worked at the ISQ as a project manager responsible for projects on the internal and external evaluation of schools. He conducts research on questions of data-based decision to support school and teaching development, including the impact of internal and external evaluation of schools.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 11

Student Feedback as a Source for Reflection in Practical Phases of Teacher Education



Kerstin Göbel, Corinne Wyss, Katharina Neuber, and Meike Raaflaub

Abstract The chapter focuses on the use of student feedback on teaching during practical phases in teacher education. After a brief introduction into the general relevance and validity of students' perceptions on teaching, and on the use of student feedback for teaching development, core findings from two comparable quasi-experimental studies from Germany and Switzerland are presented in detail. The studies focus on the change of attitudes towards student feedback and towards reflection on teaching. The chapter concludes with a discussion of challenges and opportunities for the use of student feedback as an instrument for reflection on teaching and professional development for pre-service teachers.

Keywords Teacher education · Reflection · Practical phases · Validity of student feedback · Quasi-experimental studies

1 The Relevance and Validity of Students' Perceptions

Teaching in class is a complex situation as teachers have to master many different tasks at the same time (Bromme, 2014; Stürmer et al., 2017). In this context, receiving feedback on their behaviour can be particularly helpful for teachers, as it expands their perspectives in a meaningful way and might give insights into the teaching process (Helmke, 2015). For teachers and even more for pre-service teachers, it is

K. Göbel (✉) · K. Neuber

Faculty of Educational Sciences, University of Duisburg Essen, Essen, Germany
e-mail: kerstin.gobel@uni-due.de

K. Neuber

e-mail: katharina.neuber@uni-due.de

C. Wyss

FHNW School of Education, Brugg, Switzerland
e-mail: corinne.wyss@fhnw.ch

M. Raaflaub

University of Teacher Education, Bern, Switzerland
e-mail: meike.raaflaub@phbern.ch

© The Author(s) 2021

W. Rollett et al. (eds.), *Student Feedback on Teaching in Schools*,
https://doi.org/10.1007/978-3-030-75150-0_11

173

difficult to process relevant information during teaching in class. In order to counteract restricted and possibly self-serving perspectives, student feedback may offer a specific perspective which may hold further information on teaching and learning processes relevant in the classroom (Clausen, 2002; Clausen et al., 2020; Hascher et al., 2004).

The relevance of student perceptions on teaching is apparent by the very fact that students and their learning are targets of teaching, and as such, students can refer to their experiences with different subjects and teachers. Hence, their observation of the teaching and learning process may contain highly relevant information for teachers. Concerning empirical results on student perceptions of teaching quality, studies in primary and secondary education reveal factorial validity of student ratings. They conclude that students are capable of differentiating between various aspects of teaching quality, such as classroom management, motivational quality and teaching clarity (Fauth et al., 2014; Lenske, 2016; Wagner et al., 2013). Furthermore, several studies point at the predictive validity of student ratings as student perceptions of teaching are linked to learning outcomes: Studies in mathematics reveal a correlation between classroom management, goal clarity and support for autonomy for students' mathematical learning and their self-concept or interest in mathematics (Clausen, 2002; Kunter et al., 2007; Wagner et al., 2016). A large-scale study on English learning in secondary schools shows a correlation between classroom climate, motivational quality and clarity as perceived by students with their development of listening comprehension in the course of one school year (Helmke et al., 2008). Moreover, intercultural learning outcomes in EFL (English as a Foreign Language) secondary classes could be predicted with students' perception of specific aspects of teaching quality, such as a positive error culture and classroom management (Göbel & Hesse, 2008). In some studies, the predictive validity of student ratings is even higher for the prediction of learning outcomes than expert or teacher ratings (Fauth et al., 2014; Göllner et al., 2016; Wagner, 2008).

While there is empirical evidence for predictive and factorial validity of student ratings on teaching, current studies also point at limitations when it comes to gathering information on teaching quality by student ratings. In a German interview study, 14 secondary school students were confronted with their ratings on teaching quality and asked to explain the reasons for their feedback on each of the rated items (Lenske & Praetorius, 2020). Interviews with these students revealed that they did not fully understand all items of the implemented questionnaire, although it was an instrument which had been validated in former studies. Another study by Röhl and Rollett (2020) examined data from a student survey administering student feedback questionnaires on teaching quality ($N = 860$). Their analyses on factorial validity point at halo effects of teachers' communion (community orientation) for different teaching quality ratings.

Although student feedback might be fraught with uncertainty due to problems of validity and reliability, it represents a special perspective on the teaching process and provides teachers with important orientation information on their teaching (Clausen & Göbel, 2020). Studies on the use of student feedback on teaching of in-service

teachers point at a positive impact on teaching development in terms of the teacher–student relationship and a more sophisticated view on the needs of students (Ditton & Arnoldt, 2004; Gärtner, 2013; Rösch, 2017). Analyses of in-service teachers using student feedback point at the relevance of teacher–student co-construction of the meaning of student feedback in class for a better understanding of students’ ideas (Gärtner & Vogt, 2013). Furthermore, the positive effect of student feedback seems to depend on teachers’ attitudes towards student feedback, attitudes towards cooperation, teachers’ stress experience and the quality of student feedback (Gärtner, 2013; Ditton & Arnold, 2004).

In practical phases of teacher training, student feedback has the potential to bring about changes in the attitudes of future generations of teachers, so that they can use feedback—being aware of the challenges and problems of this information—for continuous reflection and development of their teaching (Clausen & Göbel, 2020). Pre-service teachers can consider student feedback on teaching in addition to feedback from in-service teachers or lecturers during practical phases. However, the use of student feedback for learning and reflection processes during practical phases in teacher education is still rare (Hascher et al., 2004) and the students’ perspective on pre-service teachers’ teaching and professional development has been scarcely investigated empirically (Lawson et al., 2015). Therefore, the following sections seek to shed light on present research and findings in the field of student feedback in teacher education.

In the following, we present empirical results on the implementation of student feedback in teacher education. After giving an overview on international results on the topic, we present two comparable quasi-experimental studies from Germany and Switzerland which focus on the change of attitudes towards student feedback and towards reflection on teaching. The two studies are interconnected as they are similar in research design and make use of the same instruments to evaluate attitude changes in the course of student feedback use in practical phases of teacher education. At the end of the contribution, the chapter concludes with a discussion of challenges and opportunities for the use of student feedback as an instrument for reflection on teaching, and professional development of pre-service teachers.

2 Empirical Results on Student Feedback for Reflection on Teaching in Teacher Education

Although there are several hints at the relevance of student feedback for teaching improvement as well as claims for their integration into teacher education, the number of empirical studies focusing on student feedback use in teacher education is still limited (Lawson et al., 2015). The work on student feedback in teacher education started in 1942 when Porter published a paper on an exploratory study on this topic. Analyses from a questionnaire focusing on characteristics of pre-service teachers

revealed a close agreement between the ratings of students and supervisors. Pre-service teachers evaluated the feedback of their students as beneficial and their respective students reported that they appreciated being part of the evaluation process (Porter, 1942).

2.1 Systematic Settings and Measurement Problems

In 1969, Lauroesch and colleagues investigated the use of student feedback by pre-service teachers from the University of Chicago to assess the impact of student feedback on the teaching of pre-service teachers. The quality of pre-service teachers' instructional practice during the internship was measured two times using student ratings. The findings of this quasi-experimental study indicate that the provided summary of the student ratings may not be sufficient to encourage future teaching activities of the pre-service teachers. At the second time of measurement the teaching quality of those pre-service teachers who received a summary of student ratings of their lesson was rated even less positively than before (Lauroesch et al., 1969). The authors conclude that the feedback was potentially misunderstood or that pre-service teachers were overburdened to use the feedback constructively and change processes in teaching. These findings might hint at the need for implementing systematic settings for the reception and reflection of student feedback to provide pre-service teachers with concrete starting points for development in teaching. Possibly, it might not be a lack of development in teaching, but a problem of measurement. For students it might be difficult to assess changes in teaching quality. A study by Holtz and Gnamb (2017) points at the fact that student feedback could be problematic for the assessment of changes in instructional quality. They measured the teaching quality of 181 pre-service teachers in a 15-week internship at a secondary school in Thuringia (Germany) using three different rating sources (self-assessment, mentors' assessment and student ratings). The findings indicate differences in change scores between the three rating sources: Pre-service teachers themselves and their mentors perceived larger changes in instructional quality than students. Similar findings have been reported in a study by Biggs and Chopra (1979) where changes in teaching quality could not be detected by student ratings.

2.2 Constructive Feedback for Instructional Development

In the course of an exploratory study in France, Genoud (2006) implemented student feedback in the course of teacher training focusing on the classroom climate in class using the TIP—questionnaire (Trainee Interaction Profile; Wubbels & Levy, 1993). In a sample of approximately 50 pre-service teachers and their students from grade 5 and 6 a TIP questionnaire was implemented in order to show differences between

pre-service teachers' self-assessments, those of their students and their training supervisor. The intervention was evaluated positively by the pre-service teachers and their students. Pre-service teachers reported a positive perspective towards the use of student feedback on teaching for their professional development during initial teacher training.

A further exploratory study by Snead and Freiberg (2019) examined the use of Freiberg's Person-Centered Learning Assessment (PCLA; Freiberg, 1994–2017) for reflecting and developing instructional practice of 10 pre-service teachers in the United States. The pre-service teachers reported that changes in their teaching as a result of using PCLA occurred mostly in areas of planned instructional changes like engagement, levels and types of questioning, and teacher-to-student communication. Although the use of PCLA has the potential to lead to deeper levels of self-reflection and changes in teaching, further qualitative analyses of pre-service teachers' reflections on the implementation of student feedback (as a component of PCLA) showed that the quality and quantity of student feedback was heterogeneous. The authors therefore propose that in order to derive more relevant information, it would be helpful to teach students how to provide constructive feedback for instructional development.

A qualitative case study focusing on pre-service teachers' experiences with the use of feedback from different sources (teachers, faculty supervisor, peers and students in class) during their school internship was carried out by Tulgar (2019). The study examines written feedback reports from 28 pre-service teachers in Turkey. After using different sources of feedback, the participants reported development in different areas of their own professional competence, such as self-reflection, self-regulation by identifying strengths and weaknesses, evaluation of teaching performance, reflection on stress-related experiences and their planning of future lessons.

2.3 *Summary*

The presented studies in this chapter reveal a positive attitude of pre-service teachers towards student feedback, also the respective students seem to appreciate the use of student feedback. Although different instruments have been used, they all appear to have a positive impact on pre-service teachers' professional development concerning different areas of reflection on their professional actions. While student feedback is positively evaluated by pre-service teachers in general, the quality and quantity of student comments on the lesson are perceived as heterogeneous. Therefore, it is not surprising that the measurement of change in teaching quality by using student ratings is not consistent and seems problematic. In the presented studies a systematic variation in reflection settings to support reflection has not been addressed. In the following sections, two studies are presented in more detail, as they are investigating the relevance of different reflection settings when using student feedback in teacher education.

3 Studies in Germany and Switzerland

3.1 Concept and Main Findings of the ScRiPS-Study (Germany)

3.1.1 Introduction

Positive attitudes and the willingness to engage in self-reflection are considered central competences in the teaching profession; thus, an open attitude towards reflection of one's own teaching and pedagogical actions should be promoted in teacher training (Svojanovsky, 2017). The ScRiPS-study (*Schülerrückmeldungen zum Unterricht und ihr Beitrag zur Unterrichtsreflexion im Praxissemester* / The use of student feedback for reflection upon teaching during practical term) is an intervention study carried out at the University of Duisburg-Essen, North-Rhine Westphalia, Germany (Göbel & Neuber, 2017, 2019; Neuber & Göbel, 2019) and aims at supporting and analyzing the reflection on teaching with the use of student feedback in teacher training. In North-Rhine Westphalia (Germany), the first phase of teacher education is provided by universities in a Bachelor–Master structure. This first phase is mostly theoretical, addressing content knowledge, pedagogical knowledge and pedagogical content knowledge. Furthermore, two practical terms are integrated. The first practical term is an internship in schools at the beginning of the Bachelor program (duration: 5 weeks). The second internship is placed at the beginning of the Master program and lasts around 5 months. The aim of this internship in schools is to gain first experience in teaching, to reflect on practical experience and to link theoretical knowledge with practical experience. The second phase of teacher education is a mostly practical one which is realized in schools and guided by the centres for practical teacher training. The ScRiPS-study seeks to support and analyze the reflection of pre-service teachers during the 5-month practical phase of the Master program and the reflection of in-service teachers in schools when using student feedback. Changes in attitudes of pre-service and in-service teachers towards reflection and student feedback have been investigated.

3.1.2 Method

The study included 164 pre-service teachers (in the 5-month practical phase of the Master program, see above) from the University of Duisburg-Essen and 106 in-service teachers (Göbel & Neuber, 2020). The participants of the intervention groups were asked to implement student feedback on their teaching. As student feedback, a written feedback form which consisted of three open-ended questions about the quality of the lesson (What did you like about the last lesson? What did you not like about the last lesson? What could be improved for the next lesson?) was implemented. Furthermore, standardized questionnaires with a focus on either classroom management (e.g. Gruehn, 2000), classroom climate (e.g. Rakoczy et al., 2005) or cognitive

activation (e.g. Baumert et al., 2009) were provided to gather feedback from students. Both groups of teachers (pre-service and in-service) used the open-ended feedback questionnaire and could decide about the further standardized feedback questionnaire they wanted to use. The received student feedback was evaluated by the pre-service and in-service teachers individually and then discussed with the students in class.

The in-service teachers implemented student feedback on their lessons but were not further supported in the reception and reflection of the feedback. For pre-service teachers, the use of student feedback was investigated in a quasi-experimental control-group design with three intervention groups (IG) (Göbel & Neuber, 2017; Neuber & Göbel, 2019). Pre-service teachers of intervention group 1 ($n_{IG1} = 22$) obtained student feedback on their lessons but did not receive further support for reflection. Pre-service teachers of intervention group 2 ($n_{IG2} = 32$) and 3 ($n_{IG3} = 33$) received individual support for reflection in the form of a reflective journal entry which was developed in the ScRiPS-project. The reflective journal entry contains a catalogue of questions (prompts), which should enable a deeper reflection of the feedback results (Hübner et al., 2007) and refer to the lesson as well as to the results of the student feedback. The pre-service teachers of intervention group 3 also reflected on the student feedback in a collegial setting (peer reflection in tandems) at the University. To structure the collegial reflection setting, pre-service teachers could use the materials provided in the form of reflective questions and their reflective journal entries. The pre-service teachers of the control group did not use student feedback, reflective journal or collegial setting during their practical term. A total of 87 pre-service teachers were assigned to the intervention groups (use of student feedback and written or collegial setting during practical phase); 77 pre-service teachers were not assigned to any feedback-based reflection setting during practical phase (control group).

The use of student feedback was empirically investigated with regard to changes in attitudes of pre-service and in-service teachers towards reflection upon teaching. The attitudes of pre-service and in-service teachers towards reflection and student feedback were measured before and after the student feedback intervention via standardized questionnaires. The scales regarding the attitudes towards different forms of reflection, e.g. reflective journals or collegial settings, and towards the use of student feedback as a reflection stimulus, were formed by averaging the respective questionnaire items and proven to have acceptable reliability (Neuber & Göbel, 2018). All items are answered by using 4-point Likert scales which range from 1 (“I fully disagree”) to 4 (“I fully agree”). Differences between groups and changes in attitudes were analyzed with unpaired and paired t-tests and by conducting repeated measures ANOVA. In order to examine correlations between the pre-service teachers’ attitudes and their motivational preconditions, the motivation to study (Kauper et al., 2012) as well as the stress experience (Schwarzer & Jerusalem, 1999) were measured via standardized questionnaires. Furthermore, within the framework of a partial study of the ScRiPS-project, the personal experiences of the pre-service teachers with the use and reflection of student feedback on their own teaching were examined. The interviews were evaluated using qualitative content analysis.

3.1.3 Results

Looking at the results, pre-service teachers report fundamentally positive attitudes (Mean $M > 2.5$ in the 4-point Likert scale) towards reflection of teaching and student feedback (Göbel & Neuber, 2017). In addition, a high acceptance of the use of student feedback as well as the use of written and collegial forms of reflection during practical term can be shown ($M > 2.5$; Neuber & Göbel, 2020). The comparison of the different intervention groups showed that the pre-service teachers who were systematically supported in the reception and reflection of the student feedback (intervention groups 2 and 3) assessed the use of student feedback slightly more positively ($M_{IG2} = 3.29$, $SD_{IG2} = 0.41$; $M_{IG3} = 3.30$, $SD_{IG3} = 0.42$) than pre-service teachers without written or collegial reflection support ($M_{IG1} = 3.18$; $SD_{IG1} = 0.43$). However, there are no significant differences between the intervention groups in the assessment of the use of student feedback ($p = .521$). Furthermore, pre-service teachers who reflected on their own teaching both individually and in a collegial manner (intervention group 3) continue to assess the collegial form of reflection ($M = 2.86$; $SD = 0.72$) as being slightly more helpful for reflecting the student feedback than the written reflection sheet, which was used individually ($M = 2.78$; $SD = 0.62$).

In a comparative sub-study, the attitudes of 53 pre-service and 51 in-service secondary school teachers were compared (Göbel & Neuber, 2020). In the pre-test survey both pre-service ($M = 3.24$; $SD = 0.36$) and in-service teachers ($M = 3.20$; $SD = 0.50$) consider reflection on their own teaching to be important; the participants also have positive attitudes towards student feedback ($M > 2.5$). The two groups differ neither in the perceived relevance of reflection ($p = 0.605$) nor in the attitude towards student feedback ($p = 0.196$). The analysis indicates that pre-service teachers ($M = 3.04$; $SD = 0.55$) perceive structured reflection formats to be more helpful than in-service teachers ($M = 2.70$; $SD = 0.55$; $p = .002$). The same is true for collegial reflection formats; again, the analysis indicates a significant difference between the attitudes of pre-service teachers ($M = 3.42$; $SD = 0.42$) and the attitudes of in-service teachers ($M = 2.88$, $SD = 0.57$; $p < .001$). Furthermore, pre-service teachers ($M = 1.93$; $SD = 0.37$) are more critical of individual reflection settings than in-service teachers ($M = 2.29$; $SD = 0.54$; $p < .001$), although both groups tend to reject individual forms of reflection ($M < 2.5$). After using student feedback on teaching, both pre-service teachers ($M_{T1} = 3.32$; $SD_{T1} = 0.36$; $M_{T2} = 3.39$; $SD_{T2} = 0.42$) and in-service teachers ($M_{T1} = 3.20$; $SD_{T1} = 0.56$; $M_{T2} = 3.27$; $SD_{T2} = 0.56$) showed a slight increase in positive attitudes towards student feedback (within-subjects effect of time $F(1, 101) = 4.221$, $p = .043$, $\eta^2 = 0.040$). After finishing the internship ($M_{T2} = 3.34$, $SD_{T2} = 0.35$) the perceived relevance of reflection slightly increases for pre-service teachers compared to the time before the internship ($M_{T1} = 3.24$, $SD_{T1} = 0.36$, $p = .036$). Moreover, pre-service teachers are more critical regarding the use of written structured forms of reflection after finishing the internship ($M_{T1} = 3.04$, $SD_{T1} = 0.55$; $M_{T2} = 2.88$, $SD_{T2} = 0.61$, $p = .048$). For in-service teachers, however, no statistically significant changes in attitudes towards reflection are apparent.

Further analyses indicate that motivational preconditions of pre-service teachers are important for the use and reflection of student feedback (Göbel & Neuber, 2017).

Accordingly, the analyses reveal a positive correlation between pre-service teachers' attitudes towards student feedback and their motivation to study (Pearson's $r = .30, p = .008$) as well as with their positive stress experience (experience of challenge in teaching profession; $r = .40, p < .001$). The findings of the qualitative sub-study on pre-service teachers' experiences indicate that, in addition to motivational preconditions, organizational aspects of the use of feedback, e.g. arrangements with participating teachers, as well as time resources and characteristics of the students, are also important for the yield of feedback use and reflection (Neuber & Göbel, 2020). Collegial opportunities for reflection are perceived as being more helpful by pre-service teachers than individual forms of feedback reflection. In particular, the joint reflection of feedback with the students is considered as helpful by the pre-service teachers. However, pre-service teachers report differences between students of different grades in terms of their experiences with feedback and the information content of student feedback, which plays an important role in the yield of classroom reflection and thus in actual changes in teaching.

3.1.4 Summary

The findings of the ScRiPS-study show that both pre-service and in-service teachers confirm their positive attitudes towards the use of student feedback and reflection in general. The analyses for the pre-service teachers show that motivational preconditions are important for positive attitudes towards reflection. Additionally, time resources and characteristics of the student feedback seem relevant for the effective implementation of student feedback during practical phases. Collegial opportunities for reflection are perceived to be more helpful by pre-service teachers than individual forms for the reflection of feedback; in comparison in-service teachers also estimate collegial reflection positively, but not to the same extent as pre-service teachers. In future analyses differences in attitudinal changes between pre-service teachers who systematically used student feedback during practical phases and those who did not use student feedback (control group), will be examined.

3.2 *Concept and Main Findings of the Study SelFreflex (Switzerland)*

3.2.1 Introduction

In Switzerland, the training of teachers is mostly provided by universities of teacher education and is organized in a Bachelor-Master structure. The training includes different disciplines and addresses content knowledge, pedagogical knowledge and pedagogical content knowledge. Special attention is paid to a practice-oriented curriculum that combines theory and practice by allowing students to

gain practical experience from the very first semesters of study. In the practical phases, students have the opportunity to observe the teaching of in-service teachers and peers as well as to teach students in a classroom. These experiences are reflected at the university in order to link the practical experience with theoretical knowledge. In the project “Student feedback to promote teaching reflection” (Schülerrückmeldungen zur Förderung der Unterrichtsreflexion, *SelFreflex*) pre-service teachers at the Zurich University of Teacher Education in Switzerland gathered student feedback for reflection during their practical training. The intervention study was conducted with 235 students of lower secondary education (grades 7–9). The project was integrated into a 7-week practical phase which usually takes place in the 6th semester of 9 semesters. Before participating in the project, students had already completed 4 practical training phases. In the first year of study they completed two day placements and a block internship of 3 weeks duration, in the second year another block internship of 2 weeks duration. The data were collected with two samples of pre-service teachers in 2017 ($n_{2017} = 115$) and 2018 ($n_{2018} = 120$). As a reference group, the data of 20 in-service teachers were collected.

3.2.2 Method

At the beginning of the semester, pre-service teachers were asked about their attitudes towards student feedback and towards reflection by means of an online questionnaire (pre-test). The pre-test survey and other instruments used in the study were taken from the project ScRiPS (see above) and adapted for the project *SelFreflex*. After the pre-test the pre-service teachers received an input on the opportunities and goals of working with student feedback and were given the assignment to gather feedback from their students. During the practical term, pre-service teachers received feedback about their lessons from their students at two points in time. They could choose from three pre-defined questionnaires on the following aspects of teaching quality: classroom climate, classroom management and cognitive activation (see Sect. 3.1.2). In addition to the feedback received from their classes the pre-service teachers assessed their own lesson through self-evaluation. The comparison of the perspectives and the resulting consequences were expected to be discussed with students.

A group of 100 pre-service teachers reflected the findings from student feedback with an individual reflective journal entry (see Sect. 3.1.2). The reflective journal guides pre-service teachers towards a systematic reflection of a lesson while taking into account the student feedback. The reflective journal entries of all students were collected and analysed by means of qualitative content analysis (Mayring, 2015). A group of 130 pre-service teachers initially processed the student feedback together with a peer, who had observed the respective lesson, by means of collegial reflection. This group of pre-service teachers completed the individual reflective journal entry after they had received and discussed additional feedback from their peers. The feedback discussion was structured around the results of the student feedback, the pre-service teacher’s self-evaluation and the peer evaluation.

After completing the practical phase, a post-test survey was conducted using an online questionnaire. Similar to the pre-test, the post-test survey focused on the attitudes towards student feedback and reflection. In addition, items on experiences with student feedback were added to the questionnaire. Differences between groups and over time were analysed by using unpaired and paired *t*-tests. The lower secondary students were likewise asked about their experiences in a final survey. A short questionnaire was used to obtain their ratings on the usefulness of student feedback and on noticeable changes in the classroom. With selected lower secondary students, as well as pre-service teachers, semi-structured interviews were additionally conducted at the end of the practical phase.

3.2.3 Results

Based on the pre- and post-test survey of pre-service teachers ($N = 235$) it is apparent that pre-service teachers consider the engagement with student feedback to be very valuable, both before and after the practical phase. However, the agreement in the post-test survey is significantly lower ($M_{T1} = 3.29$, $SD_{T1} = 0.46$; $M_{T2} = 3.18$, $SD_{T2} = 0.46$) than in the pre-test ($p = .005$). The relevance of reflection is also rated as high whereby significant differences between the pre-test and post-test survey become visible ($p = .039$). After finishing the internship ($M_{T2} = 3.05$, $SD_{T2} = 0.45$) the perceived relevance of reflection increases for pre-service teachers compared to the time before the internship ($M_{T1} = 2.99$, $SD_{T1} = 0.49$).

The pre-service teachers consider collegial reflection to be very helpful. In the pre-test survey pre-service teachers rate the usefulness of peer reflection as high with a mean of 3.16 ($SD_{T1} = 0.49$). Interestingly, there is a difference between male and female participants in this respect. Female pre-service teachers hold more positive attitudes towards collegial reflection ($n = 133$, $M_{T1} = 3.23$, $SD_{T1} = 0.50$) than male pre-service teachers ($n = 102$, $M_{T1} = 3.08$, $SD_{T1} = 0.47$, $p = .028$). The pre-service teachers are generally open to sharing thoughts and information about their own teaching with others, rating the preference of individual reflection rather low ($M_{T1} = 2.01$, $SD_{T1} = 0.50$). However, the preference of individual reflection increases after the end of the internship ($M_{T2} = 2.09$, $SD_{T2} = 0.54$, $p = .029$).

The results of the qualitative data show that although pre-service teachers who worked with a peer highly value peer discussions, the perceived usefulness depends on various factors, such as the composition of the peer constellation. Pre-service teachers report in the interviews that collegial reflection with a peer is only beneficial if the peer shares a similar attitude towards teaching. Analyses of the peer discussions also show that critical aspects of teaching are rarely addressed (Raaflaub et al., 2019). It appears that peer discussions serve above all to positively confirm the student's own lesson reflection. In the discussion, the reflection partner serves primarily to mitigate potentially problematic aspects and to show solidarity with the pre-service teacher's problems.

In further analyses it became clear that the usefulness of student feedback also depends on the class, especially with regard to school level. In the interviews pre-service teachers report that the implementation of student feedback through questionnaires had differing outcomes depending on school level and grade. This estimation is supported by the findings of the final survey of the lower secondary students. The results show that students at a higher school level ($N = 1249$, $M = 3.19$, $SD = 0.84$) consider it significantly more important to give their opinions on lessons to their teachers than students at a lower school level ($N = 81$, $M = 2.99$, $SD = 0.92$; $p = .038$). Students at a lower school level also seem to have greater difficulty in completing questionnaires as a feedback instrument (Wyss et al., 2019). It should be noted that the different sample sizes may limit the interpretation of these results.

