Bielefelder Schriften zur Didaktik der Mathematik

RESEARCH

Milena Damrau

Understanding the Generality of Mathematical Statements

An Experimental Study at the Transition from School to University

OPEN ACCESS

Springer Spektrum

Bielefelder Schriften zur Didaktik der Mathematik

Band 15

Die Reihe Bielefelder Schriften zur Didaktik der Mathematik fokussiert sich auf aktuelle Studien zum Lehren und Lernen von Mathematik in allen Schulstufen und -formen einschließlich des Elementarbereichs und des Studiums sowie der Fort- und Weiterbildung. Dabei ist die Reihe offen für alle diesbezüglichen Forschungsrichtungen und -methoden. Berichtet werden neben Studien im Rahmen von sehr guten und herausragenden Promotionen und Habilitationen auch

- empirische Forschungs- und Entwicklungsprojekte,
- theoretische Grundlagenarbeiten zur Mathematikdidaktik,
- thematisch fokussierte Proceedings zu Forschungstagungen oder Workshops.

Die Bielefelder Schriften zur Didaktik der Mathematik nehmen Themen auf, die für Lehre und Forschung relevant sind und innovative wissenschaftliche Aspekte der Mathematikdidaktik beleuchten.

Milena Damrau

Understanding the Generality of Mathematical Statements

An Experimental Study at the Transition from School to University

Springer Spektrum

Milena Damrau
Bielefeld, Germany

Dissertation Bielefeld University, 2023.

ISSN 2199-739X ISSN 2199-7403 (electronic)
Bielefelder Schriften zur Didaktik der Mathematik
ISBN 978-3-658-43762-6 ISBN 978-3-658-43763-3 (eBook)
https://doi.org/10.1007/978-3-658-43763-3

I acknowledge support for the publication costs by the Open Access Publication Fund of Bielefeld University and the Deutsche Forschungsgemeinschaft (DFG) as well as by the Faculty of Mathematics of Bielefeld University.

This Springer Spektrum imprint is published by the registered company Springer Fachmedien Wiesbaden GmbH, part of Springer Nature.
The registered company address is: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Paper in this product is recyclable.

Foreword

While the fundamental role of proof in mathematics as a science is undisputed, the picture that is painted at school is one that often hardly does justice to the demands associated with it. This is especially true for the development after TIMSS and PISA, in which mathematical applications have been more heavily emphasized in curricula, while proving is assigned a rather subordinate role. Even less consideration is given to proof in the practical implementation in lessons, which are often strongly oriented towards the requirements of written examinations and central benchmarks, in which proof hardly plays a role. Consequently, students have difficulties at the transition from school to universities, especially those who choose mathematics as a subject. These problems are not new but have been magnified rather than diminished by developments in recent years. In the last two decades, this development has led to increased subject didactic research on the topic of argumentation and proof in schools and universities. However, many questions in this regard have not yet been adequately clarified. This applies in particular to the questions what comprehension and understanding of proof do students have when they enter university, what persuasive power mathematical proofs have for them, to what extent do they distinguish proofs from empirical arguments, to what extent does the individual persuasive power of proofs depend on the type of argument, and what is the underlying understanding of generality. Milena Damrau's doctoral dissertation seeds to find answers to these questions. The main objective of the study is to investigate the understanding of generality of mathematical statements of first-year students in connection to the reading and construction of proofs. In particular, the influence of different types of arguments (no argument, empirical argument, generic proof, formal proof) on the understanding of proof and on the respective degree of individual conviction is investigated. For this purpose, the author develops an experimental survey, which

is conducted on a sample of 430 mathematics students and student teachers. The study is based on five mathematical statements, each of which is presented to four sample groups together with different types of arguments to justify the truth of the statements. Subsequently, questions are asked about the truth value of the statements, the generality of the statement, and the persuasiveness of the arguments. Of particular interest is the relationship between (a) the assessment of the truth value of the respective statements and (b) the question of whether or to what extent there can be counterexamples in the respective cases. Frederick's Cognitive Reflection Test (CRT) is used as a control instrument. The study leads to impressive results: Regarding the central question about the relationship between the assessment of the truth value and the assessment of the existence of possible counterexamples, it turns out that about one third of the participants—and thus a comparatively high proportion—do not have an adequate understanding of the generality of mathematical statements. As expected, the extent of this understanding is directly related to students' previous mathematical education. Furthermore, the study provides a detailed analysis of correlations between the understanding of generality of statements and other proof-related activities; in particular, a positive correlation regarding the comprehension of proof and a negative correlation regarding the acceptance of empirical arguments should be mentioned. Interesting and new are also the results regarding the influence of the type of argument on students' performances in proof-related activities, which turn out to be very different depending on whether no arguments, empirical arguments, generic or formal proofs are offered. Remarkable and perhaps unexpected is the high influence of empirical arguments on the assessment of the truth value of a statement. Overall, Milena Damrau develops results in two main fields of research:

- With her conceptualization of the understanding of generality of mathematical statements and the development of instruments related to it, she has not only developed a methodological basis for her own work, but moreover a methodology that can also be used profitably for other future investigations.
- Her empirical study leads to detailed results on the understanding of generality of mathematical statements and their dependence on the type of argument, the type of statement, and other conditional characteristics, which were not known in this form before and which expand the didactic knowledge in argumentation and proof.

I hope that these results will attract the interest of the mathematics education community and contribute to further research into mathematical learning processes in the areas of reasoning, argumentation, and proof.

Bielefeld Rudolf vom Hofe
September 2023

Acknowledgements

I am extremely grateful to Prof. Dr. Rudolf vom Hofe for supervising my thesis, for giving me the freedom to follow my personal research interests, and for supporting me in acquiring the knowledge and skills needed to accomplish my goals. I especially thank him for the helpful feedback and discussions that improved my work.

I would also like to express my deepest gratitude to Prof. Dr. Andrea Peter-Koop for giving me the opportunity to join mathematics education research in the first place, for bringing me up as a member of the scientific community, for many fruitful discussions, and for her unlimited support. Thank you!

During the final stages of my doctorate, Prof. Dr. Stefan Ufer from LMU provided very helpful feedback on my work, for which I am very thankful. I would also like to express my gratitude for giving me the opportunity to present my work at the *Oberseminar* in Munich. The constructive feedback and discussions helped me a lot to reflect on and improve my work.

I would like to extend my sincere thanks to Prof. Dr. Daniel Sommerhoff from IPN Kiel for helpful discussions and feedback, which contributed to improving my thesis.

Special thanks to Dr. Guido Elsner, Dr. Daniel Frohn, and Prof. Dr. Charles Vial for giving me the opportunity to collect my data in their lectures, without which this thesis would not have been possible.

Many other people contributed to my work and the development as a researcher over the past years. I want to thank my colleagues at the IDM, especially Dr. Julia Streit-Lehmann, Dr. Daniel Barton, Dr. Valentin Katter, Dr. Sebastian Kollhoff, (again) Dr. Daniel Frohn, and Dr. Judith Huget for fruitful discussions, helpful feedback, and/or the coding of data.

Outside of Bielefeld University, I would like to acknowledge Prof. Dr. Kristina Reiss from TUM for the initial guidance, which brought me on the right track to find my topic. I'd also like to thank David Zenz for helpful discussions on statistics and Prof. Dr. Paul Wedrich from the University of Hamburg for helpful comments on the more mathematical parts of my thesis.

Finally, I want to thank my parents and my sister for their constant support and patience, and in particular my husband Hernán for numerous fruitful discussions and his tireless support and encouragement. Your belief in me has kept my motivation up.

Thank you all so much!

Milena Damrau

Abstract

The transition from school to university is very challenging for many mathematics students and preservice mathematics teachers, which consequently leads to high drop out rates (Heublein, Hutzsch, & Schmelzer, 2022). One of the main reasons identified in the literature are students' difficulties with proof-based mathematics to which they are commonly introduced when they start university (e.g., Gueudet, 2008; A. Selden, 2012). While research on first-year students' proof skills has increased significantly over the last decades, activities related to the *reading of statements,* which are to be proven—or for which a proof has to be read–have largely been neglected. I therefore adapted the framework on proof-related activities introduced by Mejía Ramos and Inglis (2009b) by distinguishing activities related to the *reading of the statement* and those related to the *reading and construction of arguments.* The comprehension of a statement, as part of reading the statement, involves understanding the *generality of the statement,* which can be defined as consistent evaluation of the truth value of the statement and the existence of counterexamples. Understanding generality is essential for the comprehension of mathematical statements and students' conceptions of proof, because it is the mathematical generality that is the defining element of mathematical proof and what distinguishes mathematics from other disciplines (Heintz, 2000).

The present thesis therefore aims at investigating the understanding of the generality of mathematical statements in first-year university students and its relation to the reading and construction of proofs. Previous studies have shown differences in students' understanding and evaluation of *different types of arguments* (e.g., Healy & Hoyles, 2000; Kempen, 2018, 2021; Tabach, Barkai, et al., 2010). Thus, I focused on the influence of reading different types of arguments (no argument,

empirical argument, generic proof, and ordinary proof) on students' understanding of generality and other proof-related activities. Additionally, I considered the familiarity with the statement and its truth value as important characteristics that might also influence students' performance in proof-related activities, as suggested in the literature (e.g., Barkai, Tsamir, Tirosh, & Dreyfus, 2002; Dubinsky & Yiparaki, 2000; Hanna, 1989; Stylianides, 2007; Weber & Czocher, 2019).

I designed an experimental study to analyze the effect of reading different types of arguments on students' understanding of the generality of statement, estimation of truth, proof comprehension, and conviction. The experiment was conducted online during two first-semester lectures in November 2020. 430 preservice teachers and mathematics students from a German university completed the questionnaire. They were randomly divided into four groups. All participants were asked questions about the same five universal statements (two of them familiar, two of them unfamiliar, and one of them false). The first group received no arguments to justify the statements, but was asked to produce justifications themselves. These justifications were analysed using a coding scheme mainly based on the categories suggested by Harel and Sowder (1998). The second group was provided with empirical arguments, the third group with so-called generic proofs, and the fourth group with ordinary proofs (those that are typically constructed by mathematicians). The data was analyzed using mainly generalized linear mixed models.

The results show that in a comparatively large percentage of observations (about one third), students lacked understanding of the generality of mathematical statements, that students with a correct *knowledge* of mathematical generality are more likely to have a correct understanding of the generality of statements, and that students' usage and conviction of empirical arguments is negatively related to their understanding of the generality of statements. Further, students' level of conviction of the truth of statements is related to the reading and construction of different types of arguments. In particular, empirical arguments (and to a lesser degree generic proofs) support students in successfully estimating the truth value of true universal statements–but ordinary proofs do not. My findings also provide evidence for the influence of familiarity with the statement and the truth value on students' understanding of the generality of statements and performance in proof-related activities.

Results confirm prior research findings that students lack basic mathematical knowledge and therefore have difficulties with proof comprehension and evaluation (e.g., Dubinsky & Yiparaki, 2000; Harel & Sowder, 1998; Healy & Hoyles, 2000; Kempen, 2019; Recio & Godino, 2001; Weber, 2001), that most students'

find generic and, in particular, ordinary proofs convincing (e.g., Kempen, 2021; Ko & Knuth, 2013; Weber, 2010), and that many students use empirical arguments to justify universal statements (Barkai et al., 2002; Bell, 1976; Healy & Hoyles, 2000; Lee, 2016; Recio & Godino, 2001).

These findings can be used to develop future university courses in a manner that eases and promotes the transition to proof-based mathematics. I present several suggestions for future research on students' proof skills and, in particular students' understanding of the generality of statements. Lastly, this thesis further highlights the benefits of and need for more experimental studies in mathematics education and in particular in research on proof and argumentation.

Contents

List of Figures

List of Tables

Introduction

1

Mathematical research generally results in statements that are–once they have been validly *proven*–true without any exceptions. In this respect, mathematics differs from other sciences: The *unrestricted* generality and *absolute* conviction of truth as well as the method, namely proof, with which respective results are achieved, are unique (e.g. Poincaré, 1952; Toulmin, 2003). Mathematics is therefore sometimes called a *proving science* (in German *beweisende Wissenschaft*), particularly in German literature (e.g., Dreher & Heinze, 2018; 2000; A. Heinze, Anderson, & Reiss, 2004; Hilbert, Renkl, Kessler, & Reiss, 2008; Kirsten, 2021; Reiss & Ufer, 2009). Because of its fundamental role in mathematics, *proof* has been a research focus in philosophy of mathematics and mathematics education for many decades, but the research area saw a particular rise of interest in the last 10–15 years (Sommerhoff & Brunner, 2021). Even though proof and argumentation are internationally seen as important learning goals in mathematics and are incorporated in many national curricula (e.g., Department of Basic Education, 2011; Kultusministerkonferenz, 2012; National Council of Teachers of Mathematics, 2000), students seemingly do not gain sufficient experience with proof during high school (e.g., Hemmi, 2008; Kempen & Biehler, 2019). Consequently, students of different school levels and forms seem to lack fundamental proof skills and understanding of proof (e.g., Dubinsky & Yiparaki, 2000; Harel & Sowder, 1998; Healy & Hoyles, 2000; Kempen, 2019; Recio & Godino, 2001; Weber, 2001). Difficulties with proof are diverse and include insufficient proof comprehension and difficulties with proof construction and validation (a good overview can be found in Reid & Knipping, 2010, pp. 59–72, for instance). The lack of proof skills is particularly relevant for students entering university, because in many countries it coincides with the transition to proof-based mathematics. Students' insufficient proof skills and understanding are in fact often identified as main reasons for students' difficulties with mathematics at the transition from school to university (e.g., Gueudet, 2008; A. Selden, 2012). High drop out rates

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_1

in mathematics, in particular compared to other fields, seem to be one of the consequences (e.g., Dieter, 2012; Heublein et al., 2022). Research on university students' proof skills has therefore increased significantly over the past 30 years, especially at the transition (e.g., Alcock, Hodds, Roy, & Inglis, 2015; Gueudet, 2008; Kempen & Biehler, 2019; Moore, 1994; Rach & Ufer, 2020; Recio & Godino, 2001; A. Selden & Selden, 2003; Sommerhoff, 2017; Stylianides & Stylianides, 2009; Stylianou, Chae, & Blanton, 2006).

While a large body of research has focused on students' proof construction, the reading of proof, which includes proof evaluation and comprehension, is still under-researched (e.g., Mejía Ramos & Inglis, 2009a; Sommerhoff, Ufer, & Kollar 2015). For instance, studies have repeatedly provided evidence that many students find *empirical arguments* convincing (e.g., Gholamazad, Liljedahl, & Zazkis, 2004; Healy & Hoyles, 2000; Knuth, 2002; Martin & Harel, 1989; Segal, 1999), but most do not consider them to be *proofs* (e.g., Healy & Hoyles, 2000; Lesseig, Hine, Na, & Boardman, 2015; Stylianou, Blanton, & Rotou, 2019; Tabach, Levenson, et al., 2010). However, it is unclear and therefore an open research question what *level of conviction* students gain by reading or constructing these types of arguments (Weber & Mejia-Ramos, 2015). Moreover, research on aspects that influence students' conviction, for instance, understanding the argument, the perceived generality of the argument, or being familiar with the type of argument, is still scarce (Ko & Knuth, 2013; Sommerhoff & Ufer, 2019). Further, only few studies have investigated students' proof comprehension and further research is needed to find ways to better assess it and to identify specific difficulties students encounter when trying to comprehend a proof (e.g., Mejía Ramos, Fuller,Weber, Rhoads, & Samkoff, 2012; Neuhaus-Eckhardt, 2022).

To make proofs more accessible to students, different types of arguments have been considered in research on proof comprehension as well as proof evaluation. In particular, *generic proofs* are considered to be potentially useful in the learning of proof and argumentation (Dreyfus, Nardi, & Leikin, 2012; Mason & Pimm, 1984; Rowland, 2001). However, little is known yet regarding the influence of the type of argument on students' understanding of proof (see, e.g., Lew, Weber, & Mejía-Ramos, 2020; Malek & Movshovitz-Hadar, 2011; Mejía Ramos et al., 2012).

In research on students' proof construction and reading, the comprehension of the statements which are to be proven–or for which a proof has to be read–has largely been neglected, even though it can be considered as a prerequisite for proof comprehension. The comprehension of a statement involves *understanding the statement's generality*, which is defined in this study as consistent evaluation of both the truth value of the statement and the existence of so-called counterexamples. Understanding generality is essential for the comprehension of mathematical statements and

555440_1_En_1_Chapter-online ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/1/2024 Pages: 207 Layout: German_A5

students' proof skills, because it is the mathematical generality that is the defining element of mathematical proof and what distinguishes mathematics from other disciplines, as already mentioned above. Further, without understanding the generality of statements it might be difficult to develop an *intellectual need* for proof (see, e.g., Harel, 2013). As such, the importance of understanding generality has repeatedly been emphasized in the literature (e.g., Conner, 2022; Ellis, Bieda, & Knuth 2012; Fischbein, 1982; Kunimune, Kumakura, Jones, & Fujita, 2009; Lesseig et al., 2019). However, virtually no studies have explicitly investigated students' or teachers' understanding of the generality of *statements*. The few studies that have been conducted so far in this direction, most of them qualitative, relate students' understanding of the generality of mathematical statements to the understanding of the generality of proof. For instance, some researchers have reported that students and (preservice) teachers, who seemed to be absolutely convinced of the truth of a statement and the correctness of its proof, were at the same time not fully convinced that no counterexample to the statement can exist (Chazan, 1993; Knuth, 2002). Others have investigated students' awareness that one counterexample *disproves* a universal statement (Buchbinder & Zaslavsky, 2019; Galbraith, 1981) or that a proof holds for any given subset of cases (Healy & Hoyles, 2000). With respect to the understanding of the generality of proof, some researchers argue that students' usage or conviction of empirical arguments may indicate an insufficient understanding of generality (e.g., Conner, 2022). However, this relation has not been explicitly investigated so far, in particular, with respect to the understanding of generality of statements. Overall, neither the extent to which students lack understanding of the generality of mathematical statements nor the specific relations to the construction, evaluation, and comprehension of proofs have explicitly been researched yet.

The main goal of the present study is therefore to investigate the understanding of the generality of mathematical statements and the relation to the reading and construction of proofs in first-year university students. Previous studies have shown differences in students' understanding and evaluation of *different types of arguments* (e.g., Healy & Hoyles, 2000; Kempen, 2018, 2021; Tabach, Barkai, et al., 2010). Thus, I focused on the influence of reading different types of arguments (no argument, empirical argument, generic proof, and ordinary proof) on students' understanding of generality and other proof-related activities. Additionally, I considered the familiarity with the statement and its truth value as important characteristics that might also influence students' performance in proof-related activities, as suggested in the literature (e.g., Barkai et al., 2002; Dubinsky & Yiparaki, 2000; Hanna, 1989; Stylianides, 2007; Weber & Czocher, 2019).

To investigate students' understanding of generality, proof comprehension, evaluation and construction, their relations, and, in particular, the influence of the type

555440_1_En_1_Chapter-online ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/1/2024 Pages: 207 Layout: German_A5

of argument and statement on students' performance in these activities, I designed an experimental study which was conducted online during two first-semester lectures in November 2020. 430 preservice teachers and mathematics students from a German university completed the questionnaire. They were randomly divided into four groups. All participants were asked questions about the same five universal statements (two of them familiar, two of them unfamiliar, and one of them false). The first group received no arguments to justify the statements but was instead asked to produce justifications themselves. The second group was provided with empirical arguments, the third group with so-called generic proofs, and the fourth group with ordinary proofs (those that are typically constructed by mathematicians). I analyzed the data using mainly generalized linear mixed models. Results confirm prior research findings that students lack basic mathematical knowledge and therefore have difficulties with proof comprehension and evaluation, and that many students use empirical arguments to justify universal statements. They further extent findings in that a comparatively large percentage of students lacks understanding of the generality of mathematical statements, that students level of conviction of the truth of statements is related to the reading and construction of different types of arguments, and how the familiarity with the statement and the truth value influence students' understanding of generality and performance in proof-related activities. Furthermore, the results of this study suggest relations between students' proof evaluation, proof comprehension, the estimation of truth, and their comprehension of statements, particularly their understanding of the generality of statements. Based on these findings, implications for future university courses, especially at the transition from school to university, and directions for future research can be derived.

The theoretical basis for the conceptualizations of understanding generality and proof is build in chapter 2 by highlighting characteristics of mathematical statements and generality, in particular how mathematical generality differs from generality in other sciences, how proof is shaped historically by so-called socio-mathematical norms, and different views and characteristics of proof. Further, this chapter also discusses usages and relations of the terms *reasoning, argumentation,* and *proving*, before different types of arguments widely used in mathematics education are introduced. The second theoretical chapter (chapter 3) provides an overview of the current state of research on students' understanding of generality of statements and proofs as well as proof-related activities relevant for this thesis. Thereby, the framework on proof-related activities proposed by Mejía Ramos and Inglis (2009b) is adapted and used for structuring the discussion of prior research findings. Resulting research desiderata are summarized in chapter 4, from which the research questions of the present thesis are then derived and specified. In chapter 5, the design of the study, the construction of instruments, and the collection and analysis of data are

555440_1_En_1_Chapter-online ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/1/2024 Pages: 207 Layout: German_A5

thoroughly described and justified. Subsequently, the results are presented comprehensively in chapter 6, guided by the research questions. Research findings are interpreted and discussed in detail with respect to prior research in chapter 7, where I also reflect on methodological decisions and the adapted framework on proof-related activities, identify limitations, and outline implications for the teaching of proof and further research. Lastly, in chapter 8, I conclude the thesis with a short summary and outlook regarding theoretical and practical implications.

555440_1_En_1_Chapter-online ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:29/1/2024 Pages: 207 Layout: German_A5

Theoretical Background of Mathematical Statements and Proof

2

Proof is fundamental in mathematical practice. In contrast to empirical sciences, mathematical research does not only provide evidence, but generally results in statements that are *proven* to be true or false in a particular axiomatic system. An understanding of statements and the underlying theories are vital for learning about proof, as Mariotti (2006) emphasizes: "It is not possible to grasp the sense of a *mathematical proof* without linking it to the other two elements: a *statement* and overall a *theory*" (p. 184). The understanding of mathematical statements, in particular their *generality*, as well as the acceptance and knowledge about the theoretical framework from which the *truth* of the statement is drawn, can both be seen as necessary prerequisites for students' understanding and learning about proof (see also Balacheff, 2010).

In this chapter I first explain the meaning of two main types of *quantified statements*: universal statements and existential statements. Secondly, the meaning of generality in mathematics, other sciences, and everyday life, as well as potential difficulties for students with the concept of *mathematical generality* in this regard are discussed. While the terms *proof, proving, argumentation,* and *reasoning* are widely and often used in mathematics education, there is no clear consensus about their specific meaning and distinction. I therefore review and discuss different notions, usages, and views of these terms as well as the relation between them. Finally, I review different types of argumentations that are commonly being distinguished in mathematics education.

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_2

2.1 Mathematical Statements

A mathematical statement is a declarative sentence that is either true or false. Thereby, it is not necessary that the truth value is known yet (or will ever be known). For example, consider the following two statements:

- Every even number greater than 2 can be represented as the sum of two primes.
- There is an even number greater than 2 that cannot be represented as the sum of two prime numbers.

Both sentences are either true or false and thus qualify as mathematical statements. The second statement is the logical contradictory statement of the first and therefore, only one of these statements can be true; it is still an open problem which one[1] (e.g., Schütte, 1977).

Like the two statements above, most mathematical theorems can take one of the following two forms[2]:

- For all objects x within a given domain D, property $P(x)$ holds.
- There is at least one object x within a given domain D, for which a property $P(x)$ holds.

Statements of the first category are called *universal statements* (sometimes also *general statements*). They can be written formally, i.e., in first-order logic using the *universal quantifier*: $\forall x \in D : P(x)$. Universal statements, when expressed informally (i.e., using natural language), often contain words like *(for) all, every, each, always* etc. Examples for universal statements include "For all odd numbers a and b, $a + b$ is even", "Prime numbers greater than 2 are always odd", and "Every human is mortal" (even though the latter might not be provable, unless we define humans as mortal). Statements of the second category are called *existential*

[1] The first statement is the famous *Goldbach's conjecture*. It has been tested in the scope a computer can currently handle, i.e., it holds for all even numbers up to at least 4×10^{18} (Oliveira e Silva, Herzog, & Pardi, 2013) and no counterexample has been found. Still, it is not known yet, if it holds *for all* even numbers (e.g., Reiss & Schmieder, 2014).

[2] Another common type of statement is called *conditional statement*, which can formally be written with *logical implication* as $P(x) \Rightarrow Q(x)$ ("If property $P(x)$ holds, then does property $Q(x)$"). Most conditional statements are also universal, i.e., of the form $\forall x \in D : P(x) \Rightarrow Q(x)$. Although conditional statements are very common and important, especially regarding proving practices, I do not go into more details, because the focus of this theses is on *quantified statements*, in particular on universal statements.

statements and can be represented formally with the *existential quantifier*: $\exists x \in D : P(x)$. Informally written, existential statements use expressions like *there is/exists, at least, (for) some* etc. For instance, "There exists a natural number n, such that $n^2 = 10$", "At least one prime number is even", and "Some dolphins are pink" are existential statements (and only one of them is false).

There are further subcategories of universal and existential statements. The property P of a universal statement can also involve an existential quantifier and vice versa, as the following examples illustrate:

- For all integers, there exists an additive inverse.
- There exists a natural number, such that all natural numbers are greater or equal to that number.

The first statement is an example for a so-called *universal existential statement* (since the property that holds for all given objects, in this case integers, is about the existence of an object), while the second example is an *existential universal statement* (e.g., Dubinsky & Yiparaki, 2000; Piatek-Jimenez, 2010). The involvement of existential quantifiers in universal statements (and vice versa) often only becomes apparent when the statement is expressed formally. For example, the universal statement "For all odd numbers a and b, $a+b$ is even" can formally be written as $(\forall a, b \in \{z \in \mathbb{Z} | \exists k \in \mathbb{Z} : z = 2k+1\})(\exists l \in \mathbb{Z})(a+b = 2l)$, and thus, it is more precisely a universal existential statement. Regardless how a statement is expressed, the potential presence of the existential quantifier in a universal statement becomes particularly relevant for proving it; in the example above, because of the definition of even (and odd) numbers:

> Since a and b are odd, there exist $k, m \in \mathbb{Z}$, such that $a = 2k + 1$ and $b = 2m + 1$. We therefore get: $a + b = (2k + 1) + (2m + 1) = 2k + 2m + 1 + 1 = 2k + 2m + 2 = 2(k + m + 1)$. Since $k + m + 1$ is a whole number (thus, we found the required $l \in \mathbb{Z}$), $a + b = 2(k + m + 1)$ is divisible by 2 and therefore even.

The potential involvement of an existential quantifier in a universal statement (and vice versa) could be a particular obstacle for students in reading (or constructing) proofs, which teachers and lecturers should keep in mind.

The thesis at hand particularly aims at investigating first-year university students' understanding of the generality of mathematical statements, more precisely, the understanding of the fact that *if a universal statement is true, it is true without any exceptions*. A universal statement would be *disproved* (or *refuted*) by finding just one *counterexample*. Thus, if someone is completely convinced that a universal

statement is true, the person—if having a correct understanding of the generality of statements—should be just as convinced that no counterexample to the statement exists.

This type of *generality* is characteristic for mathematics. In the following section, its particular meaning in contrast to other disciplines is clarified.

2.2 Generality in Mathematics

The significance of *generality* for mathematics and science in general was already recognized in ancient Greece. A corresponding criterion for generality can be found in Aristotle's *Posterior Analytics*, according to which "any truly scientific demonstration should hold 'for all' the entities it concerns" (Rabouin, 2016, p. 113). Consequently, progress in mathematics commonly has meant to achieve "ever higher levels of generality" (Chemla, Chorlay, & Rabouin, 2016, p. 3). However, there is no uniform meaning of (mathematical) generality. It can refer to a definition, a theorem, a method, or a type of reasoning, and different mathematicians have used different approaches to achieve different types of generality (Chemla et al., 2016). Since the focus of this thesis is on universal statements, I refer to mathematical generality (in German "Allgemeingültigkeit") as the property of a statement holding *for all* objects of a given domain. In most cases, the expression "for all" in a mathematical statement refers to an infinite domain; and, according to Poincaré, the ability to obtain such a generality—with "the power of the mind" (Poincaré, 1952, p. 13), that is by the construction of mathematical concepts and proof—is specific to mathematics (Ly, 2016). However, as stated by Aristotle, *generalizing* is a goal aimed at in other sciences as well. What might come closest to the form of mathematical, *unrestricted* generality is that of physical laws, e.g. *Newton's laws*. But even these "cannot be demonstrated by conclusive reasoning" (Kneale, 1949, p. 21). The form of generality researchers seek in other sciences (e.g., biology and chemistry) is fundamentally different (Toulmin, 2003).

> Biologists, for example, ... may not pursue formulations of unrestricted generality, but they are deeply committed to the search for formulations that we might describe as being of 'restricted' generality. Indeed, their science abounds in claims about general properties, and even statements that are often referred to as 'laws' .... But there is no domain in which these laws are presumed to be *exception-free* [emphasis added]. They are generalities, but not unrestricted generalities. It is evident that generality is valued in biology, but exceptions are neither a cause for alarm, nor do they necessarily send researchers back to the drawing board in search of better—exception-free—laws. Rather, they are reminders of how complex biological reality is. (Keller, 2016, p. 474)

Restricted generality is also particularly present in sciences such as sociology, where results are mostly being achieved by providing empirical (sometimes experimental) evidence for them. These conclusions hold *in general*, usually meaning they "have only high probability" (Kneale, 1949, p. 21). But there can be—and generally are—exceptions.

The assumption that it is commonly considered normal or even expected that there are exceptions to a (general) rule or statement can be further affirmed by the usage of the saying "The exception that proves the rule" (in German: *Ausnahmen bestätigen die Regel*). It "is often taken in the paradoxical sense of asserting that the presence of a counterexample establishes the *general truth of a rule* [emphasis added]" (Reid & Knipping, 2010, pp. 25–26), even though this would, in mathematics, *disprove* a rule or statement. The actual meaning of the expression might, however, be different. Since to "prove" comes from the Latin verb "probare", which means to test, the expression can be interpreted as "the exception that *tests* the rule", which suggests "that examining exceptions closely and reasoning out the way they occur can lead to a clarification and improvement of the rule." (Reid & Knipping, 2010, p. 26).

In summary, statements that hold *exception-free* in a(n infinite) domain are unique for mathematics. The fact that universal statements are very rare (or nearly non-existent) outside of mathematics and the common view that there are *typically* exceptions to a rule, might influence students' understanding and acceptance of the generality of mathematical statements. Furthermore, for obtaining these types of general results, proof is essential. Understanding this need for proof might not be obvious for students. Because of its essential role and further relevance for this project, I discuss the meaning of proof and its different notions in the following section.

2.3 What is Proof?

In mathematics, statements are only accepted as *true* after they have been proven in a way that meets certain standards. These standards vary over time and are based on so-called socio-mathematical norms (Dawkins & Weber, 2017; Yackel & Cobb, 1996), as Wilder (1981) emphasizes: "'Proof' in mathematics is a culturally determined, relative matter. What constitutes proof for one generation, fails to meet the standards of the next or some later generation" (p. 40). Even though there is no consistent definition of proof, certain characteristics and acceptance criteria can be identified.

In the following, I first outline major historical developments in the context of proof (for a detailed historical overview, see, e.g., Reid & Knipping, 2010). This provides the required background to better understand and reflect the different views

and usages of *proof* in recent literature, which I review in the following section. Furthermore, researchers have highlighted the potential relevance of historical developments of proof for the development of students' proof conceptions (e.g., Harel & Sowder, 2007). Lastly, main characteristics and acceptance criteria for mathematical proof are summarized and discussed.

2.3.1 Brief Historical Background

The idea to *prove* a statement—and not just provide evidence for it—originated in ancient Greece around 500 BCE (Reid & Knipping, 2010; Wußing, 2008):

> For the early Egyptians, Babylonians, and Chinese, the weight of observational evidence was enough to justify mathematical statements .... But the classical Greek mathematicians found this way of determining mathematical truth or falsehood less than satisfactory (Hanna & Barbeau, 2002, p. 36)

They started to agree on definitions of fundamental ideas and axioms, on which they based their reasoning. In the 4th century BCE, the Greek philosopher Aristotle formulated what we now call the (axiomatic) deductive method in his *Posterior Analytics* (Anglin, 1994). The deductive method is a process of reasoning where each argument "has to be justified either by an axiom or by a previously proved theorem or by a principle of logic." (Anglin, 1994, p. 63). The method shaped mathematics significantly, as it has become its defining characteristic (Anglin, 1994; Harrison, 2008; Reid & Knipping, 2010). Euclid's *Elements*, which is a structured collection of the fundamental mathematical ideas of that time, is in this regard often considered as the most influential work of mathematical literature (Reid & Knipping, 2010; Wußing, 2008).

European mathematics can mainly be seen as a continuation of the work of the Greeks (Reid & Knipping, 2010). The 17th century CE "saw an explosion of mathematical activity" (Anglin, 1994, p. 161), which led to the discovery of important results, for example, by Newton and Leibniz. However, the methods often did not meet the strict standards for proof (Reid & Knipping, 2010). Many mathematicians at that time were unsatisfied with these methods. But in the case of calculus, it lasted until the 19th century until a foundation based on precise definitions was established (Reid & Knipping, 2010).

In the late 19th and early 20th century, a demand for the formalization of mathematical statements and proofs arose (e.g., Ketelsen, 1994; Reid & Knipping, 2010; Wußing, 2009). A first successful step in this direction was due to Frege's *Begriff-*

sschrift from 1897, in which he developed axiomatic predicate logic (e.g., Sjögren, 2010). Further important developments include: Peano's *Formulaire de Mathématiques*, which present a purely formal structure for fundamental parts of mathematics; Russel's and Whiteheads's *Principia Mathematica*, an (unsuccessful) attempt to build a complete foundation of mathematics based on axioms and logical rules of inference; the axiomatic set theory of Zermelo and Fraenkel (which was further extended by v. Neumann); and *Hilbert's Program* (Ketelsen, 1994; Wußing, 2009). The latter was intended to completely formalize mathematics in axiomatic form and prove the consistency of this axiomatization (Wußing, 2009). Even though Kurt Gödel's incompleteness theorems proved the impossibility of this endeavor as a whole[3], it had a significant impact on the development of mathematical logic and proof (e.g., Ketelsen, 1994; Wußing, 2009). As a consequence, these developments made it possible—as intended by Hilbert—to formulate and answer metamathematical questions within mathematics itself; new mathematical fields such as proof theory emerged (Rav, 1999; Sjögren, 2010; Zach, 2019). Furthermore, a more formal view and approach to solve mathematical problems has since been established (Ketelsen, 1994). Regarding the future development of proof, Harrison (2008, p. 1395) argues that formalizing mathematics is a "natural further step ... towards greater clarity and precision."

2.3.2 Different Views and Usages of the Term *Proof*

As the short historical overview emphasizes, the concept of proof has developed over time (and will most likely do so in the future) and depends on social norms specified by the mathematical community. In this section, two main views of proof which have grown historically are discussed.

One common view about *mathematical proof* shared by many mathematicians, mathematics teachers, and students is that of a so-called *formal proof*[4] in the sense of Hilbert (Lakatos, 1978; Tall et al., 2012; Weber, 2003). In this sense, "a mathematical proof is a formal and logical line of reasoning that begins with a set of axioms and moves through logical steps to a conclusion" (Griffiths, 2000, p. 2). Formal proofs can be expressed using first-order logic (e.g., Rav, 1999) and have

[3] Gödel himself was surprised by his results, as it seems that he did not attempt to refute Hilbert's program, but to prove the consistency of arithmetic (Ketelsen, 1994).

[4] Sometimes, the term *formal proof* is used to describe "proofs used by mathematicians to communicate to each other in conversation and in journal articles" (Tall et al., 2012, p. 29). Whenever I use the term *formal proof*, I refer to a precise form of proof in the sense of Hilbert described here.

the property that no interpretation by the reader is necessary to verify their validity. In fact, it is possible to mechanically check the correctness of the proof via computer programs in finite time (Harrison, 2008; Lakatos, 1978; Rav, 1999). Because of these properties, formal proofs ensure rigor and reliability. They can be viewed as an *idealization of proof* (Hersh, 1993; Jahnke & Ufer, 2015; Manin, 2010; Sjögren, 2010; Sommerhoff, 2017). However, because of their properties, formal proofs are at the same time incredibly long and very difficult to read, which makes them generally useless for most mathematical fields (CadwalladerOlsker, 2011; Hanna, 1989; Jahnke & Ufer, 2015; Weber, 2003). In fact, most published proofs, for example, in textbooks and articles of mathematical journals, are not purely formal as they are not completely expressed symbolically in first-order logic and not all logical steps have explicitly been checked back to axioms (e.g., Rav, 1999; Tall et al., 2012). Aberdein (2009) further emphasizes that "this [referring to formalizing] is not something that mathematicians routinely do" (p. 1). The main mathematical fields in which formal proofs do play an important role are mathematical logic (specifically proof theory) and foundation of mathematics.

Because the majority of proofs are not (yet) completely formal[5], Thurston (1991) and other modern mathematicians and mathematics educators argue that it is important to explicitly "distinguish between formal proofs and proofs that mathematicians actually construct" (Weber, 2003, para. 2) and publish. The latter can be described as "social conventions by which mathematicians convince one another of the truth of theorems" (Buss, 1998, p. 2). They "are written in a way to make them easily understood by mathematicians" (Hales, 2008, p. 1371). In contrast to formal proofs, routine steps are omitted in these proofs and readers have to interpret the context and translate intuitive arguments into more rigorous ones (Hales, 2008). There is no clear consensus in the literature whether or not these proofs are or should be formalizable, at least theoretically. Bass (2009) argues that for mathematicians, an argument is convincing if their peer experts feel "empowered [...], given sufficient time, incentive, and resources, to actually construct a formal proof" (p. 3). However, Lakatos (1978) gives an example for a *proof* of Euler's theorem on simple polyhedra, about which he states that "there does not seem to be any feasible way to formalize this reasoning" (p. 64). Nevertheless, he is convinced "that mathematicians would accept this as a proof" (Lakatos, 1978, p. 64).

By renouncing rigor and complete formalism, these proofs enable mathematicians to focus on understanding and imparting the underlying mathematical concepts and ideas. However, the degree of (in)formality and the use of intuitive arguments

[5] It is possible that this might change in the future, given that computer(-assisted) proofs recently have become more relevant (Dawson, 2015; Hales, 2008; Harrison, 2008).

varies a lot in these proofs. For example, proofs in (abstract) algebra are usually more formal than those in geometry or topology, which quite often rely on intuitive arguments (Hales, 2008; Sjögren, 2010).

In the last decades, various terms have been introduced to separate formal proofs and proofs that mathematicians actually construct. For instance, Douek (2007) distinguishes between *mathematical proofs* and *formal proofs*. However, the usage of the term mathematical proof is not consistent in the literature, which can be confusing if the context is unclear. In contrast to Douek, Griffiths (2000) means formal proofs when using the term mathematical proof, as already noted above. The term mathematical proof is also often used more generally as a superset, which comprises *any kind of proof*, thus including formal proofs (e.g., CadwalladerOlsker, 2011; Lakatos, 1978; Sjögren, 2010; Tall et al., 2012).

Many classifications in the literature highlight two main contrasting purposes of formal proofs and proofs mathematicians actually construct: gaining absolute truth vs convincing others and understanding the underlying mathematics. In this sense, Davis, Hersh, and Marchisotto (2012) differentiate between proofs of *metamathematics* and *real mathematics*; and Recio and Godino (2001) between *foundation of mathematics* and *mainstream mathematics*. Proofs of the latter are sometimes called *mainstream (mathematical) proofs* (e.g., Harrison, 2008), *ordinary (mathematical) proofs*[6] (e.g., Tall et al., 2012) or *practical (mathematical) proof* (e.g., Hersh, 1997), to emphasize that these are the sort of proofs commonly produced in mathematical practice. Hales (2008) uses the term *traditional (mathematical) proofs*, which refers to the historical development of proof and the fact that purely formal proofs only recently became more relevant in mathematical research (see Section 2.3.1). To highlight the influence of socio-mathematical norms, Buss (1998) uses the term *social proof*. However, the usage of this term has not (yet) been established in the literature, especially in mathematics education.

Another way to distinguish ordinary proofs from formal proofs is the use of terms, which refer to a lesser degree of formality. The usage of these terms is not always consistent. In (philosophy of) mathematics literature, ordinary proofs are commonly called *informal proofs* as the opposite of formal proofs (Dawson, 2006; Marfori, 2010; Sjögren, 2010; Tanswell, 2015), even though ordinary proofs are usually not completely informal. Lakatos (1978) criticizes that often proofs are misleadingly called *informal*, even though "a competent logician ... can formalize any such proof without too much brain-racking" (p. 63). He suggests to call proofs that are not completely formal, but formalizable *formal proofs with gaps* or *quasi-*

[6] In the following, I use the term *ordinary proof* whenever I refer to proofs commonly constructed by mathematicians.

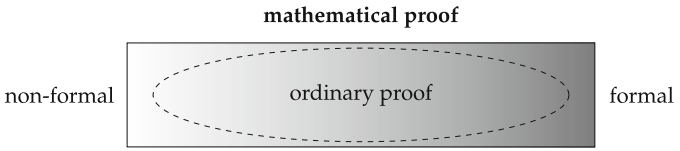


Figure 2.1 Spectrum of mathematical proof by degree of formality

formal proofs, because they are simply "incomplete formal proofs" (Lakatos, 1978, p. 63). To describe proofs that are "truly informal", Lakatos uses the term *pre-formal*[7]. To emphasize a lesser degree, but not complete absence of formality, Reid and Knipping (2010) call ordinary proofs *semi-formal*.

I understand mathematical proof as a deductive form of argumentation, accepted by the mathematical community, that comprises a spectrum of formality from purely formal to the complete absence of formality, which I refer to as non-formal. I assume non-formal proofs are rather rare and most published mathematical proofs (the *ordinary* ones) can be assumed to be neither non-formal nor formal but somewhere in between (see Figure 2.1). All these definitions of ordinary proofs are rather broad and vague. While formal proofs can be precisely defined, it is indeed not possible to exactly define what constitutes a valid ordinary proof (Buss, 1998; Davis et al., 2012; Lakatos, 1978; Sjögren, 2010). One reason already mentioned above is the significance of socio-mathematical norms and expectations of the mathematical community for proving practices, which the following well-known quote from Manin (2010) illustrates: "A proof [in the social sense] only becomes a proof after the social act of 'accepting it as a proof'" (p. 45). As already noted at the beginning of this chapter, these "standards of acceptability are changeable and subject to different constraints which vary according to different variables" (Mariotti, 2006, p. 176).

The lack of a precise definition for (ordinary) proof complicates deriving clear instructional implications for the teaching of proof. However, several acceptance criteria of the mathematical community and characteristics of proofs, which are seen to be useful for the teaching and learning of proof and argumentation, have been discussed in the literature. These are outlined in the following section.

[7] He actually divides "truly informal" proofs into two kinds: pre-formal and *post-formal* proofs. The latter are proofs of metamathematical statements, e.g., Gödel's proofs of his Incompleteness Theorems (see Lakatos, 1978; Reid & Knipping, 2010).

2.3.3 Characteristics and Acceptance Criteria for Proof

Even though a precise definition for (ordinary) proof does not exist, it is important to agree on a conceptualization within mathematics education; otherwise "it is difficult ... to meaningfully build upon each other's research and it is impossible to judge if pedagogical goals related to proof are achieved" (Weber, 2014, p. 353). However, such an agreement has not been reached yet (Balacheff, 2002; Reid & Knipping, 2010; Weber, 2014). Although there is no uniform conceptualization of proof, a shared understanding seems to be that the definition of formal proof is not very useful for mathematics education (see, e.g., CadwalladerOlsker, 2011; Hanna, 1989; Weber & Czocher, 2019), because (1) the arguments mathematicians refer to as proofs, i.e., ordinary proofs, are in general not formal (with good reasons); and (2) the underlying concepts and ideas we want students to understand are disguised in formal proofs.

In 2007, Stylianides proposed a characterization of proof[8], which is often cited and used in mathematics education research. He views proof as a mathematical argument containing...

1. ... a *set of accepted statements*, e.g., definitions, axioms, theorems, etc.;
2. ... valid and known forms of reasoning (which he calls *modes of argumentation*), e.g., application of logical rules of inferences, use of definitions, construction of counterexamples, etc.;
3. ... appropriate and known forms of expression (which he calls *modes of argument representation*), e.g., linguistic, physical, pictorial, symbolic, etc.

According to this definition, proof highly depends on the context and individuals, who construct or evaluate the proof, and thus it corresponds to a social view on proof. In particular, the familiarity with different aspects of proof (*known* forms of reasoning and expressions) seems to be an essential characteristic for the acceptance of proof in Stylianides' definition.

Weber (2014) views proof slightly differently. He argues that the frequently used approach of defining proof by identifying characteristics that are shared by *all* proofs, but not by other arguments has been unsuccessful, because there is no "consensus on which ... properties capture the essence of proof" (p. 353).

[8] Stylianides (2007) more specifically introduces a conceptualisation of proof for school mathematics. However, others (e.g., Weber & Czocher, 2019) have used his characterisation in a more general context.

Consequently, he suggests to view proof as a so-called *clustered concept* consisting of the following seven *models*: Proof as

1. a *convincing argument*.
2. a *transparent argument where a mathematician can fill in every gap*.
3. a *deductive argument*.
4. a *perspicuous argument that provides an understanding of why a theorem is true*.
5. an *argument within a representation system satisfying communal norms*.
6. an *argument that has been sanctioned by the mathematical community*.

Weber (2014) admits that these features have been stated before, but he claims it is original that according to his approach, none of "these more basic models" (p. 358) can completely characterize proof by themselves. However, he states that "it would be desirable for proofs to satisfy all six criteria". Furthermore, proofs that fit into all models should not be controversial; but some arguments that only fit into some of the models might either be disputed or can nevertheless qualify as proofs. The models 5. and 6. seem to correspond to the second and third property of the conceptualization given by Stylianides (2007). Further, the conceptualization of Weber (2014) contains several goals (or functions) of proof[9], in particular *proving as convincing* someone of the truth of a statement and *proving as explaining* (to provide insights of why the statement is true), which are both often identified as the main goals of proof (e.g., Hersh, 1993). The latter is often seen as particularly important regarding the teaching of proof and proving in school (e.g., Brunner, 2014; Hanna, 2000). In my understanding, the models are not all independent of each other. For example, models 2. to 6. can be viewed as influencing factors for being a convincing argument (i.e., for model 1).

Even though the two discussed approaches for conceptualizing proof differ, they still contain similar features. For instance, they both refer in some sense to convincing and accepted arguments. What remains unclear in this sense is what arguments *should* students find convincing or accept as proof? A common approach to answer this question is to investigate if (or to what degree) mathematicians agree on what forms of arguments and representations are valid and appropriate. In this regard, Weber and Czocher (2019) differentiate between two positions: "The consensus view on proof asserts that mathematicians agree on which inferential schemes are

[9] Because it is not of particular relevance for this thesis, I do not discuss all different functions of proof identified in the literature further. For more details see, for example, de Villiers (1990) and Hanna (2000).

permissible in a proof; the pluralistic view holds that mathematicians disagree on which inferential schemes are permissible" (p. 254). Hanna and Jahnke (1996) have argued that acceptance criteria that are shared by *all* mathematicians do not exist, which several studies confirm (e.g., Inglis & Alcock, 2012; Inglis & Mejía-Ramos, 2013; Weber, 2008, see also Section 3.2.3). However, Weber and Czocher (2019) found that there seem to be "three categories of inferential schemes": Standard methods in "typical proofs" mathematicians agree on; "invalid schemes" such as *empirical arguments*[10] on which mathematicians also agree; and "controversial schemes whose permissibility is unclear", for example, computer-assisted proofs and visual arguments (p. 264). Thus, disagreement might almost exclusively occur regarding *atypical proofs*. Apparently, the familiarity with the mathematical argument is indeed a major factor for its acceptance.

However, apart from the existence of a proof and the acceptance of it, there are other criteria that influence mathematicians' conviction of the truth of a statement. According to Hanna (1989), mathematicians accept new or unfamiliar theorems by a combination of the following criteria:

1. They understand the theorem, the concepts embodied in it, its logical antecedents, and its implications ...;
2. The theorem is significant enough to have implications in one or more branches of mathematics ...;
3. The theorem is consistent with the body of accepted mathematical results;
4. The author has an unimpeachable reputation as an expert in the subject matter of the theorem;
5. There is a convincing mathematical argument for it (rigorous or otherwise), of a type they have encountered before. (pp. 21–22)

Noteworthy, only the last criterion explicitly refers to proof. Moreover, the first three criteria highlight the importance of understanding the theorem and its implications for the acceptance of it. In line with Stylianides (2007), Hanna (1989) characterizes a convincing argument as one with which the reader is familiar, as she states that it is an argument "they have encountered before" (p. 22), i.e., in other proofs. This is of direct importance for the teaching of proof at the transition from school to university, because students usually do not gain extensive experience with proof and proving during high school (e.g., Hemmi, 2008; Kempen & Biehler, 2019). Thus, they might not have strong conceptions regarding what forms of reasoning and representations are appropriate. Research findings on students' conviction and acceptance of different types of arguments are reviewed in Section 3.2.3.

[10] Empirical arguments consist of empirical evidence for the claim, i.e., are based on examples. See Section 2.4.2 for further explanation.

The list above also emphasizes that the acceptance of an argument is indeed not a necessity for the acceptance of the truth of statement, but only one component. For instance, the fourth factor introduced by Hanna (1989), namely that of the *authors reputation* or *authority*, might also be of particular relevance regarding actual acceptance criteria of students. It can have a convincing power for them, when teachers or textbooks state that some statement is true (e.g., Harel & Sowder, 1998; Tall et al., 2012). Even more, students may ask themselves "why is it necessary to *prove* something that is *known* to be true?" (Tall, 1989, p. 29).

It seems that even mathematicians sometimes rely on *authoritarian arguments*, when estimating the truth of a statement, in particular, if the statement lies outside their field of specialty and high levels of uncertainty regarding a given argument are involved (Harel & Sowder, 1998; Inglis & Mejia Ramos, 2009). However, according to a study conducted by Heinze (2010), the reputation of an author does not often influence mathematician's conviction of the truth of a theorem (but *sometimes* it does). However, the theorem being checked and used by "other mathematicians with high standards" or the theorem existing "for a long time and no contradiction has been found" (p. 106) are both criteria mathematicians claim to employ, when deciding if a mathematical statement is true.

The different approaches and criteria reviewed in this and the previous sections illustrate the complexity of proof and the resulting challenges for teaching practices. In regard to the learning of proof and proving, *argumentation* and *reasoning* skills are often viewed as essential. Although these terms are used quite extensively in mathematics education, there is no shared understanding of their meaning. In the following section, different usages of these terms as well as their relation to each other and to *proving* are examined. In this regard, I also discuss several types of arguments that are common in mathematics education.