3.2.4 Summary

With respect to tangible results regarding the contribution of student feedback to the promotion of teaching reflection, the evaluation of the pre-service teachers' reflective journal entries shows that they predominantly evaluate their own lessons positively (Wyss et al., 2020). It is noticeable that they primarily mention aspects that can be easily observed from the outside and can therefore be positioned on the surface structure of the lesson. However, aspects that concern the deep structure of the lessons are rarely addressed. The pre-service teachers also report that the majority of the students perceive the lessons positively. When pre-service teachers were asked to compare the different perspectives, some mentioned that the perceptions were very similar whereas others noticed differences. For perceived commonality of ratings, they explain that they feel relieved that the majority of students adopted a positive attitude towards their lessons and that their self-perception is confirmed. Differences in perception are mainly attributed to different roles and interests and are thus perceived as inherent to the subject matter of teaching and to a lower extent as changeable features within lessons.

4 Discussion and Conclusions

The reported studies reveal a positive estimation of pre-service teachers towards the use of student feedback. The results support the assumption that student feedback in teacher training may be helpful to engage reflection on teaching and professional development of pre-service teachers (Tulgar, 2019). Furthermore, studies show that student feedback is evaluated positively by respective students (Porter, 1942) and may have a positive impact on teacher–student relationships (Genoud, 2006). However, pre-service teachers report that student feedback is perceived as heterogeneous (Neuber & Göbel, 2020; Snead & Freiberg, 2019; Wyss et al., 2019) and not yet treated as a valid source for the measurement of change in teaching quality

(Holtz & Gnams, 2017; Lauroesch et al., 1969). Therefore, a need for development of students' feedback competence is articulated by different authors.

The studies on student feedback in teacher education discussed in the first two sections of this chapter mostly have an exploratory design and do not address the reflection process in an explicit way. In contrast, the ScRiPS-study and the *SelfReflection*-study provide more information on different reflection settings and on the yield of student feedback for teaching reflection of pre-service teachers. In both, the German and the Swiss study, pre-service teachers positive attitudes towards the use of student feedback and towards reflection on teaching in general (Göbel & Neuber, 2017). The use of student feedback itself as well as collegial and written reflection formats are also positively evaluated (Neuber & Göbel, 2020; Raaflaub et al., 2019). The implemented collegial reflection settings and reflective journal entries seem to offer support for the reflection process. For an effective implementation of student feedback on teaching, it seems necessary that all participants (pre-service teachers and students) agree on the reflection formats to be used. Positive attitudes, motivation and volition of pre-service teachers are important for an effective implementation of student feedback. The results further point to the relevance of professional experience (in-service vs. pre-service teachers in ScRiPS) as well as gender (in *SelfReflection*). In the German sample pre-service teachers show more positive attitudes towards collegial reflection formats than in-service teachers; in the Swiss sample collegial reflection formats are more strongly preferred by female than by male pre-service teachers (Göbel & Neuber, 2020; Wyss et al., 2020).

Summing up the different findings, the use of student feedback in teacher education requires further investigation including the development of feedback instruments for different classes and school levels and furthermore concepts for reflection and time resources. For the development of pre-service teachers' reflection on student feedback, discussions between teachers and students on feedback results seem particularly promising. In these discussions open questions concerning the student feedback results can be clarified, alternative courses of action for teaching can be developed and students may get a feeling of participation and appreciation. It is important to consider that in general, both pre-service teachers and their students, might have little experience in giving and receiving feedback on teaching. Furthermore, pre-service teachers should be systematically trained and supported in the reception and reflection of student feedback while students should be trained in using the survey instruments adequately to provide helpful feedback on teaching. In the light of possible restrictions of students when giving feedback, their training of feedback competence could be a focus for further research. For future implementation of student feedback in teacher education, it is important to generate more evidence to understand better which personal prerequisites and which institutional conditions are important for a constructive use of student feedback. Furthermore, reflection on student feedback is unlikely to have an impact on classroom changes without additional support as insights gained by student feedback might not directly be translatable into teaching development. Therefore, further research is needed on different reflection concepts and settings to identify those conducive to the reflection process for pre-service teachers and their respective students.

References

- Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., Krauss, S., et al. (2009). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente*. Max-Planck-Institut für Bildungsforschung.
- Biggs, J., & Chopra, P. (1979). Pupil evaluation of teachers. *Australian Journal of Education*, 23(1), 45–57. <https://doi.org/10.1177/000494417902300105>.
- Bromme, R. (2014). *Der Lehrer als Experte: Zur Psychologie des professionellen Wissens. Standardwerke aus Psychologie und Pädagogik—Reprints, Band 7*. Waxmann.
- Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive?* Waxmann.
- Clausen, M., Diebig, K., & Jahn, S. (2020). Schülerrückmeldungen im Embedded Formative Assessment Ansatz—Wirkungen und Nebenwirkungen des Echtzeitfeedbacks mit der Becherampel. *Empirische Pädagogik*, 34(1), 46–65.
- Clausen, M., & Göbel, K. (2020). Unterrichtsrückmeldungen durch Schüler*innen. *Empirische Pädagogik*, 34(1), 6–10.
- Ditton, H., & Arnoldt, B. (2004). Wirksamkeit von Schülerfeedback zum Fachunterricht. In J. Doll, & M. Prenzel (Eds.), *Bildungsqualität von Schule: Lehrerprofessionalisierung Unterrichtsentwicklung und Schülerförderung als Strategien der Qualitätsverbesserung* (pp. 152–172). Waxmann.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>.
- Freiberg, H. J. (1994–2017). *CUN 6370 Affective instruction class syllabus: Person-centered learning assessment*. Doctoral dissertation. College of Education, University of Houston.
- Gärtner, H. (2013). Wirksamkeit von Schülerfeedback als Instrument der Selbstevaluation von Unterricht. In J. U. Hense, S. Rädiker, W. Böttcher, & T. Widmer (Eds.), *Forschung über Evaluation. Bedingungen, Prozesse und Wirkungen* (pp. 107–124). Waxmann.
- Gärtner, H., & Vogt, A. (2013). Wie Lehrkräfte Ergebnisse eines Schülerfeedbacks verarbeiten und nutzen. *Unterrichtswissenschaft*, 41(3), 252–267.
- Genoud, P. A. (2006). Influence of perceived social climate on the motivation of university students. *Studies*, 20(34), 28.
- Göbel, K., & Hesse, H.-G. (2008). Vermittlung interkultureller Kompetenz im Englischunterricht. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 398–410). Beltz.
- Göbel, K., & Neuber, K. (2017). Potenziale der Einholung von Schülerrückmeldungen für die Unterrichtsreflexion in den Phasen des Lehrerberufs. In K. Zierer (Ed.), *Allgemeine Didaktik und Lehrer/innenbildung* (pp. 88–101). Schneider.
- Göbel, K., & Neuber, K. (2019). Lernende geben Rückmeldungen zum Unterricht. Potenziale der Nutzung von Schülerfeedback und deren Bedingungen. *Friedrich Jahreshefte*, 37, 48–49.
- Göbel, K., & Neuber, K. (2020). Einstellungen zur Reflexion von angehenden und praktizierenden Lehrkräften. *Empirische Pädagogik*, 34(1), 64–78.
- Göllner, R., Wagner, W., Klieme, E., Lütke, O., Nagengast, B., Trautwein, U. (2016). Erfassung der Unterrichtsqualität mithilfe von Schülerurteilen: Chancen, Grenzen und Forschungsperspektiven. In Bundesministerium für Bildung und Forschung (Ed.), *Forschungsvorhaben in Anknüpfung an Large-Scale-Assessments* (pp. 63–82). Bundesministerium für Bildung und Forschung.
- Gruehn, S. (2000). *Unterricht und schulisches Lernen: Schüler als Quellen der Unterrichtsbeschreibung*. Waxmann.
- Hascher, T., Baillod, J., & Wehr, S. (2004). Feedback von Schülerinnen und Schülern als Quelle des Lernprozesses im Praktikum von Lehramtsstudierenden. *Zeitschrift Für Pädagogik*, 50(2), 223–243.

- Helmke, A. (2015). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (6. aktualis. Aufl.). Kallmeyer.
- Helmke, T., Helmke, A., Schrader, F.-W., Wagner, W., Nold, G., Schröder, K. (2008). Die Videostudie des Englischunterrichts. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (pp. 345–364). Beltz.
- Holtz, P., & Gnams, T. (2017). The improvement of student teachers'; instructional quality during a 15-week field experience: A latent multimethod change analysis. *Higher Education*, 74, 669–685. <https://doi.org/10.1007/s10734-016-0071-3>.
- Hübner, S., Nückles, M., & Renkl, A. (2007). Lerntagebücher als Medium des selbstgesteuerten Lernens – wie viel instruktionale Unterstützung ist sinnvoll? *Empirische Pädagogik*, 21(2), 119–137.
- Kauper, T., Retelsdorf, J., Bauer, J., Rösler, L., Möller, J., Prenzel, M., & Drechsel, B. (2012). *PaLea-Panel zum Lehramtsstudium. Skalendokumentation und Häufigkeitsauszählungen des BMBF-Projektes* (2. Welle). Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN).
- Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., Brunner, M., et al. (2007). Linking aspects of teacher competence to their instruction. Results from the COACTIV Project. In M. Prenzel (Ed.), *Studies on the educational quality of schools: The final report on the DFG Priority Programme* (pp. 39–59). Waxmann.
- Lauroesch, W. P., Pereira, P. D., Ryan, K. A. (1969). *The Use of student feedback in teacher training* (Final Report). University of Chicago.
- Lawson, T., Çakmak, M., Gündüz, M., & Busher, H. (2015). Research on teaching practicum—A systematic review. *European Journal of Teacher Education*, 38(3), 392–407. <https://doi.org/10.1080/02619768.2014.994060>.
- Lenke, L., & Praetorius, A. K. (2020). Schülerfeedback - was steckt hinter dem Kreuz auf dem Fragebogen? *Empirische Pädagogik*, 34(1), 11–29.
- Lenke, G. (2016). *Schülerfeedback in der Grundschule: Untersuchung zur Validität*. Waxmann Verlag.
- Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken*. Beltz Verlag.
- Neuber, K., & Göbel, K. (2018). *Schülerrückmeldungen zum Unterricht und Unterrichtsreflexion. Dokumentation der entwickelten Erhebungsinstrumente im Projekt „Schülerrückmeldungen zum Unterricht und ihr Beitrag zur Unterrichtsreflexion im Praxissemester (ScRiPS)“*. Aktualisierte Skalenanalysen.
- Neuber, K., & Göbel, K. (2019). Reflexion im Praxissemester – ein Forschungskonzept unter Rückgriff auf Schülerrückmeldungen zum Unterricht. In M. Degeling, N. Franken, S. Freund, S. Greiten, D. Neuhaus, & J. Schellenbach-Zell (Eds.), *Herausforderung Kohärenz: Praxisphasen in der universitären Lehrerbildung. Bildungswissenschaftliche und fachdidaktische Perspektiven* (pp. 302–311). Verlag Julius Klinkhardt.
- Neuber, K., & Göbel, K. (2020). Nutzung von Schülerrückmeldungen im Praxissemester – ein Forschungskonzept zur Förderung von Reflexivität. *Herausforderung Lehrer* innenbildung-Zeitschrift zur Konzeption, Gestaltung und Diskussion*, 3(2), 122–136. <https://www.herausforderung-lehrerinnenbildung.de/index.php/hlz/article/view/2494/3374>. Accessed 19 May 2020.
- Porter, W. A. (1942). Pupil evaluation of practice teaching. *The Journal of Educational Research*, 35(9), 700–704. <https://doi.org/10.1080/00220671.1942.10881131>.
- Raaflaub, M., Wyss, C., & Hüsler, N. (2019). Kollegiale Unterrichtsreflexion im Lehramtsstudium. *Journal Für LehrerInnenbildung*, 3, 50–57. https://doi.org/10.35468/jlb-03-2019_04.
- Rakoczy, K., Buff, A., & Lipowsky, F. (2005). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie. „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis“*. 1. Befragungsinstrumente. Deutsches Institut für nationale pädagogische Forschung.
- Röhl, S., & Rollett, W. (2020). Alles nur die Nettigkeit der Lehrkraft? Die Communion der Lehrkräfte als Erklärung für den Halo-Bias in Schülerbefragungen zur Unterrichtsqualität. *Empirische Pädagogik*, 34(1), 30–45.

- Rösch, S. (2017). *Wirkungen und Wirkmechanismen von regelmäßigem Schülerfeedback in der Sekundarstufe. Eine explorative Untersuchung im Physikunterricht*. Dissertation. Universität Basel.
- Schwarzer, R., & Jerusalem, M. (Eds.) (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Freie Universität Berlin, Humboldt-Universität Berlin.
- Snead, L. O., & Freiberg, H. J. (2019). Rethinking student teacher feedback: Using a self-assessment resource with student teachers. *Journal of Teacher Education*, 70(2), 155–168. <https://doi.org/10.1177/0022487117734535>.
- Stürmer, K., Seidel, T., Müller, K., Häusler, J., & Cortina, K. S. (2017). On what do Pre-service teachers look while teaching? An eye-tracking study about the processes of attention within different teaching settings. *Zeitschrift Für Erziehungswissenschaft*, 20(1), 74–92. <https://doi.org/10.1007/s11618-017-0731-9>.
- Svojanovsky, P. (2017). Supporting student teachers' reflection as a paradigm shift process. *Teaching and Teacher Education*, 66, 338–348. <https://doi.org/10.1016/j.tate.2017.05.001>.
- Tulgar, A. T. (2019). Four shades of feedback: The effects of feedback in practice teaching on self-reflection and self-regulation. *Alberta Journal of Educational Research*, 65(3), 258–277.
- Wagner, W. (2008). *Methodenprobleme bei der Analyse der Unterrichtswahrnehmung aus Schülersicht – am Beispiel der Studie DESI der Kultusministerkonferenz*. Dissertation. Universität Koblenz-Landau.
- Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, 28, 1–11. <https://doi.org/10.1016/j.learninstruc.2013.03.003>.
- Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology*, 108(5), 705–721. <https://doi.org/10.1037/edu0000075>.
- Wubbels, T., & Levy, J. (1993). *Do you know what you look like? Interpersonal relationships in education*. Falmer Press. <https://doi.org/10.4324/9780203975565>.
- Wyss, C., Raaflaub, M., & Hüslér, N. (2019). Schülerrückmeldungen zur Förderung der Partizipation in der Schule. In S. Hauser & N. Nell-Tuor (Eds.), *Sprache und Partizipation im Schulfeld* (pp. 181–209). hep.
- Wyss, C., Raaflaub, M., & Hüslér, N. (2020). Selbst- und Fremdwahrnehmung von Unterricht anhand von Schülerrückmeldungen. *Empirische Pädagogik*, 34(1), 79–95.

Kerstin Göbel is a Professor of Educational Research at the University of Duisburg-Essen (Germany). Previously she worked as a researcher and lecturer in the field of Educational Science at the University of Wuppertal and at the Leibniz Institute for Research and Information in Frankfurt am Main (Germany). Her research focuses on instructional development, reflection in teacher education, intercultural and interlingual learning in classroom teaching, as well as acculturation and school engagement.

Corinne Wyss is a Professor at the FHNW School of Education and Holder of the Chair for Studies on Professional Practice and Professionalism (Switzerland). From 2006 until 2015 she was a lecturer, and from 2015 until 2019 Holder of the Chair of Teacher Education Research at the Zurich University of Teacher Education (Switzerland). Her work focuses on professionalization in the teaching profession, reflection and feedback processes, video-based classroom research and teacher training, eye tracking and augmented reality in teaching and learning.

Katharina Neuber After studying educational sciences with a focus on empirical educational research and quality management, Katharina Neuber has been working as a researcher (PhD) at the University of Duisburg-Essen (Germany) since 2014. Her research focuses on the use of student feedback, reflection on teaching, practical phases during teacher education, as well as teacher stress and well-being.

Meike Raaflaub is a former teacher in lower and higher secondary education. She currently works as a researcher and lecturer for teaching methodology at the University of Teacher Education in Bern (Switzerland). Her research focuses on teaching methodology for lower secondary education, student engagement, digitalization in the language learning classroom, and the use of student feedback on teaching.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the Chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the Chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 12

Reciprocal Student–Teacher Feedback: Effects on Perceived Quality of Cooperation and Teacher Health



Jan-Erik Schmidt and Caterina Gawrilow

Abstract High lesson quality in schools is, in addition to other factors, the result of good cooperation between teachers and students. The long history of research on offer-use models of lesson quality and student–teacher relationships documents this interaction. Feedback focused on expressing the quality of cooperation can lead to higher quality of cooperation. The fact that feedback is reciprocal, from teacher to student and vice versa, helps to avoid effects of perceived injustice and rejections of feedback which otherwise are severe obstacles to the efficient use of feedback. High-frequency applications of feedback allow for the timely detection of (positive and negative) critical fluctuations of cooperation between individuals and groups and for the monitoring of processes of adaptation, as shown in other areas of applied psychology. This chapter describes the theoretical parameters of such a feedback method for students and teachers, and outlines results of an empirical study on the effects of the reciprocal method on (1) perceived quality of cooperation and (2) teacher health. Results show that, subsequent to a three-month period of reciprocal feedback, the quality of cooperation as perceived by both students and their teachers increases significantly and teacher health scores improve significantly. Reciprocal feedback techniques should be considered in teacher education and teacher training as a way to help teachers to initiate processes of improvement of lesson quality.

Keywords Feedback · Student–teacher interaction · Cooperation · Co-regulation · RCT

J.-E. Schmidt (✉)
Center for School-Quality and Teacher Education, Tübingen, Germany
e-mail: jan-erik.schmidt@zsl.kv.bwl.de

C. Gawrilow
Department of Psychology, University of Tübingen, Tübingen, Germany
e-mail: caterina.gawrilow@uni-tuebingen.de

1 Introduction

One of the most general definitions of feedback states that feedback is “information about the gap between the actual level and the reference level of a system” (Ramaprasad, 1983, p. 4). Going beyond this general definition, there is a need for a structured overview of the many different forms of feedback which have been suggested for fostering development in schools. For this, the five main characteristics of each kind of feedback should be made clear: (a) the source of feedback, (b) the recipient, (c) the topic, (d) the method, and (e) the frequency of the feedback (Hattie & Wollenschläger, 2014; Kluger & DeNisi, 1996; Mikula et al., 1990).

The most common form of feedback in the educational field is when a teacher provides feedback to a student about academic results or about their techniques of problem-solving and self-regulation in school (Hattie & Timperley, 2007). Giving feedback in the other direction—from students to teachers—seems to be an important aspect too (Hattie, 2009). Furthermore, providing feedback in both directions simultaneously could be even more powerful. This is because the teacher and student are both actors in the learning in school and both benefit from information about the transaction they create. The questions of (a) what contents and topics teachers and students should receive feedback on and (b) what kind of information is reliable are matters of intense research and are addressed in several chapters of this book, for example, Chap. 3 (Röhl and Rollett), Chap. 4 (Bijlsma), Chap. 5 (van der Lans), and Chap. 7 (Göllner et al.). This chapter refers to one specific feedback topic—the quality of cooperation between teachers and students as a class—and provides information about the views of students about their teacher and vice versa. More specifically, we also refer to a certain type of feedback—reciprocal feedback—where teacher and students send and receive feedback at the same time. Thus, the interaction between teacher and students and the dynamics of interaction can be addressed via a feedback process. The rationale and theoretical background of this kind of feedback is explained in the first part of the chapter, followed by a description of the research method used in our study. The third part presents results of a first empirical study on the effects of this form of reciprocal feedback.

1.1 Feedback Frequency

A question which—to the best of our knowledge—has so far drawn little attention from researchers is: How frequently should reciprocal feedback be provided in order to trigger practical consequences? Some research on frequency has been done in occupational settings (Ilgen et al., 1979; Kluger & DeNisi, 1996; Kuvaas et al., 2017; Park et al., 2019) and on the feedback from teachers to students (Guo & Wei, 2019; Pinter et al., 2015; Tamara et al., 2004). Also, strong support for the use of high-frequency feedback has been documented in the field of psychotherapy (Schiepek et al., 2016). However, there are no empirical studies addressing the effectiveness of

feedback frequency in the student-to-teacher direction. Furthermore, although dyadic regulation processes have already been investigated in some areas of psychology, there is no such research on the association of self-regulation and dyadic regulation processes in classroom scenarios.

The feedback introduced in this chapter is applied weekly and has been shown to be easily manageable (Schmidt, 2018). A higher frequency—e.g., daily—may be even more effective, as primacy and recency effects would be reduced. On the other hand, this would also be more difficult to realize. Weekly application thus seems a good compromise to foster co-regulation processes in the classroom between students and teachers—frequent and timely enough to be both effective and still manageable.

1.2 Interpersonal Facets of Feedback

Interpersonal facets of feedback such as “credibility” and “sender intentions” as perceived by the recipient play an important role in the acceptance and use of feedback (Umlauft & Dalbert, 2012). Those interpersonal facets can determine whether feedback information is well received and elaborated upon or is rejected. Important characteristics of persons giving feedback are their perception of being legitimated, being seen as credible, and by their motivation and intention to support the person receiving feedback. Feedback givers must also display the ability to interact in a friendly manner so that feedback information is likely to be elaborated upon. Depending on the recipient’s self-esteem and appraisal strategies, feedback carries the risk of causing negative emotions and outcomes such as lowered self-esteem and reduced effort (Leary & Terry, 2012). Feedback can be potentially perceived as unjust, and such a perception of injustice causes a variety of unwanted results, including (1) rejection of the feedback, (2) feelings of being excluded from a group, and (3) higher delinquent behavior (Mikula et al., 1990; Umlauft & Dalbert, 2012). All reported findings above focus on feedback given from instructors to their students. These mentioned risks can, however, be potentially reduced if students are involved and asked to give feedback from their perspective. When students are asked to provide feedback on cooperation with their teacher, they are implicitly addressed as competent professional partners and thus highly validated. Additionally, as teachers and students are asked to provide feedback, it is implicitly acknowledged that the views of students and teachers can differ without one being wrong or right, and that both views must be considered. Thus, perceptions of injustice could be avoided. Therefore, we see strong reasons for considering a reciprocal construction within feedback on aspects of lesson quality.

1.3 Cooperation—A Basic Ingredient for Lesson Quality

Cooperation between teachers and students addresses the fundamental characteristic of lesson quality as a transactional phenomenon, meaning that both, teachers and students, have to contribute certain activities to create a lesson. Evidence implies that feedback given from students to teachers concerning student perceptions of lesson quality can contribute to teaching effectiveness (Bill & Melinda Gates Foundation, 2012; Helmke et al., 2009; Pianta et al., 2008; Raudenbush & Jean, 2015). Still missing in this base of research is a focus on the transactional and complex character of teaching and learning (Brophy & Good, 1984; Pianta & Hamre, 2009; Pianta et al., 2003). Helmke introduces his *Angebots-Nutzungs-Modell (Offer-Uses Model of Lesson Quality)* by stating that “Good lessons are a coproduction between teachers and students” (Helmke, 2007, p. 63), suggesting that lesson quality is the result of an offer made by the teacher—as well as the result of acceptance and use of this offer by students. Moreover, the subsequent offers by the teacher are influenced by the use which students may have made of former offers. This view of teaching processes is characterized by reciprocity, irreversibility, and non-linearity as characteristics of living systems (Orsucci, 2006; Schiepek, 2009).

In other areas of psychology (as compared to educational and school psychology) the focus has changed from mere (self-) regulatory to dyadic (co-)regulation processes. Importantly, in social and health psychology, the strong claim is made that being accepted by a group and being part of a group (Forgas & Fiedler, 2020) leads to better health and a longer life. Hence, there is consistent empirical evidence that social and group relationships are protective factors for psychological and physiological health: Individuals lacking social ties are physically and mentally less healthy and more likely to die prematurely than socially integrated individuals (House et al., 1988). Transferring these social relationship results to educational and school psychology means that students and teachers who (a) work together in a cooperative and friendly manner, (b) have a productive feedback culture, and (c) feel part of the social group within the classroom and/or school, should report better well-being and maybe also better academic results. However, despite results which describe a good teacher–child relationship as a predictive factor for favorable short- and long-term outcomes in students (Hamre & Pianta, 2006), there is less research on co-regulation processes in educational contexts.

1.3.1 Cooperation and Student–Teacher Interaction

The importance of student–teacher interaction for teaching and learning has been shown across many dimensions (Hamre & Pianta, 2006; Seidel & Shavelson, 2007; Verschueren & Koomen, 2012). Students who report better relationships with their

teacher have higher academic success, as well as better social and emotional competences. In particular, “at risk” students benefit from a good student–teacher relationship (Baker, 1999; Birch & Ladd, 1998; Cheon & Reeve, 2015; Eccles & Roeser, 2011; Jennings & Greenberg, 2009; Raufelder et al., 2016; Wentzel, 2009). Specifically, student–teacher relationships include the domains of (1) organizational support, (2) academic support, and (3) social support (Eccles & Roeser, 2011; Pianta & Hamre, 2009). From a transactional point of view, such interactional support can be seen as the result of a successful process of cooperation between the teacher and their students. According to Axelrod (1984), cooperation is the willingness to abstain from maximum personal gain in favor of a common good including the willingness to seek compromises. The common good in this case can be defined as “lesson quality”, for which both students and teachers are interested in over a long-term perspective. Contributions to lesson quality by students are behaviors such as: (1) taking out their book in a good pace, (2) working silently in order not to disturb others, or (3) raising a question when feeling blocked. These behaviors are potentially hindered by students’ short-term interests in more personal gains. Example behaviors of short-term interests which may override interest in the common good of lesson quality can be: (1) making contact to a classmate, (2) taking a rest, (3) avoiding being judged by others when asking questions, or (4) low impulse control such as wishing acknowledgment for a good joke. The concept of *cooperation* serves two more benefits. Firstly, neither students nor teachers feel personally judged, since cooperation addresses an interpersonal rather than an intra-individual facet of lesson quality. Thus, feelings of humiliation are avoided—so helping to prevent withdrawal or even revenge (Furman & Ahola, 2006). Furthermore, by viewing both teachers and students as contributors to classroom success, this serves students’ need for justice as described above. Secondly, asking students how they evaluate the cooperation between themselves and their teacher implicitly conveys the message that teachers see students as capable of contributing and see their contribution as important, which supports students’ needs of self-efficacy and self-determination (Ryan & Deci, 2009). Feedback which focuses on the cooperation between students and their teacher should help the contributors reflect on their cooperation and improve it in a threefold manner: (a) by helping students to bring up ideas for the improvement of lesson quality which are from their perspective relevant, (b) by creating a situation in which teachers can learn about how students perceive lessons, tasks, and explanations and thereby receive insight into the effects of their teaching, and (c) by improving social support when listening to each other and implementing ideas developed together.

2 A Feedback Technique for Iterative Feedback About Student–Teacher Cooperation

In order to implement reciprocal feedback as described above, we developed and tested a method which focusses on the quality of cooperation between a teacher and their class as perceived by both parties. Students and teachers give their feedback weekly at the end of the last school lesson. To do so, they answer the core question of the feedback technique, “How do you evaluate the cooperation between you as a class and your teacher during the last week?”, by throwing a coin into a box with five labeled compartments (*very good*, *rather good*, *average*, *rather poor*, and *very poor*) for possible answers. The teacher answers the equivalent question, “How do you judge the cooperation between you and your class during the last week?”, by throwing a different colored coin into the box. Results of each feedback session—the distribution of the students’ answers and the teachers’ answers—were displayed on a classroom poster at the beginning of the first lesson of the next week, and the results of all weeks remained visible during the whole feedback period. Teachers and students were invited to discuss the results of the feedback each week following a solution-focused protocol in which the teachers had been trained. Thereby, the classes are guided to discuss: characteristics of weeks with higher quality of cooperation (“Why did you assess this week as having better cooperation than this other one?”); which teacher activities and which student activities contributed to good lesson quality (“What did I do to help us cooperate in this week? What did you do?”); and what could each side do to further contribute to lesson quality (“What can I do to improve our cooperation? What could you do to improve our cooperation?”).

3 Own Empirical Study

A first controlled trial study was conducted in the field of teachers’ health. We specifically investigated the effects of the reciprocal feedback method on teacher health. The rationale behind this was that the reciprocal technique could help teachers take the transactional character of lesson quality more into account by using information the students give, which would in turn foster cooperative activities which the students can participate in. The first research question was: Does teacher health improve during or after the feedback period?

To ensure an appropriate application of the feedback method it is required that teachers share the underlying idea that quality of cooperation is a core ingredient of good lesson quality and that students can contribute important information to the improvement of cooperation. Therefore, we also measured teachers’ *Resource Orientation* in respect to their students. Resource Orientation is the assumption that students have the ability to assess lesson quality and to develop ideas for the improvement of cooperation. Our hypothesis was that the experience of iterative feedback on cooperation should lead to a higher Resource Orientation among the teachers

through the experience of better cooperation, and thus reduce occupational stress which arises when teachers try to manage the class by relying primarily on their own activities. The second research question was: Does the perceived quality of cooperation as assessed by the students and by the teachers improve during the feedback period of three months?

3.1 Procedure

The sample consisted of 45 teachers from southern German mid-level schools and one of their classes between 6 and 9th grade (1022 students).

Each of the 45 teachers chose one of their classes in which they taught at least three lessons a week, and asked students to participate in the study. Teachers were randomly assigned to a treatment group ($n = 23$) or a waiting control group ($n = 22$). Resource Orientation and Teacher Health were assessed in the treatment and waiting control groups at three points of time (T_0, T_1, T_2) with 12-week intervals between each time point. After students and their parents gave written consent, the first measurement (T_0) took place. Subsequently, teachers of the treatment group received a one-day training for the feedback method and a group supervision session after four weeks. Teachers of the waiting control group received their training after T_2 . Immediately after the training, teachers and students in the treatment group applied the reciprocal feedback technique in their classes once a week for a consecutive period of 10 weeks. The supervision sessions during the feedback period were held in order to support teachers' use of the student feedback, helping them to understand the students' needs and how to lead solution-focused class talks, so that specific actions in the classroom could be derived from the feedback. For a more detailed description of the process of recruitment, random assignment, and data analysis see Schmidt (2018).

3.2 Measures

Teacher health was assessed with the General Health Questionnaire (GHQ-12) (Goldberg, 1992). The GHQ-12 is a frequently used worldwide screening instrument for detecting mental health problems. It assesses the inability to carry out one's normal healthy functions and the appearance of new phenomena of a distressing nature. The GHQ-12 asks about mental health issues during the last two weeks in comparison to the usual status of the participants. The questions include, for example, "Have you recently been feeling sad and gloomy?" Answers are coded on a four-point scale labeled e.g., less than usual, no more than usual, rather more than usual, much more than usual. Higher values indicate a higher problem level. The internal consistency of the GHQ-12 has been reported in a range of studies using Cronbach's alpha with correlations between .77 and .93.

To examine teachers' Resource Orientation, a scale called Resource Orientation Scale (ROS) was developed. The ROS consists of 12 items asking teachers how far they agree that (a) students are able to assess teacher–class cooperation and lesson quality (e.g., “My students can assess if they receive good individual support”), (b) students have useful ideas for the improvement of teacher–class cooperation and lesson quality (“My students have good ideas about what kind of support they need”), and (c) if the teacher actually uses the knowledge of students to improve lesson quality (“I use students' ideas on how to make tasks activating”). To quantify the extent of approval of the statements, answers were given on a four-point scale ranging from 1 (not true) to 4 (true). The measure's internal consistency was acceptable across time with Cronbach's alpha ranging $\alpha = .82$ at T_0 ; $\alpha = .87$ at T_1 ; $\alpha = .89$ at T_2 .

The perceived quality of cooperation was gathered by comparing the feedback of students and teachers at the beginning of the feedback process (T_1) and at the end of the process (T_2). Therefore, results of the first three weeks and results of the last three weeks of the period were averaged.

3.3 Results

To assess the effects of the training, treatment and control groups were compared with respect to changes of the outcome variables from T_0 to T_1 and from T_0 to T_2 , using regression analysis (Table 1). Therefore, outcome variables were z -standardized to T_0 means. Teachers' Resource Orientation increased significantly from T_0 to T_1 and teacher stress scores decreased significantly from T_0 to T_2 , as reported in Tables 2 and 3, respectively. The patterns of changes of the Resource Orientation Scores (ROS) and teacher health (GHQ-12) scores in treatment and control group over all three points of measurement are displayed in Fig. 1.

To assess changes in the perceived quality of cooperation, T -Tests for dependent samples have been applied. Perceived Quality of Cooperation as assessed by teachers and by students increased significantly during the three-month feedback period with

Table 1 Unstandardized scores for resource orientation and teacher health outcomes at all measurement points

| Scales | | T_0 $N_{\text{treatment}} = 23$ $N_{\text{control}} = 20$ | | T_1 $N_{\text{treatment}} = 23$ $N_{\text{control}} = 21$ | | T_2 $N_{\text{treatment}} = 21$ $N_{\text{control}} = 20$ | |
|--------|-----------|---|------|---|------|---|------|
| | | M | SD | M | SD | M | SD |
| ROS | Treatment | 2.73 | .39 | 3.02 | .39 | 2.90 | .44 |
| | Control | 2.68 | .40 | 2.68 | .46 | 2.72 | .48 |
| GHQ-12 | Treatment | 1.93 | .43 | 1.80 | .38 | 1.66 | .29 |
| | Control | 1.85 | .36 | 1.89 | .32 | 1.89 | .45 |

Note ROS = Resource Orientation Scale; GHQ-12 = General Health Questionnaire

Table 2 Regression analysis: treatment effects at T_1

| | ROS | | | GHQ-12 | | |
|-----------|----------|---------------|----------|----------|---------------|----------|
| | <i>b</i> | (<i>SE</i>) | <i>p</i> | <i>b</i> | (<i>SE</i>) | <i>p</i> |
| T_0 | .610*** | (.139) | < .001 | .563*** | (.215) | < .001 |
| Treatment | .286* | (.107) | .011 | -.119 | (.087) | .180 |
| <i>F</i> | 13,827 | | | 13,278 | | |
| <i>p</i> | < .001 | | | < .001 | | |
| R^2 | .415 | | | .399 | | |

Note ROS = Resource Orientation Scale; GHQ-12 = General Health Questionnaire
* < .05, *** < .001

Table 3 Regression analysis: treatment effects at T_2

| | ROS | | | GHQ-12 | | |
|-----------|----------|---------------|----------|----------|---------------|----------|
| | <i>b</i> | (<i>SE</i>) | <i>p</i> | <i>b</i> | (<i>SE</i>) | <i>p</i> |
| T_0 | .723*** | (.148) | < .001 | .734*** | (.137) | < .001 |
| Treatment | .079 | (.117) | .502 | -.206* | (.091) | .029 |
| <i>F</i> | 12,691 | | | 17,023 | | |
| <i>p</i> | < .001 | | | < .001 | | |
| R^2 | .420 | | | .479 | | |

Note ROS = Resource Orientation Scale; GHQ-12 = General Health Questionnaire
* < .05, *** < .001

$t(16) = 4, 24; p = .001; d = 1, 12$ for the students' view and $t(15) = 3.90; p = .001; d = 1.30$ for the teachers' view. Descriptive results for all classes of the treatment group can be seen in Fig. 2.

4 Discussion

Choosing quality of cooperation as a topic of feedback between students and their teachers and then applying reciprocal feedback repeatedly in a weekly frequency seems to be a promising approach for initiating improvement of lesson quality. Improvement in the perceived quality of cooperation from both the students' point of view and the teacher's point of view has been shown. Moreover, providing feedback about the perceived quality of cooperation to classes and inviting students to discuss cooperation in order to facilitate high lesson quality yielded improvements in teacher health. Furthermore, using such feedback for discussions between students and their teacher addresses a core process of lesson quality, since it fosters the effective use of feedback by addressing teachers and students in their role as cooperative partners.

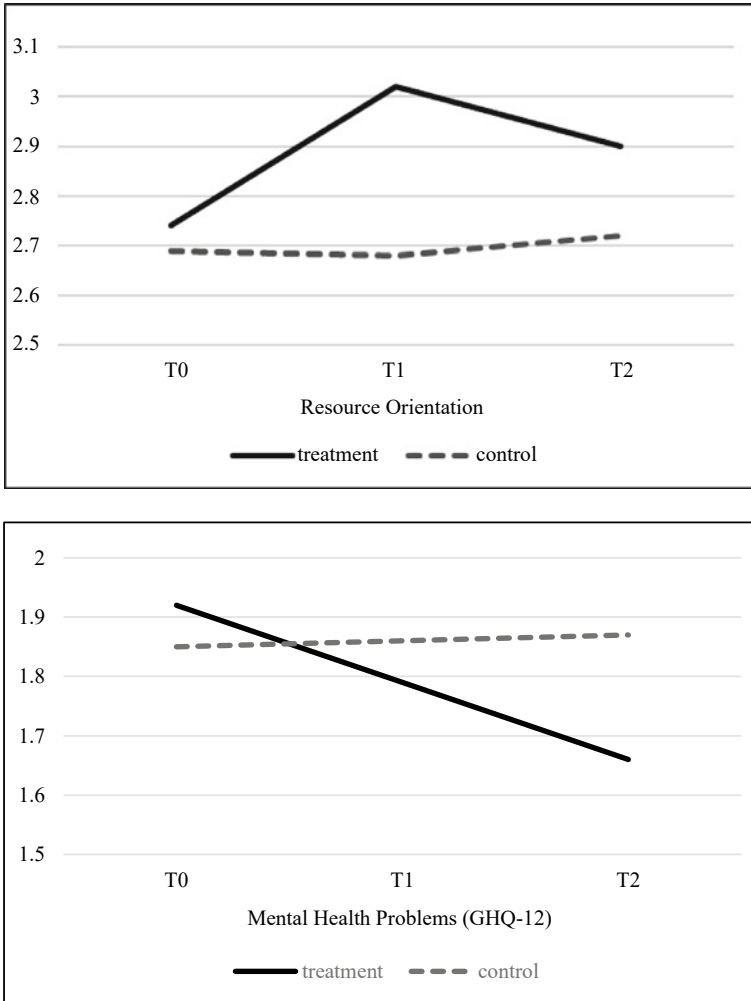


Fig. 1 Resource Orientation Scores (ROS) and teacher health (GHQ-12) scores in treatment and control group

This empirical evidence should encourage further research, as there are several limitations of the study. Firstly, the choice of classes by their teachers was deliberated. Teachers pointed out that they chose classes in which (a) improvement of cooperation between teacher and class is needed from their perspective and (b) they were confident that the group of students would be capable of using the method effectively in terms of the social relations among the students. Conflicts and a poor social climate among the students may be an obstacle to such feedback or might have to be addressed first. Secondly, future research should investigate effects of the suggested kind of feedback on other lesson quality measures than the perceived

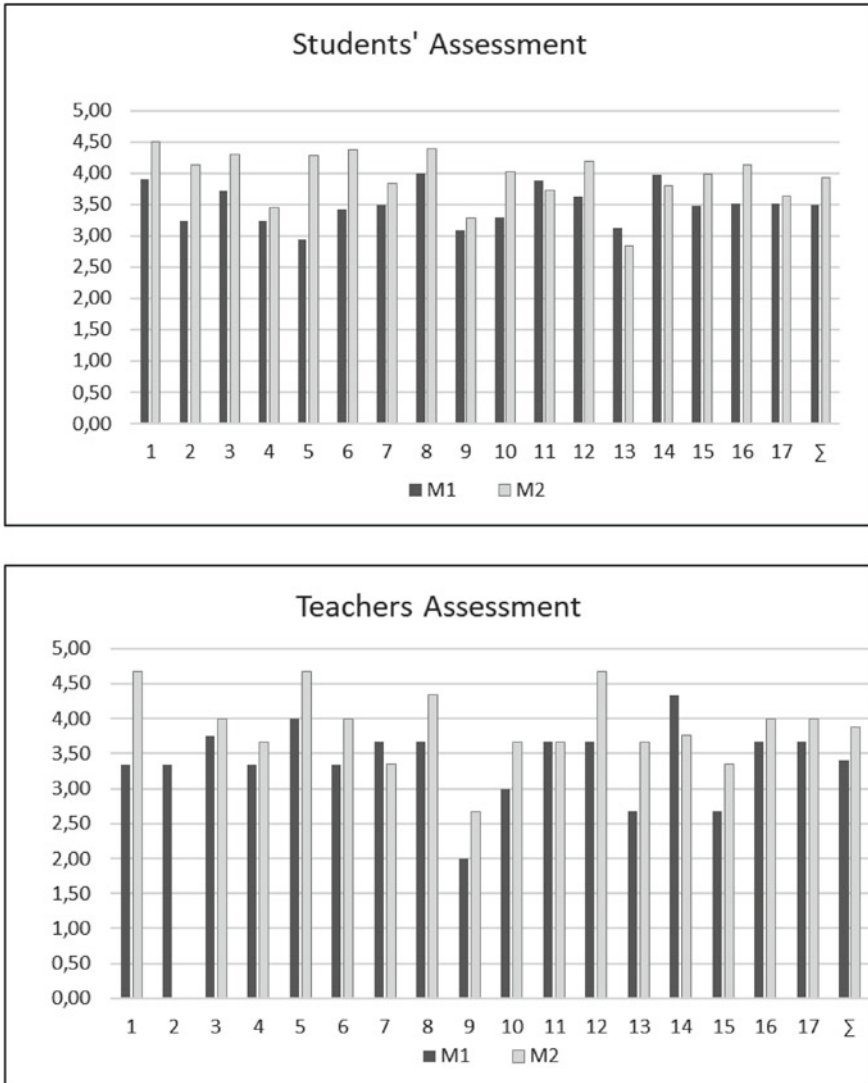


Fig. 2 Perceived quality of cooperation in each class (*Note* M1 = average of first three weeks of feedback; M2 = average of last three weeks of feedback. Due to technical barriers, not all classes provided data for all weeks of the feedback)

quality of cooperation; such measures could include time on task, cognitive activation, or emotional support, as assessed by students or external observers. We would tentatively suggest that improvement in those measures may well be as a result of improvements in cooperation between teachers and their students. Thirdly, the applied feedback technique includes several characteristics which should be further

investigated. For example, the high-frequency application of feedback could possibly be tested with other feedback topics or methods. As things can develop fast in living systems, real-time data concerning the state of a system are crucial for understanding and adapting to particular situations. The reciprocal approach—inviting students and teachers to give feedback at the same time on the same topic—can be applied to other feedback topics. Lastly, the hypothesis that the type of feedback studied here fosters student–teacher relationships should be investigated more thoroughly.

In addition, further studies are needed which examine long-term effects of the regular use of iterative and reciprocal feedback on student–teacher relationships, teacher health, and students’ academic results. Moreover, the idea that students can be viewed as partners in cooperation to improve lesson quality and that they can provide useful information to the process of cooperation should play a role in teacher education and teacher training—here teachers would develop an attitude and learn techniques to continuously strive for high lesson quality.

References

- Axelrod, R. M. (1984). *The evolution of cooperation*. Basic Books.
- Baker, J. A. (1999). Teacher-student interaction in urban at-risk classrooms: Differential behavior, relationship quality, and student satisfaction with school. *Elementary School Journal*, *100*(1), 57–70. <https://doi.org/10.1086/461943>.
- Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Policy and Practice Summary. MET Project. Bill & Melinda Gates Foundation.
- Birch, S. H., & Ladd, G. W. (1998). Children’s interpersonal behaviors and the teacher–child relationship. *Developmental Psychology*, *34*(5), 934–946. <https://doi.org/10.1037/0012-1649.34.5.934>.
- Brophy, J. E., & Good, T. L. (1984). *Teacher behavior and student achievement* (Occasional Paper No. 73). Michigan State Univ, East Lansing Inst for Research on Teaching.
- Cheon, S. H., & Reeve, J. (2015). A classroom-based intervention to help teachers decrease students’ amotivation. *Contemporary Educational Psychology*, *40*, 99–111. <https://doi.org/10.1016/j.cedpsych.2014.06.004>.
- Eccles, J. S., & Roeser, R. W. (2011). School and community influences on human development. In M. E. Lamb, M. H. Bornstein, M. E. Lamb, & M. H. Bornstein (Eds.), *Social and personality development: An advanced textbook* (pp. 361–433). Psychology Press.
- Forgas, J. C., & W., Fiedler, K. (Eds.). (2020). *Applications of social psychology: How social psychology can contribute to the solution of real-world problems*. Routledge.
- Furman, B., & Ahola, T. (2006). *The Twin Star Book: A handbook of solution focused leadership and communication*. Helsinki Brief Therapy Institute.
- Goldberg, D. P. (1992). *GHQ-12*. Nfer-Nelson.
- Guo, W., & Wei, J. (2019). Teacher feedback and students’ self-regulated learning in mathematics: A study of Chinese secondary students. *Asia-Pacific Education Researcher*, *28*(3), 265–275. <https://doi.org/10.1007/s40299-019-00434-8>.
- Hamre, B. K., & Pianta, R. C. (2006). Student–teacher relationships. In G. G. Bear, K. M. Minke, G. G. Bear, & K. M. Minke (Eds.), *Children’s needs III: Development, prevention, and intervention* (pp. 59–71). National Association of School Psychologists.
- Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>.
- Hattie, J., & Wollenschläger, M. (2014). A conceptualization of feedback. In H. Ditton, & A. Müller (Eds.), *Feedback und Rückmeldungen. Theoretische Grundlagen, empirische Befunde, praktische Anwendungsfelder* (pp. 135–149). Waxmann.
- Helmke, A. (2007). Guter Unterricht nur ein Angebot? *Friedrich Jahresheft*, 2007, 62–65.
- Helmke, A., Piskol, K., Pikowsky, B., & Wagner, W. (2009). Schüler als Experten von Unterricht. Unterrichtsqualität aus Schülerperspektive. *Lernende Schule* (pp. 98–103).
- House, J., Landis, K., & Umberson, D. (1988). Social relationships and health. *Science*, 241(4865), 540–545. <https://doi.org/10.1126/science.3399889%JScience>.
- Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4), 349–371. <https://doi.org/10.1037/0021-9010.64.4.349>.
- Jennings, P. A., & Greenberg, M. T. (2009). The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes. *Review of Educational Research*, 79(1), 491–525. <https://doi.org/10.3102/0034654308325693>.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254–284. <https://doi.org/10.1037/0033-2909.119.2.254>.
- Kuvaas, B., Buch, R., & Dysvik, A. (2017). Constructive supervisor feedback is not sufficient: Immediacy and frequency is essential. *Human Resource Management*, 56(3), 519–531. <https://doi.org/10.1002/hrm.21785>.
- Leary, M. R., & Terry, M. L. (2012). Interpersonal aspects of receiving evaluative feedback. In R. M. H. Sutton, Matthew J.; Douglis, Karen M. (Ed.), *Feedback: The communication of praise, criticism and advice* (pp. 15–28, Language as social action). Peter Lang Publishing Inc.
- Mikula, G., Petal, B., & Tanzer, N. (1990). What people regard as unjust: Types and structures of everyday experiences of injustice. *European Journal of Social Psychology*, 20(2), 133–149. <https://doi.org/10.1002/ejsp.2420200205>.
- Orsucci, F. F. (2006). The paradigm of complexity in clinical neurocognitive science. *The Neuroscientist*, 12(5), 390–397. <https://doi.org/10.1177/1073858406290266>.
- Park, J.-A., Johnson, D. A., Moon, K., & Lee, J. (2019). The interaction effects of frequency and specificity of feedback on work performance. *Journal of Organizational Behavior Management*, 39(3/4), 164–178. <https://doi.org/10.1080/01608061.2019.1632242>.
- Pianta, R. C., Hamre, B., & Stuhlman, M. (2003). *Relationships between teachers and children*. Wiley.
- Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109–119. <https://doi.org/10.3102/0013189X09332374>.
- Pianta, R. C., Karen, M., Paro, L., & Hamre, B. K. (2008). *Classroom assessment scoring system (CLASS) manual, pre-K*. Brookes Publishing Company.
- Pinter, E. B., East, A., & Thrush, N. (2015). Effects of a video-feedback intervention on teachers' use of praise. *Education and Treatment of Children*, 38(4), 451–472. <https://doi.org/10.1353/etc.2015.0028>.
- Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28(1), 4–13. <https://doi.org/10.1002/bs.3830280103>.
- Raudenbush, S. W., & Jean, M. (2015). To what extent do student perceptions of classroom quality predict teacher value added. In *Designing teacher evaluation systems* (pp. 170–202). Wiley.
- Raufelder, D., Scherber, S., & Wood, M. A. (2016). The interplay between adolescents' perceptions of teacher–student relationships and their academic self-regulation: Does liking a specific teacher matter? *Psychology in the Schools*, 53(7), 736–750. <https://doi.org/10.1002/pits.21937>.
- Ryan, R. M., & Deci, E. L. (2009). Promoting self-determined school engagement: Motivation, learning, and well-being. In K. R. Wenzel, A. Wigfield, K. R. Wenzel, & A. Wigfield (Eds.),

- Handbook of motivation at school* (pp. 171–195, Educational psychology handbook series). Routledge/Taylor & Francis Group.
- Schiepek, G. (2009). Complexity and nonlinear dynamics in psychotherapy. *European Review*, 17(2), 331–356. <https://doi.org/10.1017/S1062798709000763>.
- Schiepek, G., Aichhorn, W., Gruber, M., Strunk, G., Bachler, E., & Aas, B. (2016). Real-time monitoring of psychotherapeutic processes: Concept and compliance. *Frontiers in Psychology*, 7(604). <https://doi.org/10.3389/fpsyg.2016.00604>.
- Schmidt, J.-E. (2018). *Verborgene Kräfte im Klassenzimmer wecken: Auswirkungen iterativen Feedbacks der Qualität der Zusammenarbeit zwischen Lehrkräften und ihren Klassen auf die Lehrergesundheit und die Qualität der Zusammenarbeit*. Dissertation [Waking hidden power in the classroom: Effects of iterative feedback of the quality of cooperation between teachers and their classes on teacher health and perceived quality of cooperation]. Universität Tübingen, Tübingen.
- Seidel, T., & Shavelson, R. J. (2007). Teaching Effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. <https://doi.org/10.3102/0034654307310317>.
- Tamara, G., Nancy, L. H., Anastasia, K., & David, H. (2004). Feedback practices in a sample of children with emotional and/or behavioral difficulties: The relationship between teacher and child perceptions of feedback frequency and the role of child sensitivity. *Emotional and Behavioural Difficulties*, 9(1), 54–69. <https://doi.org/10.1177/1363275204041963>.
- Umlauf, S., & Dalbert, C. (2012). Feedback: A justice motive perspective. In R. M. Sutton, M. J. Hornsey, & K. M. Douglas (Eds.), *Feedback: The communication of praise, criticism and advice* (pp. 57–71). Lang.
- Verschueren, K., & Koomen, H. M. Y. (2012). Teacher–child relationships from an attachment perspective. *Attachment and Human Development*, 14(3), 205–211. <https://doi.org/10.1080/14616734.2012.672260>.
- Wentzel, K. R. (2009). Students' relationships with teachers as motivational contexts. In K. R. Wentzel, A. Wigfield, K. R. Wentzel, & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 301–322, Educational psychology handbook). Routledge/Taylor & Francis Group.

Jan-Erik Schmidt worked in a residential youth home, in a family-counseling-center, and is school-psychologist at the Center for School-Quality and Teacher-Education in Tübingen (Germany), since 2008. He received his doctoral degree at the University of Tübingen in 2018. His areas of interest are patterns of cooperation in educational settings and change management in complex systems. Jan-Erik is a solution-focused family therapist and board member of the European School Psychology Center for Training.

Caterina Gawrilow is Full Professor for School Psychology at the Eberhard Karls University Tübingen (Germany), since 2013. She studied psychology at the Philipps University Marburg (1997–2002), and received her Ph.D. in 2005 at the University of Konstanz. Caterina Gawrilow also worked as a Postdoc, including at the New York University and the University of Hamburg. Till 2013 she was an Assistant Professor at the Leibniz Institute for Research and Information in Education (DIPF) and the Goethe University Frankfurt. Her research focuses on cognitive, emotional, social, motivational, and neuronal correlates of ADHD, on self-regulation and self-regulation interventions—in experimental and intensive longitudinal designs.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part III
Relating to Other Fields of Research

Chapter 13

Student Voice and Student Feedback: How Critical Pragmatism Can Reframe Research and Practice



Mari-Ana Jones and Valerie Hall

Abstract This chapter recognises the diverse definitions and practices of student feedback; focussing on how student feedback can facilitate dialogue and thus contribute to the development of schools as democratic communities. Student feedback is thus positioned as a part of student voice, which has its roots in the United Nations Convention on the Rights of the Child (UNICEF, 1989). We question the ways in which schools elicit the views of students and how students' opinions are made use of, recognising the complexities arising from power relationships (Hart, 1992), the consumerisation of education (Whitty & Wisby, 2007) and the pressures of accountability. Furthermore, we consider ways in which researchers can address difficulties in the research-practice relationship (Chapman and Ainscow, 2019) and facilitate co-creation of research. We propose the perspective of critical pragmatism as a means to acknowledge the complexities of practice, whilst also highlighting the importance of critical reflection and dialogue. Critical pragmatism could move us from a “deconstructive scepticism toward a reconstructive imagination” (Forester, 2012, p. 6) in which schools and researchers collaborate to enable contextually rich practices of student feedback and student voice.

Keywords Student feedback · Student voice · Critical pragmatism · Dialogue · Reflection · Collaboration

1 Introduction

We recognise that there is a fundamental belief in the need for schools to provide safe environments in which students can speak, and for student feedback to be used to implement change (Defur & Korinek, 2010). After all, “the first claim of the school is that of its pupils for whose welfare the school exists” (Stenhouse, 1983,

M.-A. Jones (✉)
Norwegian University of Science and Technology, Trondheim, Norway
e-mail: mari.a.jones@ntnu.no

V. Hall
University of Wolverhampton, Wolverhampton, UK

p. 153). There is, however, much discussion about what we mean when we talk about “student feedback”. We have chosen to accept the premise that student feedback can be defined as “the use of formal processes to gather information from students about their perceptions of teacher practices, teacher effectiveness and the quality of educational programmes” (Mandouit, 2018, p. 756).

However, the vocabulary used around student feedback has become increasingly diverse, with concepts holding different meanings for those involved (Forrest et al., 2007). The context within which such feedback is situated varies enormously: the cultural and environmental influences; the methods and practices used to elicit such feedback; policy and regulatory frameworks; the students and staff, and their relationships; and the purposes for which such feedback is sought. The overall intent may be about improvement, but the drivers come from a broad spectrum of need: from a performativity perspective that can demonstrate accountability and effectiveness (Verhaeghe et al., 2010); to opening up a “dialogue around teaching and learning in the classroom ...[that could give].... teachers insights into the unique challenges experienced by their students” (Mandouit, 2018, p. 755). In this form, student feedback can be identified as a form of student voice, with schools aiming to serve as democratic environments in which structures can be created that enable students, teachers and the broader school family, to have “meaningful involvement in decision-making processes” (Defur & Korinek, 2010, p. 19) and for teachers’ classroom practice to be improved (Bourke & Loveridge, 2016; Mitra, 2008). Such opportunities for participation encourage the development of a student’s sense of agency and self-worth; a sense of belonging and reflection on past, present and future relationships (Thompson, 2005).