2.4 Reasoning, Argumentation, and Proving in Mathematics Education

The terms argumentation, reasoning, and proving or proof are widely used in mathematics education literature as well as in (national) curricula (e.g., Kultusministerkonferenz, 2012; National Council of Teachers of Mathematics, 2000). For instance, in the German national curricula ("Bildungsstandards") for higher secondary schools:

> This competence [referring to *mathematical argumentation*] includes both the development of independent, situation-appropriate *mathematical argumentations* [emphasis added] and conjectures and the understanding and evaluation of given mathemati-

> cal statements. The spectrum ranges from simple plausibility arguments to *operative reasonings* [emphasis added; in German *inhaltlich-anschauliche Begründungen*] to *formal proofs* [emphasis added] (Kultusministerkonferenz, 2012, p. 14, translated by the author)

As with the term "proof", there is no unique definition for the terms (mathematical) argumentation and reasoning, and a shared understanding regarding their relation to each other and to proving has not been reached yet in mathematics education (e.g., Balacheff, 1999; Reid & Knipping, 2010; Stylianides, 2016). In the following, I give a short overview of the main understandings of these terms and their relationship to emphasize differences and to avoid confusion. The terms and definitions on which the framework of this project is based are then clarified.

2.4.1 Definition of and Relation between Reasoning, Argumentation, and Proving

Among the terms reasoning, argumentation, and proving, *reasoning* may be the term least discussed in mathematics education literature. Two different understandings can nevertheless be identified: reasoning as the most fundamental activity of drawing conclusions (e.g., Duval, 1991) and reasoning as a specific form of argumentation (e.g., Hefendehl-Hebeker & Hußmann, 2003; Reiss, Hellmich, & Thomas, 2002). Following the first view, reasoning does not necessarily have to have the goal to convince someone of the truth of a (controversial) statement. Rather, it can simply be used to provide contextual explanation, e.g., to answer questions like "How have you come to be in a position to speak about [or know] this" (Toulmin, 2003, p. 199). The answer might consist of a *biographical* reason, for instance, "I know how to make toffee because my mother taught me" (Toulmin, 2003, p. 199), or a *reference to authority* (e.g., "My teacher said so").

In contrast, the second view is based on the understanding that argumentation rather than reasoning is the elementary activity of which reasoning is a specific form, namely one that is (logically) consistent (Reiss et al., 2002, p. 51). According to this view, the difference between reasoning and proving—the latter being the process of constructing a mathematical proof (e.g., Douek, 2007)—results from different modes of argumentation and degree of formality (Reiss & Ufer, 2009).

In contrast to reasoning, different views of the concept of *argumentation* have been discussed in detail (see, e.g., Kirsten, 2021; Mariotti, 2006; Pedemonte, 2007; Reid & Knipping, 2010). A shared understanding seems to be that argumentation (in general, not specifically mathematical) is a discursive activity, usually consisting of

a sequence of inferences rather than a single argument (e.g., Douek, 2007; Toulmin, 2003), being used to convince someone of the truth of a statement (Duval, 1999; Krummheuer, 1995).

There are several reasons for differences among researchers regarding their view of the relation between argumentation and proving, including different foci of characteristics of argumentation and different conceptualizations of proof. In this regard, Balacheff (1999) argues that different conceptions of argumentation can either lead to the conclusion of argumentation being an obstacle or a continuous path for the learning of mathematical proof. To "provide a system of benchmark" (Balacheff, 1999, p. 3), Balacheff compares the views of three authors: Perelman, Toulmin, and Ducrot. According to Perelman (1970), argumentation is not mainly about establishing "the validity of a statement" but about "its capacity to convince" (Balacheff, 1999, p. 3) someone. This view may contradict a *continuity position* of argumentation and proof. Because, even though it is one of its characteristics, proof—particularly in its formal sense—can most certainly not be reduced to *just* be convincing. On the other hand, in Toulmin's view, the main characteristic of argumentation is the reliance of its validity on a structure, accepted by a community (Toulmin, 2003). While Toulmin acknowledges different methods of argumentations being used in different fields (e.g., logic) and by "everyday arguers" (Toulmin, 2003, p. 37), his view of argumentation nevertheless contains main characteristics of mathematical proof. Therefore, argumentation and proof could be understood as a continuity, a view shared, for instance, by Boero, Garuti, and Mariotti (1996). Lastly, for Ducrot (1980) argumentation is the core of discourse and connecting words are essential for the imparting of an argument: "The analysis of conjunctions (connecting words) has a particular importance for Ducrot because it is they which make the information contained in a text subject to its global argumentative intention" (Balacheff, 1999, p. 3). Within Ducrot's framework, as with Perelman's, a continuous view of the relation between argumentation and proof "appears doubtful" (Balacheff, 1999, p. 4). Furthermore, such a conception might lead to the conclusion of argumentation being an obstacle for the learning of proof, as Duval (1991) highlights:

> Deductive thinking does not work like argumentation. However these two kinds of reasoning use very similar linguistic forms and propositional connectives. This is one of the main reasons why most of the students do not understand the requirements of mathematical proofs (p. 233)

While acknowledging similarities in linguistic and grammatical aspects, Duval understands argumentation and proof as two forms of reasoning—namely

argumentative reasoning and *deductive reasoning*[11]—but simultaneously as fundamentally separate activities. Regardless of whether or not one shares this view, it should not be dismissed that general argumentation (e.g., in everyday situations) might negatively effect students' understanding of argumentation in mathematics. For instance, as discussed in Section 2.1, only one counterexample disproves a universal statement. However, in other sciences and in everyday situations, not only would a counterexample *not* automatically refute the whole statement, it might even be expected that it exists (see Section 2.2).

The notion of *mathematical argumentation* is widely used in the literature and often includes not only activities regarding the verification of a statement, but a broader spectrum of mathematical activities, such as investigating conjectures and open-ended questions (e.g., Reiss & Ufer, 2009). However, the existence of *mathematical* argumentation as a particular form of argumentation being unique to mathematics, but not being part of mathematical proof, is seen controversial. It depends on the researchers view of both, argumentation and proof. Someone with a formal view of proof would most likely be able to find argumentations which do not qualify as a part of proving, but that are specific to mathematics, for example, visual arguments that are—as noted in Section 2.3.3—controversial. In contrast, someone with a broader (i.e., social or ordinary) view of mathematical proof might find it more difficult to identify such examples of argumentations. Balacheff (1999) adopts the latter position. He argues "that there is no mathematical argumentation in the frequently suggested sense of an argumentative practice in mathematics which is characterized by the fact that it escapes certain of the constraints present for mathematical proof" (p. 4). This does not imply that argumentation in mathematics does not exist, but these argumentative methods "could be used elsewhere" and would "disappear in the construction of a discourse acceptable with regard to the rules specific to mathematics" (Balacheff, 1999, p. 5). One such prominent example are plausibility arguments, in particular *empirical arguments*, which are discussed further in the following section.

In this thesis, I follow the understanding that reasoning is a fundamental activity of which argumentation is a specific form. Other than Duval (1991), I do not view proving as being fundamentally different to argumentation, but rather as a subset of argumentation that meets specific criteria (namely those discussed in Section 2.3.3. As Balacheff (1999) and in line with a broader view of mathematical proof, I take the position that *mathematical* argumentation, as an activity being completely different from proving, does not exist.

[11] This way of reducing mathematical proof to deductive reasoning is controversial, see for instance Balacheff (1999)

Another term that is often used in the context of proof and proving (e.g., Lesseig et al., 2019; Mejía Ramos & Inglis, 2009b; Yackel & Cobb, 1996), but is less discussed in mathematics education literature, is *justification* or *justify* (Staples & Conner, 2022). While it is mainly clear what is expected from students when asked to *prove* a theorem or statement, namely, to construct a(n ordinary) proof (see above), it is less obvious what is exactly meant when students are asked to justify (why) a statement (is true or false) (Dreyfus, 1999). In the context of mathematical argumentation, it usually refers to providing sufficient "mathematical evidence in support of a result" (Staples & Conner, 2022, p. 5). In this thesis, I adopt the following definition given by the National Research Council (2001):

> We use justify in the sense of 'provide sufficient reason for.' Proof is a form of justification, but not all justifications are proofs. Proofs (both formal and informal) must be logically complete, but a justification may be more telegraphic, merely suggesting the source of the reasoning. (p. 130)

In this broader sense, justification is therefore particularly relevant for the mathematics classroom and can serve as "on-the-way-to-proof reasoning practices" (Staples & Conner, 2022, p. 5, see also Dreyfus (1999)). Similar to proof, what is considered *sufficient* in this sense needs to be negotiated within the respective community.

In the following section, I describe several types of arguments that are particularly relevant for the teaching and learning of proof and proving in school and that are widely referred to in mathematics education literature.

2.4.2 Types of Arguments

Several argumentation and proof concepts[12] have been identified and discussed in mathematics education, specifically in regard to teaching proof and proving in school (e.g., Biehler & Kempen, 2016; Brunner, 2014; Reid & Knipping, 2010). In particular, the following three main types of "proofs" are often distinguished, especially in the German literature:

- experimental proof
- operative proof (in German also *inhaltlich-anschaulicher Beweis*)
- formal-deductive proof

[12] I do not discuss different forms of proof such as *direct* and *indirect* proofs, as they are not part of this project's framework.

Reference is usually made to Wittmann and Müller (1988), although they originally cite Branford (1913) (a German translation of the English publication from 1908). Branford (1908) introduces the terms "*experimental* evidence or proof", "*intuitional* evidence or proof" and "*scientific* evidence or proof" (p. 233).

Not all of these types of arguments classify as *mathematical proof*, neither in a formal nor in an ordinary view (both as defined in Section 2.3.2). *Experimental proof*, sometimes also called *experimental verification* (e.g., Kunimune et al., 2009), refers to a form of argumentation based on empirical evidence for a claim. As such, it cannot establish general truth and is therefore not a valid scheme for mathematical proof, as discussed in Section 2.3.3. Further, this type of argument is not specific to mathematics, as giving examples are a common form of argumentation in everyday situations. Other terms that are used in the literature to describe a form of reasoning where a conclusion is drawn from the observation or verification of a (small) number of cases, are *naive empiricism* (Balacheff, 1988b) or *empirical arguments* (e.g., Reid & Knipping, 2010). In this thesis, I use the latter term. Even though empirical arguments do not ensure the non-existence of counterexamples, and thus, the generality of the statement, they are nevertheless and without doubt relevant for mathematical practice, for instances, to examine patterns and to explore or test conjectures. A good overview of several functions of *experimentation* in mathematics is provided by de Villiers (2010).

Operative proofs are based on specific observations, but in contrast to empirical arguments, they reveal a structure that can be generalized to hold for a whole class of objects (Wittmann & Müller, 1988). In the understanding of Blum and Kirsch (1989), this process of generalizing should consist of correct (non-formal) inferences and the underlying idea of why it is generalizable should be grasped intuitively. The latter emphasizes Branford's usage and understanding of the term "intuitional proof". Figure 2.2 provides an example of such a proof for the statement "the sum of any two odd numbers is always even". The explanation above the figure is not necessarily required. However, some researchers argue that with regard to the generality of the argument, valid explanations should be included; otherwise it could not be classified as *proof* (e.g., Kempen & Biehler, 2019).

> Every odd number can be grouped into pairs (of twos), such that exactly one is left. By adding two odd numbers, one can group the two one's that are left, such that the sum now only consists of pairs (of twos).

Another term that is widely used in the international literature to describe a similar type of proof, is *generic proof* (e.g., Bass, 2009; Dreyfus et al., 2012; Rowland, 2001) or *generic example and thought experiment* (Balacheff, 1988b; Mason &

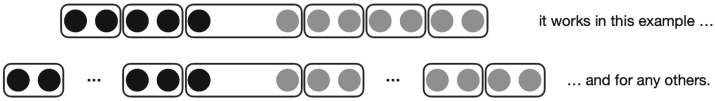


Figure 2.2 Example of an operative (or generic) proof

Pimm, 1984). Movshovitz-Hadar and Malek (1998) introduced the term *transparent pseudo proof* to highlight two main properties: That these types of arguments are not proofs in a formal sense (thus *pseudo*), but that "[one] can 'see' the formal proof through it" (Malek & Movshovitz-Hadar, 2011, p. 37), thus *transparent*—like glass. Different researchers highlight and define different features of these types of arguments (for more details, see Reid & Knipping, 2010). Regardless of the differences and terms being used, these types of arguments are often seen as an opportunity for the teaching and learning of proof to understand underlying mathematical concepts and ideas without the difficulties of (seemingly) complicated mathematical language:

> A generic proof aims to exhibit a complete chain of reasoning from assumptions to conclusion, just as in a general proof; however, ... a generic proof makes the chain of reasoning accessible to students by reducing its level of abstraction; it achieves this by examining an example that makes it possible to exhibit the complete chain of reasoning without the need to use a symbolism that the student might find incomprehensible (Dreyfus et al., 2012, p. 204)

Depending on its specific definition and with respect to a broader understanding of mathematical proof, generic arguments may qualify as valid mathematical proof and could be placed at the non-formal end of the spectrum (see Fig. 2.1 in Section 2.3.2). This would be in line with Wittmann (2014), who states that operative and formal proof (due to his explanations, I assume he actually refers to ordinary proofs) do not differ fundamentally, but only in the form of argumentation and its representation (e.g., one uses iconic, the other mainly symbolic representations). In contrast, in a formal sense, generic arguments would most certainly not qualify as proof. As noted in Section 2.3.3, visual arguments in mathematics (which are often generic) are seen controversial among mathematicians.

The term *formal-deductive proof* is sometimes used more or less synonymously for *formal proof* and should as such not be controversial. However, it appears doubtful that Wittmann and Müller (1988) explicitly wanted to refer to formal proof rather than ordinary proof, as they cite Branford (1913), who uses the term

"scientific proof". Wittmann and Müller (1988) thus describe *formal-deductive proofs* as those proofs that are constructed and published by professional mathematicians. Wittmann and Müller argue that formal(-deductive) proofs are too complex and sophisticated, and therefore an obstacle for the learning of proof and proving in school. Regarding the formal sense of proof, I agree (as most researchers do, see Section 2.3.3). However, not in a broader understanding of proof (i.e., ordinary proof) with its characteristics as discussed in Section 2.3.3. Learning about these characteristics seems to be essential to gain an appropriate understanding of proof and to ease the transition to university.

In the framework of this project, I distinguish between empirical arguments and generic and ordinary proofs, in the understanding discussed above. Further types of arguments that are used by students and are relevant in this thesis (so-called *proof schemes*) are discussed in Section 3.2.5.

State of Research

3

The previous chapter has illustrated how argumentation and proof, and in particular generality, are fundamental for mathematics, but also that defining what constitutes mathematical proof is not that simple. Because of its central role in mathematics, proof and argumentation are seen as essential for the learning of mathematics in school as well as university by many researchers (e.g., Hanna, 2000; Schoenfeld, 2009) and national curriculum frameworks around the world have specified respective learning goals (e.g., Department of Basic Education, 2011; Kultusministerkonferenz, 2012; National Council of Teachers of Mathematics, 2000). Consequently, proof and argumentation have been researched extensively in mathematics education, in particular in the last three decades (Hanna, 2000; Hanna & Knipping, 2020; Sommerhoff, Kollar, & Ufer, 2021). About 16% of all PME research reports published between 2010 and 2014 focused on proof and argumentation in secondary or tertiary education (20% in total, including primary education), as Sommerhoff et al. (2015) report. The majority of these studies is qualitative (57%) and had not more than 100 participants, which highlights the need for further large-scale quantitative studies. Research foci in mathematics education range from theoretical investigations on the importance and relevance of proof for the teaching of mathematics to empirical research on students' and teachers' proof skills regarding, for instance, the understanding, evaluation, and construction of proofs (see, e.g., Mejía Ramos & Inglis, 2009a; Sommerhoff et al., 2015). Difficulties with proof and proving related to several activities have been reported by numerous researchers (e.g., Dubinsky & Yiparaki, 2000; Harel & Sowder, 1998; Healy & Hoyles, 2000; Kempen, 2019; Recio & Godino, 2001; Weber, 2001). In particular, more attention has recently been given to empirical research on proof and argumentation in higher education with a focus on the transition from school to university and first-year university

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_3

students (e.g., Alcock et al., 2015; Gueudet, 2008; Hanna & Knipping, 2020; Kempen & Biehler, 2019; Moore, 1994; Rach & Ufer, 2020; Recio & Godino, 2001; Sommerhoff, 2017; Stylianides & Stylianides, 2009; Stylianou et al., 2006).

With respect to the present research interest, the focus of this chapter is on research findings regarding (first-year) university students and mathematics (preservice) teachers, but wherever appropriate, findings regarding secondary school students are considered as well. Thereby, this chapter is mainly divided into the following three parts. Because of its central role for this thesis, research on the understanding of the generality of mathematical statements and proofs as well as the role of (counter-)examples is first reviewed in section 3.1. In section 3.2, different activities related to proof and proving are then discussed and research findings regarding activities that are relevant for the present thesis are summarized. Lastly, in section 3.3, resources that may influence individual's performance in proof-related activties are discussed.

3.1 Understanding Generality and the Role of (Counter-)Examples

Generality as an essential characteristic of mathematical statements and proof has been repeatedly highlighted as an important learning goal in the literature on proof and argumentation (e.g., Conner, 2022; Ellis et al., 2012; Fischbein, 1982; Kunimune et al., 2009; Lesseig et al., 2019). To prove the generality of a (universal) statement (as defined in section 2.1), a *general* deductive argument needs to be constructed. The awareness and understanding of this *requirement* of proof is usually meant, when researchers refer to understanding the *generality of proof* (e.g., Conner, 2022). While most likely equally important for the learning of proof and proving, understanding the *generality of statements* has not been explicitly defined yet. Whenever I refer to the understanding of the generality of statements, I mean the understanding that true *universal statements* hold *for all* elements in the given domain–without any exceptions (see also section 2.2), i.e., no counterexamples to the statement exist. Research on students' (and teachers') understanding of generality has mainly focused on the generality of proof, occasionally in relation to understanding the generality of statements.

For instance, the framework for students' understanding of proof by Kunimune et al. (2009) consists of the two aspects *construction of proof* and *generality of proof* (see p. 442). According to Kunimune et al. (2009), one necessity for students' understanding of "generality of proof" is the understanding of the universality and generality of statements. They do not further clarify how the understanding of *uni-*

versality and *generality* of statements and proof differ. In particular, they do not explicitly define what constitutes understanding of the generality of statements. But they introduce levels of students' proof understanding. For example, regarding the generality of algebraic proof, students on level 0 "do not understand what they have to explain". It is unclear, if this corresponds to an insufficient understanding of the generality of a statement. One could argue, however, that students, who are aware that the correctness of a statement has to be explained in some way, can at the same time be unaware what it means that a universal statement is correct, namely, that no counterexamples exist. Those students would then not have a complete understanding of the generality of a statement. The study of Kunimune et al. (2009) consisted of written survey questions, which were answered by 418 lower secondary students (Grade 8 and 9) from Japan. Kunimune et al. (2009) come to the conclusion that participants who were able to construct proofs not necessarily valued the generality of proofs. They seemed to believe that empirical arguments are also an equally acceptable way to prove a statement. In this regard, Stylianides (2016) concludes that students' "inability to recognize the generality of proof may suggest that [they] do not conceive of proof as a means for establishing truth." (p. 320). If this is the case, understanding the generality of statements and proof might influence students' *intellectual need* (see, e.g., Harel, 2013) and understanding of the necessity for proof.

Lesseig et al. (2019) have also explicitly included the generality of statements and proof in their framework. They investigated preservice secondary mathematics teachers' understanding of proof. According to their framework for mathematical knowledge for teaching proof, teachers should have the following "essential proof understandings:

- *A theorem has no exceptions* [emphasis added]
- *A proof must be general* [emphasis added]
- Proof is based on previously established truth
- The validity of a proof depends on its logic structure" (p. 396)

The first aspect corresponds to an understanding of the generality of universal statements and the second aspect to the understanding of the generality of proof. In their pilot study, 34 students from the USA, Australia, and Korea completed the survey. Regarding the generality of statements and proof, Lesseig et al. (2019) found that 30% of the participants considered generality when evaluating proofs, where as 20% did so when they were asked to identify requirements of proof. Further, the researchers argue that "merely knowing that a proof must be general was not necessarily sufficient, as [the teachers] had different interpretations of what constituted generality" (p. 413).

As has been pointed out at the beginning of this section, no clear definition for understanding generality, in particular, of statements, and how to assess this understanding has been given yet. Regarding the understanding of generality of proof, Conner (2022) has recently provided a framework to identify students' (limited) understanding of generality, which she refers to as understanding the *generality requirement*, meaning "the requirement that a proof must demonstrate the claim to be true for all cases indicated within the domain of the claim" (p. 2). Her framework aims to assess students' understanding of this requirement in two proof-related activities, students' construction and evaluation of arguments. Through a case study, she identified instances that demonstrate an understanding of generality and instances that demonstrate limited understanding. For instance, with respect to the evaluation of arguments, according to Conner's framework, students who refer to the need for a general argument demonstrate understanding and students, who refer to the inclusion of examples as proof, demonstrate *limited* understanding. Further, the usage of empirical arguments to prove a universal statement was also identified as an instance, which shows limited understanding of generality. However, several researchers have argued that students' usage or conviction of empirical arguments does not necessarily relate to an insufficient understanding of generality (e.g., Healy & Hoyles, 2000; Weber, Lew, & Mejía-Ramos, 2020), which Conner herself noted. Students' conviction and usage of empirical arguments are further discussed in sections 3.2.3 and 3.2.5. Apart from the role of examples in justification and evaluation tasks, Conner (2022, p. 6) also included the usage and interpretation of variables (as "generalized numbers" vs as "placeholders for specific values"), and the notation and interpretation of diagrams (e.g., "representing all possible cases" vs. "showing a specific case") into her framework. She did not explicitly consider the understanding of the generality of statements.

Most studies that have investigated students' (or teachers') understanding of the generality of statements and proof and the respective role and usages of (counter-)examples were mainly qualitative with comparatively small sample sizes. The existing research suggests that some students and teachers do not completely understand the generality of true universal statements and proof (e.g., Balacheff, 1988b; Chazan, 1993; Galbraith, 1981; Knuth, 2002). For instance, Balacheff (1988b) observed 14 pairs of students age 13 to 14 to explore their proving processes. He found that some students understand counterexamples as exceptions from the rule and not necessarily as refutation of the statement, which indicates a limited understanding of the generality of statements. Similarly, in a study conducted by Galbraith (1981) with about 170 students age 12 to 17 from Australia, about 18% of the students thought that one (or few) counterexamples are insufficient to disprove a (universal) statement.

In other studies, participants were confronted with (different types of) arguments and proofs for a statement. Chazan (1993), for example, investigated students beliefs about empirical evidence and deductive proof through semi-structured interviews with 17 high school students from geometry classes. He found that several students were not convinced that a deductive proof ensures that no counterexamples can be found. One student explicitly stated that it is impossible "to prove a statement for everything" (p. 372)–neither with empirical arguments nor with a deductive proof–a belief shared by other students of the study as well. These students seemed to have an insufficient understanding of the generality of proof and potentially of statements. However, some students were (correctly) convinced that a deductive proof guarantees that no counterexamples can exist. Because "a substantial number of students in this study" (p. 382) seem to have beliefs that are contrary to those of the mathematical community, Chazan (1993) emphasizes the importance of discussing the characteristics of empirical arguments and deductive proof explicitly in the mathematics classrooms. However, it is not clear if teachers themselves have a complete understanding of the generality of statements and proofs and their characteristics, as some teachers seem to have similar beliefs to those of students. Knuth (2002) conducted an interview study with 16 secondary school teachers to investigate teachers' conceptions of proof. His findings suggest that several of the teachers do not have a complete understanding of the generality of statements and proofs. Six of the teachers thought that counterexamples, which would make the proof invalid, could still be found, even though they claimed to have understood (and accepted) the (ordinary) proof. Further, several teachers did not seem to be completely convinced that no counterexamples to a proven statement can exist, as some needed to verify that the argument holds for particular cases. It is not clear if this relates to limited understanding of the generality of proofs, statements, or both.

Even though 72% of the teachers who participated in a study conducted by Barkai et al. (2002) correctly justified the falsity of a universal statement by giving a counterexample, only 36% seemed to think their argument would be accepted as proof. Furthermore, several teachers gave more than one counterexample, which suggests "that they do not believe that a single counterexample is sufficient to refute a universal statement" (Reid & Knipping, 2010, p. 64).

In their influential study, Healy and Hoyles (2000) explicitly considered students' understanding of the generality of statements. They assessed understanding of the generality of a *proven* statement by asking students if the proof "automatically held for a given subset of cases" (p. 402) or if a new proof has to be constructed. About 60 % of the students correctly thought that no further proof is needed. In their study, Healy & Hoyles directly related understanding of the generality of universal

Table 3.1 Logical relationship between examples and statements, reprinted from Buchbinder and Zaslavsky (2019, p. 131), with permission from Elsevier

Classes of examples with respect to *domain* D and *property* P	Types of examples	
	For universal statements	For existential statements
$x \in D$, $P(x)$	Supporting	Confirming
$x \in D$, $\neg P(x)$	Counterexample	Non-confirming
$x \notin D$, $P(x)$ or $x \notin D$, $\neg P(x)$	Irrelevant	

statements to the generality of proof. Similar to other studies, they did not explicitly define what understanding of the generality of statements specifically consists of.

Buchbinder and Zaslavsky (2019) investigated students' understanding of the role of examples in proving. Their REP (Roles of Examples in Proving) framework highlights the relationship between (counter-)examples, statements, and proof (Table 2, p. 131). It is based on the logical relationship between examples and statements (see Table 3.1).

Because existential statements are not considered in the present thesis, only the column regarding universal statements is relevant here. It highlights two aspects of understanding the generality of universal statements in relation to proof: That examples (i.e., empirical arguments) only *support* the truth of a universal statement, but do not prove it, and that one counterexample is sufficient to refute the statement. Twelve high-attaining Grade 10 students from Israel participated in the study, which consisted of task-based semi-structured interviews. The students were interviewed in pairs "to create opportunities for verbal communication and spontaneous convincing and justifying" (p. 133). Tasks included the estimation of truth of several universal and existential statements (some of them true, some of them false) from algebra and geometry, the evaluation of arguments, and questions regarding the existence of objects with certain properties. Responses were coded according to the REP framework such that alignment with conventional mathematics was considered. They found three types of inconsistencies in students' responses: "(1) inconsistency with respect to the type of example [see Table 3.1], (2) inconsistency with respect to the type of statement [universal or existential], and (3) inconsistency with respect to the type of inference" (p. 139); for instance, regarding the "status of supporting examples [i.e., empirical arguments] in proving universal statements". The number of observations where students were aware that examples do not prove a universal

statement was equal to the number of observations where students had the contrary belief (16 observations each). Regarding the status of counterexamples in disproving universal statements, most students responded correctly (62 observations); the researchers found only 9 observations in which students though that a counterexample does not disprove a universal statement. Thus, while many participants seemed to have not fully grasped the generality of proof, most participants in this study seemed to have a correct understanding of the generality of universal statements. Interview studies with 16 secondary school students conducted by Stylianides and Al-Murani (2010) revealed similar findings. The students were selected based on their responses to an earlier survey, in which they seemed to have the belief that a proof and a counterexample can both exist to the same universal statement (as they assumed that both, a proof and a counterexamples to the same statement would get high marks from the teacher). However, Stylianides and Al-Murani found that all of these students correctly believed that a counterexample to a proven statement cannot exist.

In summary, several studies have investigated students' or teachers' understanding of the *generality of proof* and some of these considered the understanding of the *generality of statements*. However, all of these studies related the understanding of the generality of statements to that of proof. Moreover, what constitutes understanding of the generality of mathematical statements has not been explicitly defined so far and it is not clear how it relates to the understanding of (the generality of) proof as well as the usage and conviction of (counter-)examples. In particular, the existing research does not provide clear findings regarding the proportion of students and teachers with a limited understanding of the generality of statements (i.e., those who respond inconsistently regarding the estimation of truth of a universal statement and the existence of counterexamples) and aspects (such as the truth value of the statement and the type of argument) that might influence their understanding of the generality of statements.

The following section first provides an introduction to different activities that are related to proof and proving. Then, research findings regarding relevant activities with respect to the present project are discussed.

3.2 Activities Related to Proof

Research on students' proof skills and understanding of proof and argumentation focuses on different activities. Mejía Ramos and Inglis (2009b) proposed a framework to classify the respective activities. Based on a general classification of mathematical activities provided by Giaquinto (2005)—making, presenting, taking in—Mejía

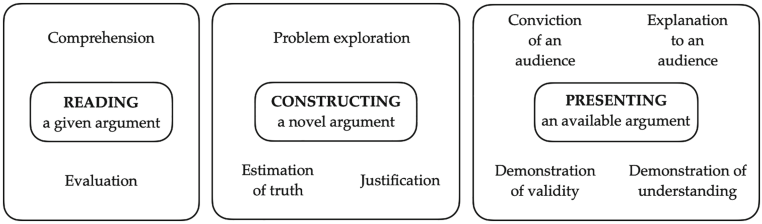


Figure 3.1 Sub-activities related to proof by Mejía Ramos and Inglis (2009b)

Ramos and Inglis distinguish three main argumentative activities: "*constructing* a novel argument, *presenting* an available argument, and *reading* a given argument [emphasis added]" (Mejía Ramos & Inglis, 2009b, p. 68), which other researchers often refer to as *proof construction, proof presentation*, and *proof reading* (e.g., Selden & Selden, 2017). In order to consider different behavior due to different contexts, they further subdivide the three activities. Based on the work of de Villiers (1990) on different goals of proof, Mejía Ramos and Inglis (2009b) propose the classification of argumentative activities as shown in Figure 3.1.

According to their framework, proof reading comprises *proof comprehension* and *proof evaluation*, depending on the goal with which a proof is being read. The goal of proof comprehension is the understanding of a given argument. Aspects of proof comprehension include knowing the meaning of definitions and terms, understanding the meaning and logical status of statements within the proof, being able to summarize main ideas of the proof, and illustrating the proof with examples (Mejía Ramos et al., 2012; Neuhaus-Eckhardt, 2022; Yang & Lin, 2008) (see further section 3.2.4). In the framework of Mejía Ramos and Inglis (2009b), proof evaluation consists of both, the validation of proofs (i.e., determining its correctness) and other evaluative tasks, such as assessing if a proof is convincing or explanatory. Similarly, Pfeiffer (2011) describes proof evaluation as "determining whether a proof is correct ... and also how **good** it is regarding a wider range of features such as clarity, context, sufficiency without excess, insight, convincingness or enhancement of understanding" (p. 5). To distinguish between proof validation and proof evaluation, A. Selden and Selden (2017) separate validation tasks from proof evaluation. Regarding proof evaluation, they put a focus on the judgement of qualitative aspects such as convincingness, clarity, context, and aesthetics (see also Inglis & Aberdein, 2015). Similar to proof reading, Mejía Ramos and Inglis (2009b) further subdivide proof presentation into sub-activities with different goals. All four of these sub-activities have in common that an argument is presented to an audience, but they

differ with respect to different *functions of proof* (e.g., the argument is presented *to convince* the audience of the truth of a statement or *to provide an explanation* why a statement is true). In the understanding of Mejía Ramos and Inglis (2009b), proof construction is not only about finding and giving arguments to justify a statement, i.e., *justification* (see also section 2.4.1); their framework equally includes *problem exploration* and the *estimation of truth* of a statement as important aspects of proof construction.

Particularly relevant for the present thesis are the two sub-activities *proof comprehension* and *proof evaluation* regarding the reading of given arguments and the sub-activities *estimation of truth* and *justification* regarding the construction of (novel) arguments. I therefore do not discuss *presenting available arguments* further. Like Mariotti (2006), I assume that it is not possible to "isolate proof from the statement to which it provides support, and from the theoretical frame within which this support makes sense" (p. 184).

Therefore, to highlight the importance of the statement itself, I propose an adapted version of the framework of Mejía Ramos and Inglis (2009b) by adding the activity reading a statement (see Fig. 3.2). The *presentation of arguments* could potentially also be included in the adapted framework, however, it seems to be more difficult to directly relate it to the other activities. In the adapted framework, the sub-activity *estimation of truth* can be viewed as part of *reading a statement* of which the truth value is initially unknown[1]. The conclusion of whether or not a statement is true can either be drawn by constructing an argument oneself or by reading a given argument (although on most occasions where an argument is provided, the reader already knows that the statement is true). Also, a deeper *comprehension of a statement* can be supported by both of these activities. Similar to proof comprehension, the comprehension of a statement contains aspects such as knowing and understanding the meaning of definitions, terms, and symbols, its logical structure, and its relation to other statements, as well as understanding the generality of the statement, i.e., understanding that there cannot be any counterexamples if the statement is universal and true (see also section 3.2.1). The comprehension of a given argument is therefore closely related to the comprehension of the statement itself. Even more: Without fully understanding the statement (including relevant definitions and terms etc.), it seems to be impossible to fully understand or construct an argument. Furthermore, the necessity for and generality of proof might only become clear when completely understanding the generality of the statement.

Before discussing main research findings related to the respective activities, quantitative results from the literature reviews of Mejía Ramos and Inglis (2009a)

[1] Mejía Ramos and Inglis (2009b) refer to such statements as *conjectures*.

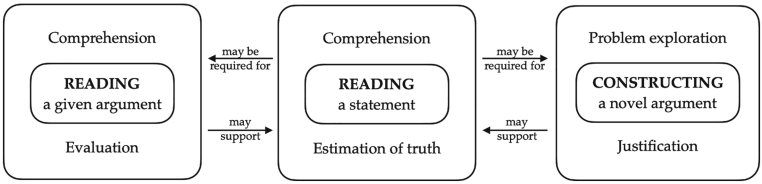


Figure 3.2 Adapted framework on proof-related activities based on Mejía Ramos and Inglis (2009b)

and Sommerhoff et al. (2015) regarding argumentative activities (Sommerhoff et al. use the term *situations*) are summarized. Both studies found that the majority of mathematics education literature is on proof construction, followed by proof reading, and (almost) no literature was found on proof presentation (see Table 3.2 for a comparing overview of the studies and the respective results). The articles in the bibliographical study of Mejía Ramos & Inglis were further allocated to argumentative sub-activities (as shown in Figure 3.1). Regarding proof construction, most articles (54%) were about problem exploration, followed by the "justification of a statement" (27%); the minority of the articles (20%) was on the "estimation of the truth of a conjecture". Articles on proof reading were mostly on the evaluation of the argument (87.5%). Only 12.5% were on the comprehension of arguments. One reason for that might be that it is more difficult to develop assessment models for proof comprehension than, for instance, proof construction, because one first has to define what proof comprehension consists of (see further section 3.2.4 on recent developments). Mejía Ramos and Inglis (2009a) argue that further research on proof comprehension and proof presentation is needed, because both of these (sub-)activities have received little attention in mathematics education so far, even though they can be viewed as two key argumentative activities.

In the following sections, research findings related to the comprehension of statements, the estimation of the truth of statements, the comprehension of arguments, the evaluation of arguments (particularly with respect to convincingness), and justification (with a focus on different types of arguments being used, i.e., so-called *proof schemes*) are reviewed with respect to the present research interest. Research on *problem exploration* is not considered, as it was not explicitly investigated in the present thesis. Lastly, relations between these activities are discussed.

Table 3.2 Comparing overview of descriptive literature reviews on argumentative activities

	Mejía Ramos and Inglis (2009a)	Sommerhoff et al. (2015)
Source of articles	Cognition & Instruction, Educational Studies in Math., For the Learning of Math., Journal of Math. Behaviour, Journal for Research in Math. Edu., Math.Thinking and Learning, and ZDM	Proceedings of IGPME
Publication type	Journal articles found in ERIC	Research Reports
Years	1966–2008	2010–2014
Education level	any level	secondary or tertiary
# of articles	131	129
Relative frequencies by argumentative activities		
Proof construction	63%	57%
Proof reading	18%	7%
Proof presentation	no articles found	1%
Multiple	not specified	7%
Not explicit	19%	29%

3.2.1 Comprehension of Statements

In literature on proof and proving, little attention has explicitly been given to the comprehension of statements themselves. Often, the understanding of the statement is seen as part of proof comprehension (see section 3.2.4), which in the framework of Mejía Ramos and Inglis (2009b) belongs to proof reading. As discussed in the previous section, in the present thesis, the comprehension of a statement is understood as (an integral) part of reading a statement which is related to both, the reading, and in particular the comprehension of proof, as well as proof construction. In section 3.1, research findings on the understanding of the generality of statements, which can be seen as part of the comprehension of statements, have already been discussed in detail. Thus, in this section, main findings on students' comprehension of statements with respect to the statements' content and logical structure are summarized.

Regarding the comprehension of mathematical statements, students seem to have difficulties with both, the logical structure of the statement (e.g. Dubinsky & Yiparaki, 2000; Moore, 1994; A. Selden, Mckee, & Selden 2010; J. Selden & Selden, 1995) as well as (basic) understanding of the content and relating it to a mathemat-

ical theory (e.g., Dubinsky & Yiparaki, 2000; Ferrari, 2002). In particular, students often seem to ignore unpacking the statement's conclusion, but solely focus on the assumption, which limits their understanding of the statement and their attempts to prove it (Moore, 1994; A. Selden et al., 2010). Correctly unpacking the logical structure of a statement is seen as an important activity for understanding the statement and a necessity to construct and evaluate proofs (e.g., J. Selden & Selden, 1995). By *unpacking* the logical structure of an *informal statement* J. Selden and Selden (1995) mean "associating with it a logically equivalent formal statement" (p. 128). J. Selden and Selden (1995) examined 61 US university students' abilities to understand and use the logical structure of mathematical statements in proof construction and validation by providing them with four informal calculus statements (two true, two false) which the participants had to unpack. They report that no student was able to unpack all informal statements into equivalent formal ones in first-order logic "which they were familiar with" (p. 139). The percentage of students who gave a correct logical unpacking of a statement was between 0 and 20.8%, depending on the statement. But students not only seem to struggle with unpacking statements into (more) formal ones, but also with the translation of terms used in informal statements into (simple) symbolic expressions, as Piatek-Jimenez (2004) reports. She illustrates this observation with the following example: One student in her study "converted her memorized definition of m being odd [the author probably means even] from 'm is an integer divisible by 2' into the symbols '$m = \frac{n}{2}$' " (p. 195). If a statement is given informally and a respective proof uses symbolic expressions, students having these kinds of difficulties would most likely struggle to follow the proof (as well as constructing one).

Dubinsky and Yiparaki (2000) also report on students' difficulties with the logical structure of statements. They conducted a study with 63 students from two US universities and one liberal arts college (mainly mathematics and mathematics education majors) in which participants had to estimate the truth value of eleven statements, nine of them in everyday context (e.g., *every pot has a cover*) and two of them mathematical statements and justify their decision. The authors of the study were in particular interested in students' understanding of universal existential statements and existential universal statements (see also section 2.1). The two mathematical statements were therefore chosen such that they basically only differed in the order of quantifiers[2]. Dubinsky and Yiparaki (2000) found that about 42% of

[2] The statements were: "For every positive number a there exists a positive number b such that $b < a$." and "There exists a positive number b such that for every positive number a $b < a$."

the participants incorrectly assumed that the two statements were equivalent. This finding indicates that students have difficulties with the interpretation of quantified statements, which was confirmed by the study conducted by Piatek-Jimenez (2004). Dubinsky and Yiparaki (2000) furthermore found that students mainly focused on the particular content of the statement and how that content relates to their experienced reality as they "referred to a world they were already familiar with and they considered that the statement described that world" (p. 53). The students were not able to understand and unpack the (logical) structure of the statements. Moreover, they had more difficulties with correctly interpreting and justifying statements when the context was mathematical, which the authors see as an indicator for students' difficulties on a *semantic* level.

Fundamental difficulties were reported by Ferrari (2002), who conducted a study with 39 Italian first-year computer science students. He identified several difficulties regarding the participants' comprehension of elementary number theory statements with respect to the content and language of the statement. For instance, students (1) had poor understanding of basic definitions; (2) lacked conceptual understanding of the content (e.g., division and fractions), as they used simplifications in a way that overemphasized procedural aspects; and (3) had difficulties with distinguishing between a number and its representation.

The findings reported above highlight the importance of (basic) content knowledge, including knowledge about (school) mathematical concepts, which also have been identified as essential prerequisites for a successful transition from school to universities (e.g., Rach & Ufer, 2020). Further, many, if not most students also seem to have difficulties with understanding and unpacking the logical structure of statements. These findings are not only relevant for the teaching and learning of proof and argumentation, but should be considered in the development of research instruments that aim to assess students' proof skills (see section 5.3).

Fully comprehending (the meaning of) a mathematical statement is not only necessary for proving it or understanding a given proof, but it is also important for getting a better intuition regarding its truth value. The following section summarizes research findings regarding students' success in estimating the truth value of statements as well as their respective strategies.

3.2.2 Estimation of Truth of Statements

Deciding wether or not a statement is true can be seen as an essential activity in mathematics. Consequently, many curricula suggest that students should be able to make, evaluate (e.g., estimating the truth value), and justify mathematical statements

(e.g., Department of Basic Education, 2011; Kultusministerkonferenz, 2012; National Council of Teachers of Mathematics, 2000). For instance, the German national curriculum for upper secondary schools states that "this competence [referring to argumentation] includes ... the understanding and evaluation of given mathematical statements" (Kultusministerkonferenz, 2012, p. 14, translated by the author). Similarly, from prekindergarten through grade 12, students in the US should be enabled to "make and investigate mathematical conjectures" (National Council of Teachers of Mathematics, 2000, p. 56).

Of all articles on proof and argumentation that were included in the literature review of Mejía Ramos and Inglis (2009a), approximately 12% were about the estimation of the truth of a conjecture. Studies have reported a wide range of percentages of participants correctly judging the statements' truth values, ranging from about 30 to 100% (Barkai et al., 2002; Dubinsky & Yiparaki, 2000; Hoyles & Küchemann, 2022; Ko, 2011; Ko & Knuth, 2009; Riley, 2003; Zeybek Simsek, 2021). In addition to investigating students' or teachers' success in estimating the truth value of mathematical statements, some studies have examined its relation to the type of statement (e.g., universal or existential, true or false) and/or strategies students or teachers use to come to a respective conclusion. For instance, Barkai et al. (2002) report on 27 elementary school teachers' correct judgements regarding the truth of three universal and three existential statements, some of which are true and some false. The percentage of teachers who correctly estimated the truth value was between 68 and 100%, thus, most were comparatively successful in estimating the truth (see Tab. 3.3). Descriptively, it seems that it was more difficult for them to correctly decide on the truth value of existential statements (the true existential statement #4 is the same as #1 and therefore true for all n, thus it can be seen as an exception), in particular, regarding statements that are true *for some* n, no matter if the statement is expressed as false universal or true existential. However, the number of participants as well as items is comparatively small and might therefore not be representative.

Comparatively high success rates were also reported by Ko (2011). She conducted semi-structured interviews with eight secondary mathematics education majors from a US university, who were either in their third-year, fourth-year, or fifths-year. The majority of participants (about 83%) correctly estimated the truth value of six statements (four of them true, two of them false; from different content areas). However, one of the false statements was correctly evaluated by only half of them (the other one by 87.5%). Moreover, most of the students (successfully) used mixed reasoning strategies in which "individuals both use examples to identify relevant patterns and structures, and manipulate (partially) correct properties, definitions, and/or theorems to identify a reasonable example to attempt to prove or disprove the statement" (p. 481), fewer participants used other strategies such as deductive

Table 3.3 Teachers' estimation of truth by truth value and domain of discourse of statements; findings reported by Barkai et al., table adapted from (Reid & Knipping, 2010, p. 70), with permission from Brill

Statement	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Type	Universal			Existential		
Truth value	True	False	False	True	False	True
Correct judgement	100%	100%	69%	100%	77%	68%
True for what set?	All *n*	No *n*	Some *n*	All *n*	No *n*	Some *n*

arguments (the author refers to *sophisticated reasoning*) or purely empirical arguments. A false statement that has been used in several studies is that *if the perimeter of a rectangle increases, the area of it also increases*. The percentage of preservice teachers who incorrectly evaluated this statement to be true has been reported to be comparatively high: About 57% of 23 preservice secondary school teachers from a US university and 72% of 50 preservice middle school teachers from a Turkish university that participated in a study conducted by Riley (2003) and Zeybek Simsek (2021), respectively, thought the statement is correct. Some researchers argue that it is no surprise that students particularly struggle with correctly estimating the truth value of false statements, because in schools (and universities), students usually have to prove true statements in contrast to disproving false ones (e.g., Buchbinder & Zaslavsky, 2007; Ko, 2011). However, no noteworthy difference was found by Ko and Knuth (2009) regarding the success rates of 36 Taiwanese mathematics undergraduates (most of them prospective secondary school teachers). For both a true and a false statement, the percentage of students who failed at estimating the truth value was about 20%. Most of these students provided an incorrect counterexample regarding the true statement and an incorrect proof regarding the false one. More research is needed to better understand the influence of the truth value on students' success in estimating the truth value of statements.

To better understand students' usage and understanding of (counter-)examples in the process of estimating the truth value of a statement, some researchers have investigated *how* students estimate the truth or falsity of statements (see also section 3.2.5 for further research on types of arguments being used by students and teachers to justify statements). In this regard, Buchbinder and Zaslavsky (2007) conducted a study to identify students' strategies in determining the truth value of statements. They found that the first step "was based on their intuition and sense

of confidence" (p. 563). Depending on their confidence regarding the truth value of the statement, students searched for evidence (either by deduction or based on empirical arguments) or directly gave empirical arguments to support their assertion. Empirical argumentations thereby occasionally resulted in "students shift to a different decision", for instance, when they "have found by chance a counterexample that contradicted [their] decision" (p. 564). Thus, as has been emphasized by other researchers, experimentation is not only common but potentially useful to explore conjectures (e.g., de Villiers, 2010 Lockwood, Ellis, & Lynch, 2016).

To estimate the degree to which mathematicians use examples to evaluate and prove universal statements, Alcock and Inglis (2008) conducted two case studies with doctoral mathematics students. The participants were asked to decide if several statements were true and to justify their decision with a proof. Both doctoral students used empirical arguments in the interviews, however, "the degrees to which they invoked examples to support their reasoning were strikingly different" (p. 126). While one participant did not seem to use empirical arguments to get a better understanding of the statement, the other one did. Alcock and Inglis (2008) conclude that these findings highlight the usefulness of skills involving experimentation for the exploration of statements.

In summary, the success rates of students' and teachers' evaluation of the truth value of mathematical statements differ substantially and seem to depend on characteristics of the statement such as its actual truth value and its domain of discourse (if it is true for all, for some, or for no entities/cases). However, more research is needed to identify specific relations, for instance, regarding the influence of the statements' truth value. Moreover, example based reasoning (or a mixed strategy including some deductive arguments) seems to be a common and useful approach–even though to different degrees–to get an understanding of the statement and a better intuition regarding its truth value. But if and how the usage of empirical arguments is related to students' success in estimating the truth value of statements and the understanding of the generality of statements is unclear and needs to be investigated.

In the following section, research findings regarding students' proof evaluation are summarized. Thereby, a particular focus is on students' conviction and acceptance regarding different types of arguments.

3.2.3 Proof Evaluation

The evaluation of given arguments is another essential sub-activity of proof reading. As already stated in section 3.2, proof evaluation can include different aspects, such as the validation of arguments, i.e., deciding whether an argument is *correct* (i.e., a

valid proof) and other evaluative activities, for instance, assessing if an argument is convincing or explanatory. Being able to decide if an argument is valid or not is seen as an important skill, for both students and teachers, and a necessity for establishing a *proper understanding of proof* (Pfeiffer, 2011; Powers, Craviotto, & Grassl, 2010; A. Selden & Selden, 2003; Sporn, Sommerhoff, & Heinze, 2021; Weber, 2010). In particular, researchers have argued that evaluative activities in general may be beneficial for the learning of proof construction, because both activities rely on the knowledge of acceptance criteria for proof (e.g., Pfeiffer, 2011; A. Selden & Selden, 2003). The focus of this section is on students' proof evaluation regarding conviction and validity of arguments, but not on other aspects, such as how explanatory an argument is. Before empirical research on proof evaluation is reviewed, the relation between conviction and validity is discussed.

Conviction and Validity

Even though some researchers define the evaluation regarding the validity of arguments as a separate activity, namely that of proof validation (e.g., A. Selden & Selden, 2017, see also section 3.2), proof validation and other evaluative activities are generally not independent of each other. The degree of conviction is influenced by several aspects, for instance, by the perceived validity of the argument (see section 2.3.3). If someone identifies an argument as not valid, they will likely also find it not (completely) convincing. On the other hand, the sole existence and acceptance of a proof does not necessarily lead to conviction (Fischbein, 1982; Segal, 1999; Weber, 2010). For instance, in a study conducted by Fischbein (1982), many of the participating high school students felt the need to verify the truth of a statement through empirical investigations, even though they had claimed that the provided argument was a valid proof.

It is not always clear what is meant by *conviction* and how students interpret questions regarding their conviction of arguments. *Conviction* is sometimes used regarding the validity of proofs (e.g., Weber & Mejia-Ramos, 2015), often to express different degrees of conviction in the validity of proofs (see also discussion on *relative* and *absolute conviction* further below). However, likely more often, conviction relates to the truth of a statement (see, e.g., Segal, 1999). In this sense, questions regarding students' conviction aim at investigating if the reading (or construction) of (certain types of) arguments lead to the conviction of the truth of a statement. As mentioned above, several aspects may influence students' conviction of the truth of a statement, in particular, the perception of the argument as *proof*.

However, one can gain high levels of conviction for the truth of a statement by reading an argument without accepting the argument as a proof (Tall, 1989; Weber, 2010) or even without the existence of a proof, as de Villiers (1990) highlights:

"Proof is not necessarily a prerequisite for conviction—to the contrary, conviction is probably far more frequently a prerequisite for the finding of a proof" (p. 18). This observation had also been made by Polya (1954), who argues that "without ... confidence [in the truth of the theorem] we would have scarcely found the courage to undertake the proof which did not look at all a routine job." (pp. 83–84).

As discussed in section 2.3, *proof* –in the social sense–is not a clearly defined concept, but depends on socio-mathematical norms. Thus, the validation of an argument depends on the respective mathematical community. To highlight the different psychological aspects that are involved in proof validation and conviction, Segal (1999) distinguishes between *personal conviction* (convincing oneself) and *public validation* (persuading others). She found that first year mathematics students from a UK university showed personal conviction regarding empirical arguments but no public validation, indicating that they assume empirical arguments do not meet the requirements of proof. Regarding ordinary proofs, however, no such difference was found. The students either found these types of arguments convincing *and* thought they were proofs or neither of those two. To highlight different degrees of conviction, Weber and Mejia-Ramos (2015) have introduced the terms *relative* and *absolute conviction*, to which I come back to further below.

Due to their relevance for the present thesis, the following sections outline research findings regarding students' evaluation of different types of arguments: empirical arguments and generic and ordinary proofs.

Findings on the Evaluation of Empirical Arguments

Healy and Hoyles (2000) found evidence that students are often convinced by empirical arguments, but at the same time recognize their limitations and thus, do not accept those arguments as valid proofs. Several studies have confirmed that most students seem to be aware of the limitations of empirical arguments (e.g., Healy & Hoyles, 2000; Lesseig et al., 2019; Stylianou et al., 2015; Tabach, Levenson, et al., 2010). Combined, these findings are in line with those of Segal, indicating that distinctions between personal conviction and public validation may be useful. However, other studies found the contrary, namely, that some students (and teachers) seem not only be convinced by empirical arguments but judge them as being valid proofs (Gholamazad et al., 2004; Knuth, 2002; Martin & Harel, 1989). For instance, in a study with 101 preservice elementary teachers from a US university conducted by Martin and Harel (1989), more than half of the participants accepted *inductive arguments*[3] as proof for both familiar and unfamiliar statements. The convincing

[3] The inductive arguments consisted of simple (numerical) examples, big numbers, several verifications that reveal a pattern (see Martin & Harel, 1989, p. 44)

power of empirical arguments is underlined by findings reported by Bieda and Lepak (2014). They conducted an interview study with 22 junior high school students from the US. The majority of the participants (15) chose an empirical argument over a general argument as being more convincing, mainly because they claimed that empirical arguments enhance their comprehension of the statement and provide more information. Four participants found the *generic argument* more convincing and two of them explicitly referred to a lack of generality of the empirical arguments.