Student voice and student feedback derived from different agendas. The expansion of interest in student voice can be traced to Article 12 in the United Nations Convention on the Rights of the Child (UNICEF, 1989), which states that children have the right to be heard. Student feedback—mainly developed in higher education institutions and intended as a quality assurance measure (Harvey, 2003)—has been used to gather views for a specific purpose. As such, their foundations are somewhat different, but there are important interconnections. At their best, they enable a collaborative dialogue and the development of consultation across all stakeholders (Nelson, 2015). At their worst, they become instrumentalist in demonstrating compliance (Charteris & Smardon, 2019), or tokenistic in positioning students as consumers of education (Hall, 2020). We need to consider whether students are being engaged as “insiders” or “outsiders” (Forrest et al., 2007, p. 26): are they informing practice from within—through collaboration and agency; or is their purpose only to help fulfil the requirements of accountability frameworks?

Within this chapter, our focus is on student voice, but we acknowledge that this concept is also evidenced within student feedback and that there are many inter-linking practices and connections between the two. We are thus interested in the ways in which researchers are in a position to support schools to critically explore how school communities espouse, enact and experience student voice (Hall, 2020). This chapter offers a critically pragmatic perspective that has the potential to enable student voice research to recognise the aspirations of student voice whilst not losing

sight of the realities of school life. Schools can be enabled to reclaim student voice. They can value their own local knowledge and experiences and their own contexts, with the “thoughtful and serious consideration of student voice” (Keddie, 2015, p. 227) having the potential to yield considerable benefits. In doing so, opportunities emerge to develop contextually relevant practices that are: enriching for students and teachers alike (Bragg & Manchester, 2012; Fleming, 2015); that take into account the diversity of concepts and contexts (Mandouit, 2018; Verhaeghe et al., 2010); and that consider the ways in which discourses interconnect and overlap within student feedback and student voice (Charteris & Smardon, 2019). The discussion within this chapter, therefore, considers ways in which we as researchers may *begin* to facilitate “co-creation” of research; how we *mediate* and *broker* knowledge through “engaging in the identification and formulation of knowledge needs” (Wollscheid et al., 2019, p. 289).

2 Situating the Chapter

Within the broader context of student voice, and by association student feedback, there are many rich discussions taking place across an international arena. Central to much of this is the acknowledgement that there are difficulties in accommodating national policies and competing priorities, differences in school contexts, and views that exist on pedagogical approaches, as evidenced by research emerging from countries who are working on collaborative European projects (Bron et al., 2018; Holcar Brunauer, 2019). We need, therefore, to appreciate the constraints and challenges imposed on schools endeavouring to meet requirements, wherever they are situated. As demonstrated by research from New Zealand (Bourke & Loveridge, 2016, p. 59) this also needs to recognise that sometimes the focus is on “what can be changed, and not what confronts practices especially if the student feedback is challenging”. The discussion within this chapter thus aligns with themes across this wider debate and seeks to broaden perceptions of student voice research and practice, highlighting some of the key drivers that influence, and sometimes hinder, the development of a more critically pragmatic approach: a “philosophy for professionals” (Ulrich, 2007, p. 1112).

To make improvements in student outcomes we know that it makes sense to go straight to the source as students can not only share opinions about their classroom experiences, but also play a significant role in school improvement efforts. But how do we best involve students in school decisions that will shape their lives and the lives of their peers? (Mitra, 2008, p. 20).

There are, however, concerns about the methods used to elicit student feedback—surveys, questionnaires, evaluation results. These relate not only to the validity of their construction and the questions asked, but also the ways in which any results are interpreted (Darwin, 2016), for “It is not just the collection of data that is important, but the value that is placed on student evaluations” (Blair & Noel, 2014, p. 881). Institutions frequently find themselves operating between two conflicting objectives,

“one which is focused on directives that accord success for meeting targets, and the other based on aspirations to enhance the community by allowing each student the possibility to be heard” (Shuttle, 2007, p. 33). A methodological quest for “authentic” student responses should be treated with caution (Spyrou, 2016), and Nelson (2015, p. 5) argues that a notion of authentic student voice “masks how power relations operate” in the production of student voice. Consideration needs to be given to *who* is assigning *value* and *worth* to such dialogue, and the emergent data, and how *equal* is the potential for all individuals to be involved (DeFur & Korinek, 2010).

Hart (1992) recognised that there would be issues of power and participation when adults in such settings attempted to work in partnership with children. His much vaulted “ladder of participation”—moving from levels of non-engagement (manipulative and tokenistic) through to levels of engagement with evidence of growing consultation, agency and the development of shared power (with the potential for transformation)—has acted as a catalyst for much discussion in the arena (Fielding, 2001, 2011; Groundwater-Smith & Mockler, 2016). Student voice has become a right and a “key aspect of youth agency” incorporating varied practices, but these require “careful, situated interpretation if we are to understand their meanings and effect” (Bragg & Manchester, 2012, p. 143). This raises some fundamental questions for both students and teaching staff. Students may feel alienated through what might be seen as a “tokenistic” approach to student voice (Fielding, 2011) or consider themselves being positioned merely as “consumers of education” (Whitty & Wisby, 2007, p. 303). Teaching staff may experience similar tensions in their understanding of the implications of student voice for teacher professionalism and whether it should be regarded as “an important element in establishing a ‘collaborative’ or ‘democratic’ professionalism, or a challenge to teachers’ authority and cement an associated ‘managerial’ model of professionalism” (Whitty & Wisby, 2007, p. 303). These discourses are linked, and even overlapping at times (Charteris & Smardon, 2019), and consequently, schools have pressed on with various student voice initiatives that might demonstrate collaboration and engagement both for compliance purposes but also undoubtedly with good intent to engage learners in constructive dialogue. Due to the constraints of complex regulatory frameworks which require evidence of both compliance and learning, however, schools are rarely able to step back; to challenge and to seek ways in which student voice can be not only a “tool for change”, but also a “tool for reflection” (Bourke & Loveridge, 2016, p. 65).

So, the challenge from our perspective is how research can work more closely with teachers and empower them to incorporate student voice as “part of their own professional learning and development” (Bourke & Loveridge, 2016, p. 66). If school leaders and teachers can become more “invested” in the creation and development of knowledge, they can participate further in the drive to identify and formulate those knowledge needs (Kauffman et al., 2017). This does not necessarily mean a call for more methodologies that facilitate the involvement of schools, or for greater engagement with action research, but rather an avoidance of a “linear dissemination from experts to practitioners” (Blackmore, 2007, p. 28). Our discussion, therefore, moves on to consider ways in which we might be able to “reframe” research within this context.

3 Reframing the Role of Student Voice Research

Hargreaves (1999, p. 125) described a “division” between researchers and practitioners and Lester et al. (2002) raised the issue of how teachers and researchers might be expected to communicate when they obviously occupied different worlds. Researchers are frustrated with simplistic, mechanistic practices whilst teachers are subject to the supposed “moral and intellectual authority” of researchers who “derive their power” from criticising at a distance (Chapman & Ainscow, 2019, p. 915). Blackmore (2007) described the failings of a linear view of the research-practice relationship in which knowledge is supposed to be passed down from academics to practitioners, pointing especially to the inaccessibility of the reporting of findings. This implies that language is the most significant barrier, however, Biesta et al. (2019) raise fundamental questions about perceptions of the relevance of research. Chapman and Ainscow (2019) criticise the ways in which knowledge is produced by research, advocating for an inclusive “messy social learning process” (p. 914) which addresses unequal power relationships between researchers and practitioners.

We contend that these issues are especially noticeable in student voice work. At times, as Mager and Nowak (2012, p. 50) suggest, student voice researchers have conducted “too little methodologically strong research”. Fielding (2011, p. 10) argues that student voice research has not paid enough attention to theoretical frameworks, due to the “corrosive nature of market-led approaches”. There are questions about the value of student voice research and its capacity to influence practice. Likewise, the emancipatory and empowerment traditions of student voice research contribute to difficulties in the enactment and experience of student voice (Hall, 2020). Whilst teachers were willing to engage with research, Bourke and Loveridge (2016) report that it was challenging for them to take account of findings which appeared to contradict their experiences and views. Harris (2010, p. 88) concurs, noting that teachers had varying responses to findings:

There were some teachers in each group who really wanted to be handed immediate ideas that they could take back to their classrooms. Others felt they came with considerable expertise and there was nothing new they did not already know. Still others were pleased to engage in reflective discussion and make their own links to classroom practice whilst being open to new ideas.

There are surprisingly few mandates for teachers to connect with educational research, despite the professionalisation of education, with teachers often seen as receiving knowledge from external sources, rather than being part of creating it (Wollscheid et al., 2019). An argument further supported by Harris (2010), who suggests that teachers are expected to receive and reproduce knowledge. Our intention, therefore, is to propose a reframing of research and the roles of researchers and practitioners, which involves a “reconstruction of relations” (Hargreaves, 1999, p. 136) in which teachers are “at the heart” (ibid.). To achieve these aims we need a “brokering” system (Wollscheid et al., 2019, p. 270) in which knowledge moves fluidly and dynamically between research and practice. For student voice research,

these suggestions would support the construction of a dialogue which is more in keeping with the democratic, inclusive and transformational traditions of the field (Fielding, 2011). In this way, the concerns raised about how research might “reach” the “practice of education ... [moving the focus and so] ... changing the location of research and the identity of the researcher” (Biesta et al., 2019, p. 2) lead us to our next consideration: where, and how, such a shift might be enabled.

4 Critical Pragmatism as a Way Forward

So, having established the tensions, constraints—and possibilities, how might we find a way forward? What has emerged from the discussion is the need to reach “beyond the confines of technical philosophy” (Dewey, 1949, p. xiv) towards a more critical approach. Such a perspective has the potential to help “bridge” the gap and facilitate discussion between research and practice and to have progressive adjustments made “in light of collective deliberation grounded in the experience of every member of society” (Curren, 2010, p. 494). Before considering its relevance and how it might be applied to student voice research and practice, it is necessary to first define critical pragmatism.

Critical pragmatism incorporates both pragmatism and critical theory. Dewey (1925) identifies Peirce (1839–1914) as the originator of pragmatism, having been inspired by Kant’s 1785 distinction between the practical and the pragmatic. Dewey (1925) explains that Peirce was interested in how concepts could be made clear, which according to Peirce, could only be achieved by their application to human experience. Dewey (1925) elaborates, arguing that action is the intermediary through which concepts gain meaning. Furthermore, because actions can be different, meanings can be different. Biesta (2006, p. 30) interprets Dewey’s thinking thus; “it is because people share in a common activity that their ideas and emotions are transformed as a result of the activity in which they participate”. When applied to student voice, this understanding of pragmatism can help to explain variations in understandings and practices between schools, as well as divisions between the conceptualisation of student voice in theory and policy and how it is practised. Put simply, the concept of student voice is actioned in many ways, leading to multiple experiences and understandings. The critical aspect is crucial; encouraging reflective practice and drawing attention to power issues. As Feinberg (2015, p. 151) explains “the distinctive task of critical pragmatism is to bring competing norms to the surface, to show how they impede experience and to encourage the formation of new ways.”

At the start of this chapter, we began to explore some of the tensions that exist between student voice used as an accountability measure (Verhaeghe et al., 2010), and student voice being part of schools’ democratic processes (Bourke & Loveridge, 2016; Defur & Korinek, 2010; Mitra, 2008). We witness teachers and schools endeavouring to meet external accountability requirements connected with education’s “marketisation and the development of a consumer culture” (Murphy & Skillen,

2013, p. 84). Keddie (2015, p. 226) describes teachers having “a sense of powerlessness and high levels of uncertainty”. Teachers express concerns about the erosion of their “ability to complete what they consider core professional tasks – dealing with the issues and concerns of pupils” (Murphy & Skillen, 2013, p. 89). In this climate, there is a danger that the potential of student voice as a reflective tool can be forgotten. For schools, critical pragmatism as a lens can be useful as a means of acknowledging the demands of accountability, whilst also encouraging critical reflection. A critical pragmatist perspective suggests compromise rather than an either-or perspective; student voice need not be either for accountability or for democracy. Rather, by providing “fertile ground on which such ideas can be questioned, refined or even transformed” (Murphy & Skillen, 2013, p. 95), it enables schools to critically reflect on their practices of student voice. For example, the ways in which they are collecting the views of students and what they are doing with the data.

For student voice researchers, critical pragmatism encourages an acknowledgement of the realities of the complex network of demands on schools and the need for action, as well as a recognition of the importance of local knowledge and understanding in the practice of student voice. Taking a critical pragmatist stance mitigates against researchers becoming overly critical of student voice practices, instead highlighting the importance of examining contextualised practice. The potential now exists for research and practitioners to recognise and acknowledge that it is no longer enough for the role of research to be rooted in production of “evidence-based practice” or “evidence-informed teaching”: the “*what works*” as discussed by Biesta et al. (2019). Rather, researchers can seek to co-construct research knowledge that is “geared towards producing *useful* knowledge which is able to answer the questions practice ask ... [whilst also acknowledging] What does it work *for*?” (Biesta et al., 2019, p. 2).

It is, therefore, time to reframe our perceptions and perspectives so that rather than “determining practice” we grasp the potential for research to ‘inform practice’, with teachers viewed not as “recipients of research and reproducers of knowledge”, but rather as “producers and interrogators of research and builders of knowledge” (Harris, 2010, p. 83) in their professional capacities. A critical pragmatist orientation could thus have the potential to foster mediation, respecting the perspectives of all those involved, and—crucially—enabling each to learn “from, [and] about, each other, so that they can work to invent creative new options for action, [and] work to produce pragmatic outcomes serving their values and interests, as well” (Forester, 2012, p. 13). Critical pragmatism can enable a dialogue between researchers and schools—mutual recognition of each other’s standpoints and encourage learning from each other.

5 Conclusion

Although the focus of our chapter is student voice, we highlight interconnections with student feedback, appreciating that in spite of their different foundations and agendas, the two concepts have much in common. We have considered the diversity of concepts

and contexts (Mandouit, 2018; Verhaeghe et al., 2010), acknowledging that there are discourses that interlink and overlap (Charteris & Smardon, 2019). The capacity for collaborative dialogue and consultation across stakeholders (Nelson, 2015) on the one hand; but also, the potential to be instrumentalist, tokenistic and compliance driven (Charteris & Smardon, 2019) on the other. Our aim in this discussion, therefore, is for student feedback and student voice research to be understood as “bounded in both the context and the culture of specific settings ... [that make it] ... complex, challenging and contradictory” (Fleming, 2015, p. 224). In doing so, we broaden the debate about the ways in which both student feedback and student voice are “espoused, enacted and experienced” (Hall, 2020, p. 125) by researchers and in schools.

Situated amidst complex regulatory frameworks, schools at times operate between conflicting objectives. It can, therefore, be difficult to see ways in which student feedback and student voice research can navigate competing priorities, institutional contexts, and pedagogical beliefs (Bourke & Loveridge, 2016; Bron et al., 2018; Holcar Brunauer, 2019). Researchers have an important role. Instead of positioning ourselves as remote experts, disseminating our findings and criticising practice from afar, we are suggesting the development of a “close-to-practice” approach (Wyse et al., 2020, p. 20). Researchers should seek collaboration with practitioners, thus encouraging an iterative process of research and application that includes “reflections on practice, research, and context” (ibid.). If there is to be change, then it needs to be through mediation of the knowledge (Wollscheid et al., 2019); and that knowledge has to have been co-constructed. A critically pragmatic perspective for both researchers and schools could facilitate the development of contextually rich practice(s)—recognising the constraints that schools operate within, whilst taking the strengths of pragmatic thought, valuing local knowledge and experiences (Keddie, 2015) and also contributing a critical lens.

To support these aspirations, we propose the following:

- Developing a philosophy of enquiry and research amongst teachers;
- Considering the initial, and continuing, professional development needed for teachers to engage meaningfully in classroom research—perhaps a “toolkit” for teachers that can help to bridge the gap;
- Building a culture that ensures research is done with, not “on”, teachers, students and the institution;
- Ensuring consensus about the educational implications of any activity and research undertaken; and
- Working collaboratively to identify and promote those forms of interaction that have the most beneficial educational outcomes.

We suggest that critical pragmatism could provide a means through which to work towards these aims, enabling us to “rethink the complexities of deliberative processes” (Forester, 2012, p. 6); for researchers to start from where schools are and at the same time enable schools to critically examine their practice. Our premise, therefore, is that critical pragmatism could move us from a “deconstructive scepticism toward a reconstructive imagination” (Forester, 2012, p. 6) where there are

possibilities for *joint* gain; and for *multi-directional* gain that may satisfy the multiple and diverse needs of all.

References

- Biesta, G. (2006). "Of all affairs, communication is the most wonderful": The communicative turn in Dewey's Democracy and Education. In D. T. Hansen (Ed.), *John Dewey and our educational prospect: A critical engagement with Dewey's Democracy and Education* (pp. 23–38). State University of New York Press. <https://doi.org/10.1111/j.1748-5959.2008.00161.x>.
- Biesta, G., Filippakou, O., Wainwright, E., & Aldridge, D. (2019). Why educational research should not just solve problems but should cause them as well. *British Educational Research Journal*, 45(1), 1–4. <https://doi.org/10.1002/berj.3509>.
- Blackmore, J. (2007). How is educational research "being framed"? In B. Somekh & T. Schwandt (Eds.), *Knowledge production* (pp. 24–41). Routledge. <https://doi.org/10.4324/9780203609156>.
- Blair, E., & Noel, K. V. (2014). Improving higher education practice through student evaluation systems: Is the student voice being heard? *Assessment and Evaluation in Higher Education*, 39(7), 879–894. <https://doi.org/10.1080/02602938.2013.875984>.
- Bourke, R., & Loveridge, J. (2016). Beyond the official language of learning: Teachers engaging with student voice research. *Teaching and Teacher Education*, 57, 59–66. <https://doi.org/10.1016/j.tate.2016.03.008>.
- Bragg, S., & Manchester, H. (2012). Pedagogies of student voice. *Revista de Educación*, 359, 143–163. <https://doi.org/10.4438/1988-592X-RE-2012-359-200>.
- Bron, J., Emerson, N., & Kákonyi, L. (2018). Diverse student voice approaches across Europe. *European Journal of Education*, 53(3), 1–28. <https://doi.org/10.1111/ejed.12285>.
- Chapman, C., & Ainscow, M. (2019). Using research to promote equity within education systems: Possibilities and barriers. *British Educational Research Journal*, 45(5), 899–917. <https://doi.org/10.1002/berj.3544>.
- Charteris, J., & Smardon, D. (2019). The politics of student voice: Unravelling the multiple discourses articulated in schools. *Cambridge Journal of Education*, 49(1), 93–110. <https://doi.org/10.1080/0305764X.2018.1444144>.
- Curren, R. (2010). Pragmatist philosophy of education. In H. Siegel (Ed.), *The Oxford handbook of philosophy of education* (pp. 489–507). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780195312881.003.0027>.
- Darwin S. (2016). *Student evaluation in higher education: Reconceptualising the student voice*. Springer International Publishing. https://doi.org/10.1007/978-3-319-41893-3_2.
- DeFur, S., & Korinek, L. (2010). Listening to student voices. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 83(1), 15–19. <https://doi.org/10.1080/00098650903267677>.
- Dewey, J. (1925). *The essential dewey. Volume 1: Pragmatism, education, democracy* (L. A. Hickman & T. M. Alexander, Eds.) (1998). Indiana University Press.
- Dewey, J. (1949). Foreword. In P. Wiener (Ed.), *Evolution and the founders of pragmatism* (pp. xiii–xiv). Harvard University Press. <https://doi.org/10.9783/9781512808483>.
- Feinberg, W. (2015). Critical pragmatism and the appropriation of ethnography by philosophy of education. *Studies in Philosophy and Education*, 34, 149–157. <https://doi.org/10.1007/s11217-014-9415-6>.
- Fielding, M. (2001). Students as radical agents of change. *Journal of Educational Change*, 2(2), 123–141. <https://doi.org/10.1023/A:1017949213447>.
- Fielding, M. (2011). Student voice and the possibility of radical democratic education. In G. Czerniawski & W. Kidd (Eds.), *The student voice handbook* (pp. 3–17). Bingley.

- Fleming, D. (2015). Student voice: An emerging discourse in Irish Education Policy. *International Electronic Journal of Elementary Education*, 8(2), 223–242. <https://www.iejee.com/index.php/IEJEE/article/view/110>. Accessed 6 Aug 2020.
- Forester, J. (2012). On the theory and practice of critical pragmatism: Deliberative practice and creative negotiations. *Planning Theory*, 12(1), 5–22. <https://doi.org/10.1177/1473095212448750>.
- Forrest, C., Lawton, J., Adams, A., Louth, T., & Swain, I. (2007). The impact of learner voice on quality improvement. In D. Collinson (Ed.), *Researching leadership in the learning and skills sector: By the sector, on the sector, for the sector: Volume 4* (pp. 13–29). https://www.researchgate.net/publication/280092665_Leadership_and_the_Learner_Voice. Accessed 6 Aug 2020.
- Groundwater-Smith, S., & Mockler, N. (2016). From data source to co-researchers? Tracing the shift from ‘student voice’ to student-teacher partnerships in educational action research. *Educational Action Research*, 24, 159–176. <https://doi.org/10.1080/09650792.2015.1053507>.
- Hall, V. J. (2020). Reclaiming student voice(s): Constituted through process or embedded in practice? *Cambridge Journal of Education*, 50(1), 125–144. <https://doi.org/10.1080/0305764X.2019.1652247>.
- Hargreaves, D. H. (1999). The knowledge-creating school. *British Journal of Educational Studies*, 47(2), 122–144. <https://doi.org/10.1111/1467-8527.00107>.
- Harris, P. (2010). Mediating relationships across research, policy, and practice in teacher education. *Studying Teacher Education*, 6(1), 75–93. <https://doi.org/10.1080/17425961003669334>.
- Hart, R. A. (1992). *Children’s participation: From tokenism to citizenship* (Innocenti Essays No. 4). UNICEF International Child Development Centre. <https://ilk.media.mit.edu/courses/mas714/fall03/unicef.pdf>. Accessed 6 Aug 2020.
- Harvey, L. (2003). Student feedback [1]. *Quality in Higher Education*, 9(1), 3–20. <https://doi.org/10.1080/13538320308164>.
- Holcar Brunauer, A. (Ed.) (2019). *Student voice in education. CIDREE Yearbook 2019*. National Education Institute Slovenia. https://www.zrssi.si/digitalnknjiznica/CIDREE_Yearbook_2019.pdf. Accessed 3 May 2020.
- Kauffman, J. M., Hallahan, D. P., & Pullen, P. C. (Eds.). (2017). *Handbook of special education* (2nd ed.). Routledge. <https://doi.org/10.4324/9781315517698>.
- Keddie, A. (2015). Student voice and teacher accountability: Possibilities and problematics. *Pedagogy, Culture and Society*, 23(2), 225–244. <https://doi.org/10.1080/14681366.2014.977806>.
- Lester, F. K., William, D., & Lester, F. K. Jr. (2002). On the purpose of mathematics education research: Making productive contributions to policy and practice. In L. D. English, M. B. Bussi, G. A. Jones, R. A. Lesh, & D. Tirosh (Eds.), *Handbook of international research in mathematics education* (pp. 489–506). Lawrence Erlbaum Associates.
- Mager, U., & Nowak, P. (2012). Effects of student participation in decision making at school: A systematic review and synthesis of empirical research. *Educational Research Review*, 7(1), 38–61. <https://doi.org/10.1016/j.edurev.2011.11.001>.
- Mandouit, L. (2018). Using student feedback to improve teaching. *Educational Action Research*, 26(5), 755–769. <https://doi.org/10.1080/09650792.2018.1426470>.
- Mitra, D. (2008). Amplifying student voice. *Educational Leadership*, 66(3), 20–25. <http://www.ascd.org/publications/educational-leadership/nov08/vol66/num03/Amplifying-Student-Voice.aspx>. Accessed 6 Aug 2020.
- Murphy, M., & Skillen, P. (2013). The politics of school regulation: Using Habermas to research educational accountability. In M. Murphy (Ed.), *Social theory and education research, understanding Foucault, Habermas, Bourdieu and Derrida* (pp. 84–97). Routledge.
- Nelson, E. (2015). Student voice as regimes of truth: Troubling authenticity. *Middle Grades Review*, 1(2) Article 3. <https://pdfs.semanticscholar.org/d32f/bbe6415fa11bfd4319639b5c6c0928a55d0e.pdf>. Accessed 6 Aug 2020.
- Shuttle, J. (2007). Learners involvement in decision making. In D. Collinson (Ed.), *Researching leadership in the learning and skills sector: By the sector, on the sector, for the sector: Volume 4* (pp. 30–48). https://www.researchgate.net/publication/280092665_Leadership_and_the_Learner_Voice. Accessed 6 Aug 2020.

- Spyrou, S. (2016). Researching children's silences: Exploring the fullness of voice in childhood research. *Childhood*, 23(1), 7–21. <https://doi.org/10.1177/0907568215571618>.
- Stenhouse, L. (1983). *Authority, education and emancipation*. Heinemann Educational.
- Thompson, C. (2005). The non-transparency of the self and the ethical value of Bildung. *Journal of Philosophy of Education*, 39(3), 519–533. <https://doi.org/10.1111/j.1467-9752.2005.00451.x>.
- Ulrich, W. (2007). Philosophy for professionals: Towards critical pragmatism. *The Journal of the Operational Research Society*, 58(8), 1109–1113. <https://www.jstor.org/stable/25681907>. Accessed 6 Aug 2020.
- UNICEF. (1989). *The United Nations Convention on the Rights of the Child*. London. https://downloads.unicef.org.uk/wp-content/uploads/2010/05/UNCRC_united_nations_convention_on_the_rights_of_the_child.pdf. Accessed 6 Aug 2020.
- Verhaeghe, G., Vanhoof, J., Valcke, M., & Van Petegem, P. (2010). Using school performance feedback: Perceptions of primary school principals. *School Effectiveness and School Improvement*, 21(2), 167–188. <https://doi.org/10.1080/09243450903396005>.
- Whitty, G., & Wisby, E. (2007). Whose voice? An exploration of the current policy interest in pupil involvement in school decision-making. *International Studies in Sociology of Education*, 17(3), 303–319. <https://doi.org/10.1080/09620210701543957>.
- Wollscheid, S., Stensaker, B., & Bugge, M. (2019). Evidence-informed policy and practice in the field of education: The dilemmas related to organizational design. *European Education*, 51, 270–290. <https://doi.org/10.1080/10564934.2019.1619465>.
- Wyse, D., Brown, C., Oliver, S., & Poblete, X. (2020). *The BERA close-to-practice research project: Research report*. British Educational Research Association. <https://www.bera.ac.uk/publication/bera-statement-on-close-to-practice-research>. Accessed 6 Aug 2020.

Mari-Ana Jones After working as a teacher and school leader for over fifteen years in the UK and Norway, Mari-Ana Jones took up a research position (Ph.D.) at the Norwegian University of Science and Technology (NTNU). Her research interests include student voice and school improvement, with a particular focus on the role of school leadership. Mari-Ana leads the Master's in Educational Leadership at NTNU and is an Academic Supervisor on the National Program for School Leaders in Norway. She is also a consultant on school improvement and leadership in several school districts in Norway.