Other studies did not find evidence that students find empirical arguments convincing. For instance, Weber (2010) conducted a study with 28 mathematics students from a US university, who had completed a transition-to-proof course. Most participants neither found the empirical arguments provided in the study convincing nor thought they constitute a proof. These findings were reproduced by D. Miller and CadwalladerOlsker (2020), who investigated 38 mathematics students from a US university, who were also enrolled in a transition-to-proof course. Thus, more advanced mathematics students do not seem to find empirical arguments convincing. Similar results were found by Ko and Knuth (2013), who investigated 55 middle school mathematics teachers' proof evaluation, and by Sommerhoff and Ufer (2019), whose findings show that most of the participating high school and university students judged the empirical arguments as being no valid proofs. Even though about 70% of more than 650 German high school students, who participated in a study conducted by Ufer, Heinze, Kuntze, and Rudolph-Albert (2009), correctly judged empirical arguments as invalid, only about one third of them could explain *why* these argument do not meet the criteria for proof. The authors argue that the students seem to be familiar with the fact that empirical arguments are insufficient to prove a universal statement, however, not in a way that they were able to explain why.

Findings on the Evaluation of Generic Proofs

Similar to empirical arguments, research findings on students' and teachers' evaluation of generic proofs seem to be inconsistent. Some studies suggest that many students do find generic proofs (in particular *diagrammatic arguments*[4]) convincing and think they constitute a proof (Ko & Knuth, 2013; Weber, 2010). Even though not all participants of the study conducted by Martin and Harel (1989) accepted generic proofs (the authors refer to *particular proofs*), those who (correctly) accepted an ordinary proof for a statement also rated the acceptance of the respective generic

[4] A diagrammatic argument is a type of visual argument. As it illustrates the statement for a particular case but reveal an underlying structure or idea, it can be seen as a form of generic proof.

proof highly. However, other studies have found the opposite. Tabach, Barkai, et al. (2010), for instance, conducted a study with 50 high school teachers and found that about half of the participants did not accept the provided generic proofs (the authors refer to *verbal justifications*) because of a perceived absence of generality. The (perceived) generality of an argument seems to be a criterion, teachers often consider when evaluating a proof (Ko & Knuth, 2013). Moreover, the mode of representation was another reason why participants in the study of Tabach, Barkai, et al. rejected correct generic proofs. Further evidence that teachers assume generic proofs do not meet the criteria for proof was provided by Lesseig et al. (2019). They found that the majority of secondary school teachers who participated in their study did not accept the presented atypical proofs (a generic proof and a visual argument). Further, some participants explicitly stated that these arguments did not convince them. The focus in both studies was on teachers' acceptance of arguments. A study conducted by Kempen (2018) aimed at investigating preservice teachers' evaluation regarding verification as well as conviction. He found that generic proofs received low ratings regarding verification, while the ordinary proof (the author refers to formal proof) received very high ratings. While ratings regarding students' conviction were higher than for verification, the generic proofs still received lower ratings than the ordinary proofs. In comparison to empirical arguments, Kempen (2021) found that students gave generic proofs higher ratings regarding both familiar and unfamiliar statements, which may indicate that students "do not mix up the idea of generic proofs with purely empirical verifications" (p. 4).

Findings on the Evaluation of Ordinary Proofs

Most students (and teachers) find ordinary proofs convincing and accept them as proof (e.g., Knuth, 2002; Ko & Knuth, 2013; Martin & Harel, 1989; Ufer et al., 2009), regardless of the familiarity with the statement (see, e.g., Martin & Harel, 1989). However, Weber (2010) observed that some students in his study did not find ordinary proofs convincing, even though they accepted them as proof, in line with findings reported by Fischbein (1982), as discussed above. Interviews conducted with the respective participants indicate that one reason for these contradictory responses is that these students seem to not have fully understood the proof (but nevertheless stated the proof was valid). Further, as discussed in section 3.1, some students believe that all arguments–ordinary proofs as well as other types of arguments such as empirical ones–can only provide evidence for a statement, but are not able to guarantee its truth (Chazan, 1993).

Several studies suggest that students also evaluate *invalid* ordinary proofs as being convincing or think they constitute a proof (Knuth, 2002; Martin & Harel, 1989; A. Selden & Selden, 2003; Weber, 2010). This indicates that students (and

teachers) tend to focus on surface features such as the use of mathematical symbols and algebraic manipulations instead of the content and logical structure of the argument (Harel & Sowder, 1998; Inglis & Alcock, 2012; Knuth, 2002; A. Selden & Selden, 2003), which researchers sometimes refer to as *ritualistic aspects* of proof (e.g., Martin & Harel, 1989). A focus on the form of an argument has also been reported by Ufer et al. (2009): Most of the German high school students (about 80%) correctly evaluated the ordinary proof (the authors refer to a "formally expressed ... solution", p. 41, in German "formal dargestellte ... Lösung"), but less students (about 66 to 68%) did so regarding a correct narrative proof, suggesting that the (perceived) formality of the proof plays a role regarding its acceptance.

Moreover, Stylianou et al. (2015) found that students' answers regarding proof evaluation and construction can be contradictory: The students in the study were asked which argument (out of four) is closest to one they would produce themselves. Most students either chose a (numeric) empirical argument, a narrative deductive proof, or a symbolic deductive proof (each was chosen by about one third of the participants). But when students were asked to construct proofs to the same statements, the majority of students constructed numeric empirical arguments (45 to 75%) or narrative deductive proofs (14 to 45%). The students' answers regarding what arguments they *would* give when asked to justify a statement did not fully reflect what types of arguments they actually construct, but they were most likely influenced by what they thought would be accepted as proof. This highlights that students are often aware that empirical arguments do not meet the standards for proofs and that general, deductive arguments are necessary. However, they are often not able to produce these types of arguments, a finding which has repeatedly been reported before, for instance, by Healy and Hoyles (2000).

Aspects that Influence Students' and Teachers' Proof Evaluation

Several aspects that may influence students' and teachers' evaluation of arguments have already been mentioned above, for instance, the form or representation of an argument (e.g., A. Selden & Selden, 2003; Tabach, Barkai, et al., 2010; Ufer et al., 2009), the perceived generality (Bieda & Lepak, 2014; Ko & Knuth, 2013; Tabach, Barkai, et al., 2010), and the comprehension of the argument (Bieda & Lepak, 2014; Weber, 2010). This section provides a summary of aspects that have been identified. Only few studies have explicitly investigated which aspects influence students' or teachers' proof evaluation[5]. Ko and Knuth (2013) have identified several characteristics that may influence teachers' conviction and judgement of validity, including

[5] Mathematicians' acceptance criteria have been outlined in section 2.3.3 and are further discussed in the following section.

the ones just mentioned. Other aspects they identified include the clarity and explanatory power of the argument, the familiarity with the type of argument (the authors refer to *similarity*), or the usage of mathematical facts (e.g., definitions or theorems). They report that the 55 participating middle school teachers most often referred to the generality of the argument regarding their conviction and to the usage of algebraic rules or mathematical symbols regarding the validity of arguments. In part, the coding scheme for acceptance criteria proposed by Sommerhoff and Ufer (2019) contains similar aspects to those identified by Ko & Knuth, such as the usage of counterexamples, understanding the argument, and *aesthetics*. But other characteristics differ. In particular, Sommerhoff & Ufer considered the structure of the proof, the proof scheme, and the logical chain in their coding, which were proposed as students' *methodological knowledge* by A. Heinze and Reiss (2003). Moreover, during their coding process, they identified additional categories, for instance, references to the argument (not) being a mathematical proof or the requirement of proofs being unambiguous. Sommerhoff and Ufer (2019) analyzed school and university students' as well as active mathematicians' justifications of why they think the purported proofs are valid or not using the coding scheme for acceptance criteria. They found that the proof structure, proof scheme, logical chain, and understanding seemed to be the most important acceptance criteria overall. They emphasize that *understanding* seemed to be particularly relevant for school and university students.

In the following section, I reflect on the research findings on students' proof evaluation, for instance, regarding their alignment with mathematicians' proof evaluation and differences in research approaches.

Reflection on Research Findings on Proof Evaluation

In the research findings outlined above, several aspects can be identified that may influence results on proof evaluation and therefore possibly limit comparability:

- the different (mathematical) background of participants (e.g., some are preservice teachers, others mathematics majors; the age, etc.)
- the different research foci and designs (e.g., conviction vs validity, phrasing of questions, etc.)
- a different (and sometimes unclear) understanding of (the level of) *conviction*.

The first two aspects provide indications of potential influencing factors and conditions under which students might find particular types of arguments convincing or accept them as proof, and how conviction and validation might be connected. These should be considered for future investigations. To address the last aspect–a different understanding of conviction–Weber and Mejia-Ramos (2015) have introduced the

notion of *relative* and *absolute conviction*: Relative conviction is thereby defined as a "subjective level of probability" regarding the truth of a claim that "exceeds a certain threshold"; absolute conviction, however, is "a stable psychological feeling of indubitability about that claim" (p. 16). Weber and Mejia-Ramos (2015) argue that mathematicians have absolute conviction in certain statements, mainly those that are well known for a long time, but also relative conviction regarding the truth of "more sophisticated claims" (p. 16) as well as the validity of proofs. As an example, they refer to Hales's proof of Kepler's Conjecture about sphere packing in the Euclidean space (a computer-assisted proof by exhaustion) that was published in 2005 in the *Annals of Mathematics* (Hales, 2005). According to the editor of the journal, "the reviewers were only 99% sure the proof was valid" (see Weber & Mejia-Ramos, 2015, p. 16). Even though the editor and reviewers only had (high) *relative conviction* and not absolute conviction in the validity of the proof, (part of the) proof got published[6]. The concept of relative and absolute conviction is related to what Duval (1990) refers to as the *epistemic value* of a statement. That is, "a personal judgement of whether and how the proposition is believed. It can take on values such as opinion, belief, certainty, principle, hypothesis, etc." (Reid & Knipping, 2010, p. 74). In theory, mathematical proof "has the function of changing the epistemic value of a statement, for example from conjecture to theorem" (p. 75), thus, leading to *absolute conviction* in the truth of the statement.

With respect to the assessment of findings on students' proof evaluation, Weber & Mejia-Ramos (2015) argue that researchers should verify if students have relative or absolute conviction regarding the truth of statements and proof. For instance, students' conviction of empirical arguments is not necessarily problematic, if these arguments only lead to *relative conviction* in the truth of the statement. Similarly, students having doubts about the truth of a statement "after reading or producing a proof of the statement" (p. 19) may also be appropriate, if students only have relative conviction in the validity of the proof. The distinction in relative and absolute conviction may therefore be useful in interpreting and comparing research findings of students' proof evaluation, regarding both, conviction in the truth of statements and validity of proofs.

To assess students' evaluation of proof (and other proof-related activities), many researchers in mathematics education refer to mathematicians' conceptions of proof and their respective acceptance criteria as a benchmark (e.g., Dawkins & Weber, 2017; Harel & Sowder, 2007; Stylianides, 2007; Weber, 2013; Weber & Czocher, 2019). Thereby, proving practices in the mathematics classrooms are not expected

[6] In 2017, a team of researchers lead by Hales published a formal proof of the conjecture in the journal *Forum of Mathematics*, see Hales et al. (2017)

"to be exact replicas of professional mathematical communities" (Weber & Czocher, 2019, p. 253), but general standards for the acceptance and understanding of proof should be consistent with those of the mathematical community (Dawkins & Weber, 2017; Harel & Sowder, 2007)[7]. Respective acceptance criteria have already been discussed in section 2.3.3. In summary, based on Weber and Czocher (2019), mathematicians seem to

- agree on the acceptance of *typical arguments*;
- agree on the non-acceptance of invalid proof schemes, for instance, empirical arguments;
- disagree on *atypical arguments* such as visual arguments and computer-assisted proofs.

Furthermore, being familiar with the form of reasoning and representation that is used in an argument can be seen as an important criterion for the acceptance of proof. As a conclusion, students should accept *typical arguments* they are familiar with–which implies that they need sufficient experience until they can be expected to accept such arguments–but not invalid proof schemes, in particular empirical arguments. Less obvious is the acceptance of *atypical arguments*, for instance, generic proofs: *Should* students accept those as proof? As there might not be a clear consensus among mathematicians regarding the validity of such arguments, it is difficult to give an absolute answer. It seems to be more useful in this regard, to let students explain *why* they accept an argument and assess if the reasoning is consistent with characteristics of proof as outlined in section 2.3.3. Moreover, as noted above, to decide wether an argument leads to (relative or absolute) conviction in the truth of the statement and wether it meets the criteria for proof (personal conviction vs public validation) might lead to different outcomes.

Even though there seems to be a consensus among mathematicians that empirical arguments should not be accepted as proof, a considerable percentage of them claim to find empirical arguments nevertheless convincing (under specific conditions), as Weber (2013) found. He conducted an experimental study in which 97 research-active mathematicians participated. The main findings of the study are that, firstly,

[7] Some researchers have argued that the imitation of research mathematics in the classroom is limited by several factors, such as the inability to replicate the process of proving as well as the fact that school mathematics deals "with theorems that have already been proven by others" (Hanna & Jahnke, 1993, p. 433).

mathematicians seem to find empirical arguments[8] more convincing if the respective statement is about integers *having* some property than if it is about integers *not having* some property. And secondly, that the mathematical domain seems to influence the persuasiveness of arguments, as an empirical argument regarding a statement about modular congruence was more convincing for the participants than an empirical argument regarding a statement about generating primes. Furthermore, about 27% of the participating mathematicians claimed they have been convinced by empirical arguments at least once in their mathematical practice. Overall, the mean ratings of persuasiveness of the empirical arguments were not very high, but Weber's study nevertheless demonstrates that under certain conditions (some) mathematicians gain *personal (relative) conviction* by empirical arguments. Because students need "to understand under what condition this type of evidence [empirical proof] might be appropriate and informative" (Weber, 2013, p. 110), mathematicians' proof evaluation should be considered, when judging students evaluation (in particular of empirical arguments) and teaching proof. Therefore, further research on the conditions under which particular types of arguments (such as empirical ones or atypical arguments, e.g., generic proofs) can provide high levels of conviction for mathematicians is needed.

In summary, research findings on students' proof evaluation of different types of arguments is at least partially ambiguous. Most students (and teachers) seem to find ordinary proofs convincing and think they are valid–even when the proof is in fact incorrect. The degree to which students find generic and empirical arguments convincing and/or think they are proofs is less clear. More experienced students seem to find empirical arguments neither convincing nor think they are proofs. Overall, the perceived generality, the representation of the argument, students' understanding of the argument, and their methodological knowledge seem to influence if and how convincing or valid students judge different types of arguments. More research is needed to investigate the degree to which students find different types of arguments convincing and what aspects influence their conviction. Distinguishing between relative and absolute conviction might not only lead to more consistent outcomes, it could also assess students' actual conviction of (empirical) arguments more accurately. Further, even mathematicians do not always agree on the validity of arguments and sometimes evaluate arguments as convincing even though they do not fully meet the criteria for proof (e.g., empirical arguments). These findings

[8] In addition to receiving empirical evidence for the statement for the first 12 relevant cases, the participants in Weber's study were provided with the information that the statements have been checked by a computer up to the first 10, 000 relevant cases.

should be taken into account regarding the assessment of students' (and teachers') evaluation of proof.

The following section provides an overview of research on students' (and teachers') proof comprehension. First, developments in the assessment of proof comprehension are discussed. Following this, empirical findings on students' and teachers' proof comprehension are outlined.

3.2.4 Proof Comprehension

So far, in comparison to proof construction, not many studies have particularly focussed on the reading comprehension of arguments and proofs (Mejía Ramos & Inglis, 2009a; Neuhaus-Eckhardt, 2022; Sommerhoff et al., 2015), as already noted in section 3.2. This is somewhat surprising, because proof comprehension can be viewed as one of the main activities in mathematics university courses. Furthermore, the goal of teaching proof and argumentation in school (and university) is not mainly to convince students of the truth of a statement, but to enable a deeper understanding (e.g., de Villiers, 1990; Hanna, 1990; Hersh, 1993).

The assessment of proof comprehension in school und university is mostly carried out by asking students to reproduce a proof or to adjust it to a different context, although researchers argue that this does not provide sufficient inside in students' actual proof comprehension, because correctly reproducing a proof can be achieved by solely memorizing it and does not necessarily require understanding (e.g., Conradie & Frith, 2000; Weber, 2012). Therefore, researchers have defined different aspects that may indicate proof comprehension (e.g., Bürger, 1979; Conradie & Frith, 2000; Kunimune et al., 2009; Pracht, 1979). Commonly, *proof (reading) comprehension* is thereby understood as the understanding of a particular (and valid) proof[9]. However, even though lists of relevant aspects have been collected, Mejía Ramos et al. (2012) point out that "what it means for a proof to be understood, and how we can tell if students comprehend a given proof remain open questions in mathematics education" (p. 4). Thus, to measure students' proof comprehension more systematically, researchers have recently begun to create assessment models, which are discussed in the following section.

[9] Whereas *proof understanding* often–but not always–also refers to students' proof conceptions, e.g., what students think a valid proof consists of (see, e.g., Harel, 1999; Reiss & Heinze, 2000).

Assessment of Proof Comprehension

Particularly noteworthy among developments in operationalising the assessment of students' proof comprehension are the works of Conradie and Frith (2000), Yang and Lin (2008), and Mejía Ramos et al. (2012). Conradie and Frith (2000) were among the first researchers, who emphasized the importance of measuring proof comprehension at university level and explicitly suggested proof comprehension tests[10]. They constructed two proof comprehension tests, which were used in a final exam for second-year university students at the University of Cape Town to measure students' understanding of two particular proofs. Apart from providing specific test questions, Conradie and Frith (2000) summarize the following aspects of proof comprehension that could be tested individually: "understanding of specific steps ..., understanding of the structure of the proof ..., understanding of concepts used in the proof ..., understanding of assumptions and conclusions ... and understanding of some of the more subtle aspects of a proof" (p. 231).

The *model of reading comprehension of geometry proof* (RCGP) proposed by Yang and Lin (2008) for the learning of proof in secondary schools was the first research based assessment model that aimed at defining and structuring relevant aspects of proof comprehension. It consists of five facets (*basic knowledge, logical status, summarization, generality, application*) which can be allocated between four different levels of understanding (*surface, recognizing elements, chaining elements, encapsulation*). For instance, the comprehension of *generality* is placed between the third and last level. In contrast to other researchers (see section 3.1), Yang and Lin (2008) define understanding of *generality* as understanding "what is really proved by this proof" (p. 70). In describing their model, Yang and Lin (2008) concentrated on the first three levels and did not further specify the highest level of *encapsulation*, explicitly stating that the RCGP model "is not aimed at diagnosing if a student has reached this top level" (p. 71). Furthermore, their model was particularly designed to measure geometry proof comprehension at secondary level. Therefore, Mejía Ramos et al. (2012) adapted the RCGP model of Yang & Lin to better fit the requirements at tertiary level. In particular, the model of Mejía Ramos et al. (2012) is supposed to expand on the highest level of the RCGP model, at which students need to understand the proof as a whole, for instance, understanding the main idea of the proof. To identify and justify relevant aspects of proof comprehension, they reviewed literature on different goals and methods of proof discussed in mathematics education and

[10] Houston (1993) has also raised the importance of comprehension tests in mathematics. He suggests several questions to assess students reading comprehension of mathematical articles in university courses on mathematical modelling. Even though these questions also address aspects of proof comprehension, he did not specifically design the tests to measure proof comprehension.

conducted interviews with mathematicians regarding their conceptions of proof comprehension. Furthermore, they drew from the proof comprehension questions suggested by Conradie and Frith (2000). Mejía Ramos et al. (2012) identified seven types of questions that each "measures a different facet of proof comprehension" (p. 5). They subdivided these facets into two groups: *local* and *holistic* aspects of proof comprehension. Local understanding thereby means the understanding of a particular statement within the given proof and its (logical) connection to other specific statements in the proof. Understanding the *meaning of terms and statements* (within the proof) is part of a local understanding, for example. In contrast, to gain holistic understanding, one has to understand the proof as a whole or at least its main parts, for instance, one has to be able to *summarize the main idea of the proof* or *transfer the proof's methods to another context*. Thus, holistic understanding relates to the highest level of the RCGP model, encapsulation. Mejía Ramos et al. (2012) specify three types of questions that relate to local understanding and four that relate to holistic understanding. The three local aspects can be summarized as

- understanding the meaning of terms and statements
- understanding the logical status of statements and proof framework
- being able to provide justifications of claims.

The four aspects of holistic understanding, that Mejía Ramos et al. have identified, were particularly valued by the interviewed mathematicians. These types of questions consist of:

- summarizing main ideas of the proof
- identifying the modular structure
- transferring the general idea or method to another context
- illustrating (parts of) the proof with examples.

Even though Mejía Ramos et al. (2012) state that they do not view their model to be hierarchical, they do not rule out the possibility that relationships between facets exist, for example, "being able to summarize the proof ... may be necessary in order to successfully transfer [the] ideas and methods to another context" (p. 16). The assessment model of Mejía Ramos et al. was used to design three reliable multiple-choice tests that "validly measure students' comprehension of the proofs that they read" (Mejía Ramos, Lew, Torre, & Weber, 2017, p. 140). Thus, they have demonstrated a useful alternative method of measuring students' proof comprehension compared to asking students to simply reproduce a proof, for example.

Another recent work on proof comprehension worth mentioning is the doctoral thesis of Neuhaus-Eckhardt (2022). She defines proof comprehension as the construction of a mental model of a valid proof in written form through processes of text comprehension (Neuhaus-Eckhardt, 2022, p. 36). Based on the assessment model of Mejía Ramos et al. (2012) and a literature review on proof comprehension, Neuhaus-Eckhardt proposes a list of aspects that indicate proof comprehension. She expanded the model of Mejía Ramos et al. (2012) by introducing a third group, namely, aspects of proof comprehension *beyond the particular proof* (in German "über den Beweis hinausgehende Aspekte", pp. 49–50). While Mejía Ramos et al. (2012) include aspects such as *transferring to another context* in the group of holistic understanding, Neuhaus-Eckhardt (2022) argues that such questions do not refer to the particular proof but to the underlying ideas and methods used in the proof. Thus, students' ability to transfer the idea of the proof to another context, for instance, indicates an understanding *beyond the particular proof.*

In the following section, I summarize empirical research findings on students' proof comprehension.

Findings on Proof Comprehension

Since literature on proof reading comprehension is rare (Mejía Ramos et al., 2012; Neuhaus-Eckhardt, 2022), findings from studies on proof reading in general and students' difficulties with proof are also considered in this section.

Studies have found that students tend to read proofs line by line–Inglis and Alcock (2012) refer to *zooming in*–focussing on local aspects (A. Selden & Selden, 2003), in contrast to mathematicians, who, as already mentioned above, claim to value a holistic understanding of proof (e.g., Mejía Ramos & Weber, 2014). Inglis and Alcock (2012) conducted an eye-tracking study, which confirmed the findings of A. Selden and Selden (2003) that students tend to focus on *surface features* of proof, such as notational and computational aspects, instead of the logical structure of the arguments. Moreover, the mathematicians in their study "made nearly 50% more between-line saccades than the undergraduates" (Inglis & Alcock, 2012, p. 380), suggesting that mathematicians tried more often to connect statements between lines on a local level. However, no evidence was found that mathematicians "engage in zooming out" (p. 380), meaning a non-sequential reading strategy to identify links between different parts of the proof on a holistic level. This can be seen as a contradiction to mathematicians' self-report on proof reading, according to which mathematicians claim to first start with skimming the proof (Mejía Ramos & Weber, 2014; Weber, 2008). Thus, mathematicians' behaviour may to be different from what they claim they do when attempting to comprehend or validate a proof.

Students' focussing on local aspects, in particular surface features, might not be surprising, because many students' seem to already lack basic knowledge. For instance, they have difficulties with knowing and understanding definitions, notations, and theorems (Conradie & Frith, 2000; Moore, 1994; Reiss & Heinze, 2000). In a study on high school students' understanding of proofs, Reiss and Heinze (2000) found that only about 9% were able to correctly define the concept of *congruence* and only about 11% were able to state a theorem involving congruence.

Furthermore, many students do not seem to understand the logical status of statements and the purpose of specific statements used in the proof, for instance, they have difficulties distinguishing between assumption and definition (regarding proof by contradiction) or between assumption and conclusion (Conradie & Frith, 2000). As already discussed in section 3.2.1, unpacking the logical structure, in particular interpreting and understanding the meaning and order of universal and existential quantifiers, seems to be another difficulty students encounter (Dubinsky & Yiparaki, 2000; J. Selden & Selden, 1995) which can be seen as an obstacle regarding proof comprehension.

Even though researchers have argued that generic proofs can improve students' proof comprehension by making the ideas more accessible to them (Dreyfus et al., 2012; Mason & Pimm, 1984; Rowland, 2001), not many empirical studies explicitly investigated the influence of different types of arguments on proof comprehension. Findings on proof comprehension of generic proofs in comparison to ordinary proofs are not consistent so far. In a qualitative study with ten first year engineering students from a university in Israel, Malek & Movshovitz-Hadar (2011) found that students, who were presented with generic proofs (or transparent pseudo proofs, as they call them, see section 2.4.2), performed better at proof comprehension than students, who received ordinary proofs. However, this was only the case for proofs involving methods with which the students were not familiar and that were based on ideas that could easily be transferred to another context. To provide more evidence for the influence of generic proofs on proof comprehension, Lew et al. (2020) employed an experimental quantitative study in which 106 mathematics students from universities in the United States and Canada participated. Students were randomly assigned to either receive a generic proof or an ordinary proof. All participants then had to complete a proof comprehension test based on the assessment model of Mejía Ramos et al. (2012). They did not find evidence that the generic proof lead to better proof comprehension than the ordinary proof. Similarly, Fuller, Weber, Mejia-Ramos, Rhoads, and Samkoff (2014) used the assessment model of Mejía Ramos et al. in a quantitative study with 300 mathematics students to investigate proof

comprehension of so-called structured proofs[11] in comparison to ordinary proofs. They could not find consistent evidence that students generally perform better on proof comprehension tests when presented with structured proofs. Even *if* generic or structured proofs do not lead to better proof comprehension for mathematics university students–for which further evidence is needed–they could still potentially improve proof comprehension of high school students or at the transition from school to university.

In summary, comparatively few studies have systematically investigated students' proof comprehension. Most students seem to focus on local aspects instead of holistic aspects. They often lack basic knowledge to comprehend particular statements or terms used in the proof and have difficulties with understanding the logical status of statements and unpacking the logical structure. These findings are not surprising, as similar difficulties have been reported regarding students' comprehension of mathematical statements (see section 3.2.1). So far, the influence of different types of proofs, such as generic proofs on students' proof comprehension is not clear. Recent experimental studies suggest that they may not improve proof comprehension in comparison to ordinary proofs. Overall, the findings discussed in this section highlight the need for further research on students' proof comprehension, in particular with respect to different types of arguments.

The following section outlines frameworks and research findings regarding the justification of statements. The focus is thereby on students' so-called *proof schemes*.

3.2.5 Justification

Students' ability to construct proofs on their own is seen as a major learning goal in mathematics (Hanna, 2000; Harel & Sowder, 1998; Stylianou & Blanton, 2015; Weber, 2001). Consequently, as noted in section 3.2, most of the research on proof and proving is on students' proof construction. Many studies have shown that students at all levels as well as (prospective) mathematics teachers have difficulties with proof construction (e.g., Barkai et al., 2002; Bell, 1976; Healy & Hoyles, 2000; Hemmi, 2008; Moore, 1994; Weber, 2001).

Several reasons for these difficulties have been discussed in the literature. These include cognitive challenges, for instance, due to the fact that proving is understood as a complex activity, which requires cognitive and other skills such as problem

[11] Structured proofs were introduced by Leron (1983) as "a novel way to present proofs in terms of levels" (Fuller et al., 2014, p. 4) to highlight the main ideas and methods used in the proof.

solving (e.g., Chinnappan, Ekanayake, & Brown, 2012; Moore, 1994; A. Selden & Selden, 2013; Sommerhoff et al., 2015; Stylianou et al., 2006; Ufer, Heinze, & Reiss, 2008; Weber, 2005). As with students' proof comprehension, basic knowledge is central for success in proof construction (e.g., Bell, 1976; Chinnappan et al., 2012; Sommerhoff, 2017; Ufer et al., 2008). Further, affective and epistemological aspects such as a lack of intellectual need for proof and missing or inappropriate conceptions of proof also seem to play a crucial role (e.g., Harel & Sowder, 1998; Tall, 1989). The latter has been explained in the literature by the fact that in the teaching of proof, not enough emphasis is put on "gradually refining students' conceptions of what constitutes evidence and justification in mathematics" (Harel & Sowder, 1998, p. 237).

In the following, the focus is on so-called *proof schemes* students (and teachers) demonstrate when asked to justify mathematical statements. These are described in the following section and different categories that have been identified in the literature are discussed. After that, findings on students' proof schemes are reviewed.

Frameworks for Students' Proof Schemes

Several studies have been conducted to identify different types of arguments school and university students use when asked to justify a mathematical statement (Balacheff, 1988b; Bell, 1976; Harel & Sowder, 1998; Recio & Godino, 2001). Reference is sometimes made to so-called *proof schemes* (Harel & Sowder, 1998; Lee, 2016; Recio & Godino, 2001) to separate these types of arguments from other distinctions, such as the distinction of *proofs that prove* and *proofs that explain* made by Hanna (1990) and classifications regarding the content and method of proof made by Usiskin (1980) (see also Harel & Sowder, 1998). Harel and Sowder (1998) emphasize that the notion of *proof schemes* should not be interpreted "in terms of mathematical proof in its conventional sense" (p. 275), but as arguments someone is convinced by or thinks others may find convincing. Investigating and understanding students' proof schemes is perceived as useful among researchers to better understand students' difficulties with proof (e.g., Balacheff, 1988b; Harel & Sowder, 1998).

An early study on students' proof schemes was conducted by Bell in 1976 with 32 pupils aged 14–15 from a grammar and two comprehensive schools. He divided students' responses to questions about the justification of statements into two main categories, empirical and deductive arguments, which both were further divided into several subcategories (see Tab. 3.4). Bell points out that the categories partly overlap, for instance, the first subcategories are both failures to provide an argument and the last subcategories are both valid proofs. Further, it should be noted that some of the statements used in the study are about finite sets, thus, those statements could validly

Table 3.4 Categories of students' proof explanations identified by Bell (1976, pp. 18–19)

Empirical arguments	**Deductive arguments**
Failure to generate correct examples or to comply with given conditions.	*Non-dependence*: One or more examples correctly worked, but not used to test the general statement; lack of awareness of connection between conclusion and details of the data.
Extrapolation: Truth of general statement inferred from a subset of the relevant cases; any apparent reasons are either assertions that the conditions have been complied with, or added fragments. The basis of the inference is clearly empirical.	*Dependence*: Attempts to make a deductive link between data and conclusion, but fails to achieve any higher category.
Non-systematic: Finds *some* of the required cases, no complete subsets, ignores the requirements to find *all*.	*Relevant, general restatement*: Makes no analysis of the situation, mentions no relevant aspects beyond what are actually in the data, but re-presents the situation as a whole, in general terms, as if aware that a deductive connection exists but unable to expose it.
Partially systematic: Finds some partially complete subsets of cases; has some awareness of the requirement to find all.	*Relevant, collateral details*: Makes some analysis of the situation, mentions relevant aspects which could form part of a proof, possibly identifies different subclasses but fails to build them into a connected argument; is fragmentary.
Systematic: Finds at least some complete subsets of cases, is clearly attempting to find all.	*Connected, incomplete*: Has a connected argument with explanatory quality, but is incomplete.
Check of full finite set of cases.	*Connected, side-step*: Failing only because it appeals to facts or principles which are no more generally agreed than the proposition itself (a 'side-step').
	Complete Explanation: Derives the conclusion by a connected argument from the data and from generally agreed facts or principles.

be proven by checking all relevant cases. Therefore, not all subcategories proposed by Bell (1976) might be relevant or useful for categorizing students' justifications of statements about infinite sets.

Table 3.5 Summary of students' main proof schemes identified by Harel and Sowder (1998, p. 245)

Students' proof schemes		
External conviction	**Empirical**	**Analytical**
Ritual: The student focusses on the appearance of the argument.	*Inductive*: The student verifies the statement by checking "*one or more* specific cases" (p. 252).	*Transformational*: The students' "observations involve operations on objects and anticipations of the operations' results" (p. 258).
Authoritarian: The student refers to a textbook or teacher.	*Perceptual*: The students' justification is based on "rudimentary mental images", thus, they do not "*anticipate results of transformations*" (p. 255) and therefore do not see the casual relationship underlying the observation.	*Axiomatic*: The student "understands that at least in principle a mathematical justifications must have started originally from undefined terms and axioms" (p. 273)
Symbolic: The student does not approach the problem with "comprehending its meaning", but with directly starting to "manipulate the symbolic expressions involved" (p. 251).		

Another well-known and often cited study, which aimed at identifying students' usage of arguments, was conducted by Harel and Sowder (1998). Through an exploratory study that consisted of classroom observations, interviews, and students' homework and tests, they identified three main categories of college students' proof schemes: *external conviction proof schemes*, *empirical proof schemes*, and *analytical proof schemes*, each with several subcategories. Table 3.5 provides a summary of the main categories and subcategories identified in the study. The empirical proof scheme subcategory of *inductive* arguments relates to what other researchers, such as Bell (1976), usually refer to as empirical arguments. In contrast to Harel and Sowder (1998), Bell (1976) did not identify *perceptual* arguments in his study. Bell mainly made distinctions within a category regarding the degree of completeness and systematization. The analytical proof schemes identified by Harel and Sowder (1998) relate to Bell's deductive arguments as Harel & Sowder (1998) describe

analytical proof schemes as those that "validate conjectures by means of logical deduction" (p. 258). However, unlike Bell (1976), who categorized students' deductive arguments mainly by successfulness and degree of completeness of the argument, Harel and Sowder (1998) identified two types of deductive/analytical proof schemes, *transformational* (which include generic proofs, for example) and *axiomatic proof schemes*.

Recio and Godino (2001) also conducted a study on students' proof schemes. They categorized first-year university students' responses into the following five categories (which were identified in an earlier study, see Recio & Godino, 1996):

1. The answer is very deficient (confused, incoherent)
2. The student checks the proposition with examples, without serious mistakes.
3. The student checks the proposition with examples, and asserts its general validity.
4. The student justifies the validity of the proposition, by using other well-known theorems or propositions, by means of partially correct procedures.
5. The student gives a substantially correct proof, which includes an appropriate symbolization.

While Recio and Godino (1996; 2001) did not explicitly define upper categories, the five categories could be divided into either empirical or deductive arguments (the first category can be viewed as *unclear*, i.e., no clear type of argument/proof scheme can be identified; the second and third category can be seen as empirical arguments; the fourth and fifths category as deductive arguments), similar to those identified by Bell (1976) and Harel and Sowder (1998) (who used the term *analytical*, or, more specific *axiomatic*).

Other researchers have build upon these three systems of categories. Kempen (2019), for instance, grounded the development of categories on Bell (1976) and Recio and Godino (2001), thus focussing on empirical and deductive proof schemes. Thereby, he introduced the category *pseudo argument*, referring to arguments that are circular, redundant or simply incorrect (see p. 118). Lee (2016), on the other hand, based his categories essentially on the three main categories identified by Harel and Sowder (1998). However, Lee's *levels* are not always strictly divided by the main categories proposed by Harel & Sowder, such as empirical and (incomplete or false) deductive proof schemes. For example, students who based their justification on examples and those, who used incorrect logical reasoning, were allocated to the same level.

In summary, the three main categories of students' proof schemes that have been identified by at least one of the studies discussed above are external proof schemes, empirical proof schemes, and deductive or analytical proof schemes. A further

distinction regarding the completeness of deductive arguments as well as other aspects, such as the subtype of proof scheme as proposed by Harel and Sowder (1998), seems to be useful for analyzing students' attempts to justify mathematical statements. In the following section, empirical findings on students' usages of different types of arguments are summarized.

Findings on Students' Proof Schemes

Several studies found that many school and first-year university students as well as teachers give empirical arguments when asked to justify a statement (Balacheff, 1988a; Barkai et al., 2002; Bell, 1976; Bieda, 2010; Healy & Hoyles, 2000; Housman & Porter, 2003; Lee, 2016; Recio & Godino, 2001; Sears, 2019; Sen & Guler, 2015; Stylianou et al., 2006). For instance, Recio and Godino (2001) report that about 40% of the 429 first-year university students, who participated in their study, gave empirical arguments to justify universal statements. Similarly, about half of the participants in a study conducted by Barkai et al. (2002) with 27 elementary school teachers used empirical arguments to *prove* a universal statement. In line with the categories proposed by Bell (1976), the nature of empirical arguments used by students (and teachers) seems to differ in that some of them only choose random examples while others search for patterns (Stylianou et al., 2006). In a study conducted by Housman & Porter (2003) with eleven above-average mathematics students, the majority of participants used perceptual arguments, only one student made also use of inductive arguments (both as defined by Harel & Sowder, 1998). But some students seem to even have difficulties to produce an empirical argument, as Bell (1976) found: About one forth of the participants were allocated to the respective first subcategory of *empirical arguments* (see Tab. 3.4). Those students were not able to generate correct examples, which Bell (1976) explains with a lack of knowledge and "an inability to coordinate all the data" (p. 34). Another 19% of the participants in Bell's study checked all relevant cases regarding a statement about a finite set, thus proving the correctness of the statement by exhaustion. However, none of the students gave a complete explanation for neither of the statements and only one student was able to give some (deductive) explanations, even though these were incomplete.

There is strong evidence that many students as well as teachers fail to construct valid deductive arguments (Barkai et al., 2002; Bell, 1976; Healy & Hoyles, 2000; Kempen, 2019; Lee, 2016; Recio & Godino, 2001; Sears, 2019; Sen & Guler, 2015; Sevimli, 2018; Stylianou et al., 2006). Most of the respective studies report that less than half of the participants justified true universal statements (and false existential statements) with a complete and correct proof (e.g., Barkai et al., 2002; Kempen, 2019; Lee, 2016; Recio & Godino, 2001). Findings suggest that many first-year

university students thereby often seem to construct *pseudo arguments* (e.g., Kempen, 2019; Stylianides & Stylianides, 2009). For instance, Kempen (2019) reports that about 26% of the 149 preservice teachers who were asked to justify a statement about the sum of two even numbers used such incorrect deductive arguments; about 23% gave an incomplete deductive argument, about 20% constructed a valid proof, about 9% gave empirical arguments, 8% no justification at all, and about 13% seemingly did not answer. The percentages differed substantially regarding the semester. For instance, first-semester preservice teachers more often gave empirical arguments (about 14%) and less often incomplete or complete deductive arguments (about 9 and 10%). Although research findings are generally consistent regarding students' difficulties with correct justifications, success at constructing proofs seems to depend on several factors such as the respective statement (universal vs existential, the truth value, mathematical context, etc.) as well as age (see also Reid & Knipping, 2010). For instance, the statements used in Barkai et al. (2002) consisted of three universal (one true, two false) and three existential statements (two true, one false), all in the context of divisibility. Depending on the statement, the percentages of teachers who gave correct justifications (either proving or disproving the statement) varied between about 23% (regarding the false existential statement) and 96% (regarding a true existential statement). The true universal statement could only be proven by about 40% of the participants; the success rates regarding the two false universal statements were significantly higher with 69 and 88%. In those cases, the statements could be disproven with providing just one counterexample, which is an easier task than producing a deductive argument that holds for an infinite number of cases. Similarly, proving a true existential statement only requires finding one example; for proving the falsity of an existential statement one needs to construct a general deductive argument.

Fewer studies have reported results on students use of external conviction proof schemes (Harel & Sowder, 1998; Sears, 2019; Sen & Guler, 2015; Sevimli, 2018; Stylianou et al., 2006). In a study conducted by Stylianou et al. (2006, p. 57) with 34 first-year mathematics students, about 20 to 35% (depending on the task) gave externally based arguments. These justifications were mainly based on symbolic manipulations or a redesign of the statement, but not on authority (e.g., a textbook or teacher). Sevimli (2018) also reports on students' usages of external arguments. Most of the justifications given by his participants (172 first-semester students from three different mathematics departments in Turkey) belonged to the external proof scheme. In contrast to the findings reported by Stylianou et al., these students particularly made reference to authority. School students also often make reference to authority, as Sen and Guler (2015) found. They conducted a study with 250 7th Grade students from Central Anatolia. Most of the participants used either external

or empirical arguments to justify mathematical statements. Thereby, the externally based arguments were mainly authoritarian and ritual (but in particular for one statement also symbolic). Sears (2019) reports on a (small-scale) study with similar results, but regarding the justification of statements given by preservice middle and secondary school teachers. The six participants mainly used external and empirical arguments, and reference was often made to authority. However, due to the small number of participants, these findings are not generalizable.

Several studies have investigated if students and teachers, who give empirical arguments to justify universal statements, are aware that these types of arguments do not meet the requirements of proof (Barkai et al., 2002; Stylianides & Stylianides, 2009). Findings suggest that many (if not most) participants are aware that a general argument is needed. For instance, in the study conducted by Barkai et al. (2002), about 20% of the 27 participating elementary school teachers stated that they know that a general proof is needed, but that they lack necessary knowledge to construct such an argument. Similarly, in a study conducted by Stylianides & Stylianides (2009) with 39 prospective elementary school teachers, most of the participants who submitted empirical arguments were aware that their arguments were not valid as proofs. Thus, as Weber and Mejia-Ramos (2015) pointed out, "behaviour on justifications tasks is [not] a sufficient warrant to establish this claim [referring to the claim that students are convinced by empirical arguments]" (p. 16). Not all students and teachers, who use empirical arguments to justify a universal statement, might do so because they think these are sufficient to prove a statement, but because they are simply not able to construct a valid proof. However, as discussed in section 3.2.3, some students and teachers *do* think empirical arguments are sufficient (see, e.g., Martin & Harel, 1989).

Overall, students as well as teachers have difficulties with the construction of ordinary proofs. Several aspects may influence students' success to construct valid arguments, for instance, the truth value of the statement. Further, most students give empirical arguments when asked to justify the truth of a statement, potentially because of their inabilities to construct ordinary proofs and not because they assume that these arguments are sufficient. Some studies have reported on (high school) students' external proof schemes such as authoritarian arguments. However, it is unclear what characteristics of the statements (e.g., truth value or familiarity) and students (e.g., age, experience,...) influence the usage of these, but also other types of arguments, which highlights the need for further research.

In the following section, potential relation between proof-related activities and respective research findings are discussed.

3.2.6 Relation Between Activities

As described in section 3.2, the three main activities reading a statement, reading an argument, and constructing an argument are related in that both reading and constructing an argument may support the comprehension of a statement as well as estimating its truth (see relations marked as **A** in Fig. 3.3). Vice versa, without *some* understanding of the statement, it is neither possible to decide if the statement is true or false nor to understand and evaluate a given argument or construct a novel one (the latter relations are marked as **B** in Fig. 3.3; see also discussions in sections 3.2.2 and 3.2.4, respectively).

Further, it seems plausible that proof evaluation, in particular with respect to conviction, and proof comprehension are related: If students do not understand an argument, they may judge it as not convincing, which in turn could influence the estimation of truth of the statement (see also Weber & Mejia-Ramos, 2015).

There are only few studies that have investigated the relation between the different (sub-)activities, as has already been highlighted by other researchers (e.g., A. Selden & Selden, 2017; Sommerhoff, 2017). As has already been discussed in section 2.3.3, the acceptance of an argument as proof is highly influenced by the context and individuals who read or construct a proof, thus, socio-mathematical norms. The respective acceptance criteria can not only influence the evaluation of given arguments, but also the construction of novel ones. This assumption is supported by research which found correlations (even though weak) between proof validation and proof construction (Ufer et al., 2009). Further, studies suggest that engaging with proof validation activities may positively influence proof construction (Pfeiffer, 2011; Powers et al., 2010; A. Selden & Selden, 2003; Yee et al., 2018). Findings reported by Sommerhoff (2017) suggest that the correlation between proof validation and proof construction is not mainly based on *methodological knowledge* (e.g., knowledge about appropriate proof schemes, see A. Heinze & Reiss, 2003), but an effect of different resources that underlie both of these activities (see also following section).

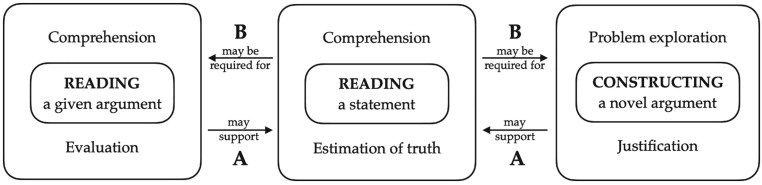


Figure 3.3 Adapted framework on proof-related activities based on Mejía Ramos and Inglis (2009b), highlighted relations

Table 3.6 Teachers' estimation of truth and correct justification by truth value and domain of discourse of statements; findings reported by Barkai et al., table adapted from (Reid & Knipping, 2010, p. 70), with permission from Brill

Statement	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
Type	Universal			Existential		
Truth value	True	False	False	True	False	True
Correct judgement	100%	100%	69%	100%	77%	68%
Correct argument	41%	88%	69%	96%	23%	64%
True for what set?	All *n*	No *n*	Some *n*	All *n*	No *n*	Some *n*

Of particular interest in the present thesis are relations between the comprehension of a statement–in particular its generality–and the estimation of truth on one hand and proof comprehension as well as justification on the other hand. Findings reported by Barkai et al. (2002) suggest a relation between the estimation of truth and justifications given by the participating teachers, in particular with respect to the truth value and domain of discourse of the statement (see Tab. 3.6, an extended version of Tab. 3.3). The first universal statement was correctly estimated as a true statement by all teachers, but only 41% constructed correct arguments to prove it. Most of the incorrect arguments were empirical (about 50%), which might imply that the empirical verifications convinced the teachers of the truth of the statement, even though they could not prove the statement (about a third of the teachers thought the empirical arguments count as proof). The truth value of the second statement (which was false) was also correctly estimated by all teachers, but significantly more teachers were able to provide a correct proof (most gave one or more counterexamples). In contrast, the third statements was correctly estimated as being false by fewer teachers (69%), however, all of these teachers were able to correctly justify their decision by providing one or more counterexamples. It seems that it was easier for the teachers do disprove a false universal statement (by providing a counterexample) than proving a true universal statement, at least for those who were able to correctly estimate the truth value. Similar observations can be made regarding the existential statements: Most of the teachers who correctly estimated the truth value of the true existential statements were able to provide a correct justification (by giving an example). But the false existential statement was proven by only 23%

of the participants, most likely because a general argument was needed (see also Reid & Knipping, 2010, p. 70).

To the author's knowledge, no studies on the relation between understanding the generality of a statement (as part of reading a statement, in particular, comprehension of a statement) and proof reading or proof construction have been conducted so far. In general, more research is needed to understand the interplay between the different (sub-)activities related to proof.

The following section aims to identify appropriate *control variables* regarding (cognitive) resources that underly students' proof skills and potentially their understanding of the generality of statements.

3.3 Resources

In research on argumentation and proof skills, several resources underlying these skills have been identified and investigated (Chinnappan et al., 2012; Sommerhoff, 2017; Ufer et al., 2008; Weber, 2001). As was pointed out by Sommerhoff (2017), no general framework or list for these resources exist so far. This section aims at giving an overview of potential resources that might be relevant for argumentation and proof skills and could therefore be used as control variables in the analysis of students' understanding of generality and other proof-related activities. Thereby, the focus is on cognitive resources rather than non-cognitive ones such as motivation and beliefs[12]

The resources identified in the literature so far can mainly be categorized into *content-specific*, *domain-specific*, and *domain-general* resources (see Sommerhoff, 2017; Ufer et al., 2008). Different terms are sometimes used to refer to similar resources (e.g., *mathematical knowledge base* and *mathematical content knowledge*), here, the notation is taken from Sommerhoff (2017). Content-specific resources refer to knowledge that belongs "to a specific mathematical content area" (Sommerhoff, 2017, pp. 45–46). It contains *conceptual* (e.g., knowledge about concepts and definitions) as well as *procedural knowledge* (e.g., knowledge about rules and procedures). Domain-specific resources, such as *mathematical strategic knowledge* (first introduced by Weber, 2001) and *methodological knowledge* (see further above), are not specific to a particular mathematical content but belong "to the [general] field of mathematics" (Sommerhoff, 2017, p. 46). In contrast, domain-general resources are

[12] No consistent results regarding affective aspects and beliefs have been shown yet (see, e.g., A. Heinze & Reiss, 2009; Sommerhoff, 2017). Sommerhoff (2017) argues that this "may result from ambiguities in the definition of the diverse affective constructs" (p. 76) as well as difficulties to measure them.

not specific to mathematics and include *problem-solving skills* and (general) *reasoning skills* (Sommerhoff specifically refers to *conditional reasoning skills*, i.e., reasoning skills needed to handle conditional statements).

According to the literature review conducted by Sommerhoff et al. (2015) (see also section 3.2), mathematical content knowledge was considered most often in PME research reports on proof and argumentation (about 47%). Methodological knowledge was studied by only 17% and problem-solving skills by 18%. Only few research reports investigated other resources such as mathematical strategic knowledge or beliefs (3–5%). Conditional or more general reasoning skills were seemingly not analyzed in any of the research reports considered in the literature review.

Overall, research findings suggest a strong impact of content- and domain-specific resources on activities related to proof and argumentation (Sommerhoff, 2017). In particular, mathematical content knowledge seems to be a main predictor for students' performance in proof construction, as several studies have shown (Chinnappan et al., 2012; Sommerhoff, 2017; Ufer et al., 2008). Further, Weber (2001) reports that a lack of mathematical strategic knowledge is "a primary cause for undergraduates' failure" (p. 115) in proof construction. The quantitative study conducted by Sommerhoff (2017) provides further evidence for the influence of mathematical strategic knowledge on students' performance in proof construction and validation.

The importance of problem-solving skills for proof competencies and their relation has been highlighted by many researchers (e.g., Chinnappan et al., 2012; Selden & Selden, 2013; Stylianou et al., 2006; Ufer et al., 2008; Weber, 2005). However, its actual influence on students' performance in activities related to proof is not clear yet. Several studies found significant correlations (Chinnappan et al., 2012; Ufer et al., 2008), while Sommerhoff (2017) reports a low and insignificant effect on students' performance in proof construction. Sommerhoff assumes that mathematical strategic knowledge, which was included in the regression model, "reduces the impact of problem-solving skills, as mathematical strategic knowledge can partially be seen as a domain-specific analogue of problem-solving heuristic" (p. 89). However, he furthermore found that there seems to be a correlation between performance in proof validation and problem-solving skills. Further research is needed to better understand the impact of students' problem-solving skills on proof performance.

Regarding mathematical reasoning skills, Chinnappan et al. (2012) report a significant influence on students' success in proof construction. However, mathematical reasoning skills were measured using a geometry test conducted at the end of Grade 10. Even though this test requires deductive reasoning, the skills needed to solve the tasks seem to overlap with other resources such as mathematical content

knowledge. According to findings reported by Sommerhoff (2017), conditional reasoning skills (measured by questions in which participants had to accept or reject logical inferences) seem to "play a minor role compared to the domain-specific resources" (p. 90).