Valerie Hall has spent almost thirty years in Further and Higher Education in the UK in senior leadership positions, teaching and supervising across Masters and Doctoral level programmes, and in the last decade having a strategic lead in Initial Teacher Education (ITE) in the post-compulsory sector. Her practice and research focus is centred on student voice, and explores individual perspectives, including self-efficacy, self-categorization, and social identities. Valerie currently holds the position of Honorary Research Fellow at the Education Observatory, University of Wolverhampton. She is also an External Academic Advisor to another ITE provider in the UK.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 14

What Can We Learn from Research on Multisource Feedback in Organizations?



John W. Fleenor

Abstract This chapter provides a review of the current state of empirical research on the use of multisource feedback (MSF) in organizations (e.g., Church et al., 2019). The review covers key topics on the research and application of MSF for developing leaders in organizations. The focus of the chapter is on how research on MSF can be applied to the implementation of student feedback to teachers in schools. Based on this research, recommendations are offered for successfully executing student feedback in schools. Topics include: (a) characteristics of effective MSF, (b) how to implement an MSF process in an organization, (c) factors that affect the reliability and validity of MSF, (d) a discussion of agreement between self-ratings and the ratings of others, (e) how to facilitate feedback to leaders, and (f) reasons why MFS processes may fail in organizations. Finally, the transferability of these findings to student-to-teacher feedback in schools is discussed.

Keywords Multisource feedback · Organizational psychology · Leadership development · Student-to-teacher feedback

1 Introduction

In organizations, feedback can have a major impact on the quality of the employees' performance. Therefore, it is important that accurate and relevant feedback is provided to the organization's leaders. Because of the hierarchical structure of organizations, feedback is usually provided only by the leaders' own managers, which is, of course, a limited perspective of the leaders' effectiveness. A valuable tool for delivering such feedback is multisource feedback (MSF; also known as 360-degree feedback). In recent times, the growth of MSF has been a significant trend in the leadership development field. Since its inception in the late 1980s, MSF has gained increasing acceptance and importance in organizations (Silzer & Church, 2009).

J. W. Fleenor (✉)
Center for Creative Leadership, Greensboro, NC, USA
e-mail: fleenorj@ccl.org

In the organizational context, feedback is defined as information provided to employees related to their behavior on the job and the impact of that behavior on others (Fleenor & Taylor, 2019). Feedback is intended to strengthen desired behaviors and to recommend changes for undesired behaviors. Under the correct conditions, feedback can be a catalyst for change (Fleenor et al., 2020).

Most employees want to know how well they are doing their jobs. If they do not receive sufficient feedback, they often seek it on their own (Fleenor et al., 2020). Receiving useful feedback is an important motivational factor that can lead to increased job satisfaction (Bracken & Rose, 2011). Feedback can enhance self-awareness by identifying strengths and can facilitate growth by highlighting areas for improvement (Nowack, 2019).

The impact of multisource feedback can be significant if it is embedded in a larger leadership development process. That is, if it is fully integrated into the human resource management (HRM) system of the organization. Research has found that MSF can improve performance and lead to behavior change over time (Smither et al., 2005; Walker & Smither, 1999). The implementation of MSF has been shown to improve the financial performance of organizations through increased knowledge sharing and employee effectiveness (Kim et al., 2016).

For student-to-teacher feedback in schools, the most relevant counterpart to MSF in organizations is upward feedback. In upward feedback, ratings are solicited from the direct reports of the leader being assessed. This is a relatively common practice in organizations because direct reports are thought to be in the best position to judge a leader's effectiveness. The same could be said for the students of the teachers in schools. However, it would also be possible to conduct the "full circle" of feedback for teachers by including self-ratings and ratings from colleagues and headteachers or principals.

There are parallels between being a leader in an organization and a teacher in a school. For much of the discussion in this chapter, "teacher" could be substituted for "leader" and "student" could be substituted for "rater."

2 Multisource Feedback

The purpose of multisource feedback is to provide accurate and useful feedback related to the effectiveness of leaders in their organizations (Fleenor & Brutus, 2001). This process includes collecting and reporting coworkers' ratings of a leader's effectiveness and providing feedback and coaching for each leader. Traditionally in organizations, feedback has come from a single source, the manager, which provides only a limited perspective of a leader's effectiveness. With MSF, the assessment of a leader's strengths and development needs is more reliable and valid. Because it uses multiple raters, MSF provides different perspectives of performance, making the feedback more accurate and useful to the leader. Additionally, the collection of feedback from several raters with different relationships to the leader will decrease the effects of the biases of the individual raters on the ratings.

There is little agreement in the literature on the terminology used in multisource feedback (Fleenor et al., 2020). In this chapter, the individual being assessed is referred to as the leader. Coworkers who provide the feedback are called raters, and usually include peers and direct reports. The leader's direct boss is referred to as the manager, who also provides feedback. The MSF survey that is completed by the raters is called the assessment. The scales on an MSF assessment represent leadership competencies that are important for success in the organization. MSF is sometimes used with employees who are not leaders; however, in that case, there are no direct-report raters.

2.1 The Multisource Feedback Process

Most MSF processes have the following features (Fleenor & Taylor, 2019):

- Multiple raters (manager, peers, direct reports) provide ratings of the leader's effectiveness using a quantitative rating scale. Leaders also provide self-ratings. The ratings are collected anonymously and reported in the aggregate; therefore, the leader does not know who provided specific ratings. Because most leaders have only one direct manager, the anonymity of the manager's ratings usually cannot be maintained.
- A report is provided to leaders that summarizes the results of their feedback. In a feedback session, leaders identify their strengths and development needs (weaknesses) and examine differences between their own and others' ratings of their effectiveness.
- Based on this feedback, leaders work with feedback coaches (or their managers) to develop an action plan to improve their effectiveness.

Typically, in an MSF process, the leader selects a number of coworkers to participate in the feedback process. Working individually, the raters and the leader complete surveys designed to collect information about the leader's specific skills, behaviors, and other attributes that are important for leader effectiveness. Leader effectiveness is defined as performance that makes leaders successful in their organizations (e.g., the leader's team successfully meets its goals for the year; Fleenor et al., 2020).

After raters complete the surveys, their ratings are electronically sent to a centralized location for scoring. A report is produced and delivered to a feedback coach, who then meets with the leader to review the report. The coach can be an internal human resource (HR) professional or the leader's manager who is trained to interpret the results of the assessment and assist the leader in understanding the report. The coach helps the leader use the feedback to create a plan to address developmental needs identified by the feedback.

Multisource feedback provides a structured means of collecting and processing data, and an opportunity to reflect on this valuable information. It may be the only opportunity some leaders have to consciously self-reflect on their effectiveness. MSF systems also guarantee the anonymity of the raters. There is evidence that anonymous

feedback is more honest than open feedback (Kozlowski et al., 1998). This appears to be particularly true when direct reports are rating their leaders. A climate of trust must be created for the MSF process—when anonymity is ensured, the feedback will be more accurate. If raters believe that anonymity was violated, then less honesty can be expected in future MSF administrations, with a corresponding loss of reliability and validity (London & Wohlers, 1991). Anonymity differs from confidentiality. Confidentiality requires that access to MSF data be limited to individuals who are permitted to see the data in accordance with organizational policy. Confidentiality is important to ensure the participants that their data are protected and will not be seen by unauthorized individuals in the organization. A lack of confidentiality may result in lower participation rates in future MSF administrations.

2.2 Using Multisource Feedback for Leader Development

Because of its structure, thoroughness, and anonymity, MSF is likely to be accepted and acted on by the leaders receiving the feedback (Atwater et al., 2007). To ensure the effectiveness of MSF, it should be implemented within a broader leadership development context. For example, MSF should be integrated into the organization's leader development and succession planning systems to help identify how leaders can become more effective in their organizations. The organization's leadership development system is responsible for providing activities, such as MSF, that will increase the effectiveness of its leaders. The succession planning system is responsible for creating a pipeline of leadership talent for the future. The integration of the leader development and the succession planning systems should create conditions that allow leaders to receive ongoing feedback along with new job assignments, thus increasing their current competencies (McCauley & Brutus, 2019).

Many organizations use MSF as an integral part of development processes for individual leaders. Even when leaders have good insights about their own strengths and development needs, they may not be fully aware of how their behaviors affect their coworkers (Fleenor et al., 2010). After they receive the results of their MSF assessment, leaders have a clearer idea of how their behaviors consistently affect others.

In addition to its use in developing individual leaders, some organizations use aggregated MSF data to determine group strengths and weaknesses for needs analysis purposes. Furthermore, the process of responding to the assessment underscores desired behaviors and creates discussion of which behaviors are valued throughout the organization. This occurs because the items on the MSF assessment indicate what leadership behaviors are considered important by the organization (Bracken & Rotolo, 2019).

2.3 Characteristics of Multisource Feedback

The characteristics of MSF can be thought of as the interactive product of both the assessment and the raters (Bracken & Rose, 2011). According to Bracken and Rotolo (2019), the most important characteristics of MSF are: (a) awareness of the feedback (including reactions and receptivity); (b) acceptance of the feedback; and (c) accountability for acting on the feedback. These characteristics are important for ensuring MSF will result in desired behavior change in the focal leaders (Bracken et al., 2001). Each of these characteristics is discussed below:

2.3.1 Awareness of the Feedback

Awareness involves bringing the information to the attention of the leaders. Thus, they must be aware of the feedback before they can act on it. Awareness of their feedback is required before leaders will recognize their weaknesses and take action to correct them. Awareness of the feedback includes reactions and receptivity to the feedback by the recipients. Reactions can range from being pleased with the feedback to feeling hurt and resentment. A leader's health and psychological well-being may be negatively affected by receiving unfavorable feedback (Nowack, 2019). Feedback coaches play an important role in helping leaders work through any emotional reactions (Fleenor et al., 2020).

Receptivity relates to a leader's psychological readiness to receive the feedback. It is positively related to both, emotional intelligence and perceptions of the feedback environment (Dahling et al., 2012). Additionally, research indicates that feedback orientation, which is the degree to which a leader is ready to receive the feedback, can predict the leader's emotional reactions to their feedback (Braddy et al., 2013).

2.3.2 Acceptance of the Feedback

Acceptance is the leaders' belief that the feedback is an accurate description of their behavior (Ilgen et al., 1979). A key event occurs when the leader decides to accept the feedback as valid and useful information. For the feedback to be accepted, a leader must be aware of and receptive to it. When the feedback is not accepted, no behavior change will result (Bracken & Rose, 2011). First-time MSF participants may experience shock, anger, and rejection of the feedback before finally accepting it (Brett & Atwater, 2001). To ensure acceptance, resources for assisting leaders in dealing with their feedback should be provided by the organization (e.g., coaches, workshops, developmental activities, etc.; Fleenor et al., 2020).

2.3.3 Accountability for Acting on the Feedback

Accountability for acting on the feedback is necessary for a sustainable MSF process. This requires organizations to ensure leaders will conduct improvement-oriented actions on their feedback. Methods for ensuring accountability include the full support of the leader's manager for the MSF process and providing access to developmental resources such as new job assignments and training (London, 2003). Accountability is the major component for moving from acceptance to improved leader effectiveness (Bracken & Rotolo, 2019).

A successful MSF process requires full accountability, not only from the leaders, but also from other groups involved, namely, raters, managers, and the organization (London et al., 1997). If raters believe leaders are not being held accountable for acting on their feedback, they will be less likely to provide effective feedback in future MSF administrations. On the other hand, when raters see their feedback is being used productively, they can be expected to continue to provide accurate, honest feedback (Bracken & Rotolo, 2019).

3 Reliability and Validity of Multisource Feedback

There are a number of factors that affect the validity of an MSF implementation (Bracken et al., 2001). These factors are directly related to the characteristics of a successful MSF process (Bracken & Rotolo, 2019). For MSF, the conceptualization of validity (e.g., content, construct, and criterion-related validity) is more complex than traditional notions of validity that arose from controlled, standardized settings such as intelligence testing. In those settings, validity was determined by a single measurement event in which an individual responds to the items on an assessment (i.e., single-source data). MSF depends on the collection of data from potentially unreliable sources (i.e., multiple raters). It is a complex process with the characteristics of both psychometric testing and large-scale data collection (Fleenor, 2019).

3.1 *Validity Factors in Multisource Feedback*

The primary factors that affect the validity of MSF are described below and summarized in Table 1 with design recommendations from Bracken et al. (2001):

Table 1 MSF Validity Factors with Design Recommendations (Adapted from Bracken et al., 2001)

| Validity factor | Design recommendations |
|-----------------|--|
| Alignment | <ul style="list-style-type: none"> Custom design content Use internal norms Require meeting with raters Align with leader development process |
| Accuracy | <ul style="list-style-type: none"> Capacity to do high volume and secure reporting Processes to ensure zero errors Pre-code important information (e.g., demographics) |
| Clarity | <ul style="list-style-type: none"> Clear instructions and readability Training sessions for providing rating instructions Test understanding of participants |
| Cooperation | <ul style="list-style-type: none"> Keep length reasonable (50 items or fewer) Limit demands on rater (number of surveys) Communicate need for rater cooperation Do on company time |
| Timeliness | <ul style="list-style-type: none"> Do as frequently as is reasonable/needed Train raters to avoid recency error Deliver results as soon as possible |
| Reliability | <ul style="list-style-type: none"> Clear, behavioral, actionable Conduct reliability analyses Use clearly defined anchors Select raters with opportunity to observe Train on proper use of rating scale Report rater groups separately |
| Insight | <ul style="list-style-type: none"> Collect item ratings (not overall competency ratings) Provide as much information as possible to participants Collect write-in comments Require meeting with raters |

3.1.1 Alignment

This is the traditional definition of content validity—the extent to which the feedback (e.g., competencies, behaviors) is important for success in the organization (Bracken & Rotolo, 2019). If the competencies being measured are not related to success, then the content validity of the process is deficient. Alignment occurs when the values and goals of the organization are translated into a set of competencies for the entire organization (Campion et al., 2019).

3.1.2 Accuracy

This includes the process of accurately collecting and processing data, and reporting the feedback. Errors in the feedback reports can negatively affect leaders' confidence in the process. Considerations for increasing accuracy include scoring systems with the capacity to handle high volumes of data with secure reporting, quality control to eliminate errors, and pre-populating of demographic data.

3.1.3 Clarity

Raters must be given instructions on how to correctly complete the assessment and return it on time. Errors typically made by raters include miscoding the person they are rating, misusing the response scale, and providing inappropriate write-in comments. To increase clarity, orientation sessions should be held with the raters to increase their understanding of the process.

3.1.4 Cooperation

The quality of MSF depends on the willingness of the raters to fully participate and provide reliable responses. Design features that affect this factor are related to the magnitude of the task, such as instrument length and the number of surveys a rater must complete. Indicators of low cooperation include unreturned or incomplete surveys and the effects of rater fatigue on the feedback. A simple metric to evaluate cooperation is the overall organization-wide response rate. If less than 75% of the surveys are completed, this should be the reason for concern (Bracken et al., 2001).

3.1.5 Timeliness

Timeliness in providing feedback is an important factor to ensure acceptance of the feedback by the participant. Delays in providing results coupled with recency effects in the ratings may result in feedback that is no longer valid. This can have implications for how effective the feedback is in addressing the needs of the participant and the organization (Bracken et al., 2001).

3.1.6 Reliability

In this context, reliability refers to how dependably or consistently MSF measures the competencies on the assessment. This factor includes the importance of reliability in MSF, how it should be measured, and what level of reliability is acceptable (see Pulakos & Rose, 2019).

Some commonly-used reliability indices may not be appropriate for MSF ratings. For example, test–retest reliabilities may be affected by changes in the raters themselves (e.g., attitudes and opportunity to observe). The raters at Time 1 are often different than the raters at Time 2—with less than a 75% overlap in raters, the results can be misleading. Therefore, it is not recommended that test–retest reliability be used with MSF (Bracken et al., 2001). Internal consistency reliability (e.g., coefficient alpha) provides evidence that items on a scale (i.e., dimension or competency) are internally reliable. Often poorly written items negatively affect the internal consistency reliability of MSF ratings. The use of double and triple-barreled items in an attempt to shorten the length of surveys can also reduce the reliabilities. Overall, low reliabilities can obscure the meaningful interpretation of the feedback (Fleenor, 2019). Other factors that affect the internal consistency of MSF ratings include the misinterpretation of the rating scale by the raters. Typically, 5-to-7 point Likert scales with clearly defined anchors are recommended (Bracken & Rotolo, 2019).

Interrater reliability is used to determine the agreement within rater groups. Moderate levels of interrater reliability within these groups have been reported (e.g., Brett & Atwater, 2001). However, direct report ratings are often found to have the lowest reliabilities (Braddy et al., 2014). To increase the reliabilities within rater groups, all eligible raters should be used, particularly all direct reports. In general, more raters will result in more reliable ratings (Fleenor, 2019).

Typically, the correlation between the ratings from the various rater groups has been found to be low (Tornow, 1993). However, the reason for conducting MSF is to bring different perspectives of a leader’s performance to the process. While the rater groups may disagree, each group may have a valid perspective of a leader’s performance because leaders often interact differently with the various groups. For example, a leader may be interpersonally warm with peers, but cold and distant with direct reports.

3.1.7 Insight

Leaders should be provided with the necessary amount of information needed to take actions that are aligned with their feedback. The format of the assessment and feedback report should be designed to maximize participants’ understanding of their results.

Feedback should be provided at the item level—not just at the competency (i.e., scale) level. With item-level feedback, leaders have a basis for determining the specific behaviors (i.e., the items) that resulted in their ratings. Processes that collect written comments cannot be expected to replace item-level feedback and may even increase the burden on the raters by requiring them to provide detailed descriptions of a leader’s behavior. This may result in “rater fatigue,” especially when raters are required to complete MSF assessments for several employees (Rose et al., 2004).

3.2 Self-other Rating Agreement in Multisource Feedback

Often with MSF, self-ratings are found to differ significantly from the ratings of others (Fleenor et al., 2010). For example, individuals with high self-esteem may over-rate themselves relative to others' ratings of them. For this reason, the use of self-ratings alone is not recommended. However, the level of agreement between self-and others' ratings, can provide important and useful information (Furnham, 2019). There appears to be a relationship between self-other agreement and leader effectiveness. In general, leaders who rate themselves similarly to others (in-agreement raters) appear to be more effective than leaders who rate themselves differently (Fleenor et al., 2010). However, the relationship between self-other rating agreement and leader effectiveness is non-linear. For example, leaders who under-rate themselves appear to be more effective than those who over-rate themselves (Braddy et al., 2014). Similarly, teachers who under-rate themselves are likely to be more effective than teachers who over-rate themselves.

For MSF, the challenge is to develop a relatively simple index of self-other rating agreement that participants can easily understand in their feedback reports. For example, such an index would categorize a leader as an under-rater, in-agreement rater, or over-rater.

4 Recommendations for Facilitating Multisource Feedback

Best practices suggest that a confidential one-on-one feedback session be conducted between the leader and a coach. The coach provides an introduction to the MSF assessment, an analysis of the individual's feedback, and assists with developmental planning. These sessions are particularly important for leaders receiving feedback for the first time. They usually appreciate discussing their feedback with an experienced coach. The coach helps the participant understand that conflicting ratings may be valid, and comparisons between the different rating sources are important (Fleenor et al., 2020).

Leaders must be given adequate time to process their feedback before the one-on-one session. Unfortunately, some organizations distribute the reports and allow the leaders only a few minutes to look over their results prior their feedback session. Without time to reflect on their report and process any immediate emotional reactions to the data, leaders may not be ready to fully accept the feedback.

The coach should prepare for the session in advance by thoroughly reviewing the feedback report. The session should be held in a private room and leaders should be given the opportunity to audio record their session, which will serve as a useful resource for participants to review progress on their development plans.

5 Why Multisource Feedback Processes Fail

While best practices are fully documented in the literature (Fleenor et al., 2020), practitioners continue to struggle with implementation issues with MSF. These issues affect the quality of the feedback (e.g., validity, reliability, accuracy) and therefore the future success of MSF processes. Many of these issues can be avoided by careful design, planning, and follow-up. A number of problems are common to failed MSF implementations (Fleenor et al., 2020):

1. **Unclear Purpose** When the business reasons for conducting MSF are unclear or key stakeholders disagree on its purpose, the process is likely to fail. Organizations need to consider how their business goals align with the goals of the MSF implementation (Campion et al., 2019). The purpose of the process should be clearly defined, and an appropriate MSF assessment selected for that purpose.
2. **Lack of Organizational Readiness** A supportive organization culture is critical to the success of an MSF process. There must be full senior management buy in and public support. All senior leaders should participate fully in the MSF process. Further, a high level of trust is needed among raters so that the feedback they provide will be used constructively by the leader and the organization (Smith & Fortunato, 2008).
3. **Selecting the Wrong MSF Assessment** The organization should have an underlying leadership competency model indicating what is important for success in their organizational context. If the purpose of MSF is to measure competencies specific to an organization (rather than general leader competencies), then a customized assessment will be needed that directly measures these competencies (Conger, 2019).
4. **Poor Design and Logistics** Reasons MSF processes fail often include inadequate planning and poorly implemented logistics. For example, if MSF is administered during an extremely busy time in the organization (e.g., during the budgeting cycle), it may result in lower participation levels. A thorough communication plan is critical, particularly for those directly involved in the process, including leaders, their managers, and all other raters. Some organizations try to compress the MSF process into an unrealistic timeframe, which results in poorly implemented processes.
5. **Leader Preparation** An appropriate amount of preparation for leaders is critical. They need to be informed why they are participating, how the process works (e.g., rater selection), and the level of confidentiality and anonymity they can expect. Raters need to be told that their input is important, and their ratings will be strictly anonymous.
6. **Poor Rater Selection** Employees who provide the most accurate ratings are those who interact with the leader on a frequent basis. This allows enough time to observe the behaviors they are rating. For most leaders, the best raters are the coworkers with whom they have frequent face-to-face interactions (Bracken & Rotolo, 2019). Selecting raters who are not fully aware of a leader's behaviors will result in less valid feedback.

7. **Post-Assessment Problems** Some issues do not become problems until after the MSF assessment has been completed (Bracken et al., 2001). A common issue is the lack of clear expectations of what leaders are responsible for doing after they receive the feedback. They should meet with their managers to discuss their feedback, create a development plan and decide on the next steps. When this is not accomplished, leaders are less likely to be accountable for acting on their feedback.
8. **Confidentiality and Anonymity Issues** Confidentiality and anonymity are critical issues in the MSF process (Macey & Barbara, 2019). There can be serious issues if rater anonymity is compromised during the process. Raters are more likely to provide valid ratings when they know their individual ratings will remain anonymous and confidential.
9. **Failure to Evaluate the MSF Process** As with any leader development process, it is important to assess the impact of MSF. An evaluation should include interviews, surveys, or focus groups with the participants to determine how the organization can improve its MSF process to increase its impact.

6 The Transferability of Multisource Feedback Research to Student-to-Teacher Feedback in Schools

There seems to be considerable overlap between leading in organizations and teaching in schools. Schools themselves are, of course, organizations with particular hierarchies and cultures relevant to the educational context. Teachers could be considered as the “leaders” of the students, and the students as the “direct reports” of the teachers. As discussed previously, the most relevant organizational counterpart to student-to-teacher feedback is upward feedback. In upward feedback, ratings are solicited from the direct reports of the leader being assessed because they are thought to be in the best position to judge the leader’s effectiveness. In the same vein, students are probably in the best position to judge a teachers’ effectiveness. The upward feedback model could be expanded to include self-ratings and the ratings of others (peers and leaders such as lead teachers or principals). This would result in full “360-degree” feedback, which may provide more valid and reliable feedback for teachers.

For teachers, MSF would provide access to structured feedback from students on their teaching quality, a source of feedback they rarely receive. Given that students have a unique perspective of teacher effectiveness, such feedback could be helpful for teachers who want to improve their teaching effectiveness. Administering MSF in the classroom has an advantage over the typical organization because teachers are likely to have mostly the same students during the school year. In organizations, a leader’s direct reports may change frequently because of reorganizations, reassignments, and turnover. Because of the relatively stable rater population, teachers can be more easily evaluated over time to determine how much they have improved.

If administered as a full 360-degree feedback model, MSF would provide teachers with multiple perspectives of their performance. They would be able to compare

their ratings from various sources (students, peers, principals) to determine if they are perceived differently by these groups. Teachers would be able to compare their self-ratings to the ratings of others to see if they have an unrealistic view of their own performance.

Following the feedback facilitation model typically used in organizations, lead teachers or principals could act as coaches for the teachers and assist them with digesting their feedback and developing plans for acting on the feedback. Teachers would create development plans for improving in areas of weaknesses and leveraging their strengths. Additionally, the use of a validated teacher competency model would inform them of what capabilities are needed to be an effective teacher. In schools, MSF should be administered on a regular basis so the teachers' performance can be evaluated over time.

There is some evidence, however, that providing MSF alone may not result in sustained behavior change for teachers. For example, Bijlsma et al. (2019) found that teachers did not improve their teaching quality in response to student feedback received via smartphones. As a result of the student feedback, however, the teachers did gain more insight into how they could improve and reported improvement-oriented efforts in response to the feedback. For student feedback to be more effective in creating sustained behavior change in teachers, Bijlsma et al. recommend that:

- The teachers have a strong improvement motivation and are willing to step out of their comfort zones and search for their weaknesses.
- Definitions of desired behaviors, improvement goals, and developmental activities are clearly defined.
- A coach is provided who understands effective teaching behaviors (e.g. quality classroom management), how these behaviors can be developed, and practices that are effective if problems arise during the development process.

The above recommendations align closely with the characteristics of effective MSF in organizations. The basic tenets of MSF in organizations, therefore, should also be applied to student-to-teacher feedback:

- It should not be implemented as a stand-alone event. In addition to the assessment, there must be developmental planning and follow-up. A plan should be created that details recommendations to help the teacher improve based on the feedback.
- The feedback assessment must reflect competencies that are important for teacher effectiveness. A fully validated teacher competency model should be used.
- The support of the top leadership of the school is critical for persuading teachers to set specific development goals. Teachers must be held accountable for acting on their feedback.
- A flawed feedback process can be fatal to future administrations. The anonymity of the students' ratings and the confidentiality of the teachers' feedback reports must be strictly maintained. Students must be convinced that their teachers will not see their individual ratings. This means that the feedback should be aggregated for each class before presenting it to the teachers.

- The timing of the feedback process should consider organizational realities that could reduce its impact. For example, the process should not be implemented during end-of-semester grading periods or other times when teachers are not able to fully focus on their feedback.
- Students should be trained to be aware that only constructive feedback will help their teachers improve. MSF is not a way of venting frustration, but meant to catalyze a learning process that will benefit both students and teachers.

In summary, if the recommendations discussed in this chapter are closely followed, then student–teacher feedback is more likely to be successful in schools.

References

- Atwater, L. E., Brett, J. F., & Charles, A. C. (2007). Multisource feedback: Lessons learned and implications for practice. *Human Resource Management, 46*, 285–307. <https://doi.org/10.1002/hrm.20161>.
- Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education, 28*(217), 236. <https://doi.org/10.1080/1475939x.2019.1572534>.
- Bracken, D. W., & Rose, D. S. (2011). When does 360 degree feedback create behavior change? And how would we know it when it does? *Journal of Business and Psychology, 26*, 183–192. <https://doi.org/10.1017/iop.2016.93>.
- Bracken, D. W., & Rotolo, C. T. (2019). Can we improve rater performance? In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback* (pp. 255–290). Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Bracken, D. W., Timmreck, C. W., Fleenor, J. W., & Summers, L. (2001). 360 feedback from another angle. *Human Resource Management, 40*, 3–20. <https://doi.org/10.1002/hrm.4012>.
- Braddy, P. W., Gooty, J., Fleenor, J. W., & Yammarino, F. J. (2014). Leader behaviors and career derailment potential: A multi-analytic method examination of rating source and self-other agreement. *Leadership Quarterly, 25*, 373–390. <https://doi.org/10.1016/j.leaqua.2013.10.001>.
- Braddy, P. W., Sturm, R. E., Atwater, L. E., Smither, J. W., & Fleenor, J. W. (2013). Validating the feedback orientation scale in a leadership development context. *Group and Organization Management, 38*, 690–716. <https://doi.org/10.1177/1059601113508432>.
- Brett, J., & Atwater, L. (2001). 360° feedback: Accuracy, reactions and perceptions of usefulness. *Journal of Applied Psychology, 86*, 930–942. <https://doi.org/10.1037/0021-9010.86.5.930>.
- Campion, E. D., Campion, M. A., & Campion, M. C. (2019). Best practices when using 360 degree feedback for performance appraisal. In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback* (pp. 19–60). Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Church, A. H., Bracken, D. W., Fleenor, J. W. & Rose, D. S. (Eds.). (2019). *The handbook of strategic 360 feedback*. Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Conger, J. A. (2019). Harnessing the potential of 360 feedback in executive education programming. In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback* (pp. 343–351). Oxford University Press.
- Dahling, J. J., Chau, S. L., & O'Malley, A. (2012). Correlates and consequences of feedback orientation in organizations. *Journal of Management, 38*, 531–546. <https://doi.org/10.1177/0149206310375467>.

- Fleenor, J. W. (2019). Factors affecting the validity of 360 feedback processes. In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback* (pp. 237–254). Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Fleenor, J. W., & Brutus, S. (2001). Multisource feedback for personnel decisions. In D. Bracken, C. Timmreck, & A. Church (Eds.), *The handbook of multisource feedback* (pp. 335–351). Jossey-Bass.
- Fleenor, J. W., Smither, J. W., Atwater, L., Braddy, P. W., & Sturm, R. (2010). Self-other rating agreement in leadership: A review. *Leadership Quarterly*, *21*, 1005–1034. <https://doi.org/10.1016/j.leaqua.2010.10.006>.
- Fleenor, J. W., & Taylor, S. (2019). Developing leadership potential through 360-degree feedback and coaching. In L. A. Berger & D. R. Berger (Eds.), *The talent management handbook* (3rd ed., pp. 201–209). McGraw-Hill.
- Fleenor, J. W., Taylor, S., & Chappelow, C. (2020). *Leveraging the impact of 360-degree feedback* (2nd ed.). Berrett-Koehler.
- Furnham, A. (2019). Rater congruency: Why ratings of the same person differ. In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback* (pp. 291–308). Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Ilgen, D. R., Fisher, C. D., & Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, *64*, 349–371. <https://doi.org/10.1037/0021-9010.64.4.349>.
- Kim, K. Y., Atwater, L., Patel, P. C., & Smither, J. W. (2016). Multisource feedback, human capital, and the financial performance of organizations. *Journal of Applied Psychology*, *101*, 1569–1584. <https://doi.org/10.1037/apl0000125>.
- Kozlowski, S., Chao, G., & Morrison, R. (1998). Games raters play: Politics, strategies, and impression management in performance appraisal. In J. W. Smither (Ed.), *Performance appraisal: State of the art in practice* (pp. 163–205). Jossey-Bass.
- London, M. (2003). *Job feedback: Giving, seeking, and using feedback for performance improvement* (2nd ed.). Lawrence Erlbaum. <https://doi.org/10.4324/9781410608871>.
- London, M., Smither, J. W., & Adsit, D. L. (1997). Accountability: The Achilles' heel of multisource feedback. *Group and Organization Management*, *22*(162), 184. <https://doi.org/10.1177/1059601197222003>.
- London, M., & Wohlers, A. J. (1991). Agreement between subordinate and self-ratings in upward feedback. *Personnel Psychology*, *44*, 375–390. <https://doi.org/10.1111/j.1744-6570.1991.tb00964.x>.
- Macey, W. H., & Barbara, K. M. (2019). The ethical context of 360 feedback. In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback* (pp. 461–478). Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- McCauley, C., & Brutus, S. (2019). Application of 360 feedback for leadership development. In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback*. Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Nowack, K. M. (2019). From insight to successful behavior change: The real impact of development-focused 360 feedback. In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback*. Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Pulakos, E. D., & Rose, D. R. (2019). Is 360 a predictor or criterion measure? In A. H. Church, D. W. Bracken, J. W. Fleenor, & D. S. Rose (Eds.), *The handbook of strategic 360 feedback*. Oxford University Press. <https://doi.org/10.1093/oso/9780190879860.001.0001>.
- Rose, D. S., Farrell, T., & Robinson, G. N. (2004). *Are narrative comments in 360° feedback useful or useless?* Research report. 3D Group.
- Silzer, R., & Church, A. H. (2009). The pearls and perils of identifying potential. *Industrial and Organizational Psychology*, *2*, 377–412. <https://doi.org/10.1111/j.1754-9434.2009.01163.x>.

- Smith, F. R., & Fortunato, V. J. (2008). Factors influencing employee intentions to provide honest upward feedback ratings. *Journal of Business and Psychology, 24*, 191–207. <https://doi.org/10.1007/s10869-008-9070-4>.
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multi-source feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology, 58*, 33–66. https://doi.org/10.1111/j.1744-6570.2005.514_1.x.
- Tornow, W. W. (Ed.) (1993). Special issue on 360-degree feedback. *Human Resource Management, 32*, 2–3. <https://doi.org/10.1002/hrm.3930320202>.
- Walker, A. G., & Smither, J. W. (1999). A five-year study of upward feedback: What managers do with their results matters. *Personnel Psychology, 52*, 393–423. <https://doi.org/10.1111/j.1744-6570.1999.tb00166.x>.

John W. Fleenor is a Senior Fellow at the Center for Creative Leadership in Greensboro, NC (USA). Additionally, he is an Adjunct Associate Professor of Psychology at North Carolina State University. For over 20 years he has conducted research on measures of leadership, focusing on multi-rater (360-degree) feedback. He has published extensively on this topic, including six books. He is a Fellow of the Society for Industrial and Organizational Psychology (SIOP).

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 15

Lessons Learned from Research on Student Evaluation of Teaching in Higher Education



Bob Uttl

Abstract In higher education, anonymous student evaluation of teaching (SET) ratings are used to measure faculty's teaching effectiveness and to make high-stakes decisions about hiring, firing, promotion, merit pay, and teaching awards. SET have many desirable properties: SET are quick and cheap to collect, SET means and standard deviations give aura of precision and scientific validity, and SET provide tangible seemingly objective numbers for both high-stake decisions and public accountability purposes. Unfortunately, SET as a measure of teaching effectiveness are fatally flawed. First, experts cannot agree what effective teaching is. They only agree that effective teaching ought to result in learning. Second, SET do not measure faculty's teaching effectiveness as students do not learn more from more highly rated professors. Third, SET depend on many teaching effectiveness irrelevant factors (TEIFs) not attributable to the professor (e.g., students' intelligence, students' prior knowledge, class size, subject). Fourth, SET are influenced by student preference factors (SPFs) whose consideration violates human rights legislation (e.g., ethnicity, accent). Fifth, SET are easily manipulated by chocolates, course easiness, and other incentives. However, student ratings of professors can be used for very limited purposes such as formative feedback and raising alarm about ineffective teaching practices.

Keywords Student evaluation of teaching · SET · Validity · Teaching effectiveness

1 Introduction

In higher education, anonymous student evaluation of teaching (SET) are used to measure the teaching effectiveness of faculty members and to make high-stakes decisions about them, such as hiring, firing, promotion, tenure, merit pay, and teaching awards (Uttl et al., 2017). If available to students, they are also used by students for course selection in the same manner as the popular website www.ratemyprofessor.com (RMP). SET have their allure: (a) SET are quick and cheap to administer; (b)

B. Uttl (✉)
Mount Royal University, Calgary, AB, Canada
e-mail: buttl@mtroyal.ca

SET means and standard deviations give an aura of precision and scientific validity; and (c) SET provide tangible seemingly objective numbers for high-stake decisions and public accountability purposes. However, a still little known legal case from Ryerson University in Toronto (*Ryerson University v. Ryerson Faculty Association*, 2018 CanLII 58446, available at www.canlii.org) is a wake-up call about the uninformed use of SET, and reminder that SET are not valid as a measure of faculty's teaching effectiveness. In this chapter, I review the evidence against SET, evidence showing that they do not measure teaching effectiveness, vary predictably across factors completely irrelevant to faculty's teaching effectiveness, and can be raised with something as small as a Hershey kiss. I will also argue that the widespread use of SET may be one of the main contributors to grade inflation, driving up grades over the past 30 years, during a time period when time-spent studying has been steadily decreasing and the proportion of high school students entering colleges and universities increasing.

Typically, within the last few weeks of classes, students are asked to rate professors on various scales. A university evaluation unit then summarizes the ratings for each class and, after the classes are over and the final grades assigned, various statistical summaries including means and standard deviations are then provided to faculty and their administrators. These summaries may include departmental, faculty, or university "norms," such as the means and standard deviations of all course means within the department, faculty, and/or university. These summaries are then used as the key, if not sole, evidence of faculty teaching effectiveness (Uttl et al., 2017).

At the same time, no standards for satisfactory SET ratings are provided to anyone. Evaluators—chairs, deans, tenure and promotion committees, provosts, and presidents—use their own individual standards to arrive at their decisions about faculty teaching effectiveness. It is not uncommon for these evaluators to believe that faculty members falling below the mean are unsatisfactory and in need of improving their teaching. Moreover, these evaluators change periodically and unpredictably, even within the typical six-year time frame between a faculty member's initial hiring and eventual decision about promotion and/or tenure.

There are three types of commonly-used SET tools—those that are developed in-house by an institution, those that are obtained for free, such as the SEEQ (Marsh, 1980, 1991), and those that are developed commercially for purchase, such as the ETS SIR-II sold by the Education Testing Service (www.ets.org), the IDEA SRI sold by IDEA Center (www.ideaedu.org), and the CIEQ sold by C.O.D.E.S Inc. (www.cieq.com). In all of these systems, faculty's SET ratings are often compared to the departmental, faculty, university, or "national norms" (i.e., the average SET ratings for all institutions that purchased a particular commercial SET system). The commercial systems also give some guidelines on interpretation of SET. For example, the C.O.D.E.S. Inc guidelines specify that faculty scoring below the 70th percentile need at least "some improvement," implying that only the top 30% of faculty with the highest SET ratings are good enough and need "no improvement" (see www.cieq.com/faq). Notably, all of the commercial systems are explicitly intended to be used for both faculty development (formative uses) and for high-stakes personnel

decisions (sumative uses) and their developers believe that they are valid measures of teaching effectiveness.

The focus on norm-referenced interpretation of SET ratings, requiring faculty to place above the 30th, 50th, or even 70th percentile, to avoid criticism of their teaching, will always, by definition, result in large proportions of unsatisfactory and in “need of at least some improvement” faculty members. Assuming few faculty members want to be labeled unsatisfactory or in “need of at least some improvement”, this type of norm-referenced interpretation of SET sets up and fuels a race among faculty members to reach as high of ratings as possible. By definition, depending on the specific percentile cut-offs, 30, 50, or 70% of the faculty will lose this race. The higher the percentile cut off, the more intense and more high-stakes the race becomes.

Regardless of the specific percentile cut-offs for “unsatisfactory” or in “need of some improvement” labels, some proponents of SET ratings also argue that SET identify faculty members who successfully match their academic standards, teaching demands, and workload to students’ abilities. For example, in response to arguments that SET are responsible for grade inflation and work deflation, Abrami and d’Apollonia (1999) argued:

academic standards that are too high may be as detrimental to the learning of students as academic standards that are too low. The arts and science of good teaching is finding the balance between what students might learn and what students are capable of learning. We believe that ratings help identify those instructors who do this well. (p. 520)

As Uttl et al. (2017) observed, in Abrami and d’Apollonia’s (1999) view, SET are an appropriate standards meter allowing professors to determine what students’ perceive to be an appropriate workload, appropriate amount to learn for specific grades, and, in short, a proxy of appropriate academic standards from the students’ perspective. Professors who get high SET ratings are appropriately matching their standards to students’ standards and professors who get low SET ratings are failing to do so.

2 SET Are an Invalid Measure of Faculty Teaching Effectiveness

Are SET a valid measure of faculty teaching effectiveness? Do students learn more from more highly rated professors? If SET are a valid measure of faculty’s teaching effectiveness, SET ought to strongly correlate with student achievement attributable to the professors’ teaching styles, and ought not to be influenced by teaching effectiveness irrelevant factors (TEIFs) such as students’ intelligence, cognitive ability, prior knowledge, motivation, interest, subject field, class size, class meeting time, etc. SET also ought not to be influenced by certain students preference factors (SPFs) such as professors’ hotness/attractiveness, age, gender, accent, nationality, ethnicity, race, disability, etc., whose consideration runs afoul to human rights legislation. Finally, SET ought not to be influenced by ill-advised or detrimental to student learning factors

(DSLFS) such as professors reducing workloads, inflating grades, and distributing chocolates and cookies. Review of the literature, however, now convincingly shows that SETs are not a valid measure of teaching effectiveness, that students do not learn more from more highly rated professors and that SET are substantially influenced by numerous TEIFs, SPFs, and DSLFS.

2.1 There Is No Widely Accepted Definition of Effective Teaching

The first fundamental problem in assessing the validity of SET as a measure of faculty teaching effectiveness is that professors, administrators, and even experts do not agree on what effective teaching is (Uttl et al., 2017). In turn, experts do not even agree on which teaching methods are effective and which specific teaching behaviors amount to effective teaching. For example, some professors, administrators, and experts believe that teaching methods such as unannounced pop quizzes, questioning students in front of their peers, and encouraging student attendance by leaving out words or phrases from lecture slides are effective teaching methods. In contrast, others believe that these same methods are insensitive, anxiety-producing, and even demeaning, disrespectful, and detrimental to student learning.

In the absence of an agreed upon definition, it is impossible to measure effective teaching. However, the experts do agree that effective teaching ought to result in student learning (Uttl et al., 2017). Accordingly, studies attempting to establish the validity of SET as a measure of effective teaching have focused on determining the correlation between professors' mean class SET ratings and student achievement.

2.2 Students Do Not Learn More from More Highly Rated Professors

For nearly 40 years, the key evidence cited to support the validity of SET as a measure of faculty teaching effectiveness have been multisection studies that examine the correlations between the mean class SET and the mean class student achievement on common exams. An ideal multisection study has several critical features: (a) it examines the correlation between SET and student achievement in a large course split into numerous smaller sections, with each section taught by a different professor, (b) professors follow the same course outline, use the same assessments, and the same final exam, (c) students are randomly assigned to the sections, and (d) SET are administered prior to the final exam at the same time to all sections. In this design, if students learn more from more highly rated professors, the sections' average SET ratings ought to be highly correlated with sections' average final exam scores. Experts have generally agreed that multisection studies are the strongest evidence for

determining the validity of SET as a measure of professors' teaching effectiveness, that is, professors' contribution to students' learning (Uttl et al., 2017).

Cohen (1981) published the first meta-analysis of 67 multisection studies available to that date and reported a small-to-moderate SET/learning correlation $r = 0.43$. Cohen concluded: "The results of the meta-analysis provide strong support for the validity of student ratings as a measure of teaching effectiveness" (p. 281) and continued: "we can safely say that student ratings of instructors are a valid index of instructional effectiveness. Students do a pretty good job of distinguishing among teachers on the basis of how much they have learned" (p. 305). Cohen's findings and conclusions have subsequently been cited over 1,000 times as evidence of SET validity as a measure of faculty teaching effectiveness (Web of Science, Google Scholar).

However, Uttl et al. (2017) recently demonstrated that Cohen's (1981) conclusions were unwarranted, and the result of flawed methods and data analyses. Most critically, Cohen disregarded the sample sizes of primary studies in his meta-analysis. In doing so, he gave equal weight to many small sample sized studies as he gave to fewer larger sample sized studies. Compounding this problem, Cohen also failed to take into account small sample size bias clearly visible from scatterplots of SET/learning correlations as a function of sample size. After taking into account small sample size bias, the best estimate of SET/learning correlation was only $r = 0.27$, substantially less than $r = 0.43$ reported by Cohen. Uttl et al. (2017) reported a new updated analysis of 97 multisection studies. Figure 1, Panel A, shows the results of Uttl et al.'s new updated meta-analysis based on 97 multisection studies. It confirms the strong small sample size bias already visible in Cohen's (1981) data set. Taking into account the small sample size bias, the best estimate of SET/learning correlation from this new meta-analysis is $r = 0.08$. Panel B shows the Uttl et al. results but only for studies that adjusted the SET/learning correlations for prior learning/ability. The best estimate of SET/learning correlations taking into account both the small sample size bias and prior learning/ability is nearly zero, $r = -0.02$. Accordingly, taking into account small sample size bias and prior learning/ability, the multisection studies demonstrate that SET/learning correlations are zero. In other words, students do not learn more from more highly rated professors.

2.3 SET Are Influenced by Many Teaching Effectiveness Irrelevant Factors

SET correlate with numerous TEIFs such as students' intelligence, cognitive ability, interest, and motivation; subject field; class size; etc.

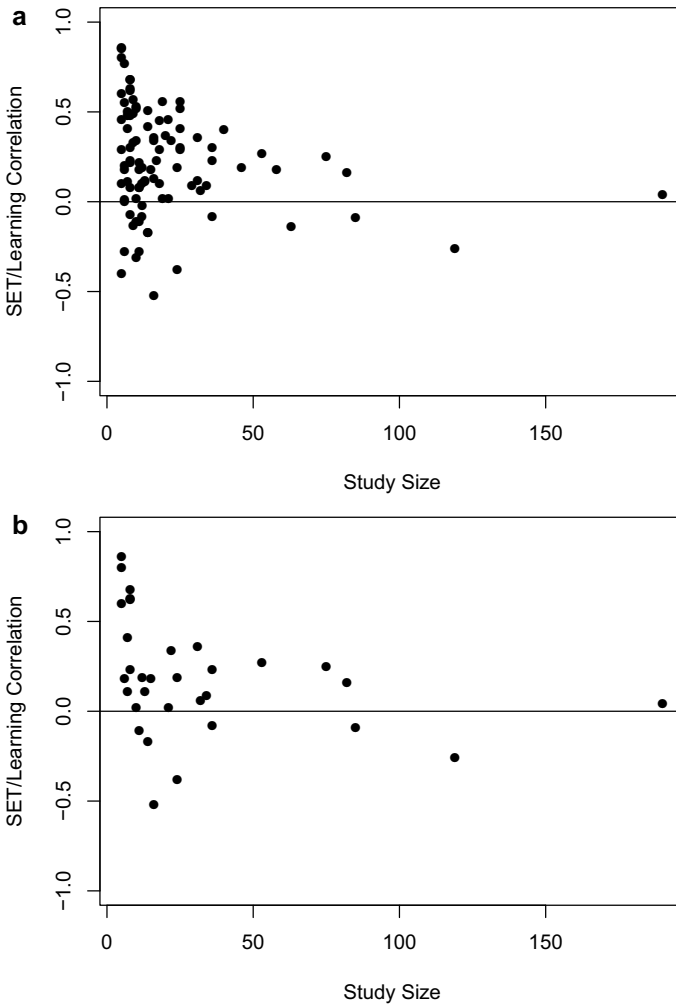


Fig. 1 The results of meta-analyses of multisection studies. Panel **A** shows the scatterplot of SET/learning correlations by study size for Uttl et al.'s (2017) new updated meta-analysis. After taking into account a small sample bias, the SET/learning correlation was only $r = 0.08$ for SET averages. Panel **B** shows Uttl et al. (2017) results but only for studies that adjusted the SET/learning correlations for prior learning/ability. After taking into account both the small sample size bias and prior/learning ability, the SET/learning correlation is nearly zero, $r = -0.02$

2.3.1 Students Intelligence, Ability, and Kruger Dunning Effect

Numerous studies have demonstrated that people are generally very poor in assessing their own cognitive abilities including attention, learning, and memory. Correlations between self-assessment of abilities and performance on objective tests of those

abilities are generally close to zero (Uttl & Kibreab, 2011; Williams et al., 2017). Yet, many SET forms ask students to rate how much they learned from their professors.

Furthermore, as Kruger and Dunning (1999) demonstrated, people's self-assessment of their abilities depends on the abilities themselves. Those scoring low on objective ability tests hugely overestimated their performance whereas those scoring high on objective ability tests tended to underestimate their own performance. Moreover, low-ability individuals were less able to distinguish superior performance from inferior performance of their peers. As Kruger and Dunning observed, the incompetent are not only incompetent but their incompetence deprives them of the ability to recognize their own incompetence as well as the competence of others. It is self-evident that students who believe that their work deserved As or Bs but received Ds or Fs are unlikely to be satisfied and unlikely to give their professors high SET ratings.

2.3.2 Student Interest and Motivation

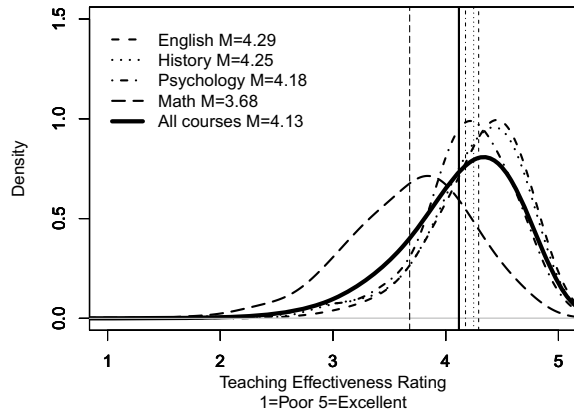
Hoyt and Lee (2002) reported SET ratings by student motivation and class size for the 20 items of the IDEA SRI. Student motivation was measured by a question "I really wanted to take this course regardless of who taught it." Collapsed across questions and class size, the least motivated students gave SET ratings that were 0.44 lower than those of the most motivated students, corresponding to an approximately 0.75 standard deviation difference. Moreover, this effect was substantial on each and every question, ranging from a 0.24 to 0.70 difference on a 1–5 rating scale.

2.3.3 Course Subject

Centra (2009) reported that the natural sciences, mathematics, engineering, and computer science courses were rated substantially lower, about 0.30 standard deviation lower, than courses in humanities such as English, history, and languages. Similarly, Beran and Violato (2009) reported that courses in natural science were rated 0.61 standard deviation lower than courses in social science. Surprisingly, Centra as well as Beran and Violato concluded that these effects were ignorable.

Using 14,872 course evaluation data from a US mid-sized university, Uttl and Smibert (2017) demonstrated that the differences in SET ratings between subjects such as English and Math are substantial (the difference between the means was 0.61 on a 5-point scale), and that professors teaching quantitative courses are far more likely to be labeled unsatisfactory when evaluated against common criteria for a satisfactory label. Figure 2 shows the distribution of SET ratings for Math, English, Psychology, History, and all courses. The distributions of math professors ratings are more normal and substantially shifted toward less than excellent ratings whereas the distribution of English, history, psychology, and all professor courses professors ratings are higher and positively skewed. Thus, if the same standards are applied to professors teaching quantitative vs. non-quantitative courses, professors teaching

Fig. 2 Smoothed density distribution of overall mean ratings for all courses and for courses in selected subjects. Figure highlights that math professors received much lower ratings than professors in English, History, Psychology, and all courses (from Uttl & Smibert, 2017, Fig. 1) (a smoothed density distribution can be thought of as a smoothed histogram with area below the curve equal to 1)



quantitative courses are far more likely to be not hired, fired, not re-appointed, not promoted, not tenured, denied merit pay, and denied teaching awards.

Of course, the simple fact that professors teaching quantitative vs. non-quantitative courses receive lower SET ratings is not evidence that SETs are biased. It may be that professors teaching quantitative vs. non-quantitative courses are simply incompetent, less effective teachers. However, as pointed out by Uttl and Smibert (2017) this incompetence explanation is unlikely. A wealth of evidence strongly suggests that the lower ratings of professors teaching quantitative vs. non-quantitative courses is due to factors unrelated to professors themselves. First, the mathematical knowledge and numeracy abilities of populations worldwide have decreased over the years. For example, half of Canadians now score below the level required to fully participate in today's society (Orpwood & Brown, 2015). Second, Uttl et al. (2013) found that fewer than 10 out of 340 undergraduate students were "very interested" in taking any one of the three statistics courses offered in the psychology department at Mount Royal University. In contrast, 159 out of 340 were "very interested" in taking the Introduction to the Psychology of Abnormal Behavior. Thus, professors teaching statistics classes vs. abnormal psychology are facing students who differ vastly on one of the best predictors of student learning: interest in the subject.

2.3.4 Class Size

Armchair theorizing suggests that class size (i.e., the number of enrolled students) ought to be inversely related to SET ratings. Small classes, with 10, 20, or even 30 students, allow each student to have a far greater opportunity to interact with their professors. In contrast, in classes beyond 20 or 30 students, professors are unlikely to learn even student names. Surprisingly, in the first meta-analysis of SET/class size relationship, Feldman (1984) concluded that the average SET/class size correlation was only $r = -0.09$ (corresponding to $d = -0.18$). Fifteen years later, Aleamoni (1999) summarily declared the notion that class size can affect student ratings to be

a myth. Another 10 years later, Gravestock and Gregor-Greenleaf (2008) concluded that “the correlation between class size and ratings is statistically insignificant and is therefore not viewed as having any impact on validity.”

Our review (Uttl et al., 2018) of over 100 studies that examined the relationship between SET and class size, including those reviewed by Feldman (1984), revealed that the vast majority of these studies did not report sufficient information to interpret their findings. For example, many studies did not report the smallest class size, did not report the largest class size, did not report the number of classes within each class size category, did not examine the linearity of SET/class size relationship, did not examine whether there was a decline in SET for classes with fewer than 20 or 30 students, did not show scatterplots of SET/class size relationships, had very small sample sizes, included extreme outliers, etc.

When only studies that reported sufficient data to plot the relationship between SET and class size and examined, the decline in SET is initially steep and then levels off for class sizes between 30 and 50 students. The overall decline is about 0.5 point on 1–5 rating scale. When each study’s data are standardized using the smallest class size group in each study as a reference group and the average standard deviation of SET means within each study, the declines in SET ratings to class sizes up to 30 or 50 students amount to about 0.5 standard deviation and that the declines continue even thereafter but at a much lower rate. Accordingly, disregarding uninterpretable studies, the evidence clearly shows that declines in SET ratings are steep as class size increases to 30–50 students, and that SET declines level off thereafter.

3 SET Are Influenced by Student Preference Factors (SPFs) Whose Consideration Violates Human Rights Legislation

A substantial body of research has also reported that SET are influenced by factors whose consideration in high-stakes personnel decisions violates human rights legislation such as professors accent, nationality, ethnicity, race, age, gender, etc.