To my knowledge, other more general cognitive measures have not been considered as potential resources underlying argumentation and proof skills so far. In section 5.3.6, the so-called *Cognitive Reflection Test* (CRT) is introduced as an instrument to control for individual differences in cognitive resources.

Derivation of Research Questions

4

Argumentation and proof are undoubtedly complex and challenging parts of mathematics. In particular, during the transition from school to university, which coincides in many countries (such as Germany) with the introduction to proof-based mathematics, several activities related to proof have been shown to be challenging for many students (e.g., Kempen, 2019; Moore, 1994; Recio & Godino, 2001; Weber, 2001). Furthermore, since proof and argumentation are also major learning goals in school mathematics (e.g., Kultusministerkonferenz, 2012), it is important to investigate the knowledge and difficulties of (preservice) teachers. When teachers have an insufficient understanding of proof—which research suggests—it is not surprising that students do so as well (e.g., Reid & Knipping, 2010).

In Chapter 3, research findings on students' and teachers' proof skills and understanding were discussed. Thereby, several gaps could be identified. So far, research on proof and argumentation has mainly focused on activities related to the construction of novel arguments and partly to the reading of given arguments (Mejía Ramos & Inglis, 2009a; Sommerhoff et al., 2015). The comprehension of statements which are to be proven—or for which a proof has to be read—and underlying principles, however, have largely been neglected in research. Understanding the generality of mathematical statements and proofs is an essential part of the comprehension of statements and students' proof skills, because it is the mathematical generality that is the defining element of mathematical proof and what makes mathematics unique (see Section 2.2). However, to my knowledge, neither the extent to which students lack understanding of the generality of statements nor the relation to reading and constructing different types of arguments have been researched yet. Therefore, this thesis particularly aims at investigating students' understanding of the generality of statements and its potential connections to activities related to proof. This seems especially relevant for studies that have reported on students' high conviction regarding empirical arguments (and their validity) (e.g., Healy &

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_4

Hoyles, 2000; Martin & Harel, 1989) and students' belief that it is impossible to prove a universal statement (for every case) *at all* (Chazan, 1993). Such an understanding of proof might be related to an insufficient understanding of generality. Without an understanding of the generality of statements, it might be difficult for students to develop an intellectual need for proof and appropriate conceptions of proof. Further, because previous studies have reported ambiguous research findings regarding students' conviction, comprehension, and construction of different types of arguments (see Sections 3.2.3, 3.2.4, and 3.2.5, respectively), the present study aims to provide more clarity on this matter and to investigate the relation between students' understanding and conviction of different types of arguments and their understanding of generality. The types of arguments that are of interest in this thesis are empirical arguments, generic proofs, and ordinary proofs, because of their prominent role in mathematics education (see Sections 2.3.2 and 2.4.2). Because research findings have provided evidence for the influence of the truth value on students' performance in several proof-related activities, such as the estimation of truth (e.g., Barkai et al., 2002), both true and false statements were considered. Moreover, the influence of the familiarity with statements (and arguments) on the understanding and acceptance of proof has often been highlighted in the literature (e.g., Dubinsky & Yiparaki, 2000; Hanna, 1989; Stylianides, 2007; Weber & Czocher, 2019). The familiarity with the statement was therefore also considered as a potential influence on students' understanding of generality of statements and their performance in other proof-related activities. In short, in this study, the *type of statement* refers to characterizing statements by their truth value and students' (expected) familiarity with the statement.

The research framework is based on the adapted framework shown in Figure 3.2. That is, it is assumed that reading a statement, in particular estimating its truth, is influenced by the reading and/or construction of arguments. Understanding the generality of a universal statement, as part of the comprehension of statements and underlying (logical) principles, is defined here as consistent responses regarding the estimation of truth and the potential existence of counterexamples (see Sections 2.1 and 2.2, as well as Section 5.3.5 for the exact definition used in this study).

To control for individual differences in students' responses, resources and background information were taken into account. Thereby, the focus was on cognitive resources, because firstly, research on non-cognitive resources such as beliefs and affects has not provided evidence for major direct effects independent of cognitive resources (e.g., Furinghetti & Morselli, 2009; Herppich et al., 2017; Semeraro, Giofrè, Coppola, Lucangeli, & Cassibba, 2020), and secondly, the scope of the survey should be reasonable for participants (see also Sommerhoff, 2017). Cognitive resources that are commonly considered (see Section 3.3) are content-specific

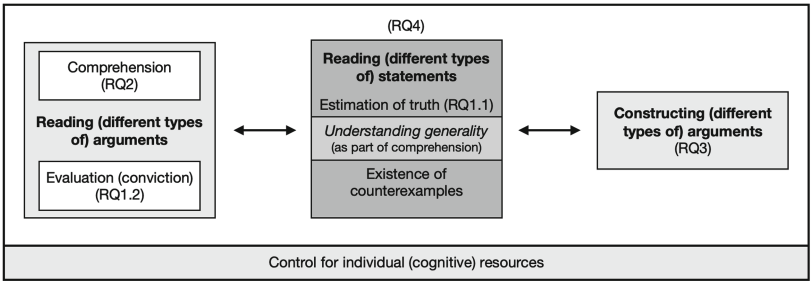


Figure 4.1 Overview and relation of research questions

knowledge, domain-specific knowledge (such as mathematical strategic knowledge), and domain-general knowledge (such as problem-solving skills and general reasoning skills). Due to its short length and positive correlation with problem-solving and reasoning skills, a cognitive reflection test (see Section 5.3.6) was used to control for individual differences in participants' general cognitive skills. Further control variables are specified in Section 5.4.1. Figure 4.1 provides an overview of the research questions and their relation.

The focus of the first set of research questions is on students' conviction of the truth of universal statements and potential relations to reading different types of arguments. Students' conviction of the truth of statements was thereby divided into students' performances in two proof-related activities: the estimation of truth and proof evaluation regarding conviction. Firstly, the potential influence of reading different types of arguments on students' estimation of truth was investigated. As previous research has provided evidence for the influence of characteristics of the statement on the estimation of truth (e.g., Barkai et al., 2002, see Section 3.2.2), the effect of the *type of statement* was also analyzed. The second research question in this set aims at investigating students' evaluation of different types of arguments. More precisely, the aim is to find out how students rate the conviction they gain regarding the truth of statements from reading different types of arguments. In contrast to other studies (e.g., Martin & Harel, 1989; Stylianou et al., 2015; Tabach, Levenson, et al., 2010; Ufer et al., 2009), the interest is not on students' validation of arguments (i.e., which arguments students identify as valid proofs). Furthermore, aspects of arguments students claim to find not convincing were identified. It was of particular interest if students refer to the (lack of) generality and what role the comprehension of the statement and proof plays regarding conviction.

RQ1: Conviction of the truth of universal statements and its relation to reading different types of arguments

RQ1.1: How do the type of argument and the type of statement influence students' estimation of the truth of universal statements?
RQ1.2: How do the type of argument, the type of statement, and the level of comprehension influence how convincing students find different types of arguments? What aspects of mathematical arguments do students identify as not convincing?

Regarding students' conviction of arguments (re. RQ1.2), I expected significant differences regarding the evaluation of different types of arguments, with ordinary and generic proofs receiving higher levels of conviction than empirical arguments. Even though findings on students' conviction of different types of arguments are ambiguous (see Section 3.2.3), many university students—which are investigated in this project—seemingly (and desirably) tend to be more convinced by deductive proofs than by empirical ones (e.g., D. Miller & CadwalladerOlsker, 2020; Weber, 2010). A positive effect was also expected regarding familiar statements, because the role of being familiar with statements and modes of argumentations for the acceptance of proof has been highlighted in the literature (e.g., Hanna, 1989). However, so far, no such influence on proof evaluation has been shown (e.g., Kempen, 2021; Martin & Harel, 1989). As research findings suggest that the comprehension of the argument is important regarding students' conviction of it (e.g., Sommerhoff & Ufer, 2019; Weber, 2010), a positive effect was expected, i.e., students' with higher levels of comprehension also have higher levels of conviction. Further, it was expected that many students would claim to have difficulties with understanding the argument when asked about aspects that they find not convincing. I also expected students to refer to the generality of the argument, in particular regarding empirical arguments, but also generic proofs, because *generality* has also been identified as an important aspect in previous research (e.g., Bieda & Lepak, 2014; Ko & Knuth, 2013; Lesseig et al., 2019; Tabach, Barkai, et al., 2010).

In contrast, regarding the influence of different types of arguments on students' estimation of truth (re. RQ1.1), no studies have been conducted so far that would lead to a respective hypothesis. If students' evaluate the proofs regarding conviction based on their *actual* conviction of the truth of the statement, I would expect a similar influence of different types of arguments on students' estimation of truth, that is, students should be more likely to evaluate (true) statements as true when reading generic and ordinary proofs than when they receive empirical arguments. Otherwise, their responses would be inconsistent. However, previous research findings suggest that students as well as mathematicians often use empirical arguments to estimate the

truth value of a statement before proving it (e.g., Alcock & Inglis, 2008; Buchbinder & Zaslavsky, 2007; Lockwood et al., 2016). Therefore, it would also be possible that empirical arguments provide students with an intuition of the truth value of the statement, and thus, make them more likely to evaluate the statement as true. As mentioned above, prior research suggests that the truth value of a statement as well as being familiar with a statement may influence students' (and teachers') estimation of truth (e.g., Barkai et al., 2002; Buchbinder & Zaslavsky, 2007; Dubinsky & Yiparaki, 2000; Hanna, 1989; Ko, 2011). In particular, researchers have argued that it is more difficult for students to correctly estimate the truth value of false statements than the truth value of true statements (Buchbinder & Zaslavsky, 2007; Ko, 2011). Therefore, I hypothesized a negative effect for the false statement compared to true statements. Regarding familiar statements, I expected a positive effect on students' estimation of truth as suggested by other researchers (Dubinsky & Yiparaki, 2000; Hanna, 1989; Stylianides, 2007; Weber & Czocher, 2019), and simply because students' should have gained (extensive) experience with these statements during school.

The second set of research questions refers to the comprehension of generic and ordinary proofs. Firstly, I investigated the effect of the type of proof and the familiarity with the statement on students' self-reported proof comprehension. Since proof comprehension relates only to the understanding of correct proofs (see, e.g., Neuhaus-Eckhardt, 2022), I excluded the false statement in the analysis. The second research question then aimed at identifying aspects of generic as well as ordinary proofs that students claim to have not understood. Of particular interest was to investigate to what extent students identify different aspects in these types of proofs that they claim to not understand.

RQ2: Proof comprehension

RQ2.1: How does students' (self-reported) proof comprehension differ between students who receive generic proofs and those who receive ordinary proofs? How does the familiarity with the statement influence students' proof comprehension?
RQ2.2: What aspects of mathematical arguments do students identify as not understandable? How do these aspects differ regarding generic and ordinary proofs?

As was pointed out in Section 3.2.4, only few studies have investigated differences in students' proof comprehension regarding different types of arguments. With respect to generic and ordinary proof, research findings are ambiguous. However, experimental studies suggest that no differences in students' comprehension of generic and ordinary proof exist (e.g., Lew et al., 2020). These studies made use of proof comprehension tests, while the present study relied on students' self-reported level of comprehension and aspects they did not understand. I nevertheless expected similar

findings, namely, no significant difference between students' comprehension of the arguments (re. RQ2.1). Similar to proof evaluation, I expected a positive influence on proof comprehension for familiar statements in contrast to unfamiliar statements. Based on previous research on proof comprehension (e.g., Conradie & Frith, 2000; Moore, 1994; Neuhaus-Eckhardt, 2022; Reiss & Heinze, 2000, see Section 3.2.4), it was expected that students mainly refer to local aspects (Mejía Ramos et al., 2012) such as surface features (A. Selden & Selden, 2003), when asked what they did not understand. In particular, I expected students to refer to not knowing definitions, expressions, and the meaning of terms, and to a lesser extent to the logical structure of the arguments (re. RQ2.2). It was further expected that the (un)familiarity with the form of generic proofs would lead to a greater proportion of students referring to the proof framework when asked what they did not understand in comparison with ordinary proof.

The focus of the third set of research questions is on the types of arguments students use to justify the truth or falsity of universal statements. At first, students' responses were allocated to different proof schemes (coding categories were based on Section 3.2.5). Then, the relation between students' proof schemes and their level of conviction regarding the truth of statements was investigated to identify differences with respect to the type of proof scheme. This research question derived from the discussion about relative and absolute conviction, initiated by Weber and Mejia-Ramos (2015) (see paragraph on the *Assessment of Research Findings on Proof Evaluation* in Section 3.2.3).

RQ3: Construction of arguments to justify the truth of universal statements (students' proof schemes)

RQ3.1: What types of arguments do students themselves use to justify the truth or falsity of a universal statement? How do students' proof schemes differ regarding the type of statement (i.e., familiarity and truth value)?
RQ3.2: What potential relation between the type of argument used by students and the level of conviction of the truth of the statement exists?

Research findings discussed in Section 3.2.5 suggest that students (and teachers) mainly have empirical proof schemes and, depending on the context, external conviction proof schemes (e.g., Barkai et al., 2002; Bell, 1976; Recio & Godino, 2001; Sevimli, 2018; Stylianou et al., 2006). Thus, it was expected that the majority of students would use empirical arguments, in particular regarding unfamiliar statements, and to a lesser degree, regarding familiar statements, external proof schemes such as authoritarian arguments (re. RQ3.1). I expected only few students to use deductive arguments. The relation between the type of proof scheme and students'

level of conviction of the truth of the statement is less clear (re. RQ3.2). To my knowledge, no studies have explicitly investigated this questions. However, students who are able to construct a (valid) proof might be more convinced of the truth of a statement—having *absolute conviction*—than students who only use empirical arguments—who might only have *relative conviction*. On the other hand, as Polya (1954) has pointed out, "without ... confidence [in the truth of the theorem] we would have scarcely found the courage to undertake the proof ..." (pp. 83–84), which suggests that a proof might not be necessary to gain high levels of conviction in the truth of a statement.

Lastly, the fourth set of research questions investigates the primary outcome variable of this study—students' understanding of the generality of mathematical statements and potential relations to proof reading and construction. Due to the scarcity of related research, the purpose of the first question in this set is to find out the proportion of first-year university students—enrolled in different study programs, namely, primary, lower secondary, and upper secondary education as well as mathematics—who have limited understanding of the generality of mathematical statements. The remaining research questions in this set then aim at identifying factors that potentially influence students' understanding of the generality of statements. Of particular interest is the influence of reading different types of arguments—empirical arguments, generic proofs, ordinary proofs, or no arguments at all (RQ4.2). Further, it was investigated how the type of statement—both its truth value and the familiarity with the statement—influences students' understanding of generality. The third question in this set aims at investigating the influence of students' proof comprehension and level of conviction on students' understanding of generality of statements. Moreover, the potential relation between types of arguments students use to justify a statement (proof schemes) and their understanding of generality was examined (RQ4.4).

RQ4: Students' understanding of the generality of mathematical statements

RQ4.1: What proportion of first-year university students have a correct understanding of the generality of statements?
RQ4.2: What is the influence of reading different types of arguments on students' understanding of the generality of mathematical statements? How does the type of statement influence students' understanding of its generality?
RQ4.3: How does students' comprehension and conviction of arguments influence their understanding of generality of statements?
RQ4.4: What potential relation exists between students' proof schemes and their understanding of the generality of statements?

Only few studies reported on students understanding of the generality of statements (e.g., Balacheff, 1988b; Buchbinder & Zaslavsky, 2019; Chazan, 1993; Galbraith, 1981). In particular, I am not aware of (large-scale) quantitative studies that have explicitly investigated students' understanding of the generality of mathematical statements without relating it to the understanding of generality of proof. Based on the findings reported in prior studies (e.g., Buchbinder & Zaslavsky, 2019), it was expected that only a minority of students would have limited understanding of the generality of statements, with a higher proportion regarding preservice primary school teachers and a lower proportion regarding mathematics majors (re. RQ4.1). Even though the study was conducted at the beginning of their first semester, which implies no influence of the study program itself yet, it can be assumed that the students' choice for a respective study program (e.g., primary education vs mathematics major) is influenced by resources that also influence their understanding of proof (and generality), for instance, their previous mathematical knowledge based on their school education.

From the studies conducted so far, it is unclear if the understanding of generality differs by the (type of) statement and/or is related to the reading and construction of (different types of) arguments, or if it is solely influenced by the *knowledge* of the meaning of mathematical generality. To my knowledge, no studies on the influence of the type of argument on students' understanding of generality (of statements) have been conducted so far. As has been mentioned above, research that has investigated the influence of the type of argument on other activities, for instance, proof comprehension, suggests no or only minor effects with respect to generic and ordinary proof (e.g., Lew et al., 2020). However, findings on students' proof evaluation reported by Kempen (2018, 2021) suggest that students are less convinced by empirical arguments than by generic and ordinary proofs. What remains unclear and is difficult to derive from these findings is if the understanding of generality of statements is affected by the *reading* of (different types of) arguments *at all*. *If* a relation exists, a positive, but weak correlation regarding analytical arguments (generic and ordinary proofs) compared to reading no arguments or empirical arguments was hypothesized (re. RQ4.2), because valid proofs should at least in theory provide the reader with certainty that no counterexamples to true universal statements exist (Reid & Knipping, 2010). A similar hypothesis was made regarding potential correlations between students' proof schemes and their understanding of generality, that is, it was expected that students with empirical proof schemes have limited understanding of the generality of statements (see also Conner, 2022) and students with deductive proof schemes are more likely to have a correct understanding (re. RQ4.4). Regarding the effect of the truth value of statements on students' performance in proof-related activities, two contradictory observations have been

reported. On the one hand, students seem to be more familiar with (proving) true statements than with (disproving) false ones, which suggests that they might be more successful in handling them (Buchbinder & Zaslavsky, 2007; Ko, 2011). On the other hand, it seems to be easier to disprove false universal statements by providing a counterexamples than to prove a true universal statement (Barkai et al., 2002). In this case, the role of a counterexample might be more apparent for students, thus, resulting in a better understanding of the (absence of) generality of the statement. Previous research findings suggest that the familiarity with a statement seems to play no or only a minor role for proof evaluation (e.g., Kempen, 2021; Martin & Harel, 1989), even though its general importance for the acceptance of proof has often been highlighted (e.g., by Hanna, 1989). However, the familiarity with the statement might nevertheless lead to a higher awareness of its generality, as students would have applied the statement to *many cases* before. Therefore, a positive effect with respect to familiar as well as false statements was expected (re. RQ4.2). The relation between students' understanding of generality and their level of comprehension as well as conviction is also difficult to derive from prior research. A positive relation regarding both activities might be expected, because higher levels of conviction and comprehension could potentially result in higher levels of certainty regarding the absence of counterexamples (re. RQ4.3).

Methodology

5

In this chapter, the research design and methods used to answer the research questions are described in detail. First, I justify the chosen approach and give an overview of my research design. Second, the data collection process is outlined. In particular, the experiment and the construction of all instruments are described. Lastly, I explain and justify the methods being used to analyze the data.

5.1 Research Design

Based on the present research interest, a quantitative, experimental research design was chosen. In this section, I first justify the decision for the particular research approach. Then, an overview of the research process and the development of the research design is given.

5.1.1 Justification of the Research Approach

Despite numerous studies being conducted on students' proof skills, surprisingly few of them have specifically focused on students' understanding of the generality of mathematical statements. Most of these studies have used qualitative methods, making their findings limited and not widely applicable (e.g., Bryman, 2012; Mat Roni, Merga, & Morris, 2020). That is why I took a different approach. I conducted a quantitative research study to determine the extent to which students lack under-

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-658-43763-3_5.

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_5

standing of generality of statements and examine how their understanding is related to reading and constructing mathematical arguments. A particular goal was to investigate the impact of reading *different types of arguments and statements* on students' performance in proof-related activities, with a specific focus on their understanding of generality. To truly uncover the cause-and-effect relationship, I designed and implemented an experimental study (e.g., Bryman, 2012). The experiment was further supplemented with open-ended questions to gain a deeper understanding of students' comprehension and conceptions on proof and generality.

To estimate the effect of different types of arguments on students' performance in proof-related activities, a between-subjects approach was taken, i.e., participants were randomly allocated to different groups, in which they received only one particular type of argument for all statements (experimental groups)—or no arguments at all (control group). Another approach would have been to provide different types of arguments to each participant. However, I decided against such a within-subjects approach regarding the type of argument, mainly because I wanted to avoid potential influences caused by participants directly comparing the types of arguments, in particular regarding proof evaluation—which would also be an interesting question, but not one I was investigating (see also discussion in Section 7.2.2).

A within-subjects approach via repeated measures was chosen to investigate the effect of different types of statements (familiarity and truth value, see Chapter 4) on students' understanding of generality of the statements and their performances in other proof-related activities. Furthermore, the inclusion of different statements was assumed to provide more reliable results (e.g., Bryman, 2012; Döring & Bortz, 2016).

A field experiment was chosen to provide high external validity (e.g., Döring & Bortz, 2016). The decision against the conduction of a laboratory experiment was mainly based on the difficulty of recruiting participants and a potentially resulting selection bias. Potential advantages of and suggestions for laboratory experiments are discussed in Section 7.2.2.

5.1.2 Overview of the Research Process

The research process consisted of two parts, the development and conduction of the pilot study, and the main study. The pilot study was conducted in October 2019 and aimed to ensure the feasibility of the chosen approach as well as to identify any modifications needed in the overall design and items. Figure 5.1 provides an overview of the timeline of the research process.

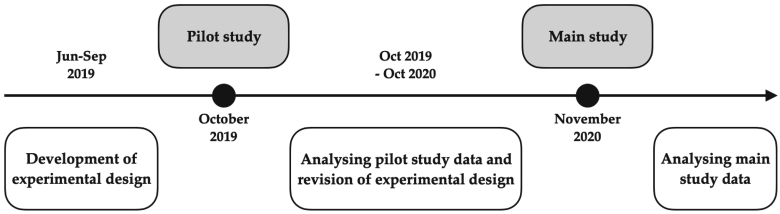


Figure 5.1 Timeline of research process

The first version of the experimental design was developed between June and September 2019. After the research questions had been specified, the experimental groups regarding the type of argument were defined. Due to their prominent role in mathematics education, the following types of arguments were chosen: empirical arguments, generic proofs, and ordinary proofs (see Section 2.4.2). The fourth group received no arguments at all, and therefore corresponds to a control group.

During the development of the experimental design, the questionnaire was given to several colleagues working in mathematics and mathematics education as well as to a small number of high school and university students. The mathematics (education) researchers were asked to evaluate the questions regarding correctness (both mathematical and language-wise), validity, completeness, structure, and other aspects they notice; the students mainly provided feedback regarding comprehensibility, completeness of answer options, and how reasonable the questionnaire is regarding its length and duration[1]. The experimental design and items were then revised accordingly. The pilot study was deliberately conducted during the second session of a first-semester mathematics lecture (with the title *Arithmetik und Algebra*) to avoid influences of university lectures on students' performance in proof-related activities, since the aim of this study is to identify students' understanding and performance in proof-related activities when they enter university. The participants consisted of students who want to become either primary (the majority) or lower secondary school teachers. They were randomly given a piece of paper with

[1] Due to the possibility of reduced test-taking motivation, "respondent fatigue" and "a greater tendency for questionnaires not to be answered in the first place" (Bryman, 2012, p. 235), overly long questionnaires should be avoided (see also Moosbrugger & Kelava, 2012). Reduced test-taking motivation is associated with lower performance (e.g., Wise & DeMars, 2005). It was therefore aimed at an average duration of about 25 minutes. This duration seemed feasible, also considering that the experiment was planned to be conducted during the second half of a lecture, which limited the maximum duration in which all participants should be able to finish the questionnaire to about 40-45 minutes.

a QR code on it with which they could access one of four different online questionnaires via their mobile phones, tablets, or laptops. The four questionnaires related to the four experimental groups. 382 students completed the questionnaire in the pilot study in October 2019.

The data of the pilot study was then analyzed and design issues regarding structure, selection of statements and arguments, as well as other necessary modification were identified. The experimental design was adapted accordingly and the questionnaire was again given to and discussed with colleagues and students. The main study was conducted in November 2020, again right at the beginning of the winter term. The data was collected in the lecture *Arithmetik und Algebra* as well as in a second lecture—*Lineare Algebra 1*—for first-semester mathematics majors and preservice higher secondary school teachers. The decision to collect data in these two lectures was made to provide a larger sample size as well as to investigate differences regarding the study program and students' (cognitive) resources.

5.2 Data Collection

In this section, the data collection of the main study is described in detail. First, background information regarding the setting in which the data collection took place is presented. Secondly, the characteristics of the sample and the participants' background are described. Lastly, the experimental design is outlined.

5.2.1 Setting

As mentioned in Section 5.1.2, the experiment was conducted in two mathematics lectures at Bielefeld University in North Rhine-Westphalia (NRW), Germany, which are aimed at two different groups of first-semester university students.

Due to the COVID-19 pandemic and respective lockdowns, university courses could not be given in person. This meant that, in contrast to the pilot study, the participants of both lectures were not present in a lecture hall, but attended the lectures via the online conference tool *Zoom*. Thus, students participated in the experiment via answering online questionnaires, which implied some lack of control over the participants' environment, one disadvantage of internet-based research. However, internet-based research seems to be as valid as more traditional methods, such as pencil-and-paper-questionnaires (e.g., Gosling, Vazire, Srivastava, & John, 2004). Moreover, due to students not being present in a lecture hall, another way of randomly assigning students to the four experimental groups had to be found.

The decision was made to use the breakout-room-function already implemented in Zoom, with which it is possible to create several sub-meetings and assign participants randomly. The four versions of the questionnaire (one for each experimental group) were implemented via the web-based survey-software *unipark*. The links were given to the participants via chat in the respective *Zoom*-breakout-rooms.

5.2.2 Participants

In total, 430 students completed the questionnaire (67.4% female, 31.2% male, and 1.4% chose not to answer). The average age of the participants was about 21 years ($SD = 4.2$) and about 96% were German native speakers. 116 of the participants received no arguments, 112 empirical arguments, 107 generic proofs, and 95 ordinary proofs. As the same number of participants had been allocated to each experimental group in the beginning (via the *Zoom*-breakout-room-function), the percentage of participants not completing the questionnaire (i.e., dropping out of the experiment) was the highest for ordinary proofs.

Figure 5.2 provides an overview of the distribution of participants with respect to the study program. The vast majority of the participants were preservice primary school teachers without mathematics as major[2] (290), followed by mathematics students[3] (70), preservice lower secondary school teachers (26), preservice higher secondary school teachers (25), and preservice primary school teachers with mathematics as major (19). This distribution corresponds roughly to the actual distribution of mathematics students at the faculty of mathematics at Bielefeld University (see Universität Bielefeld, 2020). About 94% of the participants stated to have attended the respective lectures for the first time. However, about 25% of the students were in the second semester or higher[4]. Therefore, it cannot be ruled out that these students have already gained experience with proof in other mathematics lectures. About 57% of the participants attended a transition course—a so-called *Vorkurs* (see footnote in Section 5.3.7)—prior to the lecture (58% of participants that attended the lecture *Arithmetik und Algebra* and 54% of participants that attended the lecture *Lineare Algebra 1*). Most participants got their university entrance degree in North

[2] At Bielefeld University, all preservice primary school teachers are trained in mathematics, but they can also choose it as their major. In contrast, preservice lower and higher secondary school teachers do have to choose their subjects—and all respective participants in this study had chosen mathematics.

[3] Either mathematics majors or minors.

[4] Students at Bielefeld University can begin their studies either in the winter (the more common case) or summer term.

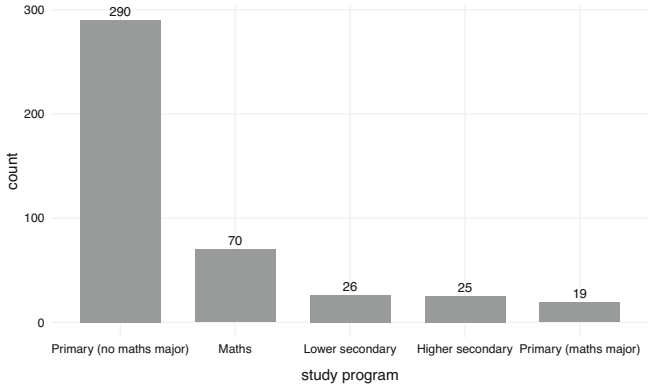


Figure 5.2 Distribution with respect to study program

Rhine-Westphalia (about 90 %), about 9% in another German state, and less than 1% in another country. The mean university entrance grade[5] of the participants ($M = 2.28$, $SD = 0.58$) corresponds approximately to the average of the university entrance grade in North Rhine-Westphalia (e.g., in 2021: $M = 2.35$; see Kultusministerkonferenz, 2022); the participants' mean final high school grade in mathematics was $M = 2.36$ with noticeably higher dispersion ($SD = 1.05$). About 37% of participants (about 25% of participants that attended the lecture *Arithmetik und Algebra* and 81% of participants that attended *Lineare Algebra 1*) specialized in mathematics during high school in a so-called *Leistungskurs* (honors course, see footnote in Section 5.3.7).

5.2.3 Structure of the Experiment

I designed an experiment, which mainly aimed at analyzing the influence of different types of arguments on students' understanding of the generality of mathematical statements, their conviction of the truth of the statements, as well as their comprehension of proof. Apart from the instructions, the experiment consisted of three main parts as shown in Figure 5.3. The instructions contained explanations on the overall goal and implications of the project—investigating students' knowledge at the transition from school to university to be able to better support future students.

[5] Grades in Germany are scaled from 1 to 6, where 1 is the best grade.

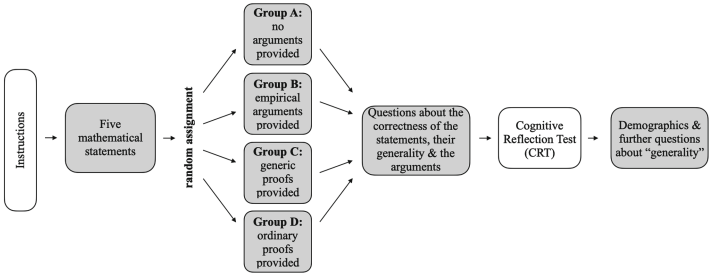


Figure 5.3 Overall experimental design

More specific research interests such as students' understanding of generality or students' proof skills were not communicated to avoid influencing the participants. In the first part of the experiment, all participants had to read five universal statements and respective arguments. The type of argument participants received depended on the experimental group they were randomly assigned to. Group A got no arguments at all, group B only got empirical arguments, group C got generic proofs, and group D got ordinary proofs. Then, all participants had to estimate the truth value of each statement and decide wether or not counterexamples might exist. In addition, participants in Group A—who received no arguments—were asked to justify the truth or falsity of each statement. The participants in the remaining groups were asked to evaluate the provided arguments regarding conviction and if they have comprehended the arguments. At the end of the first part, all participants were asked to evaluate the difficulty of the questions asked so far. In the second part, all participants had to complete a Cognitive Reflexion Test (see Section 5.3.6). In the third and last part, participants were asked to answer questions about their demographics as well as their understanding of the meaning of *mathematical generality* (in German *Allgemeingültigkeit*). The decision to put the demographic questions at the end was made because thinking about these questions can unconsciously influence the participants' answers to other questions. For instance, Steele and Ambady (2006) showed that "women who were subtly reminded of ... their gender identity ... expressed more stereotype consistent attitudes towards the academic domain of mathematics ... than participants in control conditions." (p. 428)

5.3 Instruments

In this section, I first describe and justify the selection of statements and arguments. Then, in Sections 5.3.2, 5.3.3, 5.3.4, 5.3.5, the respective items addressing the research questions are specified. In Section 5.3.6, the CRT (Cognitive Reflexion Test) used as a control instrument for individuals' cognitive resources, is introduced. Lastly, I give a summary of collected demographic information.

5.3.1 Selection of Statements and Arguments

The influence of the *type of statement* (familiarity and truth value) on students' understanding of the generality of statements, their conviction of the truth of the statement, their comprehension of the statement, and their construction of arguments were investigated in this study (see research questions in Chapter 4). Because of their prominent place in school curricula, the following two *familiar statements* were chosen: 1) the pythagorean theorem and 2) the sum of interior angles of a triangle. Both statements are explicitly listed in all NRW lower secondary curricula and it is even expected that the school students prove these statements or justify the overall idea of the proof (e.g., Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen, 2019, p. 30 and p. 34). Even though it is questionable, if the participants actually had to prove these theorems in school, it can be assumed that they had to apply them frequently in their mathematics classes.

Secondly, for the selection of *unfamiliar* mathematical statements, three main criteria were specified:

- The statement should not be explicitly mentioned in the school curricula in NRW.
- Only (little) basic content knowledge should be necessary to understand and prove the statement.
- It should be possible to support/prove the statement by the three types of arguments: empirical, generic, and ordinary.

In previous studies, statements from elementary number theory (arithmetic) have often been chosen, because they generally require comparatively few prior knowledge, are therefore mostly easy to understand, and they can quite easily be proven by generic arguments (e.g., Barkai et al., 2002; Healy & Hoyles, 2000; Kempen, 2018; Martin & Harel, 1989; Tabach et al., 2011). Therefore, two *unfamiliar statements* (both of them true) were selected: 1) the sum of two odd numbers is always even and 2) the product of two odd numbers is always odd. In particular the first statement

has been used in studies on proof and argumentation before (e.g., Healy & Hoyles, 2000; Kempen & Biehler, 2019).

To find suitable (unfamiliar) universal statements that are *false* and fulfill all three criteria turned out to be more difficult. In particular, it had to be possible to construct (non-general) arguments (generic and ordinary *proofs*) for these statements, where it is not too obvious why these arguments do *not* prove the truth of the statements *for all cases*. One such statement, which had been used in prior research (e.g., Barkai et al., 2002), was identified: The sum of three consecutive numbers is always divisible by 6. This statement proved to be suitable, because both a generic *proof* and an ordinary *proof* could be constructed rather easily on the basis that the statement is true if (and only if) the first number is odd (see Fig. 5.5 and 5.6 further below).

In summary, five statements were selected for the main study, two of them familiar, two of them unfamiliar and one of them false (and unfamiliar). The decision was made to phrase all statements using natural language, because firstly, students tend to have difficulties interpreting (more) formal, symbolic statements, in particular quantification (e.g., Dubinsky & Yiparaki, 2000; Piatek-Jimenez, piateksp-sjimenez2004; J. Selden & Selden 1995). And secondly, the form of all statements became more similar in that way and thus, more comparable. For instance, the pythagorean theorem is often presented symbolically as $a^2 + b^2 = c^2$, were a, b are the legs and c the hypothenuse of a right triangle. To avoid any influence of the statement being expressed as an equation, which students might simply associate with a *formula*, the pythagorean statement was also phrased using natural language. Further, to express and emphasize the generality of the statement, the terms *beliebig* (in English *arbitrary*) and *immer* (in English *always*) were used. The term *Behauptung* (in English *claim*) was used for all statements to express uncertainty about the truth value.

Regarding the order of the statements in the questionnaire, two possibilities were discussed: Using randomization or ordering the statements from (presumably) easiest to most difficult. Randomization would have had the advantage of considering potential influences of the order of statements. However, it was decided to order the five statements by difficulty instead, because studies have shown that this can lead to a lower percentage of participants dropping out of the study and higher performance overall (e.g., Anaya, Iriberri, Rey-Biel, & Zamarro, 2022; Kleinke, 1980). Placing the most difficult item right at the beginning of a questionnaire, which can occur

if items are randomly ordered, particularly increases the risk of high drop out rates (Anaya et al., 2022). For the assessment of difficulty of the statements, the required knowledge to understand the statement and the complexity of the proofs (number of steps, required concepts etc.) were considered. Further, several colleagues were asked to evaluate the difficulty of the statements and proofs. The *type of statement* (i.e., familiarity and truth value) was also taken into account. This process lead to the following order of statements (English translations, see Appendix A in the Electronic Supplementary Material for the original German items):

Claim 1: *The sum of two arbitrary odd numbers is always even.*
Claim 2: *In an arbitrary triangle, the sum of the interior angles is always equal to* 180°.
Claim 3: *The product of two arbitrary odd numbers is always odd.*
Claim 4: *The sum of three arbitrary consecutive natural numbers is always divisible by* 6.
Claim 5: *In an arbitrary right triangle, the sum of the areas of the squares of the legs is always equal to the area of the square of the hypothenuse.*

In the following, the selection and phrasing of the different types of arguments are described and justified.

Empirical Arguments

The empirical arguments used for the present study consisted of four examples for the false and the two unfamiliar (arithmetic) statements and three examples for both familiar (geometry) statements[6] (see Fig. 5.4 for the empirical arguments used to justify claim 1 and 2, respectively). The empirical arguments always started with the sentence *Begründung: Ich habe mir verschiedene Beispiele angeschaut und die Behauptung überprüft* (which roughly translates to *Justification: I looked at several examples and verified the claim*), followed by the respective examples. The examples were chosen in a way that they appear to cover various cases (e.g., right and equilateral triangles)—which some students seemingly consider when they evaluate or construct empirical arguments (e.g., Chazan, 1993)—and that participants could easily verify their correctness.

[6] Three examples instead of four were used for the geometry statements because of space limitations.

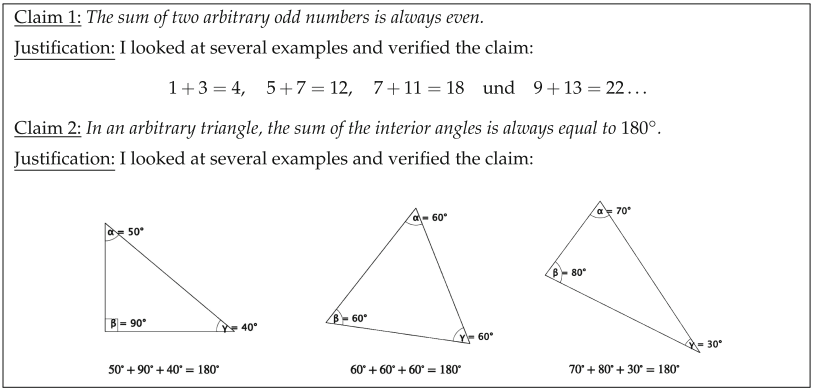


Figure 5.4 Example items for empirical arguments to justify claims 1 and 2 (translated)

Generic Proofs

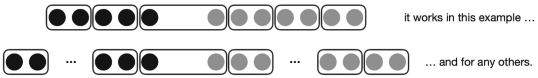
All generic proofs (and the incorrect one for statement 4) consisted of specific examples which reveal the respective underlying structure that can be generalized, and an explanation that the illustrated idea indeed works for any other example as well, as suggested by Kempen & Biehler (2019), for instance (see Fig. 5.5 for example items). The explanations were rather detailed to ensure that participants can follow the argument. For instance, in the generic proof for claim 1, it was explained that every odd number can be divided into pairs of twos such that exactly one is left. The decision to use more detailed explanations was made, because students seem to lack knowledge regarding basic concepts (Conradie & Frith, 2000; Moore, 1994; Reiss & Heinze, 2000), which was also observed in the pilot study of this project (e.g., the definition of even and odd numbers). The generic *proof* for the false claim 4 is based on the generic proof for the (true) statement *the sum of four arbitrary consecutive odd numbers is always divisible by* 8 given by Brunner (2014, p. 22). A similar argument proves claim 4, but only if the first number is odd. This fact was used to construct the generic *proof* as well as the ordinary one (see below).

Ordinary Proofs

Similar to the generic proofs, the ordinary proofs (and the incorrect one for statement 4) were rather detailed to enable participants to comprehend the arguments more easily. Further, no illustrative figures were included in the proofs for claims 3 and 5 (the familiar geometry statements), even though it might have facilitated understanding the arguments. This decision was made, because such figures always show specific

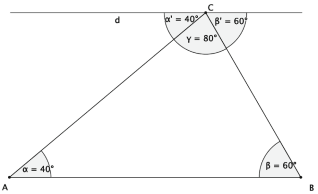
Claim 1: *The sum of two arbitrary odd numbers is always even.*

Justification: Every odd number can be grouped into pairs (of twos), such that exactly one is left (e.g., 5 consists of two pairs of twos and one). By adding two odd numbers–for instance, as in the picture below 5 and 7–one can group the two one's that are left, such that the sum now only consists of pairs (of twos) and is therefore even.



Claim 2: *In an arbitrary triangle, the sum of the interior angles is always equal to* $180°$.

Justification: We consider the pictured triangle ABC with angles $\alpha = 40°$, $\beta = 60°$ and $\gamma = 80°$. d is the line parallel to side AB that passes through the vertex C. Then $\alpha = \alpha' = 40°$ and $\beta = \beta' = 60°$, because α and α' as well as β and β' are each alternate interior angles with respect to the parallel line d. Therefore, $\alpha + \beta + \gamma = \alpha' + \beta' + \gamma = 40° + 60° + 80° = 180°$. This works for any other triangle as well.



Claim 4: *The sum of three arbitrary consecutive natural numbers is always divisible by 6.*

Justification: The first sum in the picture below is divisible by 6. For each summand, 2 are added to the respective summands in the next sum, for a total of 6, as illustrated in the picture. Therefore, the sum of three arbitrary consecutive natural numbers is divisible by 6.

$$\begin{array}{ccccccc} 1 & + & 2 & + & 3 & = & 6 \\ +2\downarrow & & +2\downarrow & & +2\downarrow & & +6\downarrow \\ 3 & + & 4 & + & 5 & = & 12 \end{array}$$

Figure 5.5 Example items for generic proofs to justify claims 1, 2, and 4 (translated)

examples and the distinction between an ordinary proof and a generic one would then not have been that clear. Figure 5.6 provides examples of the ordinary proofs for claims 1, 2, and 4. As was already mentioned above regarding the generic *proof* of claim 4, the argument used in the ordinary *proof* for claim 4 is also not general.

5.3.2 Conviction of the Truth of Statements

Two closed items were designed to investigate students' conviction of the truth of universal statements and the respective influence of different types of arguments. All participants were first asked to estimate the truth value of the statement (see Fig. 5.7).

Claim 1: *The sum of two arbitrary odd numbers is always even.*

Justification: Since even numbers are divisible by 2, they can be written as $2n$, where n is a whole number. Because the next larger number of an even number is odd, odd numbers can be written as $2n+1$. One of the odd numbers can therefore be written as $2k+1$ and the other as $2m+1$, where k and m are each whole numbers. If I add them together, I get:

$$(2k+1)+(2m+1) = 2k+2m+1+1 = 2k+2m+2 = 2(k+m+1),$$

where $k+m+1$ is a whole number. Thus, the sum of two odd numbers is divisible by 2 and therefore even.

Claim 2: *In an arbitrary triangle, the sum of the interior angles is always equal to* $180°$.

Justification: We consider an arbitrary triangle ABC with angles α, β and γ. We then draw the line d that is parallel to side AB and passes through the vertex C. We label the angle between side AC and d with α' and the angle between the side BC and d with β'. Then α and α' as well as β and β' are each alternate interior angles with respect to the parallel line d. Therefore, $\alpha = \alpha'$ and $\beta = \beta'$. Since α', β' and γ form a full rotation at line d, we get that $\alpha+\beta+\gamma = \alpha'+\beta'+\gamma = 180°$. Therefore, the sum of the interior angles of an arbitrary triangle is always $180°$.

Claim 4: *The sum of three arbitrary consecutive natural numbers is always divisible by* 6.

Justification: A natural number is either even or odd. The sum of three arbitrary consecutive natural numbers can therefore be written as $(2n-1)+2n+(2n+1)$, where n is a natural number. We get:

$$(2n-1)+2n+(2n+1) = 2n+2n+2n-1+1 = 3\cdot 2n = 6n,$$

thus, the sum of three arbitrary consecutive natural numbers is always divisible by 6.

Figure 5.6 Example items for ordinary proofs to justify claims 1, 2, and 4 (translated)

Is the claim correct?

a) Yes, I am absolutely sure the claim is correct.

b) I think it is correct, but I am not completely sure.

c) I think it is false, but I am not completely sure.

d) No, I am absolutely sure the claim is false.

e) I have no idea.

Figure 5.7 Closed item for the estimation of truth of the statements (translated)

Thereby, it was decided to give participants the opportunity to express *absolute* or *relative* conviction, as was proposed by Weber and Mejia-Ramos (2015). Further, the term *richtig* (in English *correct*) was used instead of *wahr* (in English *true*), to avoid any confusion about the meaning of *true* statements. The participants of the three experimental groups B, C, and D, who were provided with different types of arguments, were then asked if the provided argument has convinced them of the truth (correctness) of the statement (see Fig. 5.8). The decision was made, to not only ask if the participants find the argument convincing, as it might not have been clear to them, what is specifically meant by that and may have left more possibilities

Does the justification convince you of the correctness of the claim?

a) Yes, I am completely convinced by the justification.
b) I am only partially convinced by the justification.
c) No, I am not at all convinced by the justification.

Figure 5.8 Closed item for the conviction of arguments (translated)

for interpretation. Instead, it was explicitly referred to conviction regarding the truth of the statement. The response options were again chosen in a way such that absolute and relative conviction could be expressed, but the participants only had three options. This decision was made, because distinguishing between being partially convinced and being partially *not* convinced did not seem to be useful. Further, the option *I have no idea* was not provided, because this question does not assess knowledge and I wanted participants to take a stand. Participants who were not completely convinced by the provided argument were further asked to describe why the justification did not convince them of the correctness of the claim. The responses were coded based on aspects identified in the literature (see Section 5.4.3 for the coding scheme).

5.3.3 Comprehension of Arguments

To assess students' comprehension of generic and ordinary proofs, participants were first asked if they have understood the provided argument (see Fig. 5.9). Similar to the closed item regarding students' conviction, three response options were provided. Participants, who self-reportedly could not completely understand the provided argument were further asked to describe what they did not understand about the argument. These answers were coded regarding aspects of proof comprehension identified in the literature (see Section 5.4.4 for the respective coding scheme).

Can you understand the justification?

a) Yes, I can completely understand the justification.
b) I can only partially understand the justification.
c) No, I cannot understand the justification at all.

Figure 5.9 Closed item for the comprehension of arguments (translated)

Why do you think/are you convinced that the claim is correct/false?

Figure 5.10 Open item regarding students' proof schemes (translated)

5.3.4 Justification: Students' Proof Schemes

The participants in group A received no arguments justifying the truth of the statements. Instead, they were asked to explain why they think/are convinced that the claim is correct/false (see Fig 5.10). The main research question regarding students' proof construction was to investigate students' proof schemes. Participants were not asked to *prove* the statements, because it would have been unclear what they view as proof and it was not the goal of this study to investigate their respective understanding, but the aim was to analyze how students convince themselves of the truth and falsity of universal statements—and in particular how this relates to their understanding of the generality of statements (RQ4.4). The responses to this open-ended question were coded with respect to the main proof schemes (see Section 5.4.5).

5.3.5 Understanding the Generality of Statements

While students' understanding of the generality of mathematical statements has not been explicitly defined in previous studies (see also Section 3.1), related findings mostly referred to students or teachers who were seemingly convinced of the truth of the statement (and/or the correctness of a respective proof), but were at the same time not convinced that no counterexamples can exist (e.g., Chazan, 1993; Knuth, 2002), or to students' awareness that one counterexample disproves a universal statement (e.g., Buchbinder & Zaslavsky, 2019; Galbraith, 1981). Another approach has been taken by Healy and Hoyles (2000), who assessed students' understanding of the generality of a *proven* statement by asking them if the proof "automatically held for a given subset of cases" (p. 402) or if a new proof has to be constructed. In all these conceptualizations, students' understanding of generality of statements is *explicitly* related to students' understanding of the *generality of proof*.

My aim was to define and analyze students' understanding of the *generality of statements* independent of that of proof. The potential influence of and relation to proof was considered through the experimental design of my study. Therefore, for this study, the understanding of generality of mathematical statements was defined as shown in Table 5.1. The checkmarks stand for a

Table 5.1 Definition of a correct understanding of the generality of mathematical statements

	absolute conviction that no counter-examples exist	**relative conviction that no counter-examples exist**	**relative conviction that counter-examples exist**	**absolute conviction that counter-examples exist**
absolute conviction that statement is true	✓	inconsistent	inconsistent	inconsistent
relative conviction that statement is true	inconsistent	✓	inconsistent	inconsistent
relative conviction that statement is false	inconsistent	inconsistent	✓	inconsistent
absolute conviction that statement is false	inconsistent	inconsistent	inconsistent	✓

consistent estimation of truth of a statement and the existence of counterexamples and indicate a correct understanding of the generality of mathematical statements. Thus, in addition to the estimation of truth of the statement (see Fig. 5.7), the participants were asked to decide if a counterexample to the statement can exist. Thereby, the term *counterexample* was not used, as it could not be assumed that participants are familiar with the concept. Further, the usage of the term *counterexample* might even be suggestive and therefore potentially bias results. Instead, individual closed items for each statement were constructed, in which participants were indirectly asked about the existence of respective counterexamples. Figure 5.11 provides an example for such an item for claim 1. Participants again had the opportunity to express absolute or relative conviction regarding the (non-)existence of counterexamples. The variable

Are there two odd numbers whose sum is **odd**?

a) Yes, such numbers certainly exist.
b) I think such numbers could exist.
c) I don't think there are such numbers.
d) No, there are definitely no such numbers.
e) I have no idea.

Figure 5.11 Example for a closed item regarding the existence of counterexamples (translated)

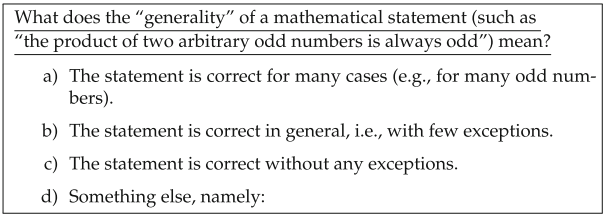


Figure 5.12 Item regarding the meaning of generality of mathematical statements (translated)

measuring students' understanding of generality (yes/no) was then defined based on Table 5.1. Observations in which participants responded "I have no idea" to both questions were treated as *missing values*, because no decision could be made regarding the understanding of generality. Observations in which participants responded "I have no idea" to only one of the questions were seen as inconsistent. These observations were therefore set to not having a correct understanding of generality.

The last item of the questionnaire (see Fig. 5.12) aimed at assessing students' (domain-specific) *knowledge* regarding the meaning of generality of mathematical statements. A variable regarding correct *knowledge* of the meaning of generality (yes/no) was defined based on correct (response option c)) and incorrect (response options a), b), and d), as the comments of those who had chosen d) were also incorrect) responses.

5.3.6 Cognitive Reflection Test

A so-called *Cognitive Reflection Test* (CRT) was used to control for individual cognitive resources. The CRT, first described by Frederick (2005), "is designed to measure a person's propensity to override an intuitive, but incorrect, response with a more analytical correct response" (Thomson & Oppenheimer, 2016, p. 99). It is assumed that the intuitive answers do not require any effort, while effortful thinking is needed to ultimately come to the correct solution (Frederick, 2005; Patel et al., 2019). Therefore, the CRT has been very influential in literature on so-called *dual-process theory*, which is based on the assumption that thinking processes can be divided into these very two types, an intuitive *System 1* and a more analytical, reflective *System 2* (Kahneman & Frederick 2002; Patel et al., 2019; Stanovich & West, 2000). A huge body of research has provided evidence that CRT performance is associated

with broad measures of rational thinking and thinking dispositions[7] (e.g., Frederick, 2005; Patel et al., 2019; Primi et al., 2016; Thomson & Oppenheimer, 2016). For instance, there seems to be a strong relation between CRT score and the so-called *need for cognition* (Frederick, 2005; Toplak, West, & Stanovich, toplak2014). A need for cognition (Cacioppo & Petty, 1982) is defined as a person's "tendency to enjoy effortful thinking" (Thomson & Oppenheimer, 2016, p. 99)—which should be quite useful in mathematical activities, in particular, in those related to proof. Further, CRT performance is also associated with mathematical abilities (Frosch & Simms, 2015), (insight) problem-solving skills, including cognitive restructuring, which "involves the ability to reinterpret a problem" (Patel et al., 2019, p. 2131), the preference for more explanatory detail (Fernbach, Sloman, Louis, & Shube, 2013), general reasoning skills (Primi et al., 2016), decision making (Frederick, 2005; Primi et al., 2016), and belief bias (e.g., Toplak et al., 2011), which is defined as "the tendency to be influenced by the believability of the conclusion when evaluating the validity of logical arguments" (Thomson & Oppenheimer, 2016, p. 99). While several studies have provided evidence (see, e.g., Shenhav, Rand, & Greene, 2012; Toplak et al., 2011) that "the CRT assesses something beyond general intelligence" (Patel et al., 2019, p. 2131), there is no consensus in the literature regarding the question if "individual differences in the disposition to overcome an initial intuition account for the predictive power of the CRT" (Baron, Scott, Fincher, & Emlen Metz, 2015, p. 279) and, in particular, if the CRT measures something completely unique from numeracy[8] (see, e.g., Pennycook, Cheyne, Koehler, & Fugelsang, 2016; Sinayev & Peters, 2015).

Despite the fact that more research is needed to fully understand what the CRT actually measures, it seems nevertheless useful for investigating students' performance in proof-related activities for two main reasons. Firstly, it can be assumed that activities related to proof and argumentation require numeracy, but also high levels of rational thinking, including problem-solving skills (e.g., Chinnappan et al., 2012; Moore, 1994; Stylianou et al., 2006; Weber, 2005), and thinking dispositions (such as need for cognition and belief bias) might play a role, in particular regarding the understanding of generality of mathematical statements. As the CRT performance correlates with these measures, it should cover several individual cognitive resources that have been discussed in Section 3.3. Secondly, the CRT requires not much (time) effort, thus, it does not significantly affect the test duration, while still providing potentially useful information regarding individuals' cognitive differences. The CRT

[7] A thinking disposition "is a tendency, propensity, or inclination to think in certain ways under certain circumstances" (Siegel, 1999, p. 209).

[8] Numeracy means the ability to do basic mathematical calculations.

1. A bat and a ball cost 1.10€ in total. The bat costs 1.00€ more than the ball. How much does the ball cost?
   [correct answer: 0.05€; intuitive answer: 0.10€]
2. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?
   [correct answer: 47 days; intuitive answer: 24 days]
3. How many cubic meters of dirt are there in a hole that is 3m deep, 3m wide and 3m long?
   [correct answer: 0; intuitive answer: 27]
4. Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are there in the class?
   [correct answer: 29 students; intuitive answer: 30 students]

Figure 5.13 CRT items used in the study. (Based on Frederick, 2005; Primi et al., 2016; Thomson & Oppenheimer, 2016)

was therefore used as a control instrument for individual differences in cognitive resources. I want to highlight that the purpose for using the CRT in the present study was *not* to identify specific cognitive resources underlying students' performance in proof-related activities. For the main study, German translations of the four items shown in Figure 5.13 were used in random order.

The first two items are based on Frederick's original version of the CRT. Because of potential prior exposure to the questions and the CRT's reliance on numeracy as well as its level of difficulty, it was decided to also include CRT items from alternative versions. The third item was taken and translated from the so-called *CRT-2* proposed by Thomson and Oppenheimer (2016), which showed to be significantly less reliant on numeracy than the original CRT. The fourth and last item is based on an item from *CRT-long* proposed by Primi et al. (2016)[9]. The CRT-long was developed as a more appropriate scale for a wider target group, because the original CRT is limited in that the item difficulty leads to floor effects in many populations, such as non-elite groups and adolescents (Frederick, 2005; Primi et al., 2016; Toplak et al., 2014). On average, 1.24 out of 3 items (about 41%) were correctly solved by participants in the CRT by Frederick and 2.95 out of 6 items (about 49%) in the CRT-long by Primi et al. About 33% and 12% of the participants scored zero in the CRT and CRT-long, respectively.

[9] The item has also been used in prior CRTs, such as the *CRT7* proposed by Toplak et al. (2014). It was seemingly provided by Shane Frederick (see Toplak et al., 2014, p. 151).

In summary, the CRT items were selected such that the risk of prior exposure, over-reliance on numeracy, and floor-effects were reduced. Furthermore, the items' context was considered regarding a meaningful translation into German.

The CRT score was calculated based on the number of correctly solved items, resulting in values between 0 (no CRT item correctly solved) and 4 (all CRT items correctly solved).

5.3.7 Demographics

In the last part of the questionnaire, the participants were asked about their demographics including university-entry grade, final high school mathematics grade, study degree program, and number of semesters. Furthermore, they were asked if they were attending the respective lecture for the first time and if they attended a so-called *Vorkurs*[10] prior to the beginning of the winter term. Lastly, they were asked if they specialized in mathematics during high school in a so-called *Leistungskurs*[11]. This information was used to (at least indirectly) consider prior mathematical knowledge of the participants. It can be assumed that, on average, students who specialized in mathematics gained more experience in mathematics at school than students attending a regular mathematics course. Attendance in a *Vorkurs* was considered because at least one of the statements used in the study had been discussed and proven in at least one of these courses. Thus, participants who attended the transition course might have been *more familiar* with the respective item.

5.4 Data Analysis

In this section, the methods and approaches used to analyze the data are described. First, the procedures used in this study are explained and justified. What follows is

[10] The two *Vorkurse* (transition courses)—one for preservice primary and lower secondary school teachers and one for preservice higher secondary school teachers and mathematics students—aim at closing the gap at the transition from school to university by recapping selected school topics and providing an introduction to proof-based mathematics. Students can optionally attend the *Vorkurs*, which takes place during the two weeks prior to the start of regular lectures.

[11] In the German school system, high school students usually have to choose two subjects as *Leistungskurse* (honors courses). These courses include more hours per week and cover more material than regular courses (so-called *Grundkurse*).

a more detailed summary of the analysis process and respective assumptions that were made.

5.4.1 Statistical Analysis

All statistical analysis was performed using the statistical software environment *R* (version 4.2.0, R Core Team, 2022). The relevant codes can be found at Damrau (2023).

To estimate the effects of the type of argument and statement (and other variables of interest) on students' understanding of generality—the primary outcome measure of this study—generalized linear mixed models (GLMM) were calculated using the *glmer* function of the R package *lme4* (version 1.1-31, Bates, Mächler, Bolker, & Walker, 2015) because understanding generality was defined as a binary variable (see Section 5.3.5). Classical linear models rely on normally distributed variables and could therefore not be used (e.g., Bolker, 2015). Similarly, to analyze students' performance in estimating the truth value of statements, students' conviction, and students' proof comprehension, respectively, cumulative link mixed models (CLMM) were fitted using the function *clmm* of the R package *ordinal* (version 2019.12-10, Christensen, 2019), because the respective dependent variables are ordinal. In both cases (GLMM and CLMM), logistic link functions were used. Further, using mixed regression models has the advantage that both random and fixed effects can be included (e.g., Bolker, 2015). The individuals participating in the study were considered as a random effect to control for individual differences in the repeated measures. All other independent variables were considered as fixed effects, such as the type of argument and statement.

For variable selection and regression model building, the following approach was taken (based on suggestions in Gelman & Hill, 2007; Harrell, 2015; G. Heinze, Wallisch, & Dunkler, 2018):

1. Theoretical background information was used to decide which independent variables (IVs) should be considered.
2. A directed acyclic graph (DAG) was drawn to illustrate the relationship between IVs.
3. Some IVs were eliminated based on the DAG (e.g., study program).
4. Backward elimination model selection was cautiously applied, where control variables with comparatively small and/or highly insignificant ($p > .5$) effects were dropped (as suggested by Harrell, 2015), while considering the theoretical background (e.g., the expected direction of the effect). Akaike and Bayesian

information criterions (AIC and BIC) were used for choosing the final model (see also Heinze et al., 2018).

The following control variables were included in the global models to consider individual (cognitive) difference of the participants:

- CRT score
- attendance of honors course in mathematics during high school (yes/no)
- attendance of transition course (yes/no)
- final mathematics grade in high school

The two continuous control variables (CRT score and final mathematics grade) were standardized by subtracting the mean and dividing by *two* standard deviation to be able to better compare the coefficient estimates to those of untransformed binary predictors, as suggested by Gelman (2008).

I decided to adjust the p-values in the final regression models whenever multiple testing was involved, which was for instance the case regarding hypotheses on the influence of the type of argument (e.g., comparisons of experimental groups against control) and statement (see following sections for specific comparisons and number of tests).

For choosing an appropriate correction method, the expected losses from Type I and II errors were considered. The consequences of Type I error would mainly consist of falsely considering variables (here, the type of argument, for instance) in practical implications and for future investigations. Type II errors, on the other hand, would result in falsely excluding relevant aspects. It can be expected that, in the context of the present study, both types of errors would not result in serious consequences. Further, no comparable research has previously been conducted, in particular regarding students' understanding of the generality of statements. Thus, to avoid prematurely excluding potential variables (Type II errors), it was decided to use Holm's correction (Holm, 1979) to control for family-wise error-rate (FWER) instead of the more conservative Bonferroni correction, which is a bit better in decreasing the probability of Type I errors, but also (highly) increases the probability of Type II errors (e.g., Aickin & Gensler, 1996). If the number of tests would have been higher, an even less conservative method such as Benjamini-Hochberg (BH) correction (Benjamini & Hochberg, 1995) that controls for false discovery rate (FDR) instead of FWER would also have been a reasonable choice. But due to the small number of tests, the results were to be expected not being much different than for Holm's correction. Holm's correction is a stepwise procedure in which the p-values/levels of significance are adjusted iteratively, from smallest to largest value.

The unadjusted p-values are first sorted in ascending order. The i-th BH adjusted p-value[12] $p.adj_i$ is then calculated by

$$\mathrm{p.adj_i} = \min\{\max_{1\leq \mathrm{j} \leq \mathrm{i}}(\mathrm{m} - \mathrm{j} + 1) \cdot \mathrm{p_j}, 1\},$$

where m is the number of tests and p_j is the j-th unadjusted p-value. The adjusted p-values are then compared to the unadjusted level of significance, which was set to the common value of $\alpha = .05$ in this thesis. The Holm's adjusted p-values were calculated with the R function *p.adjust*. In the regression summaries, unadjusted significance is reported using traditional stars and significance based on adjusted p-values is marked bold in the final models. Whenever I have corrected the level of significance, Holm's adjusted p-values are reported as $p.adj$.

5.4.2 Content Analysis

A quantitative content analysis (sometimes referred to as *structured qualitative content analysis*, see, e.g., Döring & Bortz, 2016; Mayring, 2022) was chosen to measure students' proof schemes, students' proof evaluation regarding conviction, and students' proof comprehension. The three respective open questions were analyzed following theory-based coding schemes. To refine the coding schemes, the following approach was taken (see Döring & Bortz, 2016; Krippendorff, 2004):

1. A set of a priori categories was identified based on the literature.
2. The categories were discussed and modified.
3. The set of categories was applied to a sample of the data, resulting in the deletion of categories, rephrasing of categories, differentiation of categories, and the addition of a few new categories.
4. The set of revised categories were further specified to maximize mutual exclusiveness as well as exhaustiveness.
5. The resulting coding scheme was pretested to ensure applicability.
6. After adequately refining and clarifying the categories, the coding scheme was settled upon as final.

The resulting coding schemes are described in Sections 5.4.3, 5.4.4, and 5.4.5, respectively.

[12] Alternatively, adjusted α's can be calculated to which the unadjusted p-values are then compared.

Over 20% of randomly chosen students' responses were coded by two colleagues working in mathematics education. Following suggestions in the literature (e.g., Krippendorff, 2004), the coders were chosen such that they are familiar with the respective content (e.g., with typical proof tasks and university students' attempts to solve them) and have strong backgrounds in mathematics. A detailed coding protocol (including decision trees) was provided (in German, see Appendix B in the Electronic Supplementary Material) and both coders were sufficiently trained. This lead to very good inter-coder reliabilities ($.88 < \kappa < .93$), based on Cohen's kappa (e.g., Davey, Gugiu, & Coryn, 2010; McHugh, 2012; O'Connor & Joffe, 2020).

In the following sections, more specific information on both the statistical as well as content analyses of the respective data is provided.

5.4.3 Conviction of the Truth of Statements

Students' conviction of the truth of statements was assessed by students' performance in two proof-related activities: The estimation of truth and proof evaluation regarding conviction. I first describe the analysis of students responses regarding estimation of truth and then outline how students' proof evaluation regarding conviction was analyzed.

Estimation of Truth

To be able to meaningfully compare responses regarding both true and false statements, students' responses were recoded with respect to a *correct* estimation of truth ("yes (absolutely sure)", "yes (relatively sure)", "no (relatively sure)", "no (absolutely sure)"). The answer "I have no idea" was coded as *undecided* and allocated between "yes (relatively sure)" and "no (relatively sure)", because these participants could not decide whether or not the statements were true. In that way, the responses were ordered conclusively and no information was lost in the regression analysis. As has been described above, cumulative linked mixed models were used to analyze students' (correct) estimation of truth (as an ordinal variable). The main goal was to estimate the effect of the type of argument and statement. In addition, the four control variables listed in Section 5.4.1 were considered for the global regression model. Holm's correction was used for analyzing the influence of the type of argument (three comparisons, each experimental group against the control group, which received no arguments) and the type of statement (two comparisons, true familiar and false statements against true unfamiliar statements).

Students' Conviction (Statistical Analysis)

The analysis of students' proof evaluation regarding conviction is based on the responses of groups B, C, and D, as participants in group A did not receive any justification. To analyze students' proof evaluation regarding conviction (measured via the closed item shown in Fig. 5.8), cumulative linked mixed models were again calculated. Thereby, the main goal was to analyze the effect of the type of argument and statement as well as students' proof comprehension (as an ordinal variable). In addition, the four control variables listed in Section 5.4.1 (CRT score, attendance of honors course in mathematics during high school, attendance of transition course, final mathematics grade in high school) were considered for the global regression model. Holm's correction was used for analyzing

- the influence of the type of argument (three comparisons in total: generic and ordinary proofs against empirical arguments, plus another comparison, generic vs ordinary proof),
- the influence of the type of statement (two comparisons, true familiar and false unfamiliar statements both against true unfamiliar statements),
- the influence of the level of proof comprehension (an ordinal variable) on students' conviction (two comparisons, completely understood and not at all understood both against partially understood).

Students' Conviction (Content Analysis)

The coding scheme used to analyze what aspects students' claim to not find convincing in different types of arguments is based on respective aspects identified in research on proof evaluation (see Section 3.2.3) as well as characteristics and acceptance criteria suggested by Stylianides (2007) and Hanna (1989) as discussed in Section 2.3.3. It was decided to include the aspect of *sample size/selection* as a possible category regarding the evaluation of empirical arguments, because of respective observations in prior research (e.g., Chazan, 1993). Table 5.2 gives an overview of the final coding scheme. The generated measured values were then analyzed using descriptive statistics. Responses of group B for statements 1 and 2, and responses of groups C and D for statements 1, 2, 3, and 5 were included in the analysis. The goal of including responses regarding empirical arguments was to investigate if students are aware of the limitations of empirical arguments and thus refer to a lack of generality when asked why the argument did not convince them. For this purpose, it did not seem necessary to analyze the responses regarding all statements, because responses which refer to a lack of generality of empirical arguments would become redundant. To account for potential differences regarding

Table 5.2 Coding scheme regarding argument evaluation of convincingness (examples translated by the author)

Code	Value	Description	Example
NA	No answer	The student did not submit an answer or claims to not know.	"I don't know"
0	Answer is deficient	The student claims to be convinced by the justification (after all) or the answer is otherwise unclear.	"the justification convinced me, but only because it is all explained by way of example."
1	Comprehension	The student (seemingly) did not understand the statement (e.g., terms used) or the justification.	"Probably because I don't understand it [the justification] 100%."; "Because the justification is given only for natural one-digit numbers and not for other cases, e.g., the integers or the rational numbers."
2	Correctness	The student states that the statement is false or the proof is incorrect.	"because I do not believe that it works like this for any triangle"
3	Sample size/selection	The student states that the number of cases is too small, other relevant cases have not been checked, or verifies the statement with additional examples.	"because there are not enough examples"
4	Generality	The student refers to a lack of generality of the argument, e.g., because it only consists of examples or is based on one example, or the student states that the justification is not a mathematical proof.	"Only examples are given, but it should be valid for all odd numbers."
5	Argument representation	The student refers to an inappropriate, inaccurate, or abstract form of representation/notation.	"Calculating with the rest-one sounds not very mathematical to me."
6	Familiarity	The student claims to not being familiar with the representation or type of argument.	"such proofs did not exist in school"

the familiarity with the statements, statements 1 (unfamiliar, number theory) and 2 (familiar, geometry) were included. Statement 4 was not considered at all because of the statement being false and therefore the proofs incorrect.

5.4.4 Comprehension of Arguments

In the analysis of students' proof comprehension, only responses of groups C and D (generic and ordinary proofs) were considered (as a reminder, group A did not receive any justification and group B only empirical arguments). Further, I excluded responses regarding the false statement 4 from the analyses since proof comprehension relates only to the understanding of correct proofs (see, e.g., Neuhaus-Eckhardt, 2022).

Statistical Analysis

Similar to the analysis of students' estimation of truth and conviction, cumulative linked mixed models were used to analyze students' (self-reported) proof comprehension. Thereby, the main goal was to estimate the effect of the type of argument (generic vs ordinary proof) and the familiarity with the statement. In addition, the four control variables listed in Section 5.4.1 were again considered for the global regression model. Holm's correction was neither used for the type of argument (generic vs ordinary proof) nor for the type of statement (unfamiliar vs familiar), because for each only one comparison (i.e., one test) was involved.

Content Analysis

The coding scheme used to analyze the open item on students' proof comprehension is mainly based on the local aspects introduced by Mejía Ramos et al. (2012) (see Section 3.2.4). It was decided to explicitly differentiate between students' lack of understanding the statement itself (as a prerequisite to understand the proof) and statements/terms/illustrations etc. only used in the proof. Based on prior research, it was assumed that students would generally not refer to holistic aspects and aspects beyond the particular proof when asked to identify what they did not understand (see Section 3.2.4). *Generality* was added as a possible category to the coding scheme, because of its importance for the present thesis. Furthermore, it was expected that students might (implicitly) refer to an insufficient understanding of the generality of generic proofs, as research on proof evaluation suggests that some students/teachers think generic proofs lack generality (see Section 3.2.3). The aspect of *generality* might contain holistic aspects, because students' not understanding why a generic proof is general could be related to an insufficient understanding of the main idea

of the proof, which due to Mejía Ramos et al.'s model corresponds to holistic understanding. Table 5.3 gives an overview of the final coding scheme. The generated measured values were then analyzed using descriptive statistics.

Table 5.3 Coding scheme regarding proof comprehension (examples translated by the author)

Code	Value	Description	Example
NA	No answer	The student did not submit an answer or claims to not know.	"I don't know"
0	Answer is deficient	The student claims to have understood the justification (after all) or the answer is otherwise unclear.	"I got it"
1	Unspecific	The student did not specify what they did not understand.	"I can not understand it [the justification], I lack the imagination"; "too complicated"
2	Meaning of the statement itself	The student (seemingly) did not understand the statement (e.g., terms used).	"What exactly is a product, simply the sum?"
3	Meaning of terms, statements, and illustrations used in the proof	The student did not know the meaning of terms and/or statements used in the proof or how to interpreted illustrations.	"I don't know what alternate angles are"; "why the 1 is outside the bracket"
4	Logical status and proof framework	The student (seemingly) did not understand the purpose of particular statements/concepts... and/or the connection between (parts of) the proof and the claim.	"why are alternate angles used to justify the statement"; "I did not understand the relation to the claim"
5	Generality of the proof	The student seemingly did not understand why the argument is general (e.g., because they did not understand the main idea of the proof)	"I did not understand why the justification is applicable to all products of odd numbers"

5.4.5 Justification: Students' Proof Schemes

The coding scheme used to analyze students' proof schemes is based on Harel and Sowder (1998), Bell (1976), Recio and Godino (2001), and Kempen (2019) (see Section 3.2.5). Table 5.4 gives an overview of the final coding scheme. The generated measured values were then analyzed using descriptive statistics. Further analysis was conducted to investigate the relation between students' proof schemes and their understanding of generality (see following section).

Table 5.4 Coding scheme regarding students' proof schemes (examples translated by the author)

Code	Value	Description	Example
NA	No answer	The student did not submit a justification or claim to not know.	"I don't know"
0	Answer is deficient	The student misunderstood the statement or the answer does not contain a justification or is otherwise unclear.	"I suspect it"; "because the number is divisible by 6 when multiplied three times"
1	Authority based argument	The student makes reference to school, a lecture, a teacher, etc.	"I have learned it like this"
2	Rule argument	The student states that the statement is a general rule/law/known theorem.	"This is an always valid law"
3	Empirical argument (no apparent awareness of generality)	The student provides examples or claims to have verified the statement by examples.	"by means of example"
4	Empirical argument (awareness of generality)	The student provides examples or claims to have verified the statement by examples, but seems to be aware of the necessity of a general argument, e.g., by making reference that no counterexample could be found.	"Since I am not aware of any numbers where this is not true"
5	Pseudo argument	The student re-states the statement or uses circular, redundant, irrelevant, not goal-oriented, or incorrect arguments.	"Because if the sum is not 180 degrees it is not a triangle"
6	Relevant aspects	The student mentions relevant aspects, which could form part of a proof, but does not attempt to construct a chain of arguments.	"Angles can be combined on a straight line"
7	Transformational argument (incomplete or with error)	The student gives a partially correct transformational argument, which is characterized by the consideration of particular examples (and operations on them) and the generality of the argument, but contains gaps and/or errors.	"Because it works for 1,3,5,7 and 9, so it will work for all numbers."
8	Transformational argument (complete)	The student gives a substantially correct transformational argument, which is characterised by the consideration of particular examples (and operations on them) and the generality of the argument.	no example found
9	Deductive argument (incomplete or with error)	The student gives a partially correct ordinary proof, which contains gaps and/or errors.	"The two ones that are too many combine to make an even number." (missing: sum of even numbers is even (and why))
10	Deductive argument (complete)	The student gives a substantially correct ordinary proof.	"Odd numbers can be written as $2k+1$. So if I add two odd numbers it looks like this: $(2n+1)+(2k+1) = 2*(n+k+1)$, which is always even.
11	Counterexample	(At least) one correct counterexample is given or it is claimed that counterexamples have been found.	"since $2+3+4$ is not divisible by 6"

5.4.6 Understanding the Generality of Statements

For a potentially easier interpretation and to be able to compare results to those of prior research, I first investigated participants who where absolutely convinced of the truth of the statement but not absolutely convinced that no counterexample to the statement exists (which relates to the first row shown in Tab. 5.1). Chi square tests and Cramer's V were used to analyze differences regarding the type of argument and statement.

For interpreting the effect size given by Cramer's V, the following rule of thumb introduced by Cohen (1988) was used (Table 5.5):

Table 5.5 Cohen's rule of thumb for interpreting Cramer's V

	Effect size:			
Degrees of freedom	Negligible	Small	Medium	Large
1	$< .10$	[.10, .30)	[.30, .50)	$\geq .50$
2	$< .07$	[.07, .21)	[.21, .35)	$\geq .35$
3	$< .06$	[.06, .17)	[.17, .29)	$\geq .29$
4	$< .05$	[.05, .15)	[.15, .25)	$\geq .25$
5	$< .05$	[.05, .13)	[.13, .22)	$\geq .22$

As has been described in Section 5.4.1, generalized linear mixed models were then calculated to analyze students' understanding of the generality of mathematical statements as defined in Table 5.1. In addition to the type of statement and argument, the analysis also aimed at estimating the effect of students' *knowledge* of the meaning of mathematical generality (correct knowledge or not, see Fig. 5.12 in Section 5.3.5). The four control variables listed in Section 5.4.1 were again considered for the global regression model. Holm's correction was used for analyzing the influence of the type of argument (three comparisons, the experimental groups against the control) and the type of statement (two comparisons, true familiar and false unfamiliar statements both against true unfamiliar statements).

Understanding Generality in Relation to Conviction and Proof Comprehension

Further regressions were calculated to analyze the relation between students' understanding of generality and their conviction and comprehension (see footnotes in Section 6.5 for p-value adjustment). Observations regarding the false statement were excluded in these analyses, because the effect was expected to be in the

opposite direction, in particular regarding conviction. Moreover, analyzing the relation to proof comprehension was based on the data of groups C and D (participants who received generic or ordinary proofs), because the influence of students' proof comprehension of empirical arguments did not seem to be meaningful.

Understanding Generality in Relation to Proof Schemes

To analyze the relation between students' understanding of generality and their proof schemes (group A), Chi-square test and Cramer's V were used instead of fitting generalized mixed effects models. This decision was made to increase statistical power and because setting a reference category was not possible in a meaningful way, even though using GLMM would have been preferable, because individuals could have been included as a random effect. To further increase statistical power, the categories introduced in Section 5.4.5 were summarized as follows, based on the main categories of Harel and Sowder (1998):

- **External conviction:** authority based, rule, pseudo
- **Empirical:** empirical (no apparent awareness of generality), empirical (awareness of generality)
- **Counterexamples:** counterexamples
- **Analytical:** deductive (complete and incomplete), transformative (complete and incomplete), relevant aspects
- **Unclear:** unclear

In theory, counterexamples could be seen as an empirical proof scheme—because they are indeed empirical—however, as counterexamples prove the falsity of a statement (in contrast to other empirical arguments, which usually are not able to proof the truth of a statement), it seemed oversimplifying to code them empirical. Because neither of the other main categories introduced by Harel & Sowder seemed to be appropriate either, counterexamples were treated as another main category. Further, pseudo arguments were allocated to external conviction proof schemes. Similar arguments, such as restating the statement or saying its contraposition, had been observed by Harel and Sowder (1998) as examples for authoritarian proof schemes. I decided to use the term *external conviction* for the main category, based on external proof schemes, proposed by Harel & Sowder, to distinguish pseudo arguments and references to a rule from explicit references to authorities. Responses that were coded as unclear were considered as an additional category to not lose potentially useful information.

Results

6

Before the results are presented in this chapter, further information on the sample characteristics and findings from preliminary analyses are provided. The subsequent sections are then structured according to the four sets of research questions. In section 6.6, *missing values* regarding the understanding of generality–those observations in which participants answered "I have no idea" regarding the estimation of truth and the existence of counterexamples, see section 5.3.5–are analyzed. Lastly, I summarize the main findings in section 6.7.

6.1 Preliminary Analysis

In this section, further information on the sample characteristics and findings from preliminary analyses are provided.

To be able to better interpret and compare the results with respect to the specific sample of my study, information regarding participants' CRT (Cognitive Reflexion Test) scores is first reported in the following. The most frequent CRT score was 0, which means that many of the participants (about one third) answered all four CRT questions incorrectly (see Fig. 6.1 for absolute frequencies of CRT scores). On average, participants solved 1.3 CRT items (about 33%) correctly ($SD = 1.2$) As discussed in section 5.3.6, similar floor effects in less elite populations have frequently been reported in the literature. The CRT score differed substantially by study program of the participants, as Figure 6.2 illustrates. The floor effect can only be observed regarding preservice primary school teachers without mathematics as

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-658-43763-3_6.

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_6

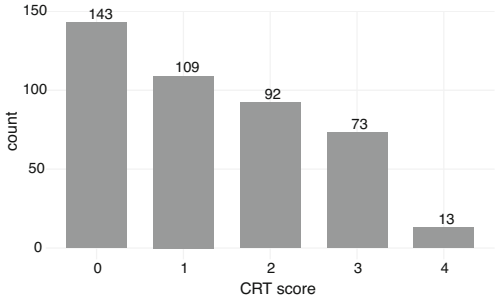


Figure 6.1 CRT scores of participants

major. About 45% of these students had a CRT score of 0. This group also had the lowest average CRT score ($M = 0.9$, $SD = 1.1$). In contrast, less than 5% of the mathematics students did not solve any of the CRT questions correctly. These students had the highest average CRT score ($M = 2.4$, $SD = 1$) compared to all other study degree programs. The preservice primary school teachers formed the largest group in the sample (see section 5.2.2), which results in the overall floor effect.

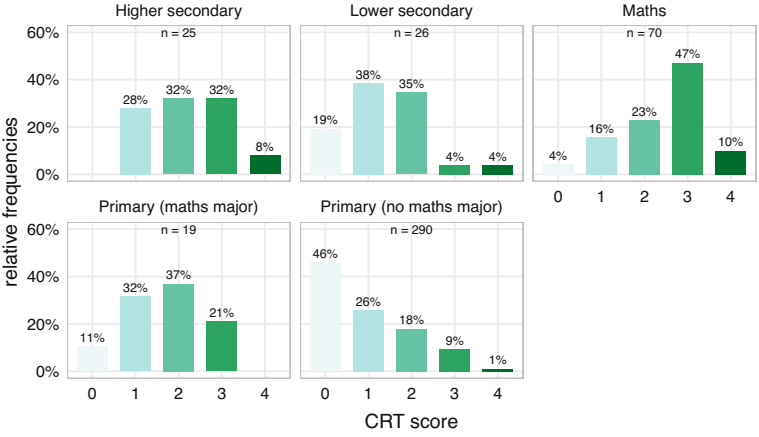


Figure 6.2 CRT scores of participants by study program

The participants were asked to rate the difficulty of questions regarding the proof-related activities to verify if respective ceiling or floor effects exist. Figure 6.3 shows the rating of difficulty of questions by the participants. Only few participants rated the questions very difficult or very easy, which indicates no ceiling or floor effects, at least regarding the *perceived* difficulty of the questions.

Even though the participants mainly received the same questions (in particular those in the experimental groups B, C, and D, who were provided with justifications for the statements), the perceived difficulty of the questions differed by the type of argument (see Fig. 6.4). The questions were perceived the most difficult by students who received ordinary proofs and the least difficult by students who received empirical arguments. Students who were not provided with any justification were asked to justify the truth/falsity of the statements. They perceived the questions to be more difficult than the participants who got empirical arguments and slightly more difficult than participants who received generic proofs, but less difficult than the participants who got ordinary proofs. Noteworthy, students who received generic proofs perceived the questions to be less difficult than students who received ordinary proofs.

On average, participants completed the questionnaire in about 24 minutes ($SD = 9$). Participants who received ordinary proofs spend the most time answering the questions, while participants who received no arguments needed the least amount of time, closely followed by participants who received empirical arguments (see Fig. 6.5). Given that participants who received no arguments were asked to justify the truth/falsity of the statements, it is surprising that they needed the least amount of time to finish the questionnaire.

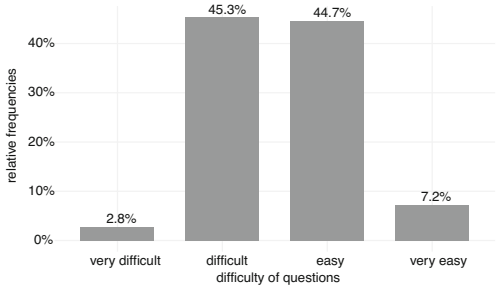


Figure 6.3 Rating of the difficulty of questions

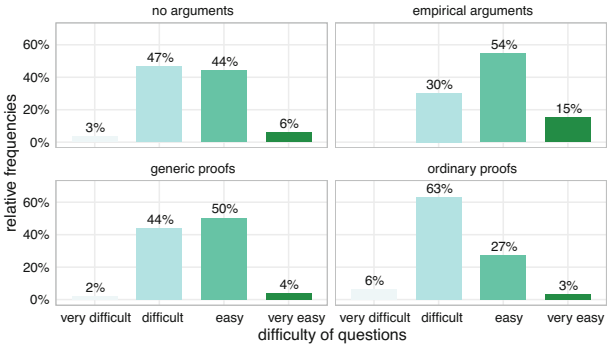


Figure 6.4 Rating of the difficulty of questions by group (i.e., type of argument)

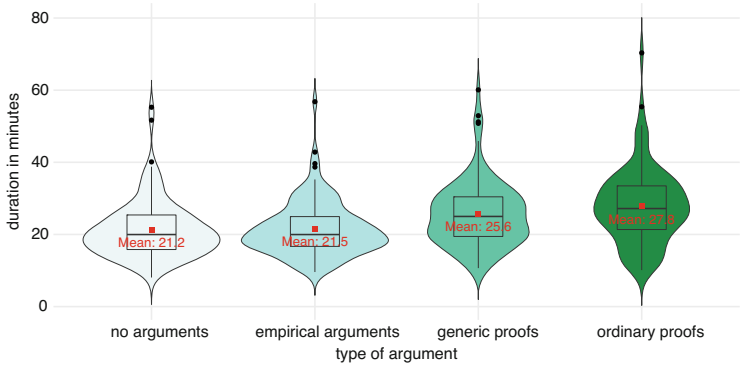


Figure 6.5 Minutes needed to answer the questionnaire by group

Mainly small to moderate intercorrelations between the considered independent variables were observed (see Tab. 6.1; because some of these variables are nominal, correlation coefficients were calculated with Cramer's V). The CRT score correlated comparatively strongly with the attendance of an honors mathematics course (LK), which should be taken into account when interpreting the results. However, both of these variables were used as controls and were other than that not of particular interest.

Overall, no severe multicollinearity was expected.

Table 6.1 Intercorrelation among variables

	meaning generality	LK	transition course	final grade maths
CRT score	0.20***	0.46***	0.10***	0.19***
meaning generality		0.18***	0.05*	0.16***
LK			0.02	0.26***
transition course				0.14***

Note: *p<.05, **p<.01, ***p<.001

6.2 Conviction of the Truth of Statements

This section reports on the results regarding the first set of research questions. It is divided into two main parts: Students' estimation of truth and students' proof evaluation regarding conviction.

6.2.1 Estimation of Truth

Between about 45 and 70% of the participants correctly estimated the truth of the statements and claimed to be absolutely sure, depending on the statement. If relative conviction of the truth of the statement is included, about 60% (regarding the false statement) to 95% (regarding one of the true unfamiliar statements, closely followed by one of the familiar statements) correctly estimated the truth of the statements. Figure 6.6 gives an overview of students' correct estimation of truth regarding the five statements (as a reminder, statements 1 and 3 were true and supposedly unfamiliar, statements 2 and 5 were familiar, and statement 4 was false). Noteworthy, almost one fourth of all participants claimed to not know if the pythagorean theorem (statement 5) is true, which seems unexpected at first. Moreover, comparatively many participants incorrectly estimated the truth values of the unfamiliar statement that the product of two odd numbers is odd (statement 3) and the false statement that the sum of three consecutive numbers is divisible by 6 (statement 4); about one third of the participants was relatively or absolutely sure that this statement is true.

Figure 6.7 shows participants' correct estimation of the truth of the statements, depending on the type of statement (familiarity and truth value) and argument (experimental group). Participants were generally more successful in estimating the truth value of familiar and unfamiliar statements–all of which are true–than of the

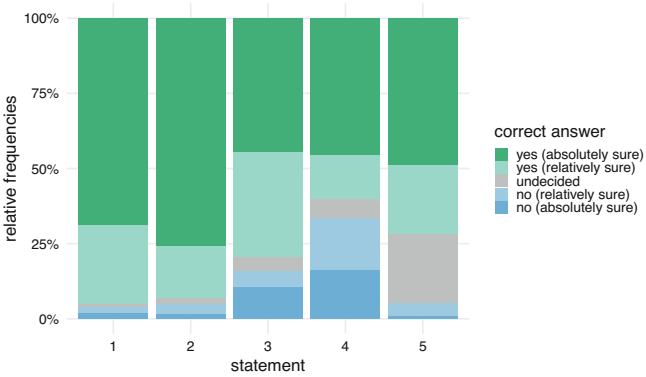


Figure 6.6 Correct estimation of truth by statement 1: sum of two odd numbers is even, 2: sum of interior angles in a triangle, 3: product of two odd numbers is odd, 4: sum of three consecutive numbers is divisible by 6, 5: pythagorean theorem

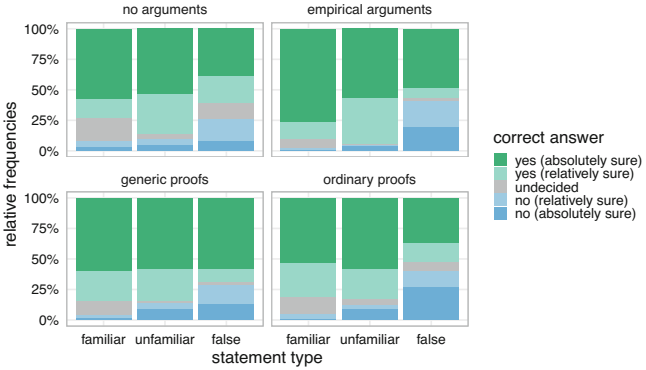


Figure 6.7 Correct estimation of truth by type of statement and argument

false (unfamiliar) statement. In particular, a substantial percentage among the participants who received empirical arguments or ordinary proofs seemed to be absolutely sure that the false statement is true. Furthermore, the graph suggests that students who received empirical arguments were overall the most successful in estimating the truth values of the true universal statements, followed by those who received generic proofs.

The effects of the type of argument and statement on students' estimation of truth was analyzed using mixed effects ordinal logistic regression (see also section 5.4.1). The results are summarized in Table 6.2. Model 2 was selected as the final model. As was expected, the familiarity and the truth value of the statement affected students' estimation of truth. Being familiar with the statement correlated positively with correctly estimating the truth value, even though not as strongly as was expected ($\beta = .20$, p.adj[1]$= .041$), while the falsity of a statement had a strong negative effect ($\beta = -.92$, p.adj $< .001$). This means that participants were more likely to correctly estimate the truth value of the familiar statements and less likely regarding the false statement, both compared to estimating the truth of the true unfamiliar statements.

Furthermore, students who received empirical arguments were more likely to correctly estimate the truth value than students who did not receive any justifications ($\beta = .44$, p.adj $= .004$). Reading generic proofs also had a positive effect on students' estimation of truth, but this effect did not reach significance regarding the adjusted p-value ($\beta = .27$, p.adj $= .088$). Ordinary proofs had no significant effect on students' estimation of truth ($\beta = -.09$, p.adj $= .537$).

Among the four control variables (CRT score, honors course, transition course, and final high school mathematics grade), only the CRT score and the participation in a mathematics honors course during high school predicted students' estimation of truth. The higher the CRT score, the more likely participants correctly estimated the truth value ($\beta = .63$, $p < .001$). Similarly, and with an even larger effect, if participants specialized in mathematics in an honors course during high school, the more likely they were successful in estimating the truth value ($\beta = .80$, $p < .001$). The effect of the final mathematics grade was comparatively smaller and did not quite reach significance ($\beta = -.20$, $p = .051$; note that in Germany, 1 is the best grade and 6 the worst, which explains the opposite sign of the estimate). The attendance of a transition course had an even smaller effect, which was highly insignificant ($\beta = .03$, $p = .798$) and therefore excluded from the models. Due to the comparatively small effect, the mathematics grade was excluded in Model 3. But because Model 3 did not have a smaller AIC value[2] than Model 2 and the influence of the mathematics grade is conclusive, Model 2 seemed to be the best choice overall.

[1] In this chapter, p.adj always refers to to the Holm's adjusted p-values and the level of significance is set to $\alpha = .05$ throughout this thesis, see section 5.4.1.

[2] AIC refers to Akaike Information Criterion. It is used to estimate the model quality relative to other models (see also section 5.4.1).

Table 6.2 CLMM comparison regarding students' estimation of truth

	Dependent variable:		
	Estimation of truth (correct)		
	Model 1	Model 2	Model 3
Threshold coefficients			
no (absolutely sure) | no (relatively sure)	−2.683***	−2.695***	−2.675***
	(0.150)	(0.143)	(0.142)
no (relatively sure) | undecided	−1.865***	−1.877***	−1.856***
	(0.135)	(0.127)	(0.126)
undecided | yes (relatively sure)	−1.244***	−1.255***	−1.234***
	(0.128)	(0.119)	(0.119)
yes (relatively sure) | yes (absolutely sure)	0.045	0.033	0.053
	(0.123)	(0.114)	(0.114)
Variables of interest			
empirical arguments	0.436**	**0.438****	0.462***
	(0.136)	(0.135)	(0.135)
generic proofs	0.269*	0.272*	0.277*
	(0.135)	(0.135)	(0.135)
ordinary proofs	−0.088	−0.085	−0.069
	(0.137)	(0.137)	(0.137)
familiar	0.200*	**0.200***	0.201*
	(0.098)	(0.098)	(0.098)
false (unfamiliar)	−0.924***	**−0.924*****	−0.924***
	(0.120)	(0.120)	(0.120)
Controls			
CRT score	0.628***	0.625***	0.681***
	(0.117)	(0.116)	(0.113)
LK	0.802***	0.804***	0.824***
	(0.119)	(0.119)	(0.119)
final grade maths	-0.197^{+}	-0.200^{+}	
	(0.104)	(0.103)	
transition course	0.026		
	(0.100)		
SD (Intercept id)	0.437	0.437	0.442
Observations	2150	2150	2150
AIC	4896.8	4894.8	4896.6
BIC	4976.2	4968.6	4964.7

Note: $^{+}$p<.1, *p<.05, **p<.01, ***p<.001. Holm's adjusted signif. marked bold

6.2.2 Proof Evaluation Regarding Conviction

About half of the participants who received generic or ordinary arguments claimed that the argument completely convinced them of the truth of the statement (see Fig. 6.8). Noticeably, in about 25% of the observations, participants claimed to be completely (!) convinced by the empirical arguments.

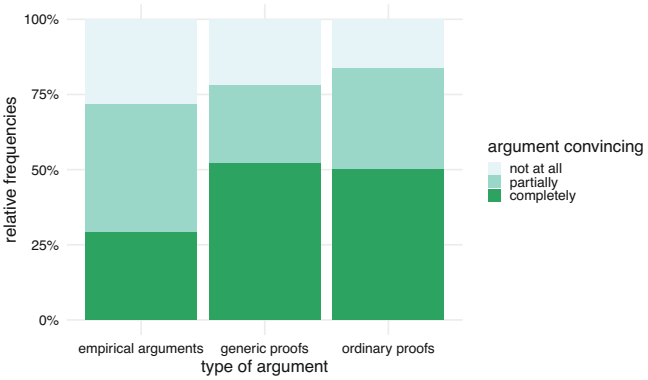


Figure 6.8 Conviction by type of argument

As would be expected, students claimed to be less convinced by the (incorrect) arguments regarding the false statement than regarding the true familiar and unfamiliar statements (see Fig. 6.9). However, over 60% of participants who received ordinary proofs were also at least partially convinced by the argument regarding the false statement. In contrast, less than 50 and 40% of participants who received empirical arguments and generic proofs, respectively, claimed to be at least partially convinced by the arguments regarding the false statement.

Figure 6.10 illustrates the relation between students' conviction and their self-reported level of comprehension of the argument (completely, partially, not at all). As would be expected, participants, who claimed to have (partially) understood the arguments were more often also (partially) convinced by them. Vice versa, participants who self-reportedly did not understand the arguments at all were in general also not at all convinced by the arguments.

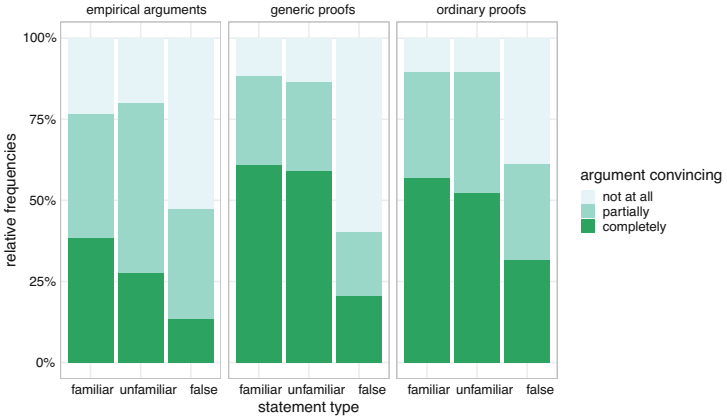


Figure 6.9 Conviction by type of argument and statement

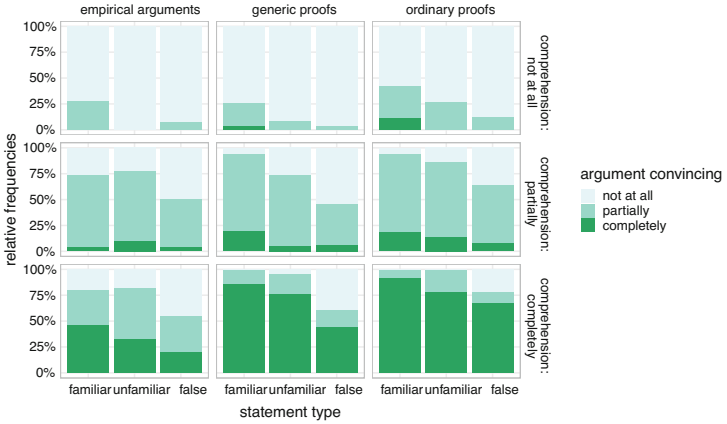


Figure 6.10 Conviction by type of argument and statement, and by comprehension of argument

To investigate the effect of the type of argument and statement as well as students' proof comprehension on their proof evaluation regarding conviction, mixed effects ordinal logistic regression was used (see Tab. 6.3). As was expected, students' rating of conviction was affected by the type of argument. Participants who received generic or ordinary proofs were more likely to claim being convinced by the argument than participants who received empirical arguments ($\beta = 1.70$, p.adj $< .001$ and $\beta = 2.20$, p.adj $< .001$, respectively). The falsity of the statement had a negative effect on students' conviction, which means that participants were less likely to be convinced by arguments regarding the false (unfamiliar) statement than regarding the true unfamiliar statements ($\beta = -1.31$, p.adj $< .001$), which was also expected. The familiarity of the statement had a comparatively smaller, but positive effect on students' conviction ($\beta = .54$, p.adj $< .001$).

Further, understanding the argument highly correlated with students' conviction with the largest effect overall. Participants who claimed to have completely understood the argument were more likely to be convinced by the argument and participants who claimed to have not understood the argument at all were less likely to be convinced by the argument, both compared to students who claimed to have partially understood the argument ($\beta = 2.73$, p.adj $< .001$ and $\beta = -2.84$, p.adj $< .001$, respectively), as would be expected.

Among the control variables, only the CRT score seemed to be predictive for students' proof evaluation. Unexpectedly, the CRT score correlated negatively with students' conviction, even though the effect was comparatively small ($\beta = -.60$, $p = .003$). It was suspected that this effect was caused by including observations regarding empirical arguments: Participants with a higher CRT score were less likely to be convinced *by empirical arguments*–but not regarding generic or ordinary proofs. To test this hypothesis, a second regression model was calculated, in which these observations were excluded (see Model 2 in Tab. 6.3). The effects reported above mainly remained[3], but the CRT score had no negative effect anymore. In fact, all controls showed only small and insignificant effects (unadjusted p-values between .475 and .850) and were therefore excluded in Model 3, in which all other effects remain significant, with .002 being the largest (adjusted) p-value.

[3] Note that the effect of ordinary proofs is now in relation to generic proofs, and therefore the effect is smaller as when compared to empirical proofs–but still significant after Holm's correction.

Table 6.3 CLMM comparison regarding students' conviction

	Dependent variable:		
	Proof evaluation (conviction)		
	Model 1	Model 2	Model 3
Threshold coefficients			
not at all | partially	3.413***	1.651***	1.632***
	(0.333)	(0.300)	(0.277)
partially | completely	6.300***	4.303***	4.283***
	(0.376)	(0.345)	(0.324)
Variables of interest			
generic proofs	**1.696*** **		
	(0.211)		
ordinary proofs	**2.203*** **	0.489**	**0.493** **
	(0.224)	(0.156)	(0.156)
familiar	**0.537*** **	0.704***	0.709***
	(0.129)	(0.167)	(0.167)
unfamiliar (false)	**−1.310*** **	−1.232***	−1.231***
	(0.165)	(0.203)	(0.202)
understood completely	**2.727*** **	2.801***	2.802***
	(0.169)	(0.194)	(0.192)
understood not at all	**−2.844*** **	−2.443***	−2.451***
	(0.273)	(0.273)	(0.272)
Controls			
CRT score	−0.589**	0.128	
	(0.201)	(0.179)	
LK	−0.329+	−0.103	
	(0.199)	(0.178)	
transition course	0.301+	0.117	
	(0.176)	(0.158)	
final grade maths	0.272	0.030	
	(0.187)	(0.156)	
SD (Intercept id)	1.101	0.326	0.333
Observations	1570	1010	1010
AIC	2429.2	1398.3	1391.3
BIC	2498.9	1457.3	1430.6

Note: $^{+}$p<.1, *p<.05, **p<.01, ***p<.001. Holm's adjusted signif. marked bold

Aspects That Influence Conviction

Participants who were not completely convinced by the arguments were asked to describe why the argument did not convince them. 353 out of 499 observations (about 71%), in which participants claimed to be not completely convinced by empirical arguments for statements 1 and 2 and generic or ordinary proofs for statements 1, 2, 3, and 5 (see section 5.4.3 for reasons why the analysis was restricted to these observations), contained explanations why (i.e., about 71% responded to the open-ended question). These were coded according to the coding scheme shown in Table 5.2 in section 5.4.3. In 9 of the responses, participants claimed to "don't know". These responses were coded as *NA* and not further considered in the analysis.

The vast majority of students who received generic or ordinary proofs referred to not having understood the argument, when they were asked why they were not (completely) convinced by the argument (see Fig. 6.11). This finding is in line with the regression analysis above (Tab. 6.3), where students' self-reported proof comprehension was highly predictive for their (lack of) conviction. The percentage was particularly high–about 81%–for students who were provided with ordinary proofs. In comparison, about 64% were not (completely) convinced by generic proofs because they did not understand them. More students referred to a lack of generality regarding generic proofs (about 12%) than ordinary proofs (about 4%). Students who received empirical arguments were mostly (about 78%) not convinced because of a lack of generality of the argument. Another 11% referred to the number

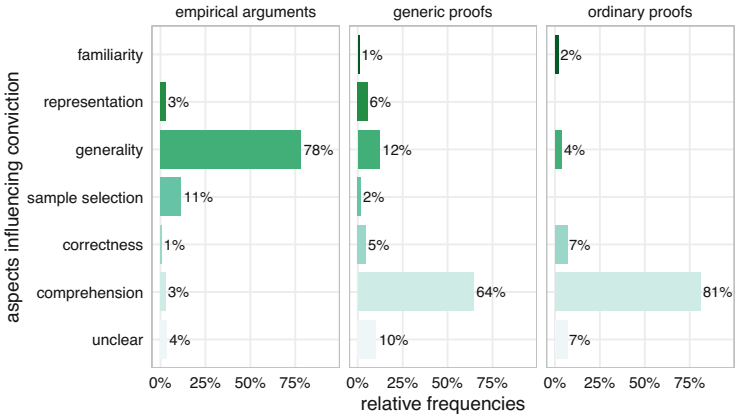


Figure 6.11 Reasons why participants did not find arguments convincing by type of argument (based on 344 observations)

or selection of examples, for instance, that too few examples were considered or that (seemingly) relevant cases were ignored. Some participants referred to the representation of the argument regarding empirical arguments and generic proofs (3 and 6% of the observations, respectively), but not regarding ordinary proofs. The familiarity with the argument was only mentioned on three occasions, once regarding a generic proof and twice regarding an ordinary proof.