3.1 Attractiveness/Hotness

Do students prefer attractive/hot young professors to unattractive/not so hot professors? Using the www.ratemyprofessor.com rating data for 6,852 US faculty, Felton et al. (2008) found that Quality (average of Clarity and Helpfulness ratings) was strongly correlated with instructor Hotness (www.ratemyprofessor.com discontinued Hotness scale in 2018 in response to a social media campaign against it), $r = 0.64$. Hotness was similarly correlated with Helpfulness, $r = 0.64$, and Clarity, $r = 0.60$, and only moderately correlated with Easiness, $r = 0.39$. Accordingly, attractive/hot

professors receive much higher ratings on Clarity, Helpfulness as well as Easiness. One may argue that www.ratemyprofessor.com is low quality data, unlike carefully designed SET. However, this argument fails for two reasons: First, www.ratemyprofessor.com Overall Quality ratings correlate highly with in class instructor SET ratings with r s ranging from 0.66 to 0.69 (Coladarci & Kornfield, 2007; Sonntag et al., 2009; Timmerman, 2008). Second, www.ratemyprofessor.com ratings are affected by various TEIFs, SPFs, and IDSLFs just as SETs are.

3.2 *Accent/Ethnicity/Nationality*

In one of the most extensive studies, Subtirelu (2015) examined the [ratemyprofessor.com](http://www.ratemyprofessor.com) ratings of 2,192 professors with US last names vs. professors with Chinese or Korean last names teaching in the USA. Subtirelu found that professors with US last names received ratings 0.60–0.80 points higher (on 5-point scale) on Clarity and 0.16–0.40 points higher on Helpfulness.

3.3 *Gender*

Hundreds of studies have examined gender differences in SET ratings. In general, gender differences in SET ratings are (a) minimal and (b) inconsistent. Moreover, most of the research has compared SET ratings of men vs. women within the university, faculty, or department. However, these studies are impossible to interpret because presence or absence of gender differences does not indicate the presence or absence of gender bias. Gender differences could arise, be reduced, or even masked by a number of different factors including but not limited to gender differences in teaching ability, gender differences in ability to satisfy students, gender differences in courses taught by men vs. women (quantitative vs. non-quantitative, nursing vs. computer science), and gender differences in ability to bake tasty treats for students (see below). However, three recent studies have claimed to show a large bias against female professors and have been widely cited for this claim: Boring (2015, 2017), MacNeill et al. (2015), and Mitchell and Martin (2018). However, a detailed review of these studies does not support their authors' conclusions as detailed below.

Boring (2015, 2017) examined gender differences in SET ratings using a French university's SET ratings of 372 fixed contract instructors teaching seminar sections of introductory courses. Boring found that male teachers received slightly higher ratings than female teachers, mainly because male students rated male teachers somewhat higher than female students (3.20 vs. 3.06 corresponding to approximately 0.2 SD). A re-analysis of Boring's (2015) data set by Boring et al. (2016) shows that the SET/Instructor Gender correlation was only 0.09, corresponding to approximately 0.2 SD. Accordingly, the Boring et al. data suggest that gender differences are small rather than large. However, Boring's data set does not allow the conclusion that

the relatively small differences in SET ratings are evidence of bias against female teachers for at least the following reasons: First, the students were not randomly assigned to seminar sections. For example, students selected whether they took early morning, mid morning, noon, mid afternoon, or late afternoon sections. Second, the students knew the grades given to them by their teachers before they completed SET. Third, there were substantial differences in the experience of female vs. male teachers. For example, a much larger proportion of male teachers had expertise in the field whereas a much larger proportion of female teachers were only PhD students. These experience differences alone could explain the small differences in SET ratings. Fourth, the seminar section teachers were free to teach their section whichever way they liked, used different assignments, etc., and thus, it is impossible to attribute the small differences in ratings to bias.

MacNell et al. (2015) examined the SET ratings of one female and one male instructor of an online course when students were either truthfully told the gender of each instructor (True Gender) or when students were misled about the instructors' genders (and told that each instructor's gender was in fact the opposite of what it was) (False Gender). Both instructors interacted with their students exclusively online, through discussion boards and emails; graded students work at the same time; used the same grading rubrics and co-ordinated their grading to ensure that grading was equitable in their sections. Based on the results of their experiment, MacNell et al. concluded that "Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender." However, MacNell et al.'s data suffer from several fundamental flaws that render them uninterpretable and MacNell et al.'s conclusions unwarranted (Uttl & Violo, 2021). First, MacNell et al.'s sample of students in each of the four conditions was extremely small, ranging from 8 to 12 students. Second, MacNell et al.'s conclusions depend on three outliers in their small data set—three students who gave their instructors the lowest possible rating on all or nearly all items. When the three outliers are removed from the data set, students rated the actual female instructor numerically higher than the actual male instructor regardless of whether the students were given the actual or false gender of the instructors. Third, MacNell et al.'s study included only one female and one male instructor. It is unwarranted to draw inferences from this small sample size in one study to how students rate female vs. male instructors in general.

Similarly, Mitchell and Martin (2018) examined SET ratings of one female (Mitchell) and one male (Martin) professor teaching different sections of the same online course and found that "a male instructor administering an identical course as a female instructor receives higher ordinal scores in teaching evaluations." Mitchell and Martin argued that their findings were evidence of gender bias as "the only difference in the courses was the identity of the instructor." However, the sections differed or may have differed in many aspects: (a) students' work was graded by different graders whose strictness varied, (b) Drs. Mitchell and Martin held face to face office hours, (c) Drs. Mitchell and Martin may have had different email styles, (c) Mitchell's ratings were based on approximately three times as many responses as

Martin's ratings, (d) Mitchell and Martin may have taught at different times of the day, etc. Moreover, Mitchell and Martin's argument that questions in Instructor/Course, Course, and Technology related to "characteristics that are specific to the course" and do not vary across the sections is simply incorrect. The questions in these categories asked, for example, what "the instructor" did and it ought to be self-evident that different instructors may do things differently, and thus, differences in ratings need not reflect gender bias. Finally, and importantly, just as with MacNell et al. (2015) study, one ought not to make sweeping conclusions about how two categories differ based on differences between two exemplars, one drawn from each of the two categories. This sample size equals one type of research is unlikely to describe what two populations are like.

4 SET Are Influenced by Chocolates, Course Easiness, and Other Incentives

SET ratings are also influenced by numerous factors whose consideration in evaluation of faculty is ill-advised or detrimental to student learning including course difficulty; distribution of chocolates, cookies, and tasty baked goods; and non enforcement of course policies including academic dishonesty and student codes of conduct policies.

4.1 Course Difficulty

Using data for 3,190 professors from US universities, Felton et al. (2004) found a moderately strong correlation between Quality and Easiness of 0.61. Moreover, the Quality/Easiness relationship became stronger as more ratings were available for each faculty member. Whereas for professors with 10–19 ratings, the Quality/Easiness correlation was 0.61, the correlation reached 0.76 for faculty with 50–59 ratings. The moderate to strong relationship between Quality and Easiness has been subsequently replicated by a number of studies including Felton et al. (2008), Rosen (2018), and Wallisch and Cachia (2019). Wallisch and Cachia (2019) confirmed a steep and accelerated decline of www.ratemyprofessor.com Overall Quality ratings (rated on 1–5 point scale) with increasing Course Difficulty (reverse of easiness) (rated on a 1–5 point scale). For each 1.0 point increase in Difficulty, Overall Quality ratings decreased by approximately 0.6 points.

4.2 *Chocolates and Cookies*

Two randomized studies demonstrate the power of chocolates and cookies in improving SET ratings. In one of the earlier randomized studies, Youmans and Jee (2007) examined whether providing small chocolate bars would result in higher SET ratings in two statistics and one research methods class. Students who were offered chocolate bars rated their instructor substantially higher than students who were not offered chocolate bars ($d = 0.33$). In another randomized study, Hessler et al. (2018) conducted a single-center randomized control group trial to determine whether the availability of chocolate cookies affects SET ratings. Relative to the no-cookie groups, the cookie groups rated teachers as well as the course material much higher, $d = 0.68$ and $d = 0.66$, respectively. Accordingly, at minimum, chocolates and chocolate cookies are both very effective ways to increase one's SET ratings.

5 SET Findings Vary with Conflict of Interest

Uttl et al. (2019) have recently shown that the correlations between SET and learning/achievement in the multisection studies discussed above depend not only on their sample size but also on their authors' degree of conflict of interest (perceived or actual). Figure 3, Panel A shows that authors with SET corporations (Corp) reported much higher SET/learning correlations than authors with no such ties, $r = 0.58$ vs. $r = 0.18$, respectively. However, as shown in Panel B, conflict of interest is not limited to authors with direct financial gains from selling SET but also extends to authors with other non financial conflicts of interest such as administrative (Admin) and evaluation units (Eval U) ties. These findings are particularly troubling; they suggest that in addition to the poor methodology employed by many SET studies (e.g., small sample sizes, insufficient method descriptions, failure to consider outliers), many SET research findings may also be the result of their authors financial and other interests, whether these biases were conscious or unconscious.

6 Discussion

SET do not measure teaching effectiveness and students do not learn more from more highly rated professors. Until recently, meta-analyses of multisection studies have been cited as the best evidence of SET validity. Those meta-analyses, however, were fundamentally flawed. The re-analyses of the previous meta-analyses as well as the new updated meta-analyses of multisection studies show that SET are unrelated to student learning in multisection designs. Accordingly, SET ought not to be used to measure faculty's teaching effectiveness.

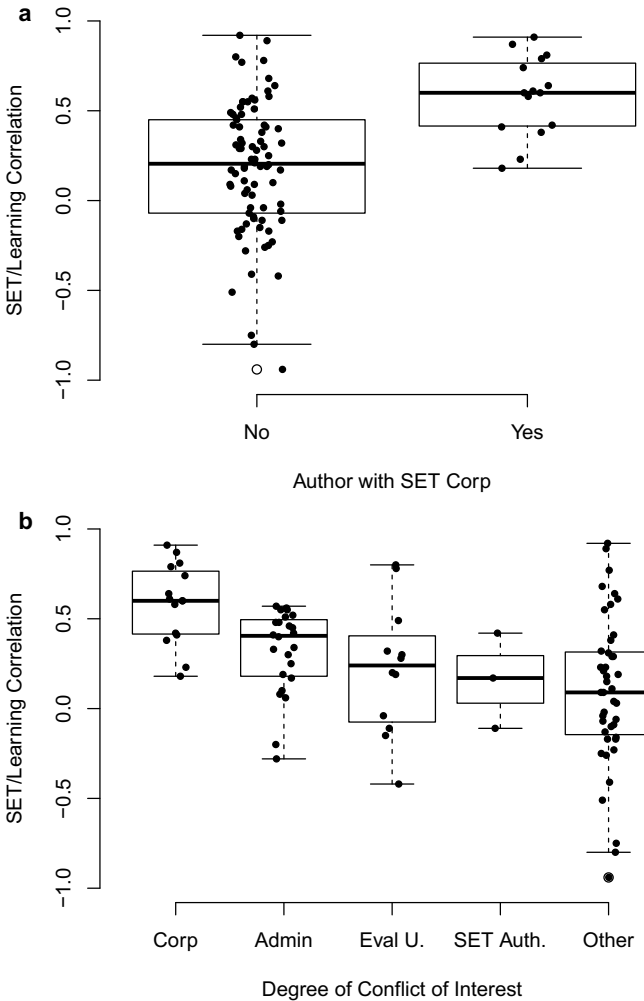


Fig. 3 SET/learning correlations and conflict of interest. Panel **A** shows that authors with SET corporations (Corp) reported much higher SET/learning correlations than authors with no such corporate ties. Panel **B** shows that authors with other conflicts of interests including administrative (Admin) and evaluation units (Eval U) also reported higher SET/learning correlations whereas authors with no identifiable conflicts of interest reported near zero SET/learning correlations (the figures are adapted from Uttl et al. 2019)

Regardless of what SET actually measure, SET are substantially influenced by (1) numerous factors not attributable to professors, including students’ intelligence and prior knowledge, students’ motivation and interest, class size, and course subject; (2) factors attributable to professors but whose consideration in high-stakes personnel decisions violates human rights legislation such as accent, race, ethnicity, national origin, age, and hotness/sexiness; and (3) factors attributable to professors but whose

consideration is at minimum unwise and/or detrimental to student learning, including course difficulty and availability of chocolates and cookies.

Although some SET systems attempt to adjust for influences of various TEIFs, SPFs, and DSLFs, these attempts are ultimately futile because no SET system can nor does adjust for all demonstrated effects of TEIFs, SPFs, and DSLFs, nor for effects of possible TEIFs, SPFs, and DSLFs. Even adjusting only for the factors reviewed above would likely be impossible. For example, to adjust for factors attributable to students, one would have to administer highly reliable and valid tests of student intelligence, prior knowledge, motivation, interest, racism, accent preference, hotness preferences, etc., then calculate average class intelligence, prior knowledge, motivation, interest, racism, accent preferences, hotness preferences, etc., and then develop some adjustment system. No one has done it so far and no one is likely to do so in the foreseeable future.

SET measure student satisfaction, that is, “a fulfillment of need or want” or “a happy or pleased feeling because of something that you did or something that happened to you” (www.m-w.com). One may argue that student satisfaction is important and that student satisfaction is properly used or ought to be used in high-stakes personnel decisions such as hiring, firing, promotion, merit pay, and teaching awards. However, the fundamental problem with using student satisfaction at all to evaluate faculty is that it depends on factors not attributable to professors.

Moreover, making high-stakes personnel decisions by comparing faculty’s SET ratings to university, faculty, or departmental norms, sets up and fuels a race among faculty members to beat at least 30, 50 or, 70% of their colleagues depending on the particular norm-referenced criteria for unsatisfactory, “in need of improvement”, etc. adopted by their institution. This race for higher and higher SET ratings is what a number of writers believe is the principal cause of run-away grade inflation and work deflation (Crumbley & Reichelt, 2009; Emery et al., 2003; Haskell, 1997; Stroebe, 2016, 2020). Although SET were relatively rare prior to 1970s, today they are used by almost all colleges and universities in North America and in many other countries to evaluate teaching effectiveness (Seldin, 1993). Accordingly, the race for higher SET pressures faculty to satisfy their students’ needs and wants, in particular, to increase grades, reduce workload, tolerate academic dishonesty, avoid topics that may antagonize some students, etc. Indeed, the grades have been increasing and have risen from C grades being the most frequently awarded grades in 1970s to A grades being the most frequently awarded grades today (Rojstaczer & Healy, 2010, 2012). At the same time, students report spending less and less time on their studies (Fosnacht et al., 2018; Rojstaczer & Healy, 2010). Whereas in the 1960s students in US were spending on average about 2 h studying outside of the class for each hour in the class, today students are spending only about 1 h. These two trends are nothing short of astonishing when one considers that the average intelligence and ability of students entering colleges and universities has declined over the last 50–100 years, as the proportions of high school graduates entering universities and colleges has increased from approximately 5% to more than 50% or even 70% depending on the country, state, and province (US Census Bureau, 2019). Notably, SET are not the only cause of grade inflation and work deflation. Other related causes include

colleges and universities' focus on high student retention; pressure on professors to limit percentages of D, F, and W (withdrawal) grades (explicitly requiring professors to increase grades); and business culture that not only strives for happy customers whose needs and wants need to be satisfied but also for as many customers as possible.

In the first public legal case of its kind, Ryerson University was forbidden from using SET as a measure of teaching effectiveness (*Ryerson University v. Ryerson Faculty Association*, 2018 CanLII 58,446, available at www.canlii.org). The arbitrator Williams stated:

That evidence, as earlier noted, was virtually uncontradicted. It establishes, with little ambiguity, that a key tool in assessing teaching effectiveness is flawed, while the use of averages is fundamentally and irreparably flawed. It bears repeating: the expert evidence called by the Association was not challenged in any legally or factually significant way. As set out above, the assessment of teaching effectiveness is critical, for faculty and the University, and it has to be done right. The ubiquity of the [SET] tool is not a justification, in light of the evidence about its potential impact, for its continuation, or for mere tinkering.

The SET ratings also run afoul of at least some codes of ethics. For example, Canadian Code of Ethics for Psychologists (Canadian Psychological Association, 2017) makes it clear that psychologist not only has a duty to not participate in incompetent and unethical behavior, such as evaluating their colleagues using invalid and biased SET tools, they also have a responsibility to call out "incompetent and unethical behavior, including misinterpretations or misuses of psychological knowledge and techniques" (Ethical Standard IV.13).

Notwithstanding the above criticisms, student surveys may continue to be useful for formative uses, that is, for improving instruction when professors themselves design or select questions relevant to their teaching methods and courses, and when SET are provided only to professors themselves to ensure that they are not misused, not used for summative uses, and used only for formative uses or to raise alarm about some ineffective teaching behaviors (e.g., not showing up for one's classes).

Finally, and importantly, this review of SET research highlights the need for transparent, replicable, and methodologically strong research, conducted by researchers with no conflict of interest and no interest in particular findings. The SET literature is replete with unsubstantiated and contradictory findings based on poor methods. As detailed above, Cohen's (1981) widely cited evidence of SET validity turned out to be an artifact of poor methods and failure to take into account small sample bias and students' prior ability. Similarly, Feldman's (1984) finding of minimal effect of class size and Aleamoni's (1999) later dismissal of the idea that class size is related to SET ratings as a myth were similarly based on poor methods and failure to adequately review the previous findings. And MacNell et al. (2015) claim of gender bias against women hinges in its entirety on three outliers, three students who disliked their instructors so much as to give them the lowest possible rating on all or nearly all items. Significantly, as shown by Uttl et al. (2019), the reported findings may be greatly influenced by a conflict of interest. It is clear that any review of this literature needs to be approached with an attitude of a detective rather than simply accepting what is written in studies' abstracts in order to ferret out true findings supported by evidence from uninterpretable and unwarranted claims.

In conclusion, continued use of SET in high-stakes personnel decision such as hiring, firing, promotion, merit pay, and teaching award is not evidence based. The evidence is that (a) students do not learn more from more highly rated professors; (b) SET are biased by a variety of factors not attributable to professors; (c) SET run afoul to human rights legislation, and (d) SET are easily manipulated by small chocolates such as Hershey's kisses, course easiness, and other factors. In short, SET do not measure faculty's teaching effectiveness and their use in high-stakes personnel decisions is improper, unethical, and ought to be discontinued immediately.

References

- Abrami, P. C., & d'Apollonia, S. (1999). Current concerns are past concerns. *American Psychologist*, 54(7), 519–520. <https://doi.org/10.1037/0003-066X.54.7.519>.
- Aleamoni, L. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2), 153–166. <https://doi.org/10.1023/A:1008168421283>.
- Beran, T., & Violato, C. (2009). Student ratings of teaching effectiveness: Student engagement and course characteristics. *Canadian Journal of Higher Education*, 39(1), 1–13.
- Boring, A. (2015). *Gender Biases in student evaluations of teachers* (No. 2015–13). Documents de Travail de l'OFCE. Observatoire Francais des Conjonctures Economiques (OFCE). <https://ideas.repec.org/p/fce/doctr/1513.html>. Accessed 4 June 2020.
- Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145, 27–41. <https://doi.org/10.1016/j.jpube.2016.11.006>.
- Boring, A., Ottononi, K., & Stark, P. B. (2016). *Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors*. The London School of Economics and Political Science. <https://doi.org/10.14293/s2199-1006.1.sor-edu.aetbzc.v1>.
- Canadian Psychological Association. (2017). *Canadian code of ethics for psychologists* (4th ed.). Canadian Psychological Association.
- Centra, J. A. (2009). *Differences in responses to the Student Instructional Report: Is it bias?* Educational Testing Service.
- Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, 51(3), 281–309. <https://doi.org/10.3102/00346543051003281>.
- Coladarsi, T., & Kornfield, I. (2007). RateMyProfessors.com versus formal in-class student evaluations of teaching. *Practical Assessment, Research & Evaluation*, 12(6), 1–15.
- Crumbley, D. L., & Reichelt, K. J. (2009). Teaching effectiveness, impression management, and dysfunctional behavior: Student evaluation of teaching control data. *Quality Assurance in Education: An International Perspective*, 17(4), 377–392. <https://doi.org/10.1108/09684880910992340>.
- Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, 11(1), 37–46. <https://doi.org/10.1108/09684880310462074>.
- Feldman, K. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education*, 21(1), 45–116. <https://doi.org/10.1007/BF00975035>.
- Felton, J., Koper, P. T., Mitchell, J., & Stinson, M. (2008). Attractiveness, easiness and other issues: Student evaluations of professors on Ratemyprofessors.com. *Assessment & Evaluation in Higher Education*, 33(1), 45–61. <https://doi.org/10.1080/02602930601122803>.
- Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: The relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, 29(1), 91–108. <https://doi.org/10.1080/0260293032000158180>.

- Fosnacht, K., McCormick, A. C., & Lerma, R. (2018). First-year students' time use in college: A latent profile analysis. *Research in Higher Education*, 59(7), 958–978. <https://doi.org/10.1007/s11162-018-9497-z>.
- Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends*. Higher Education Quality Council of Ontario. <https://deslibris.ca/ID/215362>. Accessed 22 Feb 2020.
- Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty. *Education Policy Analysis Archives*, 5, 6. <https://doi.org/10.14507/epaa.v5n6.1997>.
- Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C., et al. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education*, 52(10), 1064–1072. <https://doi.org/10.1111/medu.13627>.
- Hoyt, D. P., & Lee, E. (2002). *Technical Report No. 12: Basic data for the revised IDEA system*. The IDEA Center.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134. <https://doi.org/10.1037/0022-3514.77.6.1121>.
- MacNell, L., Driscoll, A., & Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291–303. <https://doi.org/10.1007/s10755-014-9313-4>.
- Marsh, H. W. (1980). *Students' evaluations of college/university teaching: A description of research and an instrument*.
- Marsh, H. W. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, 83(2), 285–296. <https://doi.org/10.1037/0022-0666.83.2.285>.
- Mitchell, K. M. W., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, 5(3), 648–652. <https://doi.org/10.1017/S104909651800001X>.
- Orpwood, G., & Brown, E. S. (2015). Closing the numeracy gap. CGC Educational Communications. http://www.numeracygap.ca/assets/img/Closing_the_numeracy_Executive_Summary.pdf. Accessed 20 May 2020.
- Rojstaczer, S., & Healy, C. (2010). Grading in American colleges and universities. *Teachers College Record*.
- Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940–2009. *Teachers College Record*, 114(7), 23.
- Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: A large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education*, 43(1), 31–44. <https://doi.org/10.1080/02602938.2016.1276155>.
- Seldin, P. (1993). The use and abuse of student ratings of professors. *The Chronicle of Higher Education; Washington*, 39(46), A40.
- Sonntag, M. E., Bassett, J. F., & Snyder, T. (2009). An empirical test of the validity of student evaluations of teaching made on RateMyProfessors.com. *Assessment & Evaluation in Higher Education*, 34(5), 499–504. <https://doi.org/10.1080/02602930802079463>.
- Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 11(6), 800–816. <https://doi.org/10.1177/17456916166650284>.
- Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology*, 42(4), 276–294. <https://doi.org/10.1080/01973533.2020.1756817>.
- Subtirelu, N. C. (2015). “She does have an accent but...”: Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society*, 44(1), 35–62. <https://doi.org/10.1017/S0047404514000736>.

- Timmerman, T. (2008). On the validity of RateMyProfessors.com. *Journal of Education for Business*, 84(1), 55–61. <https://doi.org/10.3200/JOEB.84.1.55-61>.
- US Census Bureau. (2019). CPS historical time series tables: Table A-1 Years of school completed by people 25 years and over, by Age and Sex: Selected years 1940 to 2019. US Census Bureau. <https://www.census.gov/data/tables/time-series/demo/educational-attainment/cps-historical-time-series.html>. Accessed 20 May 2020.
- Uttl, B., Bell, S., & Banks, K. (2018). Student evaluation of teaching (SET) ratings depend on the class size: A systematic review (No. 8110392). In *Proceedings of International Academic Conferences*. International Institute of Social and Economic Sciences. <https://ideas.repec.org/p/sek/iacpro/8110392.html>. Accessed 27 May 2020.
- Uttl, B., Cnudde, K., & White, C. A. (2019). Conflict of interest explains the size of student evaluation of teaching and learning correlations in multisection studies: A meta-analysis. *PeerJ*, 7(7), e7225. <https://doi.org/10.7717/peerj.7225>.
- Uttl, B., & Kibreab, M. (2011). Self-report measures of prospective memory are reliable but not valid. *Canadian Journal of Experimental Psychology = Revue Canadienne De Psychologie Experimentale*, 65(1), 57–68. <https://doi.org/10.1037/a0022843>.
- Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, 5(5), e3299. <https://doi.org/10.7717/peerj.3299>.
- Uttl, B., & Violo, V. (2021). Small samples, unreasonable generalizations, and outliers: Gender bias in student evaluation of teaching or three unhappy students? *ScienceOpen Research*. <https://doi.org/10.14293/S2199-1006.1.SOR-2021.0001.v1>.
- Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22–42. <https://doi.org/10.1016/j.stueduc.2016.08.007>.
- Uttl, B., White, C. A., & Morin, A. (2013). The numbers tell it all: Students don't like numbers! *PLoS ONE*, 8(12), e83443. <https://doi.org/10.1371/journal.pone.0083443>.
- Wallisch, P., & Cachia, J. (2019). *Determinants of perceived teaching quality: The role of divergent interpretations of expectations*. <https://doi.org/10.31234/osf.io/dsvqg>.
- Williams, P. G., Rau, H. K., Suchy, Y., Thorgusen, S. R., & Smith, T. W. (2017). On the validity of self-report assessment of cognitive abilities: Attentional control scale associations with cognitive performance, emotional adjustment, and personality. *Psychological Assessment*. <https://doi.org/10.1037/pas0000361>.
- Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology*, 34(4), 245–247. <https://doi.org/10.1080/00986280701700318>.

Bob Uttl is a Professor of Psychology at Mount Royal University, Calgary (Canada). Previously he held a variety of academic posts around the world including: Red Deer College, Canada; Tamagawa University, Japan; University of Tsukuba, Japan; Oregon State University, USA; Henry M. Jackson Foundation for Advancement of Military Medicine, USA; and National Institute of Health, USA. His research includes: the relationship between perception, processing resources, and prospective and retrospective memory; changes in cognitive functions due to normal and pathological aging; measurement and research methods; and evaluation of teaching in higher education.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part IV
Discussion and Future Directions

Chapter 16

Student Feedback on Teaching in Schools: Current State of Research and Future Perspectives



Wolfram Rollett, Hannah Bijlsma, and Sebastian Röhl

Abstract The aim of this volume was to give a comprehensive overview of the current state of the research on student perceptions of and student feedback on teaching. This chapter provides a resume of the important theoretical considerations and empirical evidence the authors contributed to this volume. First, evidence concerning the validity of student perceptions of teaching quality is discussed, highlighting the quality of the questionnaires used and accompanying materials provided by their authors. In the next step, empirical findings are summarized on student and teacher characteristics that can influence important processes within the feedback cycle. Subsequently, it is emphasized that the effectiveness of student feedback on teaching is significantly related to the nature of the individual school's feedback culture. Furthermore, it is argued that the efficacy of student feedback depends on whether teachers are provided with a high level of support, when making use of the feedback information to improve their teaching practices. As the literature review impressively documents, teachers, teaching, and ultimately students can benefit substantially from student feedback on teaching in schools.