6.3 Comprehension of Arguments

Overall, participants had higher levels of self-reported proof comprehension regarding the generic proofs compared to ordinary proofs. In particular, more students claimed to have completely understood the generic proofs than the ordinary proofs (see Fig. 6.12). However, the percentage of participants who claimed not having understood the provided arguments at all was comparatively small in both experimental groups (about 10%).

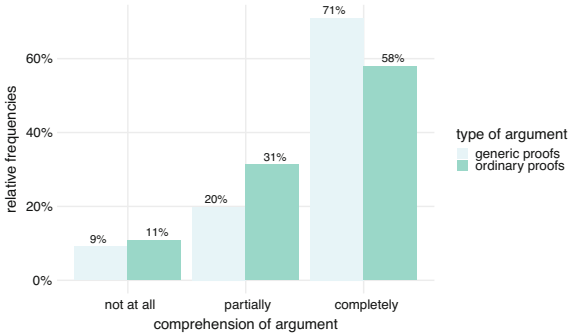


Figure 6.12 Comprehension of argument: generic vs ordinary proof

Unexpectedly, participants claimed to have comprehended the generic and ordinary proofs regarding the unfamiliar statements from elementary number theory more often than the proofs regarding the familiar geometry statements (see Fig. 6.13).

Figure 6.14 illustrates the relation between students' proof comprehension and the type of argument, familiarity with the statement, and their attendance in an honors (LK) or regular course (GK) in mathematics during high school. Students who

Figure 6.13 Comprehension of argument by familiarity with the statement: generic vs ordinary proof

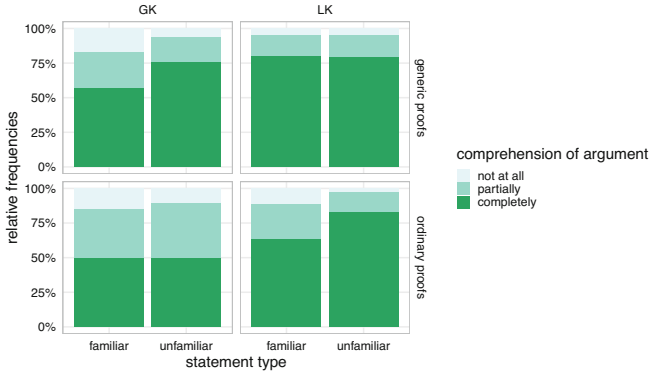


Figure 6.14 Comprehension of argument by familiarity with the statement: generic vs ordinary proof and *LK* (honors course) vs *GK* (regular course)

attended an honors course generally claimed to have comprehended the provided arguments more often than students who attended a regular course.

Mixed effects ordinal logistic regression was used to estimate the effect of the type of argument and familiarity with the statement on students' (self-reported) proof comprehension (see Tab. 6.4). Model 2 was selected as the final model (see explanation further below). Participants who received ordinary proofs were less likely to claim having understood the arguments than participants who received

generic proofs ($\beta = -.59, p < .001$). The familiarity with the statement also had a significant effect on students' self-reported proof comprehension. Participants were less likely to claim having (completely) understood the arguments regarding the familiar (geometry) statements ($\beta = -.55, p < .001$) than the true unfamiliar statements.

Table 6.4 CLMM comparison regarding students' self-reported proof comprehension

	Dependent variable:	
	Proof comprehension	
	Model 1	Model 2
Threshold coefficients		
not at all | partially	−2.865***	−2.817***
	(0.228)	(0.206)
partially | completely	−1.102***	−1.054***
	(0.191)	(0.166)
Variables of interest		
ordinary proofs	−0.593***	−0.594***
	(0.168)	(0.168)
familiar	−0.545***	−0.546***
	(0.154)	(0.154)
Controls		
CRT score	0.638**	0.648**
	(0.200)	(0.199)
LK	0.593**	0.581**
	(0.198)	(0.197)
final maths grade	-0.324^{+}	-0.312^{+}
	(0.170)	(0.168)
transition course	−0.088	
	(0.174)	
SD (Intercept id)	0.478	0.481
Observations	808	808
AIC	1326.5	1324.8
BIC	1368.8	1362.3

Note: $^{+}$p<.1, *p<.05, **p<.01, ***p<.001

Further, except for attending a transition course all other considered control variables (CRT score, honors vs regular mathematics course, final mathematics grade in high school) correlated positively[4] with students' self-reported comprehension of the arguments ($\beta = .65$, $p = .001$, $\beta = .58$, $p = .003$, and $\beta = -.31$, $p = .063$, respectively). Regarding these variables, the CRT score and the participation in an honors class had the largest effects. The mathematics grade had the smallest effect and did not reach significance.

Aspects of Students' Proof Comprehension

284 observations were made in which participants claimed not having completely understood the (correct) generic and ordinary proofs. 149 of these observations contained responses regarding the open-ended question on what participants did not understand. Because responses to the open-ended question on students' conviction also often contained information regarding aspects students (seemingly) did not understand, these responses were also considered (see coding protocol in Appendix B in the Electronic Supplementary Material). In total 208 responses (about 73% of observations in which students claimed not having completely understood the arguments) were coded according to the coding scheme shown in Table 5.3 in section 5.4.4.

The aspect students most often referred to when asked what they did not understand was local proof comprehension (32% for generic proofs and 54% for ordinary proof, see Fig. 6.15). These students claimed to have not understood particular statements, equations, or illustrations used in the proof. In particular, many participants seemed to not fully understanding the meaning of variables. Several participants referred to not understanding why two different variables are needed for the two odd numbers, if "$2n + 1$ stands for every odd number". Further, about one fourth of the participants who received generic proofs had difficulties understanding the statement itself, compared to about 14% of participants who received ordinary proofs. These students lacked knowledge of the meaning of basic terms, for instance, *divisible*, *product*, and *odd* and *even numbers*. Not understanding the proof's framework was slightly more often mentioned regarding generic proofs (about 8%) than ordinary proofs (about 5%). Further, reference was made to the generality of the proof in about 14% of the observations for generic proofs, while none of the participants who received ordinary proofs claimed to not have understood why the argument is general. The percentage of participants not being able to specify what they did

[4] Note again that in Germany, the best grade is 1 and the worst is 6, which explains the opposite sign of the estimate.

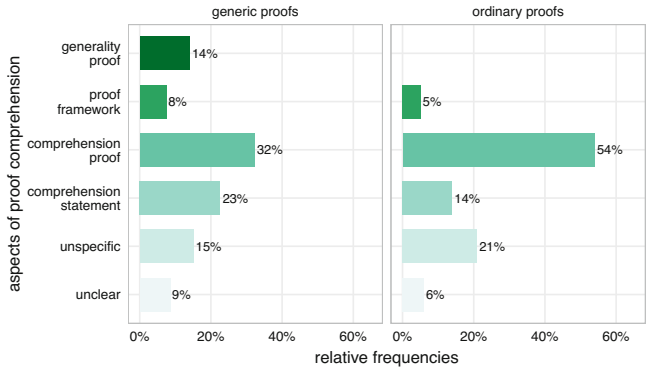


Figure 6.15 Aspects of participants' (self-reported) proof comprehension: generic vs ordinary proof (based on 208 observations)

not understand was higher for ordinary proofs (about 21%) than for generic proofs (about 15%).

Aspects of proof comprehension mentioned by participants also differed regarding the familiarity with the statement (see Fig. 6.16). Students more often made reference to not having understood the arguments regarding unfamiliar statements than regarding familiar ones (25 and 10%, respectively). Further, the proof frame-

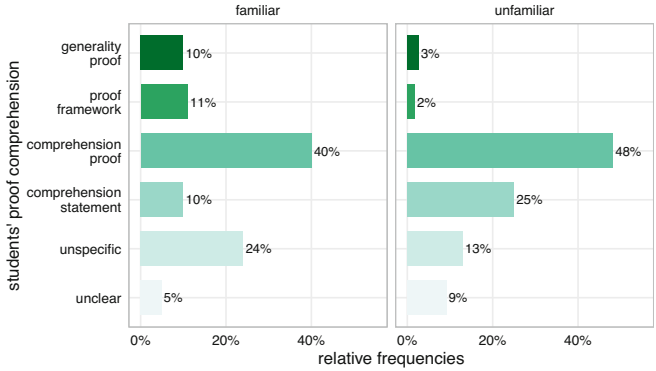


Figure 6.16 Aspects of participants' (self-reported) proof comprehension by type of statement (based on 208 observations)

work and the generality of the proof was more often (self-reportedly) not understood by the participants regarding the familiar (geometry) statements (11 and 10%, respectively) than the unfamiliar (arithmetic) statements (2 and 3%, respectively). The percentage of participants not being able to specify what they did not understand was almost twice as high for familiar statements (about 24%) than for unfamiliar ones (about 13%).

6.4 Justification: Students' Proof Schemes

Most participants responded when asked to justify why they think the statements are true/false (about 80% of all 580 potential observations–i.e., potential responses of the 116 participants in the control group A for each of the 5 statements). Overall, 467 observations were coded according to the proof schemes shown in Table 5.4. 5 of these observations were excluded from the analysis reported in this section because they included references to the geometry statements not being true on the sphere (which is of course correct, but made these responses difficult to interpret regarding the analysis of students' proof schemes). Table 6.5 provides an overview of the number of observations in each category.

As was expected, most students used empirical proof schemes (109 in total). Fewer students had a deductive proof schemes (39 in total). Transformative proof schemes were only occasionally observed (5). Many participants also showed external proofs schemes, such as referring to an authority (49) or claiming that the statement is a general rule (65). Noteworthy, 59 observations (about 13%) were coded as unclear. Most of these responses were coded as unclear because participants either seemed to have not understood the statement and/or falsely thought the statement was incorrect or were not able to give any argument, for instance, they just stated "I suspect it".

Table 6.5 Students' proof schemes (462 observations)

unclear	authority	rule	empirical	empirical (generality)	pseudo	relevant
59	49	65	94	15	78	12
transf. (incomplete)		transf. (complete)		ded. (incomplete)	ded. (complete)	counterex.
5		0		30	9	46

Students' proof schemes differed highly by the type of statement (see Fig. 6.17). Regarding familiar statements, participants most often referred to the statement being a general rule or a known theorem (about 37%), in particular regarding the pythagorean theorem. Further, one quarter of participants used pseudo arguments to justify the familiar statements. For instance, one student wrote that "if the sum is not 180 degrees it is not a triangle" (see also Tab. 5.4 in section 5.4.5). Another quarter of participants gave authoritarian arguments. These students often stated to have learned the statement in school or in a lecture. In contrast, the majority of justifications for the unfamiliar statements were coded as empirical proof schemes (about 45% in total). Only few students referred to the statement being a general rule (1%) or to any type of authority (4%). Complete or incomplete deductive arguments were given in 3 and 12% of the observations, respectively. Another 4% contained relevant aspects, but no chain of arguments, and even fewer participants attempted to give a transformative argument (2%). 16% of the observations regarding the unfamiliar statements contained pseudo arguments which often consisted of re-stating the claim. About half of the participants correctly provided one or more counterexamples to refute the false statement. Only few participants gave complete or incomplete deductive arguments to disprove the false statement (2% each). Further, 18% gave empirical arguments to (falsely) justify the truth of the statement.

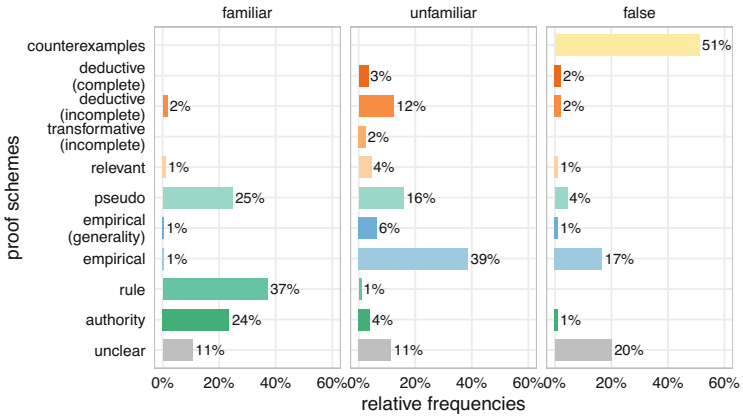


Figure 6.17 Students' proof schemes by type of statement (based on 462 observations) empirical proof schemes in shades of blue, external proof schemes in shades of green, analytical proof schemes in shades of orange, counterexamples in yellow

Figure 6.18 illustrates the relation between students' proof schemes, the type of statement, and students' (correct) estimation of truth of the statement. It should be noted that the number of observations for each of the proof schemes differed substantially (see Fig. 6.17) and the frequencies reported are based on these observations. Except for one observation regarding the false statement, participants with a (complete or incomplete) deductive proof scheme correctly estimated the truth of the statements mostly with absolute conviction. The participant who gave a complete deductive proof to refute the false statement stated that "the statement is correct, provided that one accepts fractions as a solution." The student then gives a correct proof why the statement is generally false, if fractions are not accepted as a solution. But regarding the closed item on estimating the truth value, the student chose the answer "Yes, I am absolutely sure the claim is correct." Further, most participants referring to a rule or authority correctly estimated the truth of the statements with absolute conviction. Regarding the true statements, the vast majority of participants giving empirical arguments was only relatively convinced of the truth of the statement. Similarly, most students who gave examples to justify the truth of the false statement was relatively (and not absolutely) convinced of the truth (which is of course not the correct answer in this case). About 25% of participants who provided correct counterexamples for the false statements were *not* absolutely

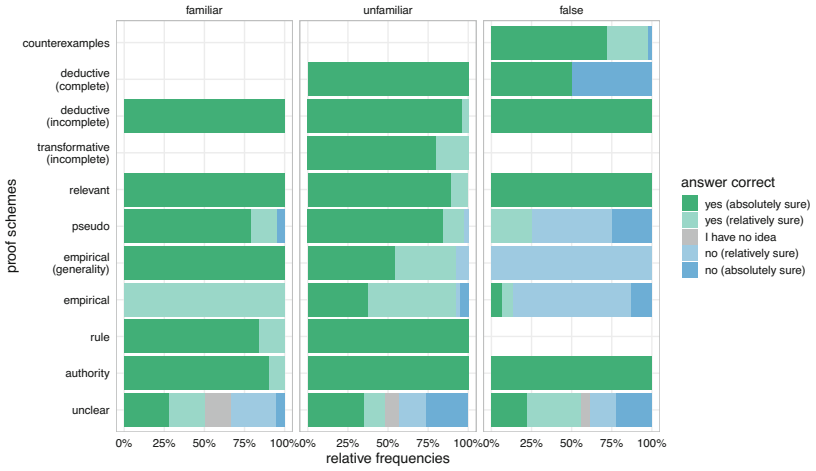


Figure 6.18 Students' proof schemes by type of statement and estimation of truth (based on 462 observations)

convinced of the falsity of the statement, but only relatively–even though they had in fact disproven the statement.

6.5 Understanding the Generality of Statements

To be able to better compare the results on students' understanding of generality with those of previous research, I first report results regarding those students who were absolutely convinced of the truth of a statement, but not absolutely sure that there cannot be counterexamples. Depending on the type of statement, about 4 to 35% of the observations in which participants correctly estimated the truth of the statement consisted of inconsistent estimations of the existence of counterexamples (see Tab. 6.6). Participants more often showed a correct understanding of generality regarding the false statement than the true statements. Moreover, the percentage of observations regarding a correct understanding of generality of statements was higher for the familiar statements than for the (true) unfamiliar statements. Based on Chi squared test, these differences were highly significant with medium effect size ($\chi^2(2) = 73.4$, $p < .001$, Cramer's V $= .25$).

The percentage of observations in which participants correctly estimated the truth of the statement (with absolute conviction) but inconsistently estimated the existence of counterexamples also differed by the type of argument (see Tab. 6.7). However, these differences were comparatively small. Moreover, in contrast to the type of statement, the differences regarding the type of argument were not significant ($\chi^2(3) = 3.5$, $p = .32$, Cramer's V $= .05$).

In the remainder of this section, results regarding students' understanding of the generality of statements as defined in Table 5.1 are reported. Overall, in about 64% of all observations, participants showed a correct understanding of generality (see Fig. 6.19). In about 6% of the observations, participants claimed to not know the answer to the two respective questions. These observations were therefore treated as *missing values* in the regression analysis (see section 5.3.5).

Table 6.6 Number/Percentages of observations in which the truth of the statement was correctly estimated (with absolute conviction) and the existence of counterexamples as well (yes) or not (no) by type of statement

	familiar	unfamiliar (true)	unfamiliar (false)	Sum
no	122 (22.7%)	170 (34.9%)	8 (4.1%)	300 (24.6%)
yes	415 (77.3%)	317 (65.1%)	188 (95.9%)	920 (75.4%)
Sum	537	487	196	1220

Table 6.7 Number/Percentages of observations in which the truth of the statement was correctly estimated (with absolute conviction) and the existence of counterexamples as well (yes) or not (no) by type of argument

	no arg.	emp. arg.	generic proofs	ordinary proofs	Sum
no	66 (21.8%)	97 (27.5%)	81 (25.6%)	56 (22.7%)	300 (24.6%)
yes	237 (78.2%)	256 (72.5%)	236 (74.4%)	191 (77.3%)	920 (75.4%)
Sum	303	353	317	247	1220

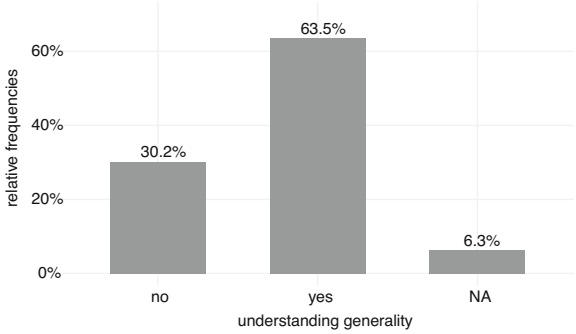


Figure 6.19 Participants' understanding of the generality of statements

The understanding of generality differed (significantly) by study program. As was expected, the higher the level of mathematics in the study program, the more participants seemed to have a correct understanding of generality (see Fig. 6.20). As most of the participants were in their first semester, influence of the study program itself on students' proof skills (including their understanding the generality of statements) is highly unlikely. Therefore, the study program was not used as a predictor in the regression models (see further below). But other control variables that were considered may (at least partially) explain differences regarding the study program. For instance, as was shown in Fig. 6.2, participants in study programs with higher levels of mathematics had higher CRT scores. Therefore not surprisingly, students with a higher CRT score also showed more often a correct understanding of generality (see Fig. 6.21). Another variable that might explain differences regarding the study program is the attendance in a mathematics honors course (LK). As was reported in section 5.2.2, students in study programs with higher levels of mathematics also participated in an honors course more often. Expectedly, the percentage of participants with a correct understanding of generality was higher for students who

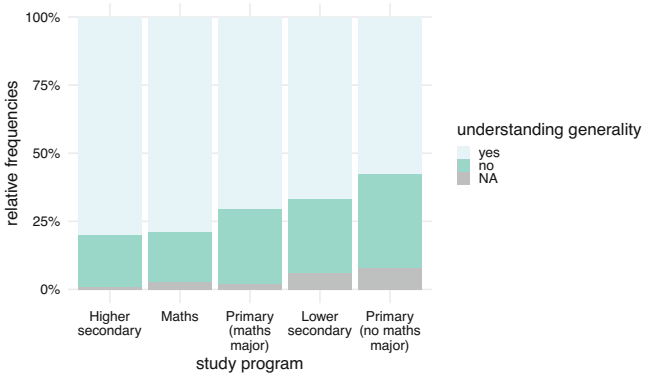


Figure 6.20 Participants' understanding of generality by study program

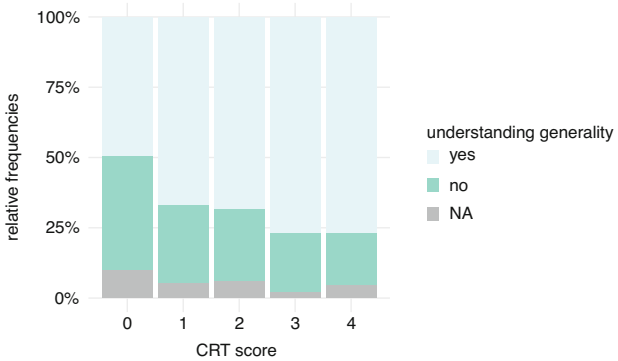


Figure 6.21 Participants' understanding of generality by CRT score

participated in an honors course than for students who had a regular mathematics course in high school (see Fig. 6.22).

Figures 6.23 and 6.24 illustrate the relation between students' understanding of generality and the type of argument and the type of statement, respectively. Similar to the results reported above regarding the more restricted assessment of students understanding of generality (see Tab. 6.6), differences regarding the type of statement can be observed. The percentage of students who showed a correct understanding of generality was the highest regarding the false (unfamiliar) statement. But differences regarding true familiar and unfamiliar statements are now less obvious. One reason for this is the comparatively high percentage of missing values regarding

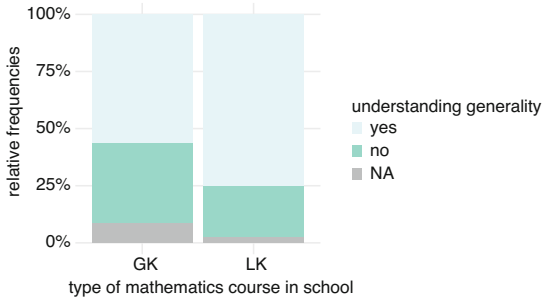


Figure 6.22 Understanding generality by type of mathematics course in high school *GK* (regular course) vs *LK* (honors course)

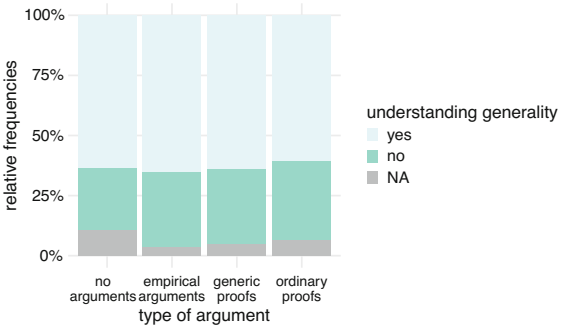


Figure 6.23 Understanding of generality by type of argument

the familiar statements. Also in line with the results regarding the more restricted assessment of students understanding of generality (see Tab. 6.7), the differences regarding the type of argument are comparatively small. However, the percentage of observations in which participants showed a correct understanding of generality was the lowest regarding ordinary proofs. Noteworthy is again the comparatively high percentage of *missing values* for participants who received no arguments.

At the end of the questionnaire, participants were asked about the meaning of generality (see Fig. 5.12 for the respective item and Fig. 6.25 for the result). The majority chose the correct answer (option 3). However, about a quarter responded incorrectly. 4 participants chose option 4. These participants then gave either incorrect meanings of generality (e.g., "it [the statement] is valid until there is a case where this statement is not true.") or claimed to not know.

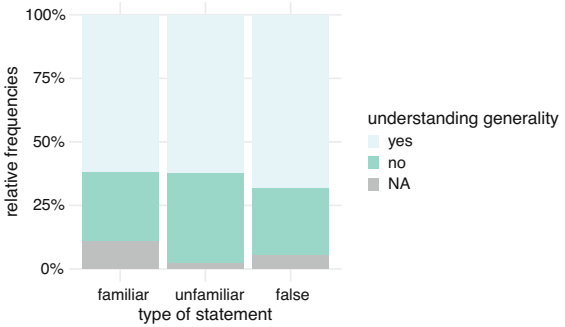


Figure 6.24 Understanding of generality by type of statement

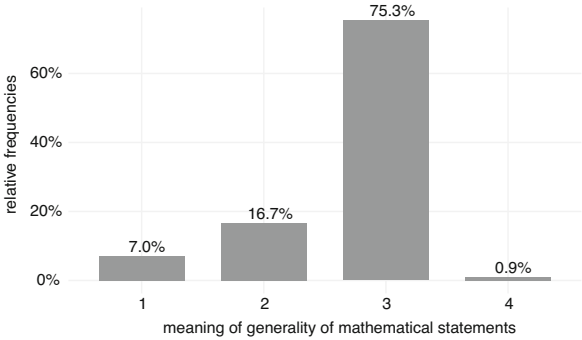


Figure 6.25 Students' knowledge of the meaning of mathematical generality 1: "The statement is correct for many cases (e.g., for many odd numbers)", 2: "The statement is correct in general, i.e., with few exceptions", 3: "The statement is correct without any exceptions", 4: "Something else, namely:..."

Figure 6.26 shows the relation between students' *actual* understanding of generality–as defined in this study via consistent responses regarding the estimation of truth and the existence of counterexamples–and their *knowledge* of the meaning of mathematical generality. The percentage of participants with an incorrect understanding of mathematical generality was higher for students who also had an incorrect *knowledge* of the meaning of mathematical generality (about 41 vs 27%). However, about 27% of participants with a correct knowledge of generality still responded inconsistently regarding their conviction of the truth of statements and the existence of counterexamples.

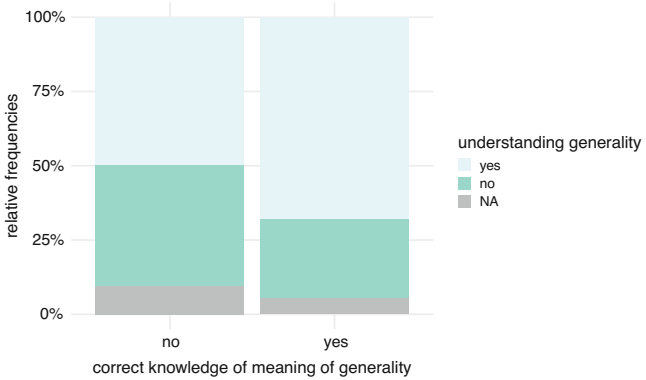


Figure 6.26 Understanding generality by knowledge of the meaning of mathematical generality

To estimate the effects of the variables of interest on students' understanding of generality, generalized linear mixed models were used (see also section 5.4.1). As can be seen in Table 6.8, the three models that were fitted do not differ much regarding AIC. Models 2 and 3 seem to be better than Model 1 regarding both AIC and BIC. Given the small difference in AIC and a better BIC value, the smaller Model 3 seemed to be the best choice overall. The GLMM results confirm the observed effects reported above. Participants who received ordinary proofs were less likely to have a correct understanding of generality than participants that received no arguments, even though this effect did not quite reach significance after Holm's adjustment ($\beta = -.41$, p.adj $= .075$). A similar, but smaller effect can be observed regarding participants who received generic proofs and empirical arguments ($\beta = -.32$, p.adj $= .148$ and $\beta = -.27$, p.adj $= .148$), also not reaching significance.

Being familiar with the statement as well as the truth value seemed to have influenced students' understanding of generality. Participants were more likely to have a correct understanding of the generality of familiar statements ($\beta = .29$, p.adj $= .011$) and the false (unfamiliar) statement ($\beta = .46$, p.adj $= .003$) compared to true unfamiliar statements. Participants who correctly answered the closed item on the meaning of mathematical generality, were also more likely to show a correct understanding of generality as defined in this study ($\beta = .68$, $p < .001$). This effect was overall the largest.

Further, among the considered control variables, only the CRT score and the participation in an honors course were predictive for students' understanding of

Table 6.8 GLMM comparison regarding students' understanding of generality

	Dependent variable:		
	Understanding generality		
	Model 1	Model 2	Model 3
(Intercept)	0.376*	0.358*	0.257
	(0.181)	(0.179)	(0.170)
Variables of interest			
empirical arguments	−0.257	−0.246	−0.267
	(0.176)	(0.176)	(0.176)
generic proofs	−0.297^{+}	−0.298^{+}	−0.319^{+}
	(0.178)	(0.178)	(0.178)
ordinary proofs	−0.387*	−0.382*	−0.408*
	(0.182)	(0.182)	(0.182)
familiar	0.291*	0.291*	**0.293***
	(0.115)	(0.115)	(0.115)
false (unfamiliar)	0.458**	0.458**	**0.457****
	(0.142)	(0.142)	(0.142)
meaning generality	0.678***	0.682***	0.675***
	(0.146)	(0.146)	(0.147)
Controls			
CRT score	0.614***	0.642***	0.655***
	(0.148)	(0.143)	(0.143)
LK	0.387**	0.394**	0.381**
	(0.147)	(0.147)	(0.147)
transition course	−0.236^{+}	−0.223^{+}	
	(0.129)	(0.128)	
final grade maths	−0.098		
	(0.134)		
SD (Intercept id)	0.707	0.708	0.715
Observations	2014	2014	2014
AIC	2398.1	2396.6	2397.7
BIC	2465.4	2458.3	2453.7

Note: $^{+}$p<.1, *p<.05, **p<.01, ***p<.001. Holm's adjusted signif. marked bold

generality. Participants with a higher CRT score were more likely to have a correct understanding of generality than participants with a lower score ($\beta = .66$, $p < .001$). Similarly, participants who attended an honors course were also more likely to have a correct understanding than participants who had a regular mathematics course in high school ($\beta = .38$, $p = .010$). The participation in the transition course had an unexpected negative effect (and was therefore excluded in the final model), however, not quite reaching significance (in Model 2, $\beta = -.22$, $p = .081$). The effect of the final mathematics grade was comparatively small and not significant (in Model 1, $\beta = -.10$, $p = .463$).

Students' Understanding of Generality in Relation to Their Conviction and Comprehension

Figure 6.27 shows the relation between students' level of conviction regarding different arguments and their understanding of generality. Regarding empirical arguments, there is a negative relation between students' understanding of generality and their level of conviction. Participants, who were convinced by empirical arguments (partially or completely), had an incorrect understanding of generality more often (about 37 and 43%, respectively) than participants who claimed to not find the empirical arguments convincing at all (about 20%).

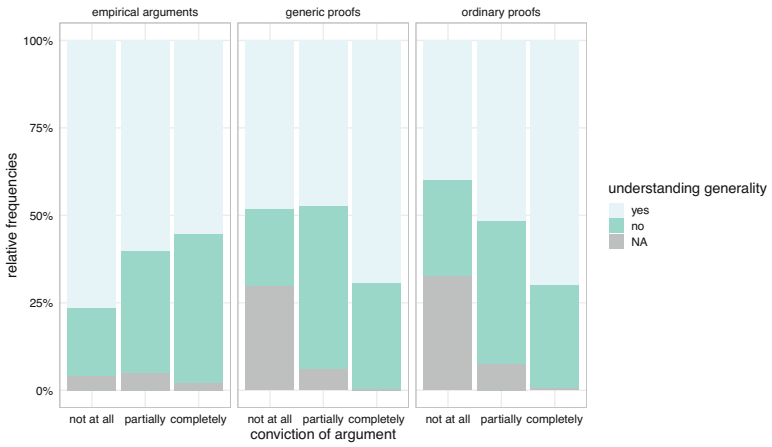


Figure 6.27 Understanding generality by type of argument and level of conviction

A simple mixed effects logistic regression[5] was calculated to analyze this relation, in which the individuals were considered as a random effect (see Model 1 in Tab. 6.9). Participants, who claimed to find the empirical arguments not at all convincing were more likely to have a correct understanding of generality than participants who claimed to be partially convinced by empirical arguments ($\beta = .90$, p.adj $= .027$). In contrast, participants who were completely convinced by empirical arguments were less likely to have a correct understanding of generality than participants who were partially convinced, but this effect was comparatively smaller and not significant ($\beta = -.26$, p.adj $= .353$).

It seems that this effect is reversed regarding generic and ordinary proof. However, more than a quarter of observations in which students found the arguments not convincing at all consisted of *missing values* for their understanding of generality (these participants responded "I have no idea" regarding the two relevant questions, see section 5.3.5). After removing these observations, the percentage of observations in which participants claimed to be only partially convinced by the generic or ordinary proofs with a correct understanding of generality was lower than for both, participants claiming to be completely convinced by the argument and participants who were not convinced at all (about 53% vs about 70 and 65%, respectively). A mixed effects logistic regression[6] was calculated with the individual participants as a random effect and the type of argument and the level of conviction as fixed effects (see Model 2 in Tab. 6.9). Participants, who claimed to find generic or ordinary proofs completely convincing were more likely to have a correct understanding of generality than participants who claimed to be only partially convinced ($\beta = .74$, p.adj $< .001$). Noteworthy, participants, who were not at all convinced by the proofs were also more likely to have a correct understanding of generality than participants who claimed to be only partially convinced, but this effect did not reach significance ($\beta = .52$, p.adj $= .202$). The effect of the type of argument (generic vs ordinary proof) was very small and highly insignificant ($\beta = .05$, $p = .788$).

Similar to the findings regarding students' conviction, the percentage of *missing values* for understanding generality is very high regarding participants who claimed to have not understood the generic and ordinary proofs at all (see Fig. 6.28). There seems to be a positive relation between students' self-reported proof comprehension and their understanding of generality, as the percentage of students with a correct understanding of generality was the highest for students claiming to have under-

[5] Regarding conviction of empirical arguments, p-values were adjusted based on two comparisons, not at all and completely convinced against partially convinced

[6] Regarding conviction of generic and ordinary proofs, p-values were adjusted based on four comparisons, not at all and completely convinced against partially convinced, both for each of the two models (Model 2 and 4).

Table 6.9 GLMM comparison regarding students' understanding of generality in relation to conviction and comprehension (Model 1 regarding empirical arguments; Models 2–4 regarding generic and ordinary proofs)

	Dependent variable:			
	Understanding generality			
	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.652**	0.145	0.161	−0.632*
	(0.202)	(0.176)	(0.194)	(0.285)
completely convincing	−0.258	**0.744*** **		0.352
	(0.278)	(0.180)		(0.226)
not at all convincing	**0.899***	0.519		**1.069****
	(0.365)	(0.316)		(0.375)
understood completely			**0.697*** **	0.423^{+}
			(0.193)	(0.245)
understood not at all			−0.288	−0.793^{+}
			(0.351)	(0.410)
ordinary proofs		0.050	0.105	0.184
		(0.184)	(0.188)	(0.183)
familiar				0.173
				(0.166)
meaning generality				0.634**
				(0.221)
CRT score				0.554**
				(0.204)
LK				0.232
				(0.207)
SD (Intercept id)	0.978	0.631	0.658	0.505
Observations	431	760	760	760
AIC	538.9	973.4	971.9	949.5
BIC	555.2	996.5	995.1	1000.5

Note: $^{+}$p<.1, *p<.05, **p<.01, ***p<.001. Holm's adjusted signif. marked bold

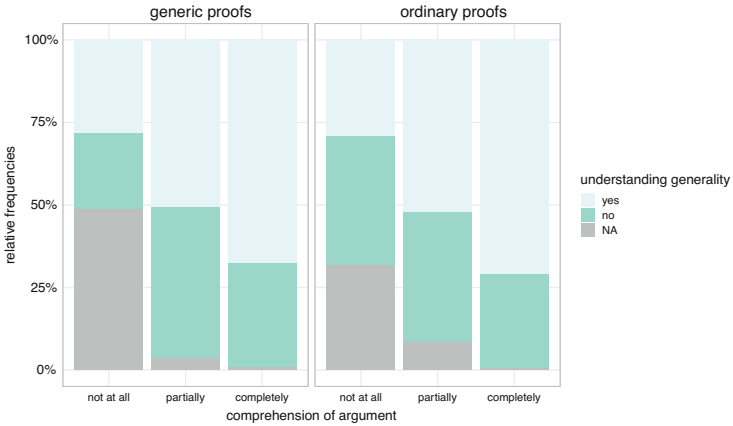


Figure 6.28 Understanding generality by type argument and level of comprehension

stood the respective arguments completely, regarding both generic and ordinary proofs (about 68 and 71%, respectively). However, due to the percentage of *missing values*, the effect is again less clear regarding students who claimed to have not understood the argument at all. After removing observations with *missing values*, for generic arguments, the percentage of students with a correct understanding of generality was slightly higher for students claiming to have not understood the respective arguments at all than for those claiming to have partially understood them (about 55 and 52%, respectively). For ordinary arguments, the positive relation between students' self-reported proof comprehension and their understanding of generality holds in general: The higher the level of proof comprehension, the higher the percentage of observations with a correct understanding of generality (about 43, 57, and 71%). A mixed effects logistic regression[7] was again calculated with the individual participants as a random effect and the type of argument and the level of self-reported comprehension as fixed effects (see Model 3 in Tab. 6.9). Participants, who claimed to have understood the generic or ordinary proofs completely were more likely to have a correct understanding of generality than participants who claimed have only partially understood the arguments ($\beta = .70$, p.adj $= .001$). In contrast, participants who claimed to have understood the proofs not at all were less

[7] Regarding comprehension of generic and ordinary proofs, p-values were adjusted based on four comparisons, not at all and completely understood against partially understood, both for each of the two models (Model 3 and 4).

likely to have a correct understanding of generality, but this effect was comparatively smaller and not significant ($\beta = -.29$, p.adj $= .411$). The type of argument was again not predictive ($\beta = .10$, $p = .578$).

A further mixed effects logistic regression was fitted, in which the predictive variables from the main analysis of students' understanding of generality (Model 3 in Tab. 6.8) were included. After controlling for these variables, the observed effects regarding the influence of students' conviction and proof comprehension are partly different. The direction of the effect of students' self-reported proof comprehension remains: Students who claimed to have understood the proofs completely were more likely to have a correct understanding of generality and students who self-reportedly understood the proofs not at all were less likely to have a correct understanding of generality, both compared to students who claimed to have only partially understood the proofs ($\beta = .42$, p.adj $= .169$ and $\beta = -.79$, p.adj $= .159$, respectively), even though these effects did not reach significance. The effect regarding students being completely convinced compared to students being only partially convinced by the proofs remains positive, however, not reaching significance anymore ($\beta = .35$, p.adj $= .202$). The effect that students who were not convinced by the argument at all were more likely to have a correct understanding of generality than students who claimed to be partially convinced is larger after controlling for other variables ($\beta = 1.07$, p.adj $= .013$). The positive effects of the CRT score and students correct knowledge of the meaning of mathematical generality on students understanding of generality mainly remain. However, the participation in an honors course (LK) and the familiarity with the statement were not predictive anymore after students' proof comprehension and conviction were included in the model ($\beta = .23$, $p = .263$ and $\beta = .17$, $p = .299$, respectively).

Students' Understanding of Generality in Relation to Their Proof Schemes
Figure 6.29 gives an overview of students' proof schemes in relation to their understanding of generality. The percentage of students with a correct understanding of generality was the highest for students with (complete and incomplete) deductive proof schemes (about 90%) and the lowest for students with purely empirical (no awareness of generality) or incomplete transformative proof schemes) about 60%). The percentage of participants having a correct understanding of generality was similar for students referring to relevant aspects, authorities, a rule, or giving pseudo arguments, namely about 75 to 80%.

To analyze the statistical significance of these differences, Chi square test was used. To increase the power of the test, the categories were summarized as explained in section 5.4.6. Overall, a relation between students' proof schemes and their understanding of generality seems to exist (see Tab. 6.10). The percentage of students

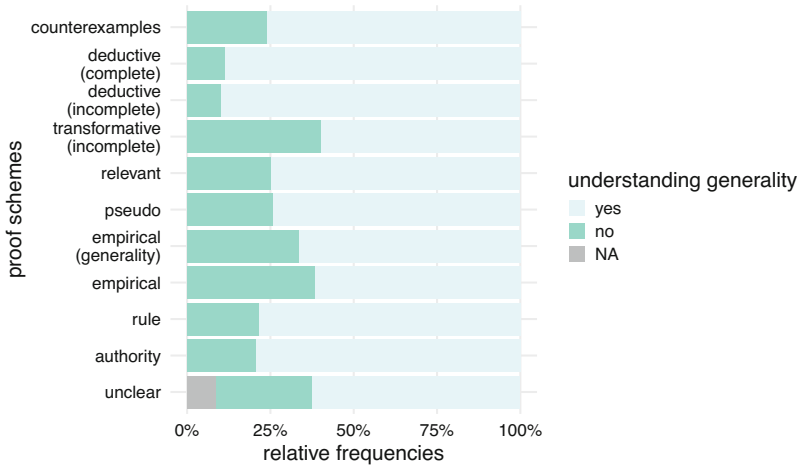


Figure 6.29 Understanding of generality by students' proof schemes (based on 462 observations)

Table 6.10 Number/Percentages of observations in which students had a correct (yes) or incorrect (no) understanding of generality by proof schemes

	unclear	external	empirical	counterexamples	analytical	Sum
no	17 (31.5%)	44 (22.9%)	41 (37.6%)	11 (23.9%)	9 (16.1%)	122 (26.7%)
yes	37 (68.5%)	148 (77.1%)	68 (62.4%)	35 (76.1%)	47 (83.9%)	335 (73.3%)
Sum	54	192	109	46	56	457

with a correct understanding of generality was the highest for students with an analytical proof scheme (about 84%), followed by students with an external proof scheme or one that consists of giving correct counterexamples (about 77 and 76%, respectively). The lowest percentage was observed for students having empirical proof schemes (about 62%). The percentage of students with a correct understanding of generality whose justifications were coded as unclear was lower than the average in group A (about 69% vs 73%). The differences reported are statistically significant with medium effect size ($\chi^2(4) = 12.0$, $p = .017$, Cramer's V = .16).

6.6 Analysis of *Missing Values*

The results presented above have shown a comparatively high percentage of students who responded "I have no idea" to both questions used to determine their understanding of generality. Even though these observations are not true missing values (because the participants in fact chose an answer regarding the two relevant questions), a decision regarding the understanding of generality of statements could not be made for these observations. Therefore, they were treated as *missing values* in the regression analyses. This section aims at identifying any patterns among these observations by calculating mixed effects logistic regression models. A dummy variable *dropout generality* was defined as follows:

- yes (1), for *missing values* in the variable understanding generality, and
- no (0), if a value for understanding generality was observed (either yes or no).

The individuals were again used as a random effect. All variables that were considered in the regression models above were used as fixed effects, to analyze the potential relation between these variables and observations with *missing values* (in the sense described above) regarding the understanding of generality. Table 6.11 shows the regression results. To estimate the effect of the type of argument participants received, Model 1 excluded the variables regarding students' conviction and comprehension, because this data was not collected for participants in group A, who received no arguments. Participants who received any type of argument were less likely to "drop out" (i.e., answering "I have no idea" regarding both the estimation of truth and the existence of counterexamples) than participants who received no arguments at all. This effect was particularly large and highly significant regarding empirical arguments ($\beta = -1.23$, $p < .001$) and generic proofs ($\beta = -.97$, $p < .001$) , but also present for ordinary proofs ($\beta = -.57$, $p = .031$). Compared to the true unfamiliar statements, participants were significantly more likely to answer "I have no idea" regarding the familiar statements ($\beta = 1.80$, $p < .001$) and the false statement ($\beta = .96$, $p = .003$). Participants who attended a mathematics honors course in high school were less likely to drop out than participants who attended a regular mathematics course ($\beta = -1.22$, $p < .001$). Noteworthy, the CRT score seemed only have a minor effect on participants' likelihood of choosing "I have no idea" regarding both questions, not reaching significance ($\beta = -.47$, $p = .063$). Furthermore, the higher the math grade (which in Germany means a worse grade), the more likely participants dropped out ($\beta = .44$, $p = .036$). However, similar to the CRT score, this effect was comparatively small.

Table 6.11 GLMM results of the dropout variable regarding students' understanding of generality

	Dependent variable:	
	Dropout generality	
	Model 1	Model 2
(Intercept)	−3.100***	−1.506*
	(0.318)	(0.611)
empirical arguments	−1.231***	
	(0.295)	
generic proofs	−0.965***	
	(0.278)	
ordinary proofs	−0.572*	0.192
	(0.265)	(0.396)
familiar	1.800***	2.066**
	(0.265)	(0.715)
false (unfamiliar)	0.957**	
	(0.323)	
CRT score	-0.466^{+}	−0.368
	(0.251)	(0.501)
LK	−1.218***	−0.918
	(0.295)	(0.580)
transition course	−0.179	
	(0.204)	
final grade maths	0.442*	0.359
	(0.211)	(0.391)
partially understood		−1.565**
		(0.579)
completely understood		−2.524**
		(0.838)
partially convinced		-1.061^{+}
		(0.617)
completely convinced		−3.063**
		(1.052)
SD (Intercept id)	0.606	0.065
Observations	2150	808
AIC	893.9	225.8
BIC	956.3	282.2

Note: $^{+}$p<.1, *p<.05, **p<.01, ***p<.001

In Model 2, only observations from groups C and D (generic and ordinary proofs) were considered and the false statement was again excluded. Because the participation in a transition course showed no significant effect, it was excluded in Model 2, also, because of the otherwise large number of variables. The participation in an honors course was not predictive for missing values in the variable understanding generality, once comprehension and conviction were included ($\beta = -.92$, $p = .114$). The level of self-reported comprehension and conviction were both predictive for the missing values regarding understanding of generality. In particular, students who claimed to have partially or completely understood the proofs were less likely to drop out than students who claimed to have not understood the proofs at all ($\beta = -1.57$, $p = .007$ and $\beta = -2.52$, $p = .003$). Similarly, students who found the proofs partially or completely convincing were less likely to drop out than students who claimed to be not convinced by the proofs at all ($\beta = -1.06$, $p = .085$ and $\beta = -3.01$, $p = .004$).

Overall, these results indicate that the *missing values*–observations in which students responded with "I have no idea"–substantially depended on other variables, such as the (type of) statement, the type of argument (or more general, receiving any argument at all), the self-reported comprehension of the proofs, and how convincing students evaluated the proofs.

6.7 Summary of Main Results

This section provides an overview of the main results of the present study, in particular regarding the influence of the type of argument and statement on proof-related activities, and students' understanding of the generality of mathematical statements and the relation to proof reading and construction.

6.7.1 Influence of the Type of Argument

The study was mainly designed to experimentally analyze the influence of the *type of argument*–receiving no arguments, empirical arguments, generic proofs, or ordinary proofs–on students' understanding of the generality of statements and other proof-related activities (see section 5.2.3). In summary, the type of argument significantly influenced:

- Students' estimation of truth: Participants who received empirical arguments were more likely to correctly estimate the truth value of the statements than

participants who got no arguments. Reading generic proofs had a similar, but smaller effect and did not reach significance after Holm's correction was applied.

- Students' proof evaluation regarding conviction: Participants who received generic or ordinary proofs were more likely to claim being convinced by these arguments than participants who received empirical arguments. The reasons why participants claimed not to be convinced by the arguments also differed by the type of argument. The reason most often referred to by participants who received empirical arguments was a lack of generality of these arguments (78% of observations), while participants who received generic or ordinary proofs were mainly not convinced by these arguments because they did not (completely) understand them (64 and 81%, respectively).
- Students' proof comprehension: Participants who received ordinary proofs were less likely to have self-reportedly understood the arguments than those participants, who received generic proofs. Further, aspects that participants claimed to have not understood differed between generic and ordinary proofs. For instance, the generality of the proof was not mentioned at all by participants who received ordinary proofs, but by those who received generic proofs (14%).
- The probability of *missing values*: Participants who received *any* type of argument were less likely to answer "I have no idea" regarding the estimation of truth and the existence of counterexamples than participants who received no arguments. This effect was particularly strong for participants who got empirical arguments.

The type of argument did not have a large effect on students' understanding of the generality of statements, but participants who received ordinary (and with a smaller effect generic) proofs were less likely to have a correct understanding than students who got no arguments. However, these effects did not reach significance after Holm's correction.

6.7.2 Influence of the Type of Statement

To analyze the influence of the type of statement (truth value and familiarity), all participants received five statements of different types: Two (true) familiar statements, two (true) unfamiliar statements, and one false (unfamiliar) statement. Overall, the truth value of the statement and the familiarity with the statement both significantly influenced students' performance in all considered activities. In summary, compared to true, unfamiliar statements, participants were

- less likely to correctly estimate the truth value of the false statement,
- less likely to be convinced by (incorrect) arguments regarding the false statement,
- *more* likely to show a correct understanding of generality regarding the false statement,
- more likely to answer "I have no idea" regarding the estimation of truth and the existence of counterexamples for the false statement.

Regarding the familiar (geometry) statements, participants were

- more likely to correctly estimate the truth value,
- more likely to be convinced by the arguments,
- *less* likely to claim to have understood the arguments,
- more likely to show a correct understanding of generality (with a comparatively smaller effect than regarding the false statement),
- much more likely to answer "I have no idea" regarding the estimation of truth and the existence of counterexamples,

all compared to the true, unfamiliar statements from elementary number theory.

Moreover, students' proof schemes also differed by the type of statement: Participants mainly gave counterexamples to refute the false statement (51%), empirical arguments to justify the true unfamiliar statements (45%), and external arguments (pseudo, rule based, authority) to justify the true familiar statements (25, 37, and 24%, respectively).

6.7.3 Students' Understanding of Generality and the Relation to Proof

The focus of the present thesis was to analyze students' understanding of the generality of mathematical statements. Most predictive for students' (correct) understanding of generality was their *knowledge* of the meaning of mathematical generality (measured via the closed item shown in Fig. 5.12). Furthermore, the truth value and the familiarity with the statement both influenced students' understanding of generality (see above). While reading different types of arguments significantly affected the probability of participants answering "I have no idea" to the two relevant questions (estimation of truth and existence of counterexamples), it did not seem to have a large effect on students' understanding of the generality of statements. The analysis of the relation between students' understanding of generality and their performance in other proof related activities suggests:

- There is a positive relation between students' self-reported proof comprehension and their understanding of generality of statements: Students who claimed to have completely understood the proofs were more likely to have a correct understanding of generality and students who claimed to have not understood the proofs at all were less likely to have a correct understanding of generality both compared to students who claimed to have partially understood the proofs.
- There is a negative relation between students' conviction of empirical arguments and their understanding of generality: Students who claimed to be *not at all* convinced by the empirical arguments were more likely to have a correct understanding of the generality of statements than those who claimed to be partially (or completely) convinced by the arguments.
- There is no clear relation between students' evaluation of generic and ordinary proofs and their understanding of generality. Students' who claimed to be completely convinced by these arguments might be more likely to have a correct understanding of generality than students, who were only partially convinced. However, after considering other (predictive) variables (such as knowing the meaning of generality and the CRT score), this effect did not reach significance. Even more, participants who claimed to be not at all convinced by the arguments were more likely to have a correct understanding of generality.
- Students' proof schemes are related to their understanding of the generality of statements: Participants with empirical proof schemes most often had an incorrect understanding of generality, followed by participants with external proof schemes. Among the participants with deductive proof schemes, the percentage of participants with an incorrect understanding of the generality of statements was the lowest.

6.7.4 Predictive Power of Control Variables

The CRT (Cognitive Reflection Test) score and the attendance of an honors mathematics course in high school (LK) were overall the most predictive control variables. In particular, participants with a higher CRT score were more likely

- to have a correct understanding of generality,
- to correctly estimate the truth value,
- to claim to have (completely or partially) understood the proofs.

Attendance of an honors course had similar effects, however, after proof comprehension (and conviction) was considered, it was not predictive for a correct under-

standing of generality anymore (the CRT score still was). Unexpectedly, the final mathematics grade was only predictive for students' self-reported proof comprehension (with a smaller effect when compared to the CRT score and LK participation) and the estimation of truth (even though not reaching significance) but not for understanding the generality of statements.

Discussion

7

The main purpose of the present thesis is to investigate first-year university students' understanding of the generality of mathematical statements and the relation to proof reading and construction (see Fig. 4.1 in Chapter 4). The respective research questions were structured through my adapted version of the framework on proof-related activities introduced by Mejía Ramos and Inglis (2009b), in which I suggest to explicitly consider the reading of the statement that has to be proven or for which a proof has to be read. The reading of a statement involves the comprehension of the statement, among other aspects, its generality. I defined understanding the generality of statements as consistent responses regarding the estimation of truth and the existence of counterexamples, which was then used to operationalize this understanding. To investigate the relation to proof reading and construction, students' performances in the relevant activities—estimation of truth, proof evaluation regarding conviction, proof comprehension, and proof construction—were considered. Moreover, the experimental design of my study particularly aimed at analyzing the influence of the type of argument as well as the type of statement (truth value and familiarity) on students' understanding of the generality of statements and their proof skills.

In the following, the results presented in the previous chapter are interpreted and discussed in the context of prior research. Thereby, I follow the structure of the four sets of research questions derived in Chapter 4. Further, the adapted framework on proof-related activities, methodological decisions, and potential limitations of this study are discussed in Section 7.2. Lastly, main implications of the results for the

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-658-43763-3_7.

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_7

learning and teaching of proof at the transition from school to university and future research are presented in Sections 7.3 and 7.4, respectively.

7.1 Interpretation

In the following, the research questions are answered one by one and the results are interpreted and discussed in relation to prior research.

7.1.1 Estimation of Truth and Proof Evaluation Regarding Conviction

The first set of research questions focused on students' performance in estimating the truth of statements and proof evaluation regarding conviction:

RQ1: Conviction of the truth of universal statements and its relation to reading different types of arguments

RQ1.1: How do the type of argument and the type of statement influence students' estimation of the truth of universal statements?
RQ1.2: How do the type of argument, the type of statement, and the level of comprehension influence how convincing students find different types of arguments? What aspects of mathematical arguments do students identify as not convincing?

As researchers and prior studies have suggested (e.g., Barkai et al., 2002; Buchbinder & Zaslavsky, 2007; Dubinsky & Yiparaki, 2000; Hanna, 1989; Ko, 2011), the type of statement, in particular the statements' truth value but also the familiarity with the statement affected students' estimation of truth. The falsity of the statement had a negative effect and being familiar with a statement had a (smaller) positive effect. These results were not surprising, but provide clear experimental evidence on what has already been suggested in the literature. The comparatively small effect of familiarity can mainly be explained by students' estimation of truth of the pythagorean theorem. A comparatively large percentage of participants was unsure about the truth value of this statement, even though they should be very familiar it. The fact that the pythagorean theorem was expressed in natural language and not as an equation is most likely the reason why some students seemed to not have recognized the statement and therefore had difficulties with estimating the truth value.

This can be seen as a limitation (see also Section 7.2), but also provides information regarding students' content-specific knowledge and level of comprehension of mathematical statements they have been taught in school.

Reading empirical arguments (and generic proofs) supports students to estimate the truth value of statements. The type of argument affected students' estimation of truth. Participants who received empirical arguments (and generic proofs, with a smaller effect and not reaching significance after Holm's correction) were more likely to correctly estimate the truth value of (true) statements than participants who received no arguments. No prior studies on the influence of different types of arguments on students' estimation of truth had been conducted before, which made it difficult to formulate a hypothesis. However, students as well as professional mathematicians use empirical arguments to estimate the truth value of statements (e.g., Alcock & Inglis, 2008; Buchbinder & Zaslavsky, 2007; Lockwood et al., 2016), possibly because these experimental investigations provide better understanding of the statement and a better intuition regarding its truth value (see also de Villiers, 2010). This is in line with findings reported by Bieda and Lepak (2014) that empirical arguments provide students with more information and enhance their comprehension of the statement in comparison to ordinary proofs. My findings now provide strong evidence that empirical arguments (and to a lesser degree generic proofs) may indeed help to better understand mathematical statements and therefore lead to better performance in the estimation of truth. Furthermore, reading *any* of the considered types of arguments seems to make participants more likely to choose an answer different from "I have no idea", in particular the reading of empirical arguments. This strengthens the assumption that empirical arguments may provide participants with a better understanding of the statement—or at least give them the feeling to better understand it.

The second research question in this set focused on students' evaluation regarding conviction. Similar to the findings on students' estimation of truth, the type of statement also affected students' conviction. Participants were less likely to find the arguments convincing regarding the false (unfamiliar) statement than regarding the true (unfamiliar) statements, which would be expected, because the respective proofs were in fact incorrect. Further, participants were more likely to be convinced by the arguments regarding the familiar statements than the unfamiliar (true) statements, even though this effect was comparatively smaller than the effect of the truth value. A positive effect of familiarity with the statement on students' conviction was also expected, because the role of familiarity for the acceptance of proof (which most likely influences conviction) has been highlighted in the literature, as

already mentioned (e..g., Hanna, 1989). But prior studies had not found respective evidence (e.g., Kempen, 2021; Martin & Harel, 1989).

The type of argument also affected students' conviction. In line with prior research (e.g., Kempen, 2021; D. Miller & CadwalladerOlsker, 2020; Weber, 2010), participants who received generic or ordinary proofs were more likely to find these arguments convincing than students who received empirical arguments. However, in a comparatively high percentage of observations, participants nevertheless claimed to be completely (!) convinced by empirical arguments (about 25%). Those who were not (completely) convinced by the empirical arguments were asked to explain why they are not convinced by these arguments. Participants most often referred to a lack of generality (78% of observations) of the arguments, which indicates that the majority of these students is not only aware of the limitations of empirical arguments, but understands—at least to some extent—why. In contrast, in a study conducted by Ufer et al. (2009), only about one third of the participating high school students could "adequately" explain why an empirical argument is not valid. However, in the present study, the 78% reported above only refer to those participants, who were not or only partially convinced by these arguments *and* who responded to the open question. Further, the responses were not thoroughly coded regarding *adequacy* of their responses, but with respect to mentioned aspects (here *generality*). Thus, the comparability of the results may be limited, also because of the differences regarding age and experience of participants (high school students vs first-year university students).

As discussed in Section 3.2.3, prior research findings on students' and teachers' evaluation of generic proofs has been ambiguous. Some studies found that many teachers are not convinced by generic proofs, for instance, because of a perceived lack of generality and modes of representations that do not meet the criteria for proof (e.g., Lesseig et al., 2019; Tabach, Levenson, et al., 2010). The majority of participants in the present study seemed to be (at least partially) convinced by generic proofs; participants claimed to not find these arguments convincing at all in less than 25% of the observations. Moreover, a lack of generality was indeed mentioned more often as a reason for not finding the arguments (completely) convincing by participants who received generic proofs than by those who received ordinary proofs (12% vs 4%). In contrast to the findings reported by Tabach, Barkai, et al. (2010), the mode of representation, was only mentioned occasionally regarding generic proofs (6%) (and not at all regarding ordinary proofs). Further, participants in the present study were more convinced by the ordinary proofs than the generic proofs, which experimentally confirms findings reported by Kempen (2018), for instance.

The results of the present study furthermore clearly confirm the influence of students' (self-reported) proof comprehension on their (self-reported) conviction (as has been reported by Weber, 2010, for instance): Participants with higher levels of (self-reported) proof comprehension were also more likely to claim being convinced by the arguments. Moreover, the results of the content analysis of aspects students identified as not convincing also highlight these findings: The comprehension of the statement or proof was the reason why most of the participants claimed to be not convinced by generic or ordinary proofs (64 and 81%, respectively). While these findings may not be surprising and confirm prior research findings (Ko & Knuth, 2013; Sommerhoff & Ufer, 2019), they nevertheless emphasize the strong relation between proof comprehension and proof evaluation regarding conviction.

Self-reported conviction of arguments does not reflect actual conviction of the truth of statements. While participants showed higher levels of self-reported conviction regarding generic and ordinary proofs compared to empirical arguments, the participants who received empirical arguments (and generic proofs) were more likely to correctly estimate the truth value of (true) universal statements than those participants who received no arguments. Differences regarding ordinary proofs were not significant. This suggests that participants assume that ordinary (and generic) proofs *should* generally be convincing, in particular compared to empirical arguments, but that empirical arguments (and potentially generic proofs) *actually* provide higher levels of conviction regarding the truth of the statement. This finding highlights the gap between self-report—which can potentially be influenced by social desirability, for instance—and reality (Golke, Steininger, & Wittwer, 2022) and is particularly relevant for the construction of future questionnaires. I come back to this finding and its implications in Section 7.4.

7.1.2 Comprehension of Arguments

The second set of research questions aimed at investigating students' (self-reported) proof comprehension, in particular regarding differences between generic and ordinary proofs:

RQ2: Proof comprehension

RQ2.1: How does students' (self-reported) proof comprehension differ between students who receive generic proofs and those who receive ordinary proofs? How does the familiarity with the statement influence students' proof comprehension?
RQ2.2: What aspects of mathematical arguments do students identify as not understandable? How do these aspects differ regarding generic and ordinary proofs?

Participants show higher levels of proof comprehension regarding generic proofs than regarding ordinary proofs. Based on prior experimental studies (e.g., Lew et al., 2020), it was hypothesized that no significant differences between students' comprehension of generic and ordinary proofs exist. Therefore, the finding that participants who received ordinary proofs were less likely to claim to have understood these arguments than participants who received generic proofs was unexpected. In contrast to other studies, for instance, by Lew et al., the present study relied on students' self-report on their proof comprehension. Thus, the participants who received generic proofs might not actually have better understood these proofs, but these types of arguments might just have appeared more comprehensible to them. Previous research has indeed found that mathematics students often inaccurately assess how well they have understood a proof (A. Selden & Selden, 2003). However, it could also be the case that generic proofs *do* provide students with better understanding, as other researchers have suggested (Dreyfus et al., 2012; Malek & Movshovitz-Hadar, 2011; Mason & Pimm, 1984; Rowland, 2001). Given that generic proofs supported students' correct estimation of truth (see finding above), they may indeed also help with the comprehension of proof. Further (experimental) studies, which do not solely rely on students' self-reports are needed to definitely answer this question.

Unexpectedly, the familiarity with the statement had a negative effect on students' proof comprehension. This result is surprising, because the participants have encountered these statements and potentially the proofs before during school and should also be more familiar with the underlying theories of these statements. However, the statements did not only differ with respect to familiarity, but also regarding their content domains. The familiar statements were from geometry and the unfamiliar statements from elementary number theory. Thus, most likely, participants have perceived the proofs regarding the familiar statements *from geometry* to be more difficult, not because of the familiarity.

With respect to the second research question in this set, in line with prior research (e.g., Conradie & Frith, 2000; Moore, 1994; Neuhaus-Eckhardt, 2022; Reiss & Heinze, 2000) and therefore expected, participants mainly referred to local aspects such as not having understood the terms, statements, equations, and/or illustrations used in the proof. Moreover, in a comparatively high percentage of observations, participants seemed to have not understood the statements themselves, for instance, the meaning of simple terms such as the meaning of odd or even numbers, product, or the square of the legs (in German *Kathetenquadrat*), and the explanations included in the proofs seemed to have not clarified these terms for the participants. This observation is of practical relevance for the teaching of proof, because it emphasizes the need to first focus on sufficiently understanding the statements and relevant terms before a proof is presented and discussed or before students are asked to prove a claim. One would assume that this is obvious, but lecturers might not be aware of the extent to which students have difficulties with simple terms and the comprehension of statements.

In contrast to ordinary proofs, participants stated to not having understood why the generic proofs are general on several occasions (14%), which is in line with the findings on aspects participants identified as not convincing reported above. As was expected, participants who received generic proofs also referred to the proof framework slightly more often than participants who received ordinary proofs when asked what they did not understand about the argument (8% vs 5%). But overall, this aspect was not mentioned that frequently. Most likely students have generally limited experience with proof and proving (as suggested by prior research, for instance, Hemmi, 2008; Kempen & Biehler, 2019) and therefore do not often consider the general proof idea, but focus on surface features, as has been reported in the literature (e.g., A. Selden & Selden, 2003). Further, a smaller percentage of participants referred to not having understood particular statements, equations, illustrations used in the generic proofs than in the ordinary proofs (33 vs 58%), but comparatively more participants did not understand the statements themselves regarding the reading of generic proofs than the reading of ordinary proofs (24 vs 13%). This does not necessarily mean that participants who received ordinary proofs had actually better understood the statements. It could also mean that the reading of generic proofs more often *reveals* an insufficient understanding. Further research would be needed to investigate this hypothesis.

7.1.3 Justification: Students' Proof Schemes

The first experimental group did not receive any arguments but instead had to justify why they think the statements are true or false. This group therefore served as a control group regarding the influence of reading arguments and provided data to answer the third set of research questions, which aimed at analyzing students' proof schemes:

RQ3: Construction of arguments to justify the truth of universal statements (students' proof schemes)

RQ3.1: What types of arguments do students themselves use to justify the truth or falsity of a universal statement? How do students' proof schemes differ regarding the type of statement (i.e., familiarity and truth value)?
RQ3.2: What potential relation between the type of argument used by students and the level of conviction of the truth of the statement exists?

As was expected based on prior research findings (e.g., Barkai et al., 2002; Bell, 1976; Recio & Godino, 2001; Sevimli, 2018; Stylianou et al., 2006), empirical proof schemes could be observed most often, when participants were asked to justify the truth of the unfamiliar statements (45%). In contrast, participants mainly showed external proof schemes regarding familiar statements (86%). These participants often made reference to authorities (24%), such as school or university, claimed the statement is a general rule (37%), or gave pseudo arguments (25%). Moreover, the majority of participants used counterexamples to correctly refute the false statement (about 51%). Expectedly, deductive proofs schemes were much rarer, but could be observed more often regarding unfamiliar statements from elementary number theory than familiar statements from geometry. Most likely, not only the (un)familiarity with the statements but also the different content domains account for these differences (see Section 7.2 for a further discussion). Transformative arguments, such as generic proofs, were only used 5 times, and these were all incomplete. I want to highlight again that the participants were not explicitly asked to *(dis)prove* the statements, but to justify why they think the statements are true or false, similar to what has been done by Barkai et al. (2002), for instance. Thereby, the aim was to gain insights about the types of arguments that convince students of the truth or falsity of universal statements, in the sense proof schemes were defined by Harel and Sowder (1998). Thus, these results might not be comparable to those of other studies, in which students were explicitly asked to construct a proof, for instance, in

the studies conducted by Recio and Godino (2001) and Stylianou et al. (2006), even though these studies also found that many students fail to construct valid deductive proofs and often give empirical arguments instead.

A relation between students' proof schemes and their level of conviction was identified. Participants who gave empirical arguments were generally only *relatively* convinced (100% regarding the familiar statements and more than 60% regarding the unfamiliar statements) of the truth of the respective statements. Thus, as argued by Weber and Mejia-Ramos (2015), one should not automatically worry about students' usage of empirical arguments, if they *do not* gain absolute conviction by these arguments. There was still a comparatively large percentage of participants with empirical proof schemes who seemed to have gained absolute conviction of the truth of the statements. Thus, the usage of empirical arguments might be problematic for *some* students. However, that fact that these participants gave empirical arguments does not necessarily mean that these arguments were the only source for their conviction in the truth of the statement. Further research is needed to investigate why some students seem to gain absolute conviction in the truth of a statement by empirical arguments and what other factors may influence students' (level of) conviction. The findings of the present study suggest that students who construct (complete or incomplete) deductive arguments have high levels of conviction of the truth of statements. Participants with deductive proof schemes were in fact almost all *absolutely* convinced of the truth of the true familiar and unfamiliar statements. However, it cannot be derived from these findings that the construction of deductive arguments (automatically) leads to absolute conviction, because other factors might have played a role as well. The respective participants might have been convinced by the truth of the statements before they even attempted to prove them (for instance, because they were familiar with the statements), as has been pointed out by Polya (1954). Noteworthy, external proof schemes seemed to provide most students with absolute conviction as well. But given that participants mainly had external proof schemes regarding familiar statements, the familiarity with these statements may be mainly responsible for the high levels of conviction.

7.1.4 Understanding the Generality of Statements

Finally, the last set of research questions build the focus of the present thesis, which is on students' understanding of the generality of mathematical statements and the relation to proof reading and construction:

RQ4: Students' understanding of the generality of mathematical statements

RQ4.1: What proportion of first-year university students have a correct understanding of the generality of statements?
RQ4.2: What is the influence of reading different types of arguments on students' understanding of the generality of mathematical statements? How does the type of statement influence students' understanding of its generality?
RQ4.3: How does students' comprehension and conviction of arguments influence their understanding of generality of statements?
RQ4.4: What potential relation exists between students' proof schemes and their understanding of the generality of statements?

In 64% of all observations (about 68 % if "don't knowers" are excluded), participants showed a correct understanding of the generality of mathematical statements. The percentage of students having a correct understanding of generality of statements thereby differed with respect to the study program. Overall, the higher the level of mathematics in the chosen study program, the higher the percentage of students with a correct understanding of generality. This can mainly be explained by differences in prior knowledge/experience (e.g., attendance of an honors course) and general cognitive skills (e.g., CRT score).

Understanding the generality of statements is not *solely* determined by students' knowledge of the meaning of mathematical generality but positively related to it.

The most predictive for students' understanding of generality of statements was their *knowledge* of the meaning of mathematical generality. However, since a comparatively large percentage of students with a correct knowledge of the meaning of generality still responded inconsistently regarding the estimation of truth and the existence of counterexamples, solely *knowing* what mathematical generality means is not sufficient for a consistent correct understanding of the generality of statements.

The percentage of observations in which participants responded inconsistently regarding the two relevant questions also differed by the type of statement, which further indicates that the understanding of generality of statements is not solely determined by students' knowledge of the meaning of generality. Participants were more likely to have a correct understanding of generality of the false and the familiar statements than of the (true) unfamiliar statements, as was expected. The content analysis of students proofs schemes further suggests that most students who correctly refuted the false statement most likely did so, because they found one or more counterexamples. These participants therefore *knew* that a counterexample exists,

which proves the falsity of the statement, and consequently responded more often consistently regarding the truth of the statement and the existence of counterexamples. Similarly, as has been argued in Chapter 4, familiar statements have most likely been applied by the participants to many *arbitrary* cases before, which might have made them more confident in the non-existence of counterexamples and therefore more likely to have a correct understanding of generality for these statements.

The reading of *any* type of argument mainly influenced students' responding behavior in that they were less likely to answer "I have no idea" regarding the estimation of truth and the existence of counterexamples. Thus, reading an argument may at least give them the feeling of knowing *enough* to make a decision. However, reading generic or ordinary proofs did not lead to a higher probability of having a correct understanding of generality, as was hypothesized. On the contrary. If at all, it made participants *less* likely to respond consistently to the respective questions. Reading empirical arguments seemed to have no significant effect on students' understanding of generality in comparison to reading no arguments at all. The findings reported above indicate that reading ordinary proofs did not support students' correct estimation of truth, potentially because students lack knowledge to gain information and certainty from proofs. Because of their limited knowledge, reading proofs might actually make them more uncertain regarding the existence of counterexamples, which might explain the higher likelihood of an incorrect understanding of generality of the statement. However, the significance of this effect is unclear.

With respect to the third research question, there seems to be a positive relation regarding students' (self-reported) proof comprehension and their understanding of generality of statements. After controlling for other individual resources, such as the CRT score and students' knowledge of the meaning of generality, this effect was however smaller and did not reach significance. Thus, the relation between proof comprehension and students' understanding of generality might at least partially be explained by other variables which influence both proof comprehension and students' understanding of generality (for instance, the CRT). In contrast, the participation in an honors mathematics course during high school, which was predictive for students' understanding of generality before proof comprehension was considered, had no significant effect after it was included. Thus, the participation in honors courses most likely provides students with better comprehension or these are both simultaneously influenced by a further variable. As the present study relied on students' self-report on their understanding of the arguments, findings are limited (see Section 7.2.3 for a further discussion on this). Measuring students' proof comprehension through assessment tests, as suggested by Mejía Ramos et al. (2012),

for instance, might provide further insights into the relation between students' proof comprehension and their understanding of the generality of statements.

The conviction of empirical arguments is related to students' understanding of the generality of statements. The findings reported in Section 6.5 on the influence of conviction on understanding of generality are inconclusive. A clear negative effect was found regarding students' conviction of empirical arguments on their understanding of generality. Participants who were *not at all* convinced by empirical arguments were more likely to have a correct understanding of the generality of statements than participants who were (at least partially) convinced by these arguments. A relation between students conviction of empirical arguments and their understanding of the generality of *proof* was suggested by some researchers (e.g., Conner, 2022). My findings now provide clear evidence for a relation between students' conviction of empirical arguments and their understanding of the generality of *statements*—which assumably is related to understanding the generality of proofs. The effect of conviction regarding generic or ordinary proofs, however, is less clear. There seemed to be a positive relation between students' conviction by the argument and their correct understanding of generality, however, after including other (control) variables, this effect diminished and—even more—participants who claimed to be not at all convinced by generic or ordinary proofs were then more likely to have a correct understanding of generality than those who claimed to by partially convinced by the proofs. Among the participants who claimed not to be convinced by the arguments, a high percentage answered "I have no idea" regarding the relevant questions for measuring students' understanding of generality. These were treated as missing values, which might have affected the estimates of the regressions. Another explanation for this unexpected relation could be that participants with an overall good understanding of proof—and potentially correct understanding of generality—tend to either be completely convinced or not at all, but not partially. Or vice versa, participants, who do not have a good understanding of proof and generality might tend to answer that they are partially convinced, simply because they have no reference of what (should) convince(s) them.

Students with empirical proof schemes are less likely to have a correct understanding of the generality of statements than students with deductive proof schemes. In contrast to the effect of reading different types of arguments on understanding the generality of statements, a relation between students' proof schemes and their understanding of generality was found. Participants with empirical proof schemes had an incorrect understanding of generality most often, compared to participants with any other proof scheme. The percentage of participants with a

correct understanding of generality was the highest among those with analytical proof schemes, in particular deductive ones. The percentage of participants with external proof schemes and a correct understanding of generality was between these two groups. Overall, these differences were significant with medium effect size. On the one hand, these results make sense in that participants who were able to construct a proof gain absolute conviction (as discussed above) and simultaneously make them more aware that no counterexamples exist, thus leading to a correct understanding of generality. On the other hand, it is interesting that students who give empirical arguments to justify universal statements respond inconsistently most often. It is conclusive that they only gain relative conviction (as was found in this study), but this does not explain that their estimation regarding the existence of counterexamples is then inconsistent. These findings combined with those regarding the influence of (or absence thereof) reading different types of arguments, could suggest that students who give empirical arguments in general have an insufficient understanding of proof which also affects their understanding of the generality of statements, but that it is not the reading or construction itself that explains this relation.

7.2 Reflections and Limitations

The present study provides many new insights into students' proof skills, in particular students' understanding of the generality of statements. In the following sections, I reflect on my adapted framework on proof-related activities and several methodological decisions that were made in this study. In addition, I outline specific limitations of this study as well as more general limitations of empirical (field) research.

7.2.1 The Adapted Framework on Proof-Related Activities

The present study was based on the framework for proof-related activities presented in Section 3.2, which is an adapted version of the framework introduced by Mejía Ramos and Inglis (2009b). I chose to distinguish between activities that are related to the *statements*, which are to be proven or for which a proof is to be read, and activities that are related to the *arguments* that aim to justify the statements. Moreover, I proposed potential relationships between the activities (see Fig. 7.1; *problem exploration* was not explicitly considered in the present study).

The adapted framework has been shown to be very useful and conclusive in this study. Further, my findings mainly confirm or at least highlight the presumed

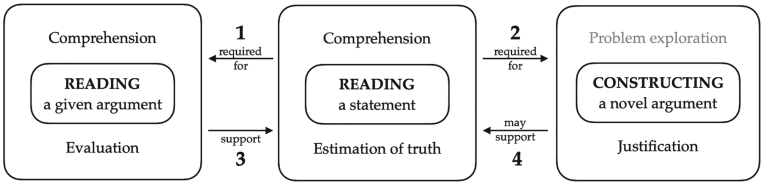


Figure 7.1 Adapted framework on proof-related activities based on Mejía Ramos and Inglis (2009b), numbers refer to identified relationships

relationships. The content analysis of students' responses regarding aspects they did not understand showed that comparatively many participants insufficiently comprehended the statements for which they received and read arguments. These participants were therefore not able to understand the arguments and they would also not have been able to decide if the arguments are valid proofs. Thus, not surprisingly, reading the statement with respect to comprehension is required for activities regarding the reading of given arguments (relationship 1 in Fig. 7.1). Similarly, without understanding the statement, students were not able to justify its truth or falsity, which mainly resulted in *unclear* responses regarding students' proof schemes (relationship 2 in Fig. 7.1). Reading (particular types of) arguments affected students' success in estimating the truth of the statements (relationship 3 in Fig. 7.1). In particular, reading empirical arguments (and to a lesser degree generic proofs) supported students in deciding if the statements are true or false, likely because they helped them to better understand the statements. But reading ordinary proofs did not have this effect. Even more, it seemed that—if at all—reading ordinary proofs negatively influenced students' understanding of the generality of statements (as part of statement comprehension). Moreover, my findings provide evidence that strong relations between different proof reading activities exist as well, for instance, between comprehension of the arguments and evaluation regarding conviction. The influence of constructing (different types of) arguments to justify the truth or falsity of statements may also support students' comprehension of statements and their success in estimating the truth (relationship 4 in Fig. 7.1). Participants who were not able to give any argument—not even empirical ones—to justify the statements (coded as *unclear* regarding the proof scheme) were also most often unsuccessful in estimating the truth value of the statements. Vice versa, most students who provided arguments correctly estimated the truth of the statements, at least with relative conviction, and the type of proof scheme was related to their level of conviction as well as their understanding of generality of the statements. Further research would be

needed to explicitly investigate if and how the construction of arguments supports the comprehension of statements and students' success in estimating the truth value.

Overall, my adapted framework, in particular the distinction between activities related to statements and arguments, provides a useful basis for further research on proof-related activities and their relations.

7.2.2 Overall Research Design

In the present study, participants were randomly assigned to experimental groups (types of arguments), which was methodologically desirable. However, there might still have been a selection bias induced by the different types of arguments participants received: The percentage of students who chose not to answer or complete the questionnaire was higher for generic and ordinary proofs than for empirical arguments and no arguments, possibly due to the perceived difficulty of reading these proofs. Moreover, the percentage of "I don't knowers" was lower among the participants who received any type of argument in comparison to participants who were not provided with arguments. Thus, responding behavior of participants who did not drop out of the experiment early on was influenced by reading the arguments, in that it made them more confident in choosing an answer different from "I have no idea". Therefore, while the *missing values* might limit my results to some extent, they also provide information on aspects that make participants less likely to be "I don't knowers"—for instance the reading of arguments—which can be useful for future studies.

Furthermore, some of the questions might have been redundant for participants who received empirical arguments, in particular, regarding the open-ended question on why they thought these arguments are not convincing, as some students mentioned this in their responses to the open-ended questions. This could have resulted in participants perceiving the questions as too easy and getting bored, thus a reduced test-taking motivation, which can lead to lower performance (e.g., Asseburg & Frey, 2013). To avoid this, an alternative approach for investigating the influence of the type of argument could have been taken. For instance, instead of providing participants with the same type of argument, each participant could have been given different types of arguments (as was shortly discussed in Section 5.1.1). However, such a design also comes with its downsides. First, participants' responses, particularly regarding conviction, could be influenced by the possibility of comparing the different types of arguments, which I did not aim for. Second, the number of items in such a questionnaire should be larger for such an approach, because more than one mathematical statement for each kind of argument would be necessary to be able to

draw conclusions about the influence of the type of argument. Because otherwise, the differences might not result primarily from the type of argument but the specific statement. Such an approach could be used best in a laboratory experiment, where the conditions are more controllable (e.g., Döring & Bortz, 2016). In addition, laboratory experiments with financial rewards could increase test-taking motivation, even though respective research is not consistent (e.g., Baumert & Demmrich, 2001; Braun, Kirsch, & Yamamoto, 2011; O'Neil, Sugrue, & Baker, 1995), resulting in better answer quality (e.g., Wise & DeMars, 2005) and the possibility to increase testing time. Alternatively, one could also design different versions of questionnaires, such that all combinations of statements and types of arguments are considered. This might, however, increase the needed sample size, because the specific combination of type of argument and (type of) statement could also influence students' responses. Given the framework of this study, the chosen experimental design seemed to be the best approach regarding the research questions of this study, even considering the described limitations.

7.2.3 Conceptualization and Operationalization

Only few prior studies have investigated students' understanding of the generality, and to my knowledge, no studies have explicitly analyzed understanding of generality of statements and the specific relation to proof construction and reading. Further, prior studies on students' understanding of generality have mainly reported on students or teachers who were convinced of the correctness of the statement and/or proof but not convinced that no counterexample exists (Chazan, 1993; Knuth, 2002) or regarding students' awareness that one counterexample disproves a universal statement (Buchbinder & Zaslavsky, 2019; Galbraith, 1981). I decided to consider participants' responses regarding the estimation of truth and relate them to the responses regarding the existence of counterexamples. An incorrect understanding of generality was then defined as inconsistent responses. This conceptualization might have limited the comparability of my results to the few prior studies that have been conducted. Moreover, other reasons for inconsistent responses cannot be ruled out and should also be discussed. For instance, participants responding inconsistently might lack logical reasoning skills, as the question regarding the existence of counterexamples was expressed implicitly via the negation of the statement. However, these students would then nevertheless have an insufficient understanding of the particular statement, specifically regarding the appearance of respective counterexamples. Moreover, the high correlation between students' understanding of generality and their *knowledge* of the meaning of mathematical generality, which

was assessed via a closed item—further indicates that the chosen conceptualization and operationalization of students' understanding of the generality of statements indeed provided valid results. Therefore, the chosen approach does not only provide new results of students' understanding of the generality of statements and the relation to proofs, but also builds a new basis for future research on students' understanding of generality.

As has been highlighted by other researchers (e.g., Sommerhoff, 2017), studies on students' proof skills have generally not been using exactly the same definitions, conceptualizations, and operationalizations. Even though I have tried to identify and implement the essential commonalities, decisions were made regarding conceptualization and operationalization, which may limit the generalizability and comparability of my results. For instance, in contrast to other studies, in which proof comprehension was measured via assessment tests, I decided to rely on students' self-reports on their comprehension of the arguments. Assessing participants' proof comprehension via tests for five statements was assumed to unreasonably increase the test duration, risking reduced test-taking motivation and mental fatigue effects (e.g., Ackerman & Kanfer, 2009; Möckel, Beste, & Wascher, 2015; van der Linden, Frese, & Meijman, 2003). Further, proof comprehension was measured via a three-level scale—completely, partially, or not at all understood. More nuanced measures might have provided further insights into students' proof comprehension and its relation to understanding generality. To ensure comparability to some extent, the coding scheme used to analyze participants' responses regarding aspects they claimed to have not understood was based on the assessment model developed by Mejía Ramos et al. (2012). However, relying on students' self-reports nevertheless limits generalizability and comparability of the research findings, because self-reports are subject to several biases. For instance, students' might not be able to assess themselves accurately (see, e.g., A. Selden & Selden, 2003), which implies that self-reports on students' proof comprehension measure what participants *think* they have (not) understood or how well they *believe* they have understood an argument and not their actual comprehension of the specific proof, which limits the validity of self-reported data on students' comprehension.

To analyze students' conviction by the argument, participants were first asked if the presented justification has convinced them of the truth of the statement and, if they did not claim to be completely convinced by the argument, why the justification did not convince them. I relied again on self-reports, which implies limitations already discussed above. In particular, the combined findings on students' estimation of truth and conviction suggest that asking participants questions about how convinced they are by an argument might reveal what arguments participants assume they *should* find convincing—not necessarily what types of arguments actually

convince them the most of the truth of the statements. As conviction and acceptance criteria for proof are subjective and influenced by prior experience, for instance, in the classroom (e.g., Hanna, 1989; Stylianides, 2007), they may be susceptible to biases such as social desirability. Students' might *know* that empirical arguments do not constitute a proof (e.g., Ufer et al., 2009) and therefore assume that they *should* not find these convincing. However, empirical arguments *can* in fact be convincing and lead to high levels of conviction in the truth of a statement (e.g., Weber, 2013), which my findings experimentally confirm. Even though participants were not asked if they thought the arguments are valid proofs, they might have nevertheless—consciously or not—taken this into account when asked if the argument has convinced them of the truth of the statement. In this regard, as has been discussed by other researchers, for instance Inglis and Mejía-Ramos (2013), participants might interpret questions about how convinced or persuaded they are by an argument differently. To reduce this risk, I decided to explicitly ask participants if the argument convinces them *of the truth of the statement*. While this has hopefully led to better comparability of responses of participants in *this* study, the comparability of my results to those of other studies might be limited. Furthermore, participants were not provided with pre-defined criteria for convincing arguments, as suggested by Mejía Ramos and Inglis (2009b), for instance, and it is not fully clear what criteria participants based their decision on. However, this limitation was at least partially overcome by asking students to explain why they were not (completely) convinced by the arguments. The respective findings of the content analysis further strengthen the assumption that participants may have considered acceptance criteria for proof when asked why the arguments did not convince them of the truth of the statements.

To investigate students' proof schemes, participants were not explicitly asked to *(dis)prove* the statements, but to justify why they think the statement is correct or false. While some prior studies chose a similar approach (e.g., Barkai et al., 2002; Harel & Sowder, 1998), others explicitly asked the participants to construct a proof (e.g., Recio & Godino, 2001; Stylianou et al., 2006). Thus, these different approaches might limit the comparability to some of the prior studies on proof schemes.

7.2.4 Number, Selection, and Order of Statements

Significant effects of the type of statement were found across all analyzed activities and students' understanding of generality. A larger number of statements would still have been desirable to increase validity and reliability of the findings. But due to

testing time, the number of statements included in this study had to be limited to avoid fatigue effects (as has been mentioned above). Moreover, the defined criteria for the selection of statements also limited the number of suitable items, in particular false statements. To further increase the validity of findings, it would have been beneficial to include additional statements, in particular false ones.

The selection and allocation was mainly based on theoretical considerations, such as the extent to which the statements are present in text books and school curricula (see Section 5.3.1). However, because the two *unfamiliar* statements only require basic knowledge and should be known by teachers, it is possible that teachers teach and discuss these statements with their students, even though the statements are not part of the school curriculum. The content analysis of the data on students' proof schemes provides evidence that the allocation made in this study is generally conclusive. Regarding the (true) unfamiliar statements, only few participants used authority arguments or claimed that the statement is a general rule (4 and 1%, respectively). Instead, they mainly used empirical arguments. Thus, it can be assumed that the vast majority did not seem to have gained (much) experience with the *unfamiliar* statements during high school. In contrast, most participants used these types of arguments (authority and rule) to justify the truth of the *familiar* statements (24 and 37%), which indicates at least some degree of familiarity with these statements. This interpretation might be limited by the fact that the statements do not only differ by familiarity but also content-wise: The familiar statements were taken from geometry and the unfamiliar statements from elementary number theory. This choice was based on the respective criteria defined in Section 5.3.1. (NRW) School curricula only mention geometry statements explicitly with respect to proving; and statements from elementary number theory are assumed to require comparatively few knowledge to understand and prove them, which was one of the main criteria defined for the selection of unfamiliar statements. Few studies have explicitly reported on the effect of the content domain on students' performance in proof-related activities (e.g., Ko & Knuth, 2013). The fact that the *familiarity*-or in other words, geometry—unexpectedly had a negative effect on participants' (self-reported) proof comprehension suggests that the content area indeed plays a role. For future studies, it would therefore be desirable to consider familiar and unfamiliar statements from both content domains to specifically identify what characteristics—content domain and/or familiarity—contribute to the observed effects, even though the defined criteria for the selection of statements would make a respective implementation no simple task (at least in Germany).

Further, to ensure comparability, all statements were mainly expressed in natural language. As a consequence, participants had difficulties understanding and recognizing the pythagorean theorem, as mentioned several times in this thesis. In

particular, a relatively high percentage claimed to not know the truth value of the statement and if counterexamples exist, which resulted in *missing values* in the variable understanding of generality. This could have biased the findings regarding the effect of familiarity (because this statement was assumed to be known to the participants), as participants for whom a value for understanding of generality was measured might not only have had better content knowledge, but generally a better understanding of proof and generality.

The order of items can also influence participants' responding behavior and performance, even though research on this is ambiguous (e.g., Anaya et al., 2022; Bresnock, Graves, & White, 1989; Kleinke, 1980; Newman, Kundert, Jr, & Bull, 1988; Şad, 2020). I have decided to order the statements from easiest to most difficult, based on pre-tests and expert opinions (see Section 5.3.1). The percentage of participants who claimed to not know the truth value of the statement and if counterexamples exist increased over the course of the experiment and was the highest for the last statement, the pythagorean theorem. Other reasons for the high percentage of *missing values* regarding this statement have already been discussed. But due to the fixed order of statements, it cannot be ruled out that the position of the statement in the questionnaire also affected participants' (non-)responses. Therefore, a randomization of the statements might have been the better choice, even though this could have resulted in a higher percentage of participants dropping out of the experiment early on, if at random the first statement would have been the most difficult one (Anaya et al., 2022). Randomization might nevertheless have been preferable, because potential order effects could then have been analyzed and verified.

7.2.5 Open-Ended Questions and Content Analysis

In general, the collection and analysis of responses to open-ended questions have several limitations. One already mentioned is related to the sample size and potential selection bias, because some participants might perceive open-ended questions as too time-consuming or they lack interest in the topic and decide to not answer them (e.g., Holland & Christian, 2009; A. L. Miller & Lambert, 2014). Another general limitation concerns the texts that are being analyzed. In the present study, participants were asked open-ended questions regarding their proof schemes, proof comprehension, and conviction. Participants might not be able to fully express their thinking, identify all aspects they did not comprehend or find convincing, resulting in incomplete responses. However, some studies have shown that most people are

generally capable of articulating themselves in their answers to open-ended questions (e.g., Geer, 1988), but (more recent) research on this seems to be scarce.

It should be noted that content analyses almost always involve interpretation to some extend (e.g., Bryman, 2012). The coding of complete and incomplete arguments was specifically difficult, because mathematically it is not clear where to draw the line (unless formal proofs would have been considered, which was not done for good reasons), even mathematicians do not always agree, and it was not always clear if participants did not include specific steps in their arguments because they assumed them to be obvious or because they did not think about them. To overcome these limitations and to ensure reliability, the coding schemes were based on previous frameworks, of which some have been used extensively (e.g., Harel & Sowder, 1998), and I provided detailed coding protocols as well as tried to be as transparent as possible (see Section 5.4.2 and paragraphs on content analysis in Sections 5.4.3, 5.4.4, and 5.4.5 as well as Appendix B in the Electronic Supplementary Material). This resulted in very high inter-coder reliabilities after coders were sufficiently trained.

Overall, analyzing students' responses to the open-ended questions provided important insights into their understanding of generality and proof.

7.2.6 Control Variables

Prior knowledge was only indirectly considered by including participation in an honors mathematics course (LK) and, to a lesser degree, by the participation in a transition course (Vorkurs). The participation in an honors course proved to be a useful predictor for most of the activities and for students' understanding of generality. However, it is not completely clear what it actually controls for: Mainly content-specific resources such as conceptual and procedural knowledge and domain-specific resources such as mathematical strategic and methodological knowledge, or other (domain-general) resources or even something else. Similarly, the CRT score was considered as a control variable to account for (domain-general) cognitive resources. Overall, the CRT score was the most significant predictor regarding almost all proof-related activities (with the exception of conviction) and students' understanding of generality. Again, these findings do not provide information regarding the influence of more nuanced (domain-general) resources such as problem-solving and (general) reasoning skills. Furthermore, it can be seen as a limitation that there is even a debate about what it is that the CRT measures—cognitive reflection, rational thinking, numeracy, insight problem solving, and/or something

else (e.g., Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Patel et al., 2019; Pennycook et al., 2016; Toplak et al., 2014). However, as a control variable for individual cognitive differences, the CRT score still seems to be a useful and easy to measure control variable. Further research is needed to investigate the relation between CRT score and students' proof skills (see Section 7.4 for a further discussion).

As it was not the purpose of this study to identify specific predictive resources for students' proof skills, the limitation regarding the informative value of the considered control variables is acceptable. Moreover, the findings are nevertheless useful in that they provide evidence for differences on an individual level and suggest influences of resources that have not been considered in previous research in that way. While the assessment and inclusion of content- and domain-specific resources would have been beneficial to contribute to the existing research on the influence of individual resources (e.g., Chinnappan et al., 2012; Sommerhoff, 2017), it would not have been reasonable to consider these in this study, for instance, due to limitations regarding the test length and the focus of this thesis.

7.2.7 Sample

The overall sample size was generally satisfactory. It would nevertheless have been beneficial to have larger samples regarding some of the research questions, in particular regarding those in which only one or two experimental groups were considered, for instance, the analysis of the influence of students' conviction and proof comprehension on understanding of generality. Larger sample sizes would have increased statistical power and more robust estimates of the coefficients for the respective variables of interest. Moreover, the content analyses would have benefited from larger sample sizes as well, because the sample size was not only limited by the selection of experimental groups, but also by students' responses to prior questions and their willingness and ability to answer the open-ended questions (see also discussion further below). Further, the sample was unbalanced regarding the study program, with a large number of preservice primary school teachers and a much smaller number of preservice secondary school teachers, for example. Even though the effect of study program was not directly analyzed, individual resources (for instance, the CRT score and participation in honors mathematics courses) and most likely students' proof skills differ with respect to the study program, which could potentially have biased research findings. Therefore, a more balanced sample would have been desirable, even though the distribution of study program roughly corresponded to the actual

distribution at Bielefeld university. Using more advanced statistical tools such as generalized linear mixed models contributed to overcoming these limitations by including respective control variables such as the CRT score and the attendance of an honors course.

7.3 Implications for the Learning and Teaching of Proof at the Transition from School to University

The results of my thesis contribute to the existing research on university students' proof skills and understanding at the transition from school to university. My findings confirm those of prior studies that many students have limited knowledge and understanding of mathematical concepts and proof when they enter university (e.g., Gueudet, 2008; Kempen & Biehler, 2019; Recio & Godino, 2001). In particular, my findings suggest that many students have no sufficient knowledge of the meaning of basic terms and concepts, such as divisibility, even and odd numbers, product, and the meaning of variables. Further—and most important for this study—many students also seem to lack sufficient understanding of the generality of mathematical statements, which seems to be related to their conviction and usage of different types of proofs, in particular, empirical arguments. It is therefore no surprise that students' have difficulties with proof and proving when they enter university, and potentially even lack an intellectual need for proof (see also directions for future research further below). Most (German) universities already strive to close the gap at the transition from school to university by offering transition (to proof) courses (see, for instance, Gerdes, Halverscheid, & Schneider, 2022). However, so far, the effect of these courses is unclear (e.g., Greefrath, Koepf, & Neugebauer, 2017; Tieben, 2019) and dropout rates in mathematics remain high at German universities Heublein et al., 2022. The findings of my study further suggest that the attendance of a transition course had no significant effect on students' performance in the considered proof-related activities and their understanding of generality. While one reason might be that these courses are simply too short—two weeks at Bielefeld university—other reasons might include that the content currently included in these courses does not fully meet students'—or lecturers'—needs. In this regard, the findings of the present thesis may be useful to revise the content of transition courses. It particular, they provide a basis for (intervention) studies that 1) aim at improving students' understanding of the generality of statements, and 2) analyze if and how an improved understanding affects students' intellectual need for proof and their proof skills (see also Section 7.4).

When planning university courses for first-year students, lecturers should take into account that many students currently have an insufficient knowledge of basic mathematical terms and understanding of generality. In general, more emphasis

should be put on sufficiently understanding a theorem first—and definitions, concepts, etc. involved—before students are confronted with its proof. Moreover, reading, constructing, and potentially discussing examples (i.e., empirical arguments) can support students' understanding of the statements, which may ease comprehension and construction of proofs. Generic proofs may also be useful in this regard, as my findings suggest.

These results are particularly important in lectures for preservice teachers to break the cycle of teachers not having sufficient knowledge, therefore school students not learning sufficiently about proof and argumentation, which consequently leads to a gap in students' knowledge at the transition from school to university. The present thesis did not investigate or aim at identifying respective new teaching methods. However, as mentioned, my findings suggest that students lack basic knowledge and understanding and may therefore benefit from activities that particularly aim at assessing and improving their understanding of theorems—including their understanding of mathematical generality. In general, assessing students' knowledge and understanding can help both lecturers and students, in that it would be more transparent to the students what is expected from them and what they need to know to follow along, and lecturers would get a better picture of what their students actually know—and what not. In this respect, (real-time) quizzes can be an effective way of assessing students' knowledge and understanding (e.g., Cohn & Fraser, 2016; Méndez Coca & Slisko, 2013; Plump & LaRosa, 2017). Questions could not only assess students' (prior) knowledge of terms and statements being used in a theorem, but also consist of questions that more specifically aim at their understanding of the generality of (particular) theorems. For instance, after introducing a *new* theorem—which students most likely assume being true—lecturers could ask their students if counterexamples may exist. Thereby, I would avoid using the term *counterexample* and phrase the respective question as suggested in this study (see Section 5.3.5). The responses of the students could provide an opportunity for informative discussions about the theorem itself—including its generality—but also about the purpose and intellectual need for proof. However, the effectiveness of such instructions would need to be investigated.

7.4 Directions for Future Research

Finally, the findings and limitations combined give rise to directions for future research of which some have already been identified in Section 7.2. In this section, I first discuss potential research questions regarding the investigation of students' understanding of generality, before other, partially more general implications for future research are identified.

7.4.1 Further Investigating Students' Understanding of the Generality of Statements

To further generalize the findings on students' understanding of the generality of mathematical statements, replication studies at other (German) universities would be valuable. Moreover, it would be beneficial to include additional or different statements from other content domains to analyze if the findings of this study are content-specific or generalizable to other areas. To analyze the influence of familiarity on students' understanding of generality—but also on other proof skills—it would be valuable to include familiar and unfamiliar statements from the same content domain, as has been highlighted before.

While the focus of the present study was on first-year university students' understanding of generality, a replication of the study with experienced university students would enable to investigate potential developments of students' understanding of generality of statements throughout their studies. I would expect more experienced students to have a more consistent understanding of generality, due to more experience with higher mathematics and proof in particular, but this hypothesis needs to be tested.

The findings of the present thesis furthermore suggest a relation between students' understanding of generality and other proof skills, such as proof comprehension and evaluation. However, results were ambiguous. Several reasons have been identified, such as sample size, but also relying on students' self-reports on their proof comprehension. Future studies could replicate the study with an even higher number of students and/or consider measuring students' proof comprehension via assessment tests (see also discussion further below), as suggested by Mejía Ramos et al. (2012), for instance.

Moreover, the correlations between students' proof schemes and their understanding of generality found in this study need to be further investigated. For instance, it is not clear by the results, if students' with empirical proof schemes showed an incorrect understanding of generality more often than students' with deductive proof schemes because they were not able to produce a general argument, or if other characteristics of theses students explain the correlation. Future studies should consider this when investigating the relation between students' proof schemes and their understanding of the generality of statements.

Further, while I had chosen to investigate students' proof comprehension, conviction, and proof schemes and their relation to understanding generality, future studies could consider relations to other proof-related concepts or aspects. For instance, it may be worthwhile to analyze potential relations between students' understanding of generality and their ability of logical inferences. Even though research suggests

that logical reasoning skills, in particular conditional reasoning skills only play a minor role regarding students' proof skills (e.g., Sommerhoff, 2017), a relation to students' understanding of generality might nevertheless exist. As mentioned before, understanding logical negation should—at least in theory—be particularly relevant for the understanding of generality of statements as defined in this study. Because questions regarding the existence of counterexamples were expressed via the negation of the respective statement (see Section 5.3.5). However, this hypothesis would need to be investigated. Further, the influence of participants' CRT score on their understanding of generality suggested by my findings implies a potential relation between students' (logical) reasoning skills and rational thinking and their understanding of generality (e.g., Liberali et al., 2012; Primi et al., 2016; Toplak et al., 2014). Further research is needed to analyze 1) what the CRT particularly measures and 2) how this relates to students' understanding of the generality of statements and proof skills. Moreover, researchers who want to use the CRT score as a control instrument in future studies may want to consider alternative CRT items to avoid an overemphasis on numerical abilities and floor effects in non-elite population or younger students, for instance (e.g., Sirota, Dewberry, Juanchich, Valuš, & Marshall, 2021; Young, Powers, Pilgrim, & Shtulman, 2018).

Further, the awareness and correct understanding that no counterexample to universal statements exist might be related to or even increase students' appreciation of proof and their *intellectual need for certainty* (introduced by Harel, 2013). Because it is the mathematical generality that is the defining element of mathematical proof, the reason why a deductive proof is indeed necessary and empirical arguments are not sufficient to rule out the existence of any counterexamples. The findings of the present thesis suggest that participants with empirical proof schemes more often have an incorrect understanding of generality than students with deductive proof schemes, for example, and further, that students' who are convinced by empirical arguments are more likely to have an incorrect understanding of generality than students who are not convinced of these arguments. As some researchers have emphasized the relation between students' usage of and satisfaction with empirical arguments and their lack of intellectual need for proof (e.g., Zaslavsky, Nickerson, Stylianides, Kidron, & Winicki-Landman, 2012), investigating the potential relation between students' understanding of generality and their appreciation and intellectual need for proof could be valuable.

Lastly, the findings of my study give rise to a potential intervention study. The effect of students' *knowledge* of the meaning of generality on students' consistent responses (i.e., their actual understanding of generality) was highly significant. An intervention study that investigates the effect of explicitly teaching the meaning of generality of mathematical statements on students' understanding of generality

could be promising in this regard. However, given that the sole knowledge of the meaning of generality was also not sufficient for consistently having a correct understanding of generality, other factors (such as the familiarity with the statement and logical reasoning skills) also play a role and should be considered.

7.4.2 Self-Reported Data and Reality

Several limitations discussed in Section 7.2 concern the relation between self-reported data regarding students' conviction and proof comprehension and students' *actual* conviction and proof comprehension. The findings of the present study indicate that participants might not always be able to assess themselves accurately. For instance, as has been discussed, the question "Does the justification convince you of the correctness of the claim?" does not necessarily provide information about students' *actual* conviction of the truth of a statement by different types of arguments, but about which types of arguments they think *should* be convincing to them, thus, their conceptions of *convincing mathematical arguments* or *proof*. In general, I find it questionable what studies on students' *conviction* actually assess—most likely not students' actual conviction regarding the truth of a statement, but more likely acceptance of the argument and respective criteria in the sense of *social proof*. The gap between self-reported conviction and actual conviction was only revealed by the experimental design of this study. Thus, further investigating students' conviction or other proof skills experimentally would be very valuable. In particular, one should be careful in interpreting results solely based on students' self-reported data. The observation that self-reports do not always reflect the reality is not new (e.g., Maki & McGuire, 2002; Thiede, Griffin, Wiley, & Redford, 2009), however, only few researchers have explicitly investigated this in in the context of proof and to my knowledge, the extent of this phenomenon has not explicitly been researched yet. Further, existing studies on mathematicians' conviction of arguments have also relied on self-reports. It would be valuable to conduct a similar experimental study, in which the relation between mathematicians' estimation of truth—based on different types of arguments—and their self-reported level of conviction of the truth of statements by the arguments is analyzed. Different statements that are not too simple (i.e., that mathematicians are not familiar with) should be selected for such a study. Another potential future study concerns the relation between students' self-reported proof comprehension and their actual proof comprehension, assessed via comprehension tests (e.g., Mejía Ramos et al., 2012). Several studies on *text* comprehension have provided evidence that many learners fail to accurately judge their text comprehension in that they often overestimate but also underestimate their comprehension

(e.g., Golke et al., 2022; Maki & McGuire, 2002; Prinz, Golke, & Wittwer, 2020; Thiede et al., 2009). Similarly, A. Selden and Selden (2003) have reported that mathematics students indeed overestimate their understanding of a proof (even though the focus of their study was on proof validation). However, to my knowledge, no studies have explicitly investigated differences in students' self-reported and actual proof comprehension. Proof comprehension tests have the advantage of providing more valid, reliable, and nuanced results of students actual proof comprehension. However, their construction as well as the conduction of such tests is time consuming and not always feasible. Self-reports provide a much simpler way of measuring students' proof comprehension, which is why their validity needs to be investigated.

7.4.3 Question Order Effects

As has been discussed, the statements included in the questionnaire have been ordered from most easiest to most difficult. While this is assumed to have benefits such as a lower percentage of participants abandon the questionnaire and better performance (e.g., Anaya et al., 2022; Kleinke, 1980), it might also lead to potential order effect, such as less motivation to answer more difficult questions at the end of the questionnaire. Therefore, future studies on students' proof skills may want to consider random ordering of statements. Moreover, it would be beneficial to investigate respective order effects experimentally, because research on this is still ambiguous and studies in the context of mathematic education, in particular proof, seem to be scarce. In such a study, several questionnaires with different order of statements could be designed, for instance, from easiest to hardest, from hardest to easiest, and/or random.

Conclusions

8

It is well-known in mathematics education that the transition from school to university is very challenging for many mathematics students and preservice mathematics teachers. One of the main reasons identified in the literature is that students face difficulties with proof-based mathematics to which they are commonly introduced when they start university (e.g., Gueudet, 2008; A. Selden, 2012). Over the past 30 years, research on first-year students' proof skills has increased significantly. However, several proof-related activities are still under-researched and the specific relations between these activities are not fully understood yet. In particular, no prior research on students' understanding of the generality of mathematical statements—which can be seen as an essential part of the comprehension of statements and students' understanding of proof—and the specific relations to proof reading and construction had been conducted. Therefore, the main goal of the present study was to close this gap. Since no definition of understanding the generality of statements could be identified in the literature, I provided a clear definition for understanding the generality of statements myself by relating students' estimation of truth to that of the existence of counterexamples. A correct understanding was then defined as consistent responses.

Further, to highlight the relevance of the statement itself for proof-related activities, I suggested an adapted version of the framework on proof-related activities by Mejía Ramos and Inglis (2009b), which distinguishes activities related to the *reading of the statement*, for which a proof needs to be constructed or a proof has to be read, and activities related to the respective *reading and construction of arguments*. The research questions were then mainly guided by this framework.

Since previous studies have shown differences in students' understanding and evaluation of *different types of arguments* (e.g., Healy & Hoyles, 2000; Kempen, 2018, 2021; Tabach, Barkai, et al., 2010), I analyzed the influence of reading different types of arguments (no argument, empirical argument, generic proof, and ordinary

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_8

proof) on students' understanding of generality and other proof-related activities. Additionally, I considered the familiarity with the statement and its truth value as important characteristics that might also influence students' performance in proof-related activities, as suggested in the literature (e.g., Barkai et al., 2002; Dubinsky & Yiparaki, 2000; Hanna, 1989; Stylianides, 2007; Weber & Czocher, 2019).

Through the experimental design of my study, I provided detailed results on the influence of the type of argument and statement on students' understanding of the generality of statements, proof reading and construction, and on relations between these activities. The data was thereby analyzed using mainly generalized linear mixed models. My results extend prior research in that

- in a comparatively large percentage of observations (about one third), students lacked understanding of the generality of mathematical statements,
- students with a correct *knowledge* of mathematical generality are more likely to have a correct understanding of the generality of statements,
- students' usage and conviction of empirical arguments is negatively related to their understanding of the generality of statements,
- students' level of conviction of the truth of statements is significantly related to the reading and construction of different types of arguments; in particular, empirical arguments (and to a lesser degree generic proofs) support students in successfully estimating the truth value of (true) universal statements—but ordinary proofs do not,
- the familiarity with the statement and the truth value influence students' understanding of the generality of statements and performance in proof-related activities.

Further, my results (experimentally) confirm prior research findings that

- first-year university students lack basic mathematical knowledge and therefore have difficulties with the comprehension of statements (e.g., Dubinsky & Yiparaki, 2000; Ferrari, 2002) and proofs (e.g., Conradie & Frith, 2000; Dubinsky & Yiparaki, 2000; Moore, 1994; Reiss & Heinze, 2000) as well as with proof evaluation (e.g., Harel & Sowder, 1998; Healy & Hoyles, 2000; Kempen, 2019; Recio & Godino, 2001; Weber, 2010),
- most students find generic and, in particular, ordinary proofs convincing (e.g., Kempen, 2018, 2021; Ko & Knuth, 2013; Weber, 2010),
- about half of the students used empirical arguments (Barkai et al., 2002; Bell, 1976; Healy & Hoyles, 2000; Lee, 2016; Recio & Godino, 2001) to justify *unfamiliar* universal statements, while most of them used external arguments

(i.e., based on authorities or a rule) (Harel & Sowder, 1998; Sen & Guler, 2015; Sevimli, 2018; Stylianou et al., 2006) to justify *familiar* universal statements.

My findings can be used to develop future university courses in a manner that eases and promotes the transition to proof-based mathematics. Particular attention should be put on students' comprehension of statements, including understanding their generality, before they are confronted with proof reading or construction. Empirical arguments and potentially generic proofs can support students' understanding of statements and their success in estimating the truth value, but their limitations and the necessity of proof should be made clear.

Lastly, based on my findings, I provided several suggestions for future research on students' proof skills and students' understanding of the generality of statements. In particular, further research on students' self-reports regarding proof-related activities and their actual understanding and proof skills would be very valuable. My study further highlights the benefits of and need for more experimental studies in mathematics education and in particular in research on proof and argumentation.

References

Aberdein, A. (2009). Mathematics and Argumentation. *Foundations of Science*, 14, 1–8. https://doi.org/10.1007/s10699-008-9158-3

Ackerman, P. L., und Kanfer, R. 2009. Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied15*2163–181. https://doi.org/10.1037/a0015719

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health, 86*(5), 726–728. Retrieved 2023-02-02, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380484/

Alcock, L., Hodds, M., Roy, S., & Inglis, M. (2015). Investigating and improving undergraduate proof comprehension. *Notices of the American Mathematical Society, 62*(7), 742–752. https://doi.org/10.1090/noti1263

Alcock, L., & Inglis, M. (2008). Doctoral students' use of examples in evaluating and proving conjectures. *Educational Studies in Mathematics, 69*(2), 111–129. https://doi.org/10.1007/s10649-008-9149-x

Anaya, L., Iriberri, N., Rey-Biel, P., & Zamarro, G. (2022). Understanding performance in test taking: The role of question difficulty order. *Economics of Education Review, 90*, 102293. https://doi.org/10.1016/j.econedurev.2022.102293

Anglin, W. S. (1994). *Mathematics: A concise history and philosophy* (S. Axler, F. W. Gehring, & K. A. Ribet, Eds.). Springer. https://doi.org/10.1007/978-1-4612-0875-4

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92–104. Retrieved from https://psycnet.apa.org/record/2013-18917-006

Balacheff, N. (1988a). Aspects of proof in pupils' practice of school mathematics. In D. Pimm (Ed.), *Mathematics, teachers and children.* Hodder and Stoughton. Retrieved 2022-09-05, from https://www.bibsonomy.org/bibtex/115705530b6fdb2bb86964db5204840e1

Balacheff, N. (1988b). A study of students' proving processes at the junior high-school level. In *Second UCSMP international conference on mathematics education.* NCTM. Retrieved from https://hal.science/hal-01652045

Balacheff, N. (1999). *Is argumentation an obstacle? Invitation to a debate.* Retrieved 2022-02-24, from https://www.researchgate.net/publication/234597638_Is_Argumentation_an_Obstacle_Invitation_to_a_Debate

M. Damrau, *Understanding the Generality of Mathematical Statements*,
Bielefelder Schriften zur Didaktik der Mathematik 15,
https://doi.org/10.1007/978-3-658-43763-3_1

Balacheff, N. (2002). The researcher epistemology: A deadlock for educational research on proof. *Proceedings of the 2002 international conference on mathematics education: Understanding proving and proving to understand,* 23–44.

Balacheff, N. (2010). Bridging knowing and proving in mathematics: A didactical perspective. In G. Hanna, H. N. Jahnke, & H. Pulte (Eds.), *Explanation and Proof in Mathematics: Philosophical and Educational Perspectives* (pp. 115–135). Springer. https://doi.org/10.1007/978-1-4419-0576-5_9

Barkai, R., Tsamir, P., Tirosh, D., & Dreyfus, T. (2002). Proving or refuting arithmetic claims: The case of elementary school teachers. In A. Cockburn & E. Nardi (Eds.), *Proceedings of the 26th Meeting of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 57–64). PME.

Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265–284. https://doi.org/10.1016/j.jarmac.2014.09.003

Bass, H. (2009). *How do you know that you know? Making believe in mathematics.* Retrieved 2022-02-14, from https://deepblue.lib.umich.edu/bitstream/handle/2027.42/64280/Bass-2009.pdf?sequence=1&isAllowed=y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462. https://doi.org/10.1007/BF03173192

Bell, A. W. (1976). A study of pupils' proof-explanations in mathematical situations. *Educational Studies in Mathematics,* 7(1/2), 23–40. Retrieved 2022-04-25, from https://www.jstor.org/stable/3481809

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289–300. Retrieved 2022-12-08, from https://www.jstor.org/stable/2346101

Bieda, K. N. (2010). Enacting proof-related tasks in middle school mathematics: Challenges and opportunities. *Journal for Research in Mathematics Education, 41*(4), 351–382. Retrieved 2022-04-22, from https://www.jstor.org/stable/41103880

Bieda, K. N., & Lepak, J. (2014). Are you convinced? Middle-grade students' evaluations of mathematical arguments. *School Science and Mathematics, 114*(4), 166–177. https://doi.org/10.1111/ssm.12066

Biehler, R., & Kempen, L. (2016). Didaktisch orientierte Beweiskonzepte – Eine Analyse zur mathematikdidaktischen Ideenentwicklung. *Journal für Mathematik-Didaktik, 1*(37), 141–179. https://doi.org/10.1007/s13138-016-0097-1

Blum, W., & Kirsch, A. (1989). Warum haben nicht-triviale Lösungen von f'= f keine Nullstellen. In H. Kautschitsch & W. Metzler (Eds.), *Anschauliches Beweisen* (pp. 199–209). B.G. Teubner.

Boero, P., Garuti, R., & Mariotti, M. A. (1996). Some dynamic mental processes underlying producing and proving conjectures. In L. Puig (Ed.), *Proceedings of the 20th Conference*

of the International Group for the Psychology of Mathematics Education (Vol. 2, pp. 121–128).

Bolker, B. M. (2015). Linear and generalized linear mixed models. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Eds.), *Ecological Statistics: Contemporary theory and application* (pp. 309–333). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199672547.003.0014

Branford, B. (1908). *A study of mathematical education, including the teaching of arithmetic.* Clarendon Press.

Branford, B. (1913). *Betrachtungen über mathematische Erziehung vom Kindergarten bis zur Universität* (R. Schimmack & H. Weinreich, Trans.). B.G. Teubner.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-Grade NAEP reading assessment. *Teachers College Record, 113*(11), 2309–2344. https://doi.org/10.1177/016146811111301101

Bresnock, A. E., Graves, P. E., & White, N. (1989). Multiple-choice testing: Question and response position. *The Journal of Economic Education, 20*(3), 239–245. https://doi.org/10.2307/1182299

Brunner, E. (2014). *Mathematisches Argumentieren, Begründen und Beweisen.* Springer.

Bryman, A. (2012). *Social research methods* (Fourth ed.). Oxford University Press.

Buchbinder, O., & Zaslavsky, O. (2007). How to decide? Student's ways of determining the validity of mathematical statements. In D. Pitta-Pantazi&G. Philippou (Eds.), *Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education* (pp. 561–570).

Buchbinder, O., & Zaslavsky, O. (2019). Strengths and inconsistencies in students' understanding of the roles of examples in proving. *The Journal of Mathematical Behavior, 53,* 129–147. https://doi.org/10.1016/j.jmathb.2018.06.010

Bürger, H. (1979). Beweisen im Mathematikunterricht: Möglichkeiten der Gestaltung in der Sekundarstufe I und II. In W. Dörfler & R. Fischer (Eds.), *Beweisen im Mathematikunterricht* (pp. 103–134). Hölder-Pichler-Tempsky.

Buss, S. R. (1998). *Handbook of proof theory.* Elsevier.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

CadwalladerOlsker, T. (2011). What do we mean by mathematical proof? *Journal of Humanistic Mathematics, 1*(1), 33–60. https://doi.org/10.5642/jhummath.201101.04

Chazan, D. (1993). High school geometry students' justification for their views of empirical evidence and mathematical proof. *Educational Studies in Mathematics, 24*(4), 359–387. https://doi.org/10.1007/BF01273371

Chemla, K., Chorlay, R., & Rabouin, D. (2016). Prologue: Generality as a component of an epistemological culture. In *The Oxford handbook of generality in mathematics and the sciences.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198777267.013.1

Chinnappan, M., Ekanayake, M. B., & Brown, C. (2012). Knowledge use in the construction of geometry proof by Sri Lankan students. *International Journal of Science and Mathematics Education, 10*(4), 865–887. https://doi.org/10.1007/s10763-011-9298-8

Christensen, R. H. B. (2019). *Ordinal—Regression models for ordinal data.* Retrieved from https://CRAN.R-project.org/package=ordinal

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second ed.). Lawrence Erlbaum Associates.

Cohn, S. T., & Fraser, B. J. (2016). Effectiveness of student response systems in terms of learning environment, attitudes and achievement. *Learning Environments Research, 19*(2), 153–167. https://doi.org/10.1007/s10984-015-9195-0

Conner, K. A. (2022). A multi-faceted framework for identifying students' understanding of the generality requirement of proof. *International Electronic Journal of Mathematics Education, 17*(4), em0702. https://doi.org/10.29333/iejme/12270

Conradie, J., & Frith, J. (2000). Comprehension tests in mathematics. *Educational Studies in Mathematics, 42*(3), 225–235. https://doi.org/10.1023/A:1017502919000

Damrau, M. (2023). First-year university students' understanding of the generality of mathematical statements and its relation to proof reading and construction. *Open Science Framework.* https://doi.org/10.17605/OSF.IO/JS2F8

Davey, J. W., Gugiu, P. C., & Coryn, C. L. S. (2010). Quantitative methods for estimating the reliability of qualitative data. *Journal of MultiDisciplinary Evaluation, 6*(13), 140–162. Retrieved from https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/266/254

Davis, P. J., Hersh, R., & Marchisotto, E. A. (2012). *The mathematical experience* (Study ed.). Birkhsäuser. https://doi.org/10.1007/978-0-8176-8295-8

Dawkins, P. C., & Weber, K. (2017). Values and norms of proof for mathematicians and students. *Educational Studies in Mathematics, 95*(2), 123–142. https://doi.org/10.1007/s10649-016-9740-5

Dawson, J. W. (2006). Why do mathematicians re-prove theorems? *Philosophia Mathematica, 14*(3), 269–286. https://doi.org/10.1093/philmat/nkl009

Dawson, J. W. (2015). *Why prove it again? Alternative proofs in mathematical practice.* Springer. https://doi.org/10.1007/978-3-319-17368-9

de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras, 24,* 17–24. Retrieved from https://www.researchgate.net/publication/264784642_The_Role_and_Function_of_Proof_in_Mathematics

de Villiers, M. (2010). Experimentation and proof in mathematics. In G. Hanna, H. N. Jahnke, & H. Pulte (Eds.), *Explanation and proof in mathematics: Philosophical and educational perspectives* (pp. 205–221). Springer. https://doi.org/10.1007/978-1-4419-0576-5_14

Department of Basic Education. (2011). *Curriculum and Assessment Policy Statement. Grades 10-12. Mathematics.* Author.

Dieter, M. (2012). *Studienabbruch und Studienfachwechsel in der Mathematik. Quantitative Bezifferung und empirische Untersuchung von Bedingungsfaktoren* (Doctoral dissertation, Universität Duisburg-Essen). Retrieved 2023-02-01, from https://d-nb.info/1193651182/34

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften.* Springer. https://doi.org/10.1007/978-3-642-41089-5

Douek, N. (2007). Some remarks about argumentation and proof. In P. Boero (Ed.), *Theorems in school: From history, epistemology and cognition to classroom practice* (Vol. 2, pp. 163–181). Brill. Retrieved from https://doi.org/10.1163/9789087901691_010

Dreher, A., & Heinze, A. (2018). Mathematicians' criteria for accepting theorems and proofs—An international study. In E. Bergqvist, M. Österholm, C. Granberg, & L. Sumpter (Eds.), *Proceedings of the 42nd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 363–370). PME.

Dreyfus, T. (1999). Why Johnny can't prove. *Educational Studies in Mathematics, 38*(1), 85–109. https://doi.org/10.1023/A:1003660018579

Dreyfus, T., Nardi, E., & Leikin, R. (2012). Forms of proof and proving in the classroom. In G. Hanna & M. de Villiers (Eds.), *Proof and proving in mathematics education. The 19th ICMI study* (pp. 191–214). Springer. https://doi.org/10.1007/978-94-007-2129-6_8

Dubinsky, E., & Yiparaki, O. (2000). On student understanding of AE and EA quantification. In E. Dubinsky, A. H. Schoenfeld, & J. Kaput (Eds.), *Research in Collegiate Mathematics Education. IV* (Vol. 8, pp. 239–289). American Mathematical Society.

Ducrot, O. (1980). *Les Échelles argumentatives.* Les Édition de Minuit.

Duval, R. (1990). Pour une approche cognitive de l'argumentation. *Annales de didactique et de sciences cognitives, 3,* 195–221. Retrieved 2022-09-06, from https://publimath.univ-irem.fr/biblio/IST90008.htm

Duval, R. (1991). Structure du raisonnement deductif et apprentissage de la demonstration. *Educational Studies in Mathematics, 22*(3), 233–261. https://doi.org/10.1007/BF00368340

Duval, R. (1999). *Questioning argumentation.* Retrieved 2022-03-03, from http://www.lettredelapreuve.org/OldPreuve/Newsletter/991112Theme/991112ThemeUK.html

Ellis, A. B., Bieda, K. N., & Knuth, E. J. (2012). *Developing essential understanding of proof and proving for teaching mathematics in grades 9-12.* The National Council of Teachers of Mathematics.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research, 39*(5), 1115–1131. https://doi.org/10.1086/667782

Ferrari, P. L. (2002). Understanding elementary number theory at the undergraduate level: A semiotic approach. In S. R. Campbell & R. Zazkis (Eds.), *Learning and teaching number theory: Research in cognition and instruction* (pp. 97–115). Greenwood Publishing Group, Inc.

Fischbein, E. (1982). Intuition and proof. *For the Learning of Mathematics, 3*(2), 9–24. Retrieved 2022-07-20, from https://www.jstor.org/stable/40248127

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frosch, C. A., & Simms, V. (2015). Understanding the role of reasoning ability in mathematical achievement. In G. Airenti, B. G. Bara, & G. Sandini (Eds.), *EuroAsianPacific Joint Conference on Cognitive Science* (pp. 633–638).

Fuller, E., Weber, K., Mejia-Ramos, J. P., Rhoads, K., & Samkoff, A. (2014). Comprehending structured proofs. *Jornal Internacional de Estudos em Educação Matemática, 7*(1). Retrieved 2022-08-11, from https://jieem.pgsskroton.com.br/article/view/84

Furinghetti, F., & Morselli, F. (2009). Every unsuccessful problem solver is unsuccessful in his or her own way: Affective and cognitive factors in proving. *Educational Studies in Mathematics, 70*(1), 71–90. https://doi.org/10.1007/s10649-008-9134-4

Galbraith, P. L. (1981). Aspects of proving: A clinical investigation of process. *Educational Studies in Mathematics, 12,* 1–28. https://doi.org/10.1007/BF00386043

Geer, J. G. (1988). What do open-ended questions measure? *Public Opinion Quarterly, 52*(3), 365-367. https://doi.org/10.1086/269113

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Gerdes, A., Halverscheid, S., & Schneider, S. (2022). Teilnahme an mathematischen Vorkursen und langfristiger Studienerfolg. Eine empirische Untersuchung. *Journal für Mathematik-Didaktik, 43*(2), 377–403. https://doi.org/10.1007/s13138-021-00194-3

Gholamazad, S., Liljedahl, P., & Zazkis, R. (2004). What counts as proof? Investigation of pre-service elementary teachers' evaluation of presented 'proofs'. In D. E. Mc-Dougall & J. A. Ross (Eds.), *Proceedings of the twenty-sixth annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 640–647). OISE/UT.

Giaquinto, M. (2005). Mathematical activity. In P. Mancosu, K. F. Jørgensen, & S. A. Pedersen (Eds.), *Visualization, Explanation and Reasoning Styles in Mathematics* (pp. 75–87). Springer. https://doi.org/10.1007/1-4020-3335-4_5

Golke, S., Steininger, T., & Wittwer, J. (2022). What makes learners overestimate their text comprehension? The impact of learner characteristics on judgment bias. *Educational Psychology Review, 34*(4), 2405–2450. https://doi.org/10.1007/s10648-022-09687-0

Gosling, S., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *The American psychologist, 59,* 93–104. https://doi.org/10.1037/0003-066X.59.2.93

Greefrath, G., Koepf,W., & Neugebauer, C. (2017). Is there a link between preparatory course attendance and academic success? A case study of degree programmes in electrical engineering and computer science. *International Journal of Research in Undergraduate Mathematics Education, 3*(1), 143–167. https://doi.org/10.1007/s40753-016-0047-9

Griffiths, P. A. (2000). Mathematics at the turn of the millennium. *The American Mathematical Monthly, 107*(1), 1–14. https://doi.org/10.2307/2589372

Gueudet, G. (2008). Investigating the secondary-tertiary transition. *Educational Studies in Mathematics, 67*(3), 237–254. https://doi.org/10.1007/s10649-007-9100-6

Hales, T. (2005). A proof of the Kepler conjecture. *Annals of Mathematics, 162*(3), 1065–1185. https://doi.org/10.4007/annals.2005.162.1065

Hales, T. (2008). Formal proof. *Notices of the American Mathematical Society, 55*(11), 1370–1380.

Hales, T., Adams, M., Bauer, G., Dang, T. D., Harrison, J., Hoang, L. T., . . . Zumkeller, R. (2017). A formal proof of the Kepler conjecture. *Forum of Mathematics, Pi, 5.* https://doi.org/10.1017/fmp.2017.1

Hanna, G. (1989). More than formal proof. *For the Learning of Mathematics, 9*(1), 20–23. Retrieved from https://www.jstor.org/stable/40247941

Hanna, G. (1990). Some pedagogical aspects of proof. *Interchange, 21*(1), 6–13. https://doi.org/10.1007/BF01809605

Hanna, G. (2000). Proof, explanation and exploration: An overview. *Educational Studies in Mathematics, 44*(1/2), 5–23. https://doi.org/10.1023/A:1012737223465

Hanna, G., & Barbeau, E. (2002). What is proof? In *History of modern science and mathematics* (Vol. 1, pp. 36–48). Charles Scribner's Sons.

Hanna, G., & Jahnke, H. N. (1993). Proof and application. *Educational Studies in Mathematics, 24*(4), 421–438. https://doi.org/10.1007/BF01273374

Hanna, G., & Jahnke, H. N. (1996). Proof and proving. In A. J. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 877–908). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-009-1465-0_24

Hanna, G., & Knipping, C. (2020). Proof in mathematics education, 1980-2020: An overview. *Journal of Educational Research in Mathematics.* https://doi.org/10.29275/jerm.2020.08.sp.1.1

Harel, G. (1999). Students' understanding of proofs: A historical analysis and implications for the teaching of geometry and linear algebra. *Linear Algebra and its Applications, 302-303,* 601–613. https://doi.org/10.1016/S0024-3795(99)00139-1

Harel, G. (2013). Intellectual need. In K. R. Leatham (Ed.), *Vital directions for mathematics education research* (pp. 119–151). Springer. https://doi.org/10.1007/978-1-4614-6977-3_6

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research in Collegiate Mathematics Education III* (Vol. 7, pp. 234–283).

Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (Vol. 2, pp. 805–842). Information Age Publishing.

Harrell, F. E. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis.* Springer. https://doi.org/10.1007/978-3-319-19425-7

Harrison, J. (2008). Formal proof—theory and practice. *Notices of the American Mathematical Society, 55*(11), 1395–1406. Retrieved from https://www.ams.org/notices/200811/tx081101395p.pdf

Healy, L., & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education, 31*(4), 396–428. https://doi.org/10.2307/749651

Hefendehl-Hebeker, L., & Hußmann, S. (2003). Beweisen—Argumentieren. In T. Leuders (Ed.), *Mathematik-Didaktik* (pp. 93–106). Cornelsen.

Heintz, B. (2000). *Die Innenwelt der Mathematik. Zur Kultur und Praxis einer beweisenden Disziplin.* Springer. Retrieved 2023-02-01, from https://link.springer.com/book/9783211829615

Heinze, A. (2010). Mathematicians' individual criteria for accepting theorems and proofs: An empirical approach. In G. Hanna, H. N. Jahnke, & H. Pulte (Eds.), *Explanation and proof in mathematics: Philosophical and educational perspectives* (pp. 101–111). Springer. https://doi.org/10.1007/978-1-4419-0576-5_8

Heinze, A., Anderson, I., & Reiss, K. (2004). Discrete mathematics and proof in the high school. Introduction. *ZDM, 36*(2), 44–45. https://doi.org/10.1007/BF02655757

Heinze, A., & Reiss, K. (2003). Reasoning and proof: Methodological knowledge as a component of proof competence. In M. A. Mariotti (Ed.), *Proceedings of the Third Conference of the European Society for Research in Mathematics Education.*

Heinze, A., & Reiss, K. (2009). Developing argumentation and proof competencies in the mathematics classroom. In D. A. Stylianou, M. L. Blanton, & E. J. Knuth (Eds.), *Teaching and learning proof across the grades: A K-16 perspective* (pp. 191–203). Routledge.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal, 60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

Hemmi, K. (2008). Students' encounter with proof: The condition of transparency. *ZDM, 40*(3), 413–426. https://doi.org/10.1007/s11858-008-0089-9

Herppich, S., Praetorius, A.-K., Hetmanek, A., Glogger-Frey, I., Ufer, S., Leutner, D., . . . Südkamp, A. (2017). Ein Arbeitsmodell für die empirische Erforschung der diagnostischen Kompetenz von Lehrkräften. In A. Südkamp & A.-K. Praetorius (Eds.), *Diagnostische Kompetenz von Lehrkräften—Theoretische und methodische Weiterentwicklungen* (pp. 75–93). Waxmann.

Hersh, R. (1993). Proving is convincing and explaining. *Educational Studies in Mathematics, 24*(4), 389–399. https://doi.org/10.1007/BF01273372

Hersh, R. (1997). *What is mathematics, really?* Oxford University Press.

Heublein, U., Hutzsch, C., & Schmelzer, R. (2022). *Die Entwicklung der Studienabbruchquoten in Deutschland* [DZHW Brief 05|2022]. Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW). Retrieved 2023-02-01, from https://doi.org/10.34878/2022.05.dzhw_brief

Hilbert, T. S., Renkl, A., Kessler, S., & Reiss, K. (2008). Learning to prove in geometry: Learning from heuristic examples and how it can be supported. *Learning and Instruction, 18*(1), 54–65. https://doi.org/10.1016/j.learninstruc.2006.10.008

Holland, J. L., & Christian, L. M. (2009). The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. *Social Science Computer Review, 27*(2), 196–212. https://doi.org/10.1177/0894439308327481

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. Retrieved 2022-12-08, from https://www.jstor.org/stable/4615733

Housman, D., & Porter, M. (2003). Proof schemes and learning strategies of aboveaverage mathematics students. *Educational Studies in Mathematics, 53,* 139–158. https://doi.org/10.1023/A:1025541416693

Houston, S. K. (1993). Comprehension tests in mathematics. *Teaching Mathematics and its Applications, 12*(2), 60–73. https://doi.org/10.1093/teamat/12.2.60

Hoyles, C., & Küchemann, D. (2002). Students' understandings of logical implication. *Educational Studies in Mathematics, 51*(3), 193–223. https://doi.org/10.1023/A:1023629608614

Inglis, M., & Aberdein, A. (2015). Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica, 23*(1), 87–109. https://doi.org/10.1093/philmat/nku014

Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education, 43*(4), 358–390. https://doi.org/10.5951/jresematheduc.43.4.0358

Inglis, M., & Mejía-Ramos, J. P. (2013). How persuaded are you? A typology of responses. In A. Aberdein & I. J. Dove (Eds.), *The Argument of Mathematics* (pp. 101–117). Springer. https://doi.org/10.1007/978-94-007-6534-4_7

Inglis, M., & Mejia Ramos, J. (2009). The effect of authority on the persuasiveness of mathematical arguments. *Cognition and Instruction, 27*(1), 25–50. https://doi.org/10.1080/07370000802584513

Jahnke, H. N., & Ufer, S. (2015). Argumentieren und Beweisen. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme, & H.-G. Weigand (Eds.), *Handbuch der Mathematikdidaktik* (pp. 331–355). Springer. https://doi.org/10.1007/978-3-642-35119-8_12

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment, 49,* 49–81. https://doi.org/10.1017/CBO9780511808098.004

Keller, E. F. (2016). Practices of generalization in mathematical physics, in biology, and in evolutionary strategies. In K. Chemla, R. Chorlay, & D. Rabouin (Eds.), *The Oxford handbook of generality in mathematics and the sciences.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198777267.013.17

Kempen, L. (2018). How do pre-service teachers rate the conviction, verification and explanatory power of different kinds of proofs? In A. J. Stylianides & G. Harel (Eds.), *Advances in mathematics education research on proof and proving: An international perspective* (pp. 225–237). Springer. https://doi.org/10.1007/978-3-319-70996-3_16

Kempen, L. (2019). *Begründen und Beweisen im Übergang von der Schule zur Hochschule: Theoretische Begründung, Weiterentwicklung und Evaluation einer universitären Erstsemesterveranstaltung unter der Perspektive der doppelten Diskontinuität.* Springer. https://doi.org/10.1007/978-3-658-24415-6

Kempen, L. (2021). Investigating the difference between generic proofs and purely empirical verifications. In *14th International Congress on Mathematics Education.* Shanghai, China.

Kempen, L., & Biehler, R. (2019). Fostering first-year pre-service teachers' proof competencies. *ZDM, 51*(5), 731–746. https://doi.org/10.1007/s11858-019-01035-x

Ketelsen, C. (1994). *Die Gödelschen Unvollständigkeitssätze. Zur Geschichte ihrer Entstehung und Rezeption.* Steiner.

Kirsten, K. (2021). *Beweisprozesse von Studierenden: Ergebnisse einer empirischen Untersuchung zu Prozessverläufen und phasenspezifischen Aktivitäten.* Springer. https://doi.org/10.1007/978-3-658-32242-7

Kleinke, D. J. (1980). Item order, response location and examinee sex and handedness and performance on a multiple-choice test. *The Journal of Educational Research, 73*(4), 225–229. Retrieved 2023-01-23, from https://www.tandfonline.com/doi/abs/10.1080/00220671.1980.10885240

Kneale, W. (1949). *Probability and induction.* Oxford University Press.

Knuth, E. J. (2002). Secondary school mathematics teachers' conceptions of proof. *Journal for Research in Mathematics Education, 33*(5), 379–405. https://doi.org/10.2307/4149959

Ko, Y.-Y. (2011). True or false? Pre-service secondary mathematics teachers' strategies for evaluating statements. In L. R. Wiest & T. Lamberg (Eds.), *Proceedings of the 33rd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 478–486). University of Nevada, Reno.

Ko, Y.-Y., & Knuth, E. (2009). Problems manifested in prospective secondary mathematics teachers' proofs and counterexamples in differentiation. In F.-L. Lin, F.-J. Hsieh, G. Hanna, & M. de Villiers (Eds.), *Proceedings of the ICMI study 19 conference: Proof and proving in mathematics education* (Vol. 1, pp. 262–267). The Department of Mathematics, National Taiwan Normal University.

Ko, Y.-Y., & Knuth, E. J. (2013). Validating proofs and counterexamples across content domains: Practices of importance for mathematics majors. *The Journal of Mathematical Behavior, 32*(1), 20–35. https://doi.org/10.1016/j.jmathb.2012.09.003

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (Second ed.). Sage.

Krummheuer, G. (1995). The ethnography of argumentation. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning: Interaction in classroom cultures* (pp. 229–269). Lawrence Erlbaum Associates, Inc.

Kultusministerkonferenz (Ed.). (2012). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife: Beschluss der Kultusministerkonferenz vom 18.10.2012.*

Kultusministerkonferenz (Ed.). (2022). *Schnellmeldung Abiturnoten 2021 an Gymnasien, Integrierten Gesamtschulen, Fachgymnasien, Fachoberschulen und Berufsoberschulen -vorläufige Ergebnisse-*. Retrieved 2022-10-14, from https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Dokumentationen/Schnellmeldung_Abiturnoten_2021.pdf

Kunimune, S., Kumakura, H., Jones, K., & Fujita, T. (2009). Lower secondary school students' understanding of algebraic proof. In M. Tzekaki, M. Kaldrimidou, & T. Fujita (Eds.), *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 441–448). PME.

Lakatos, I. (1978). *Mathematics, science and epistemology* (J. Worrall & G. Currie, Eds.). Cambridge University Press. https://doi.org/10.1017/CBO9780511624926

Lee, K. (2016). Students' proof schemes for mathematical proving and disproving of propositions. *The Journal of Mathematical Behavior, 41,* 26–44. https://doi.org/10.1016/j.jmathb.2015.11.005

Leron, U. (1983). Structuring Mathematical Proofs. *The AmericanMathematical Monthly, 90*(3), 174–185. https://doi.org/10.2307/2975544

Lesseig, K., Hine, G., Na, G. S., & Boardman, K. (2019). Perceptions on proof and the teaching of proof: A comparison across preservice secondary teachers in Australia, USA and Korea. *Mathematics Education Research Journal, 31*(4), 393–418. https://doi.org/10.1007/s13394-019-00260-7

Lew, K., Weber, K., & Mejía-Ramos, J. P. (2020). Do generic proofs improve proof comprehension? *Journal of Educational Research in Mathematics,* 229–248. https://doi.org/10.29275/jerm.2020.08.sp.1.229

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25*(4), 361–381. https://doi.org/10.1002/bdm.752

Lockwood, E., Ellis, A. B., & Lynch, A. G. (2016). Mathematicians' example-related activity when exploring and proving conjectures. *International Journal of Research in Undergraduate Mathematics Education, 2,* 165–196. https://doi.org/10.1007/s40753-016-0025-2

Ly, I. (2016). Generality, generalization, and induction in Poincaré's philosophy. In *The Oxford handbook of generality in mathematics and the sciences.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198777267.013.5

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (First ed., pp. 39–67). Cambridge University Press. https://doi.org/10.1017/CBO9780511489976.004

Malek, A., & Movshovitz-Hadar, N. (2011). The effect of using Transparent Pseudo-Proofs in linear algebra. *Research in Mathematics Education, 13*(1), 33–58. https://doi.org/10.1080/14794802.2011.550719

Manin, Y. I. (2010). *A course in mathematical logic for mathematicians.* Springer. https://doi.org/10.1007/978-1-4419-0615-1

Marfori, M. A. (2010). Informal proofs and mathematical rigour. *Studia Logica: An International Journal for Symbolic Logic, 96*(2), 261–272. Retrieved 2022-02-10, from https://www.jstor.org/stable/40927692

Mariotti, M. A. (2006). Proof and proving in mathematics education. In Á. Gutiérrez & P. Boero (Eds.), *Handbook of research on the psychology of mathematics education* (pp. 173–204). Sense Publishers.

Martin, W. G., & Harel, G. (1989). Proof frames of preservice elementary teachers. *Journal for Research in Mathematics Education, 20*(1), 41–51. https://doi.org/10.2307/749097

Mason, J., & Pimm, D. (1984). Generic examples: Seeing the general in the particular. *Educational Studies in Mathematics, 15*(3), 277–289. https://doi.org/10.1007/BF00312078

Mat Roni, S., Merga, M. K., & Morris, J. E. (2020). *Conducting quantitative research in education.* Springer. https://doi.org/10.1007/978-981-13-9132-3

Mayring, P. (2022). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (13. ed.). Beltz.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. Retrieved 2023-01-29, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

Mejía Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics, 79*(1), 3–18. Retrieved 2022-04-26, from https://www.jstor.org/stable/41413095

Mejía Ramos, J. P., & Inglis, M. (2009a). Argumentative and proving activities in mathematics education research. In F.-L. Lin, F.-J. Hsieh, G. Hanna, & M. de Villiers (Eds.), *Proceedings of the ICMI study 19 conference: Proof and proving in mathematics education* (Vol. 2, pp. 88–93). National Taiwan Normal University, The Department of Mathematics Taipei.

Mejía Ramos, J. P., & Inglis, M. (2009b). What are the argumentative activities associated with proof? *Research in Mathematics Education, 11,* 77–78. https://doi.org/10.1080/14794800902732258

Mejía Ramos, J. P., Lew, K., Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education, 19,* 130–146. https://doi.org/10.1080/14794802.2017.1325776

Mejía Ramos, J. P., & Weber, K. (2014). Why and how mathematicians read proofs: Further evidence from a survey study. *Educational Studies in Mathematics, 85*(2), 161–173. https://doi.org/10.1007/s10649-013-9514-2

Méndez Coca, D., & Slisko, J. (2013). Software socrative and smartphones as tools for implementation of basic processes of active physics learning in classroom: An initial feasibility study with prospective teachers. *European Journal of Physics Education, 4*(2), 17–24. Retrieved 2023-01-27, from https://files.eric.ed.gov/fulltext/EJ1052308.pdf

Miller, A. L., & Lambert, A. D. (2014). Open-ended survey questions: Item nonresponse nightmare or qualitative data dream? *Survey Practice, 7*(5). https://doi.org/10.29115/SP-2014-0024

Miller, D., & CadwalladerOlsker, T. (2020). Investigating undergraduate students' view of and consistency in choosing empirical and deductive arguments. *Research in Mathematics Education, 22*(3), 249–264. https://doi.org/10.1080/14794802.2019.1677489

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen. (2019). *Kernlehrplan für die Sekundarstufe I Gymnasium in Nordrhein-Westfalen: Mathematik.* Retrieved

2023-01-27, from https://www.schulentwicklung.nrw.de/lehrplaene/lehrplan/195/g9_m_klp_3401_2019_06_23.pdf

Möckel, T., Beste, C., & Wascher, E. (2015). The effects of time on task in response selection—An ERP study of mental fatigue. *Scientific Reports, 5*(1), 10113. https://doi.org/10.1038/srep10113

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics, 27*(3), 249–266. https://doi.org/10.1007/BF01273731

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (2. ed.). Springer.

Movshovitz-Hadar, N., & Malek, A. (1998). Transparent Pseudo-Proofs – a bridge to formal proofs. In *Proceedings of the International Conference on the Teaching of Mathematics* (pp. 221–223). John Wiley & Sons.

National Council of Teachers of Mathematics (Ed.). (2000). *Principles and standards for school mathematics.* National Council of Teachers of Mathematics.

National Research Council. (2001). *Adding it up: Helping children learn mathematics* (J. Kilpatrick, J. Swafford, & B. Findell, Eds.). National Academies Press. https://doi.org/10.17226/9822

Neuhaus-Eckhardt, S. (2022). *Beweisverständnis von Studierenden: Zusammenhänge zu individuellen Merkmalen und der Nutzung von Beweislesestrategien* (Vol. 42). Waxmann.

Newman, D. L., Kundert, D. K., Jr, D. S. L., & Bull, K. S. (1988). Effect of varying item order on multiple-choice test scores: Importance of statistical and cognitive Difficulty. *Applied Measurement in Education, 1*(1), 89–97. https://doi.org/10.1207/s15324818ame0101_8

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19,* 1–13. https://doi.org/10.1177/1609406919899220

Oliveira e Silva, T., Herzog, S., & Pardi, S. (2013). Empirical verification of the even Goldbach conjecture and computation of prime gaps up to $4{\cdot}10^{18}$. *Mathematics of Computation, 83* (288), 2033–2060. https://doi.org/10.1090/S0025-5718-2013-02787-1

O'Neil, J., Harold F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the national assessment of educational progress mathematics performance. *Educational Assessment, 3*(2), 135–157. https://doi.org/10.1207/s15326977ea0302_2

Patel, N., Baker, S. G., & Scherer, L. D. (2019). Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. *Journal of Experimental Psychology: General, 148,* 2129–2153. https://doi.org/10.1037/xge0000592

Pedemonte, B. (2007). How can the relationship between argumentation and proof be analysed? *Educational Studies in Mathematics, 66,* 23–41. https://doi.org/10.1007/s10649-006-9057-x

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods, 48,* 341–348. https://doi.org/10.3758/s13428-015-0576-1

Perelman, C. (1970). *Le champ de l'argumentation.* Presses Universitaires de Bruxelles.

Pfeiffer, K. (2011). *Features and purposes of mathematical proofs in the view of novice students: Observations from proof validation and evaluation performances* (Doctoral dissertation, National University of Ireland, Galway). Retrieved from https://aran.library.nuigalway.ie/bitstream/handle/10379/1862/Pfeiffer%20Thesis.pdf?isAllowed=y&sequence=1

Piatek-Jimenez, K. L. (2004). *Undergraduate mathematics students' understanding of mathematical statements and proofs* (Doctoral dissertation, The University of Arizona). Retrieved from https://www.proquest.com/openview/ceafc84462d4b812a0dc4418a754e746/1?pq-origsite=gscholar&cbl=18750&diss=y

Piatek-Jimenez, K. L. (2010). Students' interpretations of mathematical statements involving quantification. *Mathematics Education Research Journal, 22*(3), 41–56. https://doi.org/10.1007/BF03219777

Plump, C. M., & LaRosa, J. (2017). Using Kahoot! in the classroom to create engagement and active Learning: A game-based technology solution for eLearning novices. *Management Teaching Review, 2*(2), 151–158. https://doi.org/10.1177/2379298116689783

Poincaré, H. (1952). *Science and Method.* Dover Publications.

Polya, G. (1954). *Mathematics and plausible reasoning* (Vol. 1). Princeton University Press.

Powers, R. A., Craviotto, C., & Grassl, R. M. (2010). Impact of proof validation on proof writing in abstract algebra. *International Journal of Mathematical Education in Science and Technology, 41*(4), 501–514. https://doi.org/10.1080/00207390903564603

Pracht, E. (1979). Beweisverständnis und dessen Überprüfbarkeit. In W. Dörfler & R. Fischer (Eds.), *Beweisen im Mathematikunterricht* (pp. 349–356). Hölder-Pichler-Tempsky.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making, 29*(5), 453–469. https://doi.org/10.1002/bdm.1883

Prinz, A., Golke, S., & Wittwer, J. (2020). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review, 31,* 100358. https://doi.org/10.1016/j.edurev.2020.100358

R Core Team. (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rabouin, D. (2016). The problem of a "general" theory in mathematics: Aristotle and Euclid. In K. Chemla, R. Chorlay, & D. Rabouin (Eds.), *The Oxford handbook of generality in mathematics and the sciences* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198777267.013.4

Rach, S., & Ufer, S. (2020). Which prior mathematical knowledge is necessary for study success in the university study entrance phase? Results on a new model of knowledge levels based on a reanalysis of data from existing studies. *International Journal of Research in Undergraduate Mathematics Education, 6*, 375–403. https://doi.org/10.1007/s40753-020-00112-x

Rav, Y. (1999). Why do we prove theorems? *Philosophia Mathematica, 7*(1), 5–41. https://doi.org/10.1093/philmat/7.1.5

Recio, A., & Godino, J. (1996). Assessment of university students' mathematical generalization and symbolization capacities. In *Proceedings of the 20th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 1–231). PME.

Recio, A., & Godino, J. (2001). Institutional and personal meanings of mathematical proof. *Educational Studies in Mathematics*(48), 83–99. https://doi.org/10.1023/A:1015553100103

Reid, D. A., & Knipping, C. (2010). *Proof in mathematics education: Research, learning and teaching.* Sense Publishers.

Reiss, K., & Heinze, A. (2000). Begründen und Beweisen im Verständnis von Abiturienten. In *Beiträge zum Mathematikunterricht 2000* (pp. 520–523). div-Verlag Franzbecker.

Reiss, K., Hellmich, F., & Thomas, J. (2002). Individuelle und schulische Bedingungsfaktoren für Argumentationen und Beweise im Mathematikunterricht. *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwissenschaftlicher und überfachlicher Kompetenzen,* 51–64. https://doi.org/10.25656/01:3938

Reiss, K., & Schmieder, G. (2014). *Basiswissen Zahlentheorie.* Springer. https://doi.org/10.1007/978-3-642-39773-8

Reiss, K., & Ufer, S. (2009). Was macht mathematisches Arbeiten aus? Empirische Ergebnisse zum Argumentieren, Begründen und Beweisen. *Jahresbericht der Deutschen Mathematiker-Vereinigung, 111,* 155–177.

Riley, K. J. (2003). *An investigation of prospective secondary mathematics teachers' conceptions of proof and refutations* (Doctoral dissertation, Montana State University). Retrieved from https://scholarworks.montana.edu/xmlui/bitstream/handle/1/8385/31762103856330.pdf;sequence=1

Rowland, T. (2001). Generic proofs: Setting a good example. *Mathematics Teaching, 177,* 40–43. Retrieved from https://nrich.maths.org/content/id/7831/Generic_proofs.pdf

Şad, S. N. (2020). Does difficulty-based item order matter in multiple-choice exams? (Empirical evidence from university students). *Studies in Educational Evaluation, 64,* 100812. https://doi.org/10.1016/j.stueduc.2019.100812

Schoenfeld, A. H. (2009). Series editor's foreword: The soul of mathematics. In D. A. Stylianou, M. L. Blanton, & E. J. Knuth (Eds.), *Teaching and learning proof across the grades: A K-16 perspective* (pp. xii-xvi). Routledge.

Schütte, K. (1977). *Proof theory.* Springer.

Sears, R. (2019). Proof schemes of pre-service middle and secondary mathematics teachers. *Investigations in Mathematics Learning, 11*(4), 258–274. https://doi.org/10.1080/19477503.2018.1467106

Segal, J. (1999). Learning about mathematical proof: Conviction and validity. *The Journal of Mathematical Behavior, 18*(2), 191–210. https://doi.org/10.1016/S0732-3123(99)00028-0

Selden, A. (2012). Transitions and proof and proving at the tertiary level. In G. Hanna & M. de Villiers (Eds.), *Proof and proving in mathematics education* (Vol. 15, pp. 391–420). Springer. https://doi.org/10.1007/978-94-007-2129-6_17

Selden, A., Mckee, K., & Selden, J. (2010). Affect, behavioural schemas and the proving process. *International Journal of Mathematical Education in Science and Technology, 41,* 199–215. https://doi.org/10.1080/00207390903388656

Selden, A., & Selden, J. (2003). Validations of proofs considered as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education, 34*(1), 4–36. https://doi.org/10.2307/30034698

Selden, A., & Selden, J. (2013). Proof and problem solving at university level. *The Mathematics Enthusiast, 10*(1-2), 303–334. https://doi.org/10.54870/1551-3440.1269

Selden, A., & Selden, J. (2017). A comparison of proof comprehension, proof construction, proof validation and proof evaluation. In R. Göller, R. Biehler, R. Hochmuth, & H.-G. Rück (Eds.), *Didactics of Mathematics in Higher Education as a Scientific Discipline* (pp. 339–345). Universitätsbibliothek Kassel.

Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics, 29*(2), 123–151. https://doi.org/10.1007/BF01274210

Semeraro, C., Giofrè, D., Coppola, G., Lucangeli, D., & Cassibba, R. (2020). The role of cognitive and non-cognitive factors in mathematics achievement: The importance of the quality of the student-teacher relationship in middle school. *PLoS ONE, 15*(4), e0231381. https://doi.org/10.1371/journal.pone.0231381

Sen, C., & Guler, G. (2015). Examination of secondary school seventh graders' proof skills and proof schemes. *Universal Journal of Educational Research, 3*(9), 617–631. https://doi.org/10.13189/ujer.2015.030906

Sevimli, E. (2018). Undergraduates' propositional knowledge and proof schemes regarding differentiability and integrability concepts. *International Journal of Mathematical Education in Science and Technology, 49*(7), 1052–1068. https://doi.org/10.1080/0020739X.2018.1430384

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General, 141*(3), 423–428. https://doi.org/10.1037/a0025391

Siegel, H. (1999). What (good) are thinking dispositions? *Educational Theory, 49*(2), 207–221. https://doi.org/10.1111/j.1741-5446.1999.00207.x

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology, 6,* 532. Retrieved 2022-09-22, from https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00532

Sirota, M., Dewberry, C., Juanchich, M., Valuš, L., & Marshall, A. C. (2021). Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *Journal of Behavioral Decision Making, 34*(3), 322–343. https://doi.org/10.1002/bdm.2213

Sjögren, J. (2010). A note on the relation between formal and informal proof. *Acta Analytica, 25*(4), 447–458. https://doi.org/10.1007/s12136-009-0084-y

Sommerhoff, D. (2017). *The individual cognitive resources underlying students' mathematical argumentation and proof skills: From theory to intervention* (Doctoral dissertation, Ludwig-Maximilians-Universität München). https://doi.org/10.5282/EDOC.22687

Sommerhoff, D., & Brunner, E. (2021). Forschungsstand Mathematisches Argumentieren und Beweisen vom Elementar- bis zum Hochschulbereich. *GDM-Mitteilungen, 111,* 74–82.

Sommerhoff, D., Kollar, I., & Ufer, S. (2021). Supporting mathematical argumentation and proof skills: Comparing the effectiveness of a sequential and a concurrent instructional approach to support resource-based cognitive skills. *Frontiers in Psychology, 11,* 572165. https://doi.org/10.3389/fpsyg.2020.572165

Sommerhoff, D., & Ufer, S. (2019). Acceptance criteria for validating mathematical proofs used by school students, university students, and mathematicians in the context of teaching. *ZDM, 51*(5), 717–730. https://doi.org/10.1007/s11858-019-01039-7

Sommerhoff, D., Ufer, S., & Kollar, I. (2015). Research on mathematical argumentation: A descriptive review of PME proceedings. In K. Beswick, T. Muir, & J. Fielding-Wells (Eds.), *Proceedings of 39th Psychology of Mathematics Education conference* (Vol. 4, pp. 193–200). PME.

Sporn, F., Sommerhoff, D., & Heinze, A. (2021). Beginning university mathematics students' proof understanding. In M. Inprasitha, N. Changsri, & N. Boonsena (Eds.), *Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 105–112). PME.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Staples, M., & Conner, A. (2022). Introduction: Conceptualizing argumentation, justification, and proof in mathematics education. In K. N. Bieda, A. Conner, K. W. Kosko, & M. Staples (Eds.), *