Keywords Student perception · Student feedback · Teaching quality · Teaching improvement · Validity

W. Rollett · S. Röhl (✉)
Institute for Educational Sciences, University of Education, Freiburg, Germany
e-mail: sebastian.roehl@ph-freiburg.de

W. Rollett
e-mail: wolfram.rollett@ph-freiburg.de

H. Bijlsma
Section of Teacher Professionalization, University of Twente,
Enschede, the Netherlands
e-mail: h.j.e.bijlsma@utwente.nl

1 Introduction

Although there exists a vast and differentiated literature about teachers' feedback to students in schools and the ways to make productive use of it (Hattie, 2009), feedback from students to teachers has received far less attention. The aim of this volume, therefore, was to present an informative overview of state-of-the-art research in this area and important neighboring scientific fields. Central topics discussed in this volume are whether student perceptions of teaching in school are reliable and valid, what has to be considered to obtain valid information, and how to successfully make use of it for the professional development of teaching and teachers.

As Hattie points out in his foreword to this volume, the knowledge of variables which may influence the success and effectiveness of feedback is rather critical. The Process Model of Student Feedback on Teaching (SFT) suggested by Röhl, Bijlsma, and Rollett (Chap. 1 of this volume) is an attempt to provide a framework which describes the feedback cycle in such a way that it can provide an orientation for research on the efficacy of student feedback as well as for the effective implementation of intervention measures. Particular focus is put on variables which characterize the affective and cognitive processing of students' feedback by the teachers and their readiness for considering improvement-orientated actions. A professional implementation of student feedback on teaching clearly has the potential to enrich the feedback and learning culture of schools substantially and, above and beyond that can contribute to their democratic culture. A corresponding approach is elaborated by Jones and Hall (Chap. 13 of this volume) advocating school and teaching practices of involving the "student voice," i.e., involving students in the planning and implementation of their own education. But, as Uttl (Chap. 15 of this volume) summarizes, findings from higher education show how potential dangers arise when student perceptions of teaching are collected with an evaluative focus.

In this final chapter, we summarize the findings and conclusions drawn from the chapters in this volume to give an overview of what we have achieved in research on student feedback, what needs to be considered when implementing student feedback in practice and where we see room for improvement. First, we discuss the validity of student perceptions of teaching quality and characteristics of survey instruments. Next, we highlight characteristics of students and teachers affecting and impacting teachers' feedback processes. We then discuss the organizational context of the evaluation and the presentation of the feedback information to stakeholders. Finally, we suggest directions forward for researchers, policymakers, and schools.

2 Validity of Student Perceptions of Teaching Quality and Characteristics of Survey Instruments

Regarding the question of the validity of measurements and tests in educational contexts, it has been emphasized that validity can only be assessed in regard to the intended interpretation and subsequent actions (AERA, 2014; Kane, 2012). In this sense we discuss the validity of student perceptions of teaching in terms of their value for improving teaching practices in a formative setting and, at the same time, we disregard a purely evaluative use of student ratings on teaching (see Chap. 8 by Wisniewski and Zierer, and Chap. 15 by Uttl in this volume).

The literature review provided across the chapters of this volume impressively illustrates how teachers and teaching can benefit from making use of formative student feedback. Nevertheless, there are still many researchers and practitioners raising concerns about the accuracy and fairness of student ratings of teaching quality, which—if tenable—would considerably limit their value for the proposed usage in the development of teaching and teaching skills. Indeed, there are good reasons to be skeptical about the results of student ratings of teaching used in the field, and several contributions to this volume address the topic of a valid measurement of student perception of teaching quality.

An important issue in this context is how well students are able to evaluate teaching practices. The referenced literature on the prognostic validity of student feedback measures point to the result that student evaluations on teaching do indeed capture aspects of the teaching quality which are relevant for students' learning and development (e.g., Fauth et al., 2014; Praetorius et al., 2018; Wallace et al., 2016). As the analyses presented in this volume indicate, there is much which can be done to improve the measurement procedures and to increase the accuracy of student ratings. For example, Bijlsma et al. (Chap. 2 of this volume) point out that the underlying psychometric theory is determining how the rater's perception is conceptualized and captured. Göllner, Fauth, and Wagner (Chap. 7 of this volume) emphasize impressively that we have to be cautious and more aware about the way we ask students about their experiences in class. Different combinations of item referents (e.g., "I / We / The class understood the subject matter well") and item addressees (e.g., "The teachers explained the subject matter clearly to me / the class") are likely to induce different evaluation processes and different results, thus affecting reliability and validity of the measurements. Accordingly, Schweig and Martínez (Chap. 6 of this volume) call for evaluating within-classroom variability of student experiences as an indicator for disparate instructional experiences and unequal participation opportunities of the students. The authors strongly argue that evaluating within-classroom variability should be considered as a defining strength of the approach of using student-survey-based measures for the improvement of teaching.

One intensely discussed topic in the literature is the agreement or disagreement of the evaluations of students and observers (e.g., Clausen, 2002; Gitomer et al., 2014;

Kuhfeld, 2017). van der Lans (Chap. 5 of this volume) provides important findings which may even have the potential to end this discussion. In his analyses, the results from students and observers converge when 25 students' and seven different observers' views are related to each other. Interestingly, the ordering of the item difficulties or teaching competences were very consistent across students or observer ratings, and could also be calibrated on the same continuum of instructional effectiveness (van der Lans et al., 2019). These findings indicate that the disagreement of students and observers often reported in the literature may be largely attributed to an insufficient number of observers.

But it is indisputable that the question has to be raised of how well students perceive different aspects of teaching quality, and how well they comprehend the corresponding items in the questionnaire they are processing. Unfortunately, little research has been done on these topics. Accordingly, Göllner and colleagues (Chap. 7 of this volume) call for studies on the students' cognitive processing of survey items and its influence on their evaluation of teaching practices, while highlighting the necessity of age- and development-appropriate survey instruments. A closer look at the topic of how well students comprehend the items in a questionnaire would improve the survey instruments substantially and subsequently enhance the validity of the feedback information. As a consequence, for example, Bijlsma et al. (under review) and Lenske (2016) intensely discuss the content of the items of their student perception questionnaires with students to make sure that they understood and interpreted the items well.

In their review of the literature, both Göllner et al. (Chap. 7 of this volume) and Röhl and Rollett (Chap. 3 of this volume), raise further questions of (1) whether students in school are actually able to distinguish between different teaching dimensions and (2) how reliable the ratings of different dimensions are. Students' lack of ability to differentiate between different teaching dimensions would lead to empirically simpler factorial structures (see also Kuhfeld, 2017). Indeed, it is quite typical to see a two-factor structure, where a general factor covers all theoretically distinguishable teaching quality constructs with the exception of classroom management (e.g., Wallace et al., 2016). As Röhl and Rollett (Chap. 3 of this volume) demonstrate, students' social perceptions of their teachers explain an important part of the common variance of different teaching quality dimensions in a second order factor model. These results indicate that students' evaluations of teaching quality might be influenced by their social perceptions of their teachers and so may lead to biased assessments. Their findings suggest to emphasize using items which are less likely to be confounded by the students' social perception of their teachers (e.g., by addressing the individual experiences in a specific lesson) and controlling for the impact of how students socially perceive their teachers can be counteracted by administering suitable scales. The literature, nevertheless, offers reliable information that the students' assessments of teaching quality dimensions show characteristics of differential predictive validity, indicating the existence of a meaningful unique variance (e.g., Klieme & Rakoczy, 2003; Raudenbush & Jean, 2014; Yi & Lee, 2017).

In order to make use of student perceptions of teaching quality, the design of the survey instruments is crucial and critically determines the nature of the perceptions. In her informative systematic review, Bijlsma (Chap. 4 of this volume) analyzes the quality of 22 student perception questionnaires on teaching quality, which leads to an extensive literature research. Overall, most of the instruments were evaluated positively regarding their theoretical foundation, their design, and the information about their statistical quality. The review revealed, however, weaknesses concerning norm information, sampling specifications, and the availability of more detailed information on the features of the instruments (e.g., by providing a user manual). The analyses illustrate how to put more emphasis on the quality of the presentation of the survey instruments to potential users.

As Röhl's review of the research (Chap. 9 of this volume) shows, there is a substantial amount of evidence for the effects of student feedback on teachers' behavior—e.g., initiating reflective thinking processes, learning about students' perspectives, reviewing their goal setting, and changing their teaching practices accordingly. When provided with student ratings or feedback on their teaching, teachers also tend to engage more in communication with their classes on teaching practices and the changes which follow student feedback. An important pattern of results from a meta-analysis of intervention studies in schools presented by Röhl (*ibid.*) shows a mean weighted effect size of $d = 0.21$ for the impact of student feedback on student's perception of the subsequent lessons. But, as an in-depth analysis showed, this effect size underestimates the potential of student feedback: A high level of support provided to the teachers when making use of feedback information yielded a significantly larger positive effect of $d = 0.52$. Medium or low levels of support, though, did not result in a better outcome. This pattern of findings, therefore, highlights how crucial an adequate level of support is for the effectiveness of student feedback measures.

3 Student and Teacher Characteristics Influencing the Feedback Process

As the research presented in this volume shows, student feedback on teaching can indeed provide a valuable basis for evaluating and improving teaching practices. It is not unusual for school students to welcome and value the opportunity to give teachers feedback regarding their teaching and to find their "student voice" recognized (Jones and Hall, Chap. 13 of this volume). Nevertheless, student ratings of teaching can be affected by a variety of student and class characteristics (Bijlsma et al., *under review*). For example, high performing students rate their teachers' teaching quality significantly higher than low and middle performing students. Students from socio-economically or educationally more privileged families tend to be more critical of teaching practices (Atlay et al., 2019). Male students seem to be more critical than female students (Kuhfeld, 2017). Moreover, differences in students' language

comprehension can affect whether items in the survey instrument are understood (Lenske, 2016). The perception or evaluation of teaching quality may differ by age or development stage (see Chap. 7 of this volume by Göllner et al.). Student ratings of teaching can also be influenced by certain individual teacher characteristics which are not systematically associated with differences in teaching performance—such as gender, age, or physical appearance; in more sophisticated evaluation contexts, procedures can be implemented to correct scores accordingly, but this is not typical. The expectation of erroneous results can be considered as the most frequent reason why teachers are reluctant to use feedback. Although, as Schweig and Martínez (Chap. 6 of this volume) conclude, “these biases are generally small in magnitude and do not greatly influence comparisons across teachers or student groups, or how aggregates relate with one another and with external variables.” But users should, of course, be aware of the biases which might occur, especially as minor differences may have severe consequences in evaluative contexts.

Ways to counteract these undesirable effects and prepare students to use the questionnaires appropriately—and thereby develop students’ feedback competences—are indeed advisable. Accordingly, Göbel et al. (Chap. 11 of this volume) call for training students to use the survey instruments adequately. Unfortunately, the authors of these survey instruments frequently do not provide the users with clear guidelines for implementing the instruments (Bijlsma, Chap. 4 of this volume).

It has been well documented that individual characteristics of teachers influence whether and how effectively teachers use student feedback measures to improve their teaching and teaching skills, as Röhl and Gärtner (Chap. 10 of this volume) document in their literature review. In their discussion, the authors particularly emphasize teachers’ attitudes toward students as a feedback providers (e.g., regarding their trustworthiness or competence) and whether teachers perceive the function of the feedback as an opportunity to develop their teaching. The effectiveness of student feedback is also influenced by the teachers’ attitudes toward the measuring process of feedback. In general, teachers tend to show a positive attitude toward formative forms of student feedback on teaching (e.g., Göbel et al., Chap. 11 of this volume). But it is not uncommon that accuracy and trustworthiness are questioned, especially when it comes to feedback from younger students. Pre-service teachers, on the other hand, tend to be more positive when considering using student feedback on teaching than in-service teachers, as the findings of Göbel et al. (*ibid.*) indicate. In their insightful investigation, they demonstrate that experiences with student feedback can have a further positive impact on pre-service teachers’ attitudes concerning student feedback in general and on their willingness to reflect on and modify teaching practices. These results thereby illustrate the potential of a widespread implementation of student feedback on teaching within the practical parts of teacher education.

As the results in this volume show, the ways in which teachers perceive, process, and make use of the feedback information determines its impact on their teaching. Accordingly, the SFT Model (see Chap. 1 of this volume by Röhl et al.) puts an emphasis on teachers’ processes and handling of feedback information. At present, research on the ways in which teachers perceive and interpret feedback information affectively and/or cognitively, how this influences them, and how they deal with

it is still rather scarce. But as the investigations of Röhl and Rollett (2021) show, teachers can very much differ in why and how they make use of student feedback on teaching. Their analysis evinced four paths of utilization: (1) Direct Formative Use (identifying aspects to improve, setting goals, evaluating target achievement); (2) Direct Communicative Use (discussing the results and looking for improvements in class); (3) Indirect Use (enabling a positive emotional experience, gathering of information); and (4) Symbolic Process Orientated Use (signaling a democratic or student-orientated attitude and an openness to criticism in classes). These results point to the importance of paying more attention to the goals individual teachers pursue when they ask their students for feedback on their teaching.

4 Organizational Context of the Evaluation and the Presentation of Feedback Information to Stakeholders

The conditions of the organizational context in schools can vary largely, and these differences can influence the effectiveness of student feedback (see Chap. 10 by Röhl and Gärtner; Chap. 11 by Göbel et al.; and Chap. 8 by Wisniewski and Zierer in this volume). The relevance of organizational characteristics for the effects of feedback is also evident from research on multisource feedback in business enterprises (Chap. 14 of this volume by Fleenor). The school setting can provide resources which strongly support teachers within the student feedback cycle, thus fostering its effectiveness (see Chap. 9 of this volume by Röhl). Schools may offer team structures which intensely accompany the process of reflecting on the feedback as well as the subsequent professional development and changes in teaching practices. Furthermore, it depends very much on the learning culture within the organization whether the feedback is considered as an opportunity to learn and whether sustainable support is provided to act on the results. Correspondingly, Röhl and Gärtner (Chap. 10 of this volume) highlight—in their literature review on the conditions of effectiveness of feedback—the importance of a positive feedback culture, organizational safety, and a focus on the professional development of teachers in contrast to a focus on control. In their approach, the feedback culture is considered as a crucial moderator for the effectiveness of feedback measures. Accordingly, the school management and leadership have a special responsibility for the success of student feedback measures by ensuring a safe learning environment and shaping a positive feedback culture within schools. Elstad and colleagues (2015, 2017) report a higher appreciation of the results of student feedback on teaching when a developmental purpose is perceived by the teachers, whereas perceiving a control purpose is linked to a rejecting attitude to the feedback measures and a lower recognition of the feedback information. Furthermore, other findings point out that its effects on teaching practices differ depending on whether teachers are intrinsically or extrinsically motivated to engage in using

student feedback, but that both motivational paths are related to positive changes in the classroom (Gärtner, 2014).

How much the organizational context matters is well illustrated by a project described by van der Lans (Chap. 5 of this volume). It followed a very sophisticated data-driven procedure: (1) determining reliable diagnostic results of individual teaching practices (from ineffective to effective); (2) allocating teachers on an empirically validated continuum of teaching effectiveness; (3) identifying the most effective development measures; and (4) tailoring the feedback procedure accordingly. Taken together, these measures provide a viable basis for the teacher's further education and professional development. This is aligned to the research field of data-based decision-making, which suggests that data (in this case student feedback data) can help improve teaching and further outcomes for students (Poortman & Schildkamp, 2016; Schildkamp, 2019; van Geel et al., 2016).

A critical issue for the effectiveness of student feedback is how it is presented to the teachers. Problems concerning accuracy and comprehensibility of feedback have been addressed for a long time (e.g., Frase & Streshly, 1994). Thereby, the designing of feedback and of support measures has to take into account the level of data literacy of teachers in order to overcome the typical struggles of making use of the data (e.g., Kippers et al., 2018). One way to reduce the complexity of the gathered student data is by condensing the information into a smaller number of performance levels, which makes it considerably easier to communicate individual strengths and weaknesses. However, this requires an adequate level of support to prepare the data accordingly. One should, nevertheless, be careful not to disregard the potentially meaningful variance of the student ratings within classes (see Chap. 6 of this volume by Schweig and Martínez).

Another important prerequisite for the effectiveness of student feedback is that the communication within the feedback cycle is performed in an appreciative and constructive manner. Yet the ability to formulate and provide feedback as well as the ability to receive and respond to feedback can vary considerably. These differences can significantly influence the cognitive and emotional dynamic within the feedback process and its effectiveness. Effective feedback can be characterized as task orientated, specific, clear, development orientated, and distinct in its implications for action (Cannon & Witherspoon, 2005). Röhl and Gärtner (Chap. 10 of this volume) discuss how characteristics of the feedback may influence its effectiveness in terms of the information format (e.g., means, boxplots), the timing of the feedback, its specificity, valence, and positivity. Unfortunately, there are only few studies in the field of student feedback on teaching addressing these issues.

5 Concluding Remarks

The present volume is the first providing a comprehensive overview of the current state of the research on student perceptions of and student feedback on teaching in schools. Its aim was to coherently present to a wider audience the extensive and important international research which has been done using student perception of teaching for improving teaching practices. The authors contributing to this volume agree in granting student feedback a high potential for the improvement of teaching in schools. The empirical evidence for this claim, which is addressed across the chapters of this book, is impressive. If set up professionally, the implementation of student feedback on teaching can be indeed a very effective way to improve teaching quality.

But there are, of course, requirements which have to be met to achieve these positive results. On the one hand, a high quality of the survey instruments and the accompanying material provided by their authors is indispensable. On the other hand, the provision of a high quality of support within the school setting is needed gathering and evaluating the data, interpreting and reflecting on the information, and putting results effectively into action. Accordingly, it has been shown that the availability of an adequate level of support is an important moderating variable concerning the effectiveness of student feedback on teaching (Röhl, Chap. 9; and Göbel et al., Chap. 11 in this volume).

In order to make better use of the potential of student feedback on teaching, different paths should be followed: Authors of student perception questionnaires should put more emphasis on providing users with sound and easily accessible information (concerning e.g., theoretical basis, measurement quality, reference norms, and guidelines for implementing and working with the instruments). Practitioners should be encouraged to use student feedback by establishing sustainable support structures within schools, which include powerful technical solutions for implementing, evaluating and acting on student feedback. Researchers should intensify investigations on teachers' ways of processing feedback information, on how a professional and ongoing implementation of student feedback in schools affects the longitudinal development of teachers, teaching, and last but not least students.

In several respects, the findings presented in this volume, indicate that the summative use of student feedback for teacher accountability to supervisors is hardly appropriate. Thus, when teachers perceive a control function of the feedback, they tend to be more resistant to the developmental use of it (see Chap. 10 of this volume by Röhl and Gärtner). Further, student perceptions of teaching quality are subject to many idiosyncrasies or biasing factors, requiring highly expert and cautious interpretation (see Chaps 3 and 7 of this volume by Röhl and Gärtner). Therefore, the use of student feedback for accountability purposes should be avoided in order to prevent damaging its developmental potential.

A promising development to support the capturing, evaluating, and scrutinizing of student feedback data are online or smartphone-based survey instruments (e.g., the Impact! tool, Bijlsma et al., 2019; FeedbackSchule, Wisniewski et al., 2020).

If set up accordingly, they can provide users with easily accessible differentiated feedback information on individual-, group-, and class-level or even might provide corrected scores for known biases. Digital solutions can be an excellent way to gather and evaluate student feedback on teaching, thus reducing the needs on resources significantly. Of course, they cannot in any way substitute the professional reflection of individual teachers within collegial settings on the feedback results, but they can help schools substantially in creating the informational basis for these processes and help them to make better use of the information and their typically limited time and staff resources.

Another very inspiring perspective for the future of student feedback is provided by Schmidt and Gawrilow (see Chap. 12 of this volume) when advocating the implementation of measures for systematic reciprocal feedback between students and teachers, thereby addressing teachers and students as cooperative partners. The potential of combining approaches of feedback from students to teachers with those of feedback from teachers to students provides an important outlook for developments to come.

An important point arising from the research overview presented in this volume is the question of why some countries and regions seem to be more reluctant to use student feedback to improve teaching and professionalize teachers than others (see e.g. Chap. 3 by Bijlsma, and Chap. 9 by Röhl). Here, cultural aspects like the role of teachers and students, but also characteristics of the school systems could provide an explanatory framework. These issues should be elaborated on in further research in order to develop appropriate and effective forms of making use of student feedback on teaching for these cultural contexts and countries.

As outlined throughout this book, student feedback on teaching is a highly beneficial and—from our point of view indispensable—way to improve teaching practices. Based on the extensive body of research on the benefit and effectiveness of student feedback on teaching presented in this volume the authors hope to contribute to a wide and systematic use of student feedback in schools to sustainably improve teaching quality and the learning experiences of students.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. <https://doi.org/10.1037/e577932014-003>.
- Atlay, C., Tieben, N., Fauth, B., & Hillmert, S. (2019). The role of socioeconomic background and prior achievement for students' perception of teacher support. *British Journal of Sociology of Education, 40*(7), 970–991. <https://doi.org/10.1080/01425692.2019.1642737>.
- Bijlsma, H. J. E., Glas, C. A. W., & Visscher, A. J. (under review). *The factors influencing digitally measured student perceptions of teaching quality*. Paper presented at the EARLI conference in Aachen.
- Bijlsma, H. J. E., Visscher, A. J., Dobbelaer, M. J., & Veldkamp, B. P. (2019). Does smartphone-assisted student feedback affect teachers' teaching quality? *Technology, Pedagogy and Education, 28*(2), 217–236. <https://doi.org/10.1080/1475939x.2019.1572534>.

- Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *The Academy of Management Executive*, 19(2), 120–134. <https://doi.org/10.5465/ame.2005.16965107>.
- Clausen, M. (2002). *Unterrichtsqualität: eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstrukt- und Kriteriumsvalidität* [Teaching quality: A matter of perspective? Empirical analyses of agreement, construct and criterion validity]. Waxmann. <https://doi.org/10.1080/0267152980130104>.
- Elstad, E., Lejonberg, E., & Christophersen, K.-A. (2015). Teaching evaluation as a contested practice: Teacher resistance to teaching evaluation schemes in Norway. *Education Inquiry*, 6, 375–399. <https://doi.org/10.3402/edui.v6.27850>.
- Elstad, E., Lejonberg, E., & Christophersen, K.-A. (2017). Student evaluation of high-school teaching: Which factors are associated with teachers' perception of the usefulness of being evaluated? *Journal for Educational Research Online*, 9(1), 99–117.
- Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, 29, 1–9. <https://doi.org/10.1016/j.learninstruc.2013.07.001>.
- Frase, L. E., & Streshly, W. (1994). Lack of accuracy, feedback, and commitment in teacher evaluation. *Journal of Personnel Evaluation in Education*, 8(1), 47–57. <https://doi.org/10.1007/bf00972709>.
- Gärtner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, 42, 91–99. <https://doi.org/10.1016/j.stueduc.2014.04.003>.
- Gitomer, D. H., Bell, C. A., Qi, Y., McAffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, 116(6), 1–32.
- Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. <https://doi.org/10.1007/s1159-011-9198-8>.
- Kane, M. T. (2012). Validating score interpretations and uses. *Language Testing*, 29, 3–17. <https://doi.org/10.1177/0265532211417210>.
- Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention? *Studies in Educational Evaluation*, 56, 21–31. <https://doi.org/10.1016/j.stueduc.2017.11.001>.
- Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In Deutsches PISA-Konsortium & J. Baumert (Eds.), *PISA 2000—Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 333–359). Springer. https://doi.org/10.1007/978-3-322-97590-4_12.
- Kuhfeld, M. R. (2017). When students grade their teachers: A validity analysis of the Tripod student survey. *Educational Assessment*, 22, 253–274. <https://doi.org/10.1080/10627197.2017.1381555>.
- Lenske, G. (2016). *Schülerfeedback in der Grundschule: Untersuchungen zur Validität* [Student feedback in primary schools: Studies of validity]. Münster: Waxmann.
- Poortman, C. L., & Schildkamp, K. (2016). Solving student achievement problems with a data use intervention for teachers. *Teaching and Teacher Education*, 60, 425–433. <https://doi.org/10.1016/j.tate.2016.06.010>.
- Praetorius, A.-K., Klieme, E., Herbert, B., & Pinger, P. (2018). Generic dimensions of teaching quality: The German framework of Three Basic Dimensions. *ZDM Mathematics Education*, 50, 407–426. <https://doi.org/10.1007/s11858-018-0918-4>.
- Raudenbush, S. W., & Jean, M. (2014). To what extent do student perceptions of classroom quality predict teacher value added? In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the measures of effective teaching project* (1st ed., pp. 170–202). Jossey-Bass. <https://doi.org/10.1002/9781119210856.ch6>.
- Röhl, S., & Rollett, W. (2021). Jenseits von Unterrichtsentwicklung: Intendierte und nicht-intendierte Nutzungsformen von Schülerfeedback durch Lehrpersonen [Beyond teaching development: Teachers' intended and unintended ways of student feedback use]. In K. Göbel, C. Wyss,

- K. Neuber, & M. Raaflaub (Eds.), *Quo vadis Forschung zu Schülerrückmeldungen?* Springer VS. <https://doi.org/10.1007/978-3-658-32694-4>.
- Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, 1–17. <https://doi.org/10.1080/00131881.2019.1625716>.
- van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2019). Same, similar, or something completely different? Calibrating student surveys and classroom observations of teaching quality onto a common metric. *Educational Measurement: Issues and Practice*, 38, 55–64. <https://doi.org/10.1111/emip.12267>.
- van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal*, 53(2), 360–394. <https://doi.org/10.3102/0002831216677346>.
- Wallace, T. L., Kelcey, B., & Ruzek, E. A. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal*, 53, 1834–1868. <https://doi.org/10.3102/0002831216671864>.
- Wisniewski, B., Zierer, K., Dresel, M., & Daumiller, M. H. (2020). Obtaining students' perceptions of instructional quality: Two-level structure and measurement invariance. *Learning and Instruction*, 66. <https://doi.org/10.1016/j.learninstruc.2020.101303>.
- Yi, H. S., & Lee, Y. (2017). A latent profile analysis and structural equation modeling of the instructional quality of mathematics classrooms based on the PISA 2012 results of Korea and Singapore. *Asia Pacific Education Review*, 18, 23–39. <https://doi.org/10.1007/s12564-016-9455-4>.

Wolfram Rollett is a Professor of Empirical Educational Research at the University of Education Freiburg and the Freiburg Advanced Center of Education (FACE). Previously he worked as a researcher and lecturer in the field of Educational Science and Psychology at the Universities of Potsdam, Braunschweig, Dortmund, and Wuppertal. His research focuses on school development processes, the quality of extra- and co-curricular activities, educational effectiveness, and classroom composition effects.

Hannah Bijlsma is a researcher (Ph.D.) at the section of Teacher Professionalization at the University of Twente (the Netherlands) and a primary school teacher (grade 1). Her research focuses on measuring teaching quality and on the use of student perceptions of teaching quality in school contexts. In 2016 she founded a professional association for academic primary school teachers, of which she has been chairman for about five years. She is now a board member of the International Congress for School Effectiveness and Improvement (ICSEI) and a board member of the EARLI SIG on School Effectiveness and Improvement.

Sebastian Röhl has been an Academic Assistant in the Department of Educational Science at the University of Education Freiburg and is currently a Postdoctoral Researcher in the Institute of Education at Tübingen University (Germany). Before that he worked for more than 10 years as a grammar school teacher, school development consultant, and in teacher training. Among other areas, he conducts research in the fields of teaching development and teacher professionalization through feedback, social networks in inclusive school classes, as well as teachers' religiosity and its impact on professionalism. In addition, he is the Director of an in-service professional master's study program for teaching and school development.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

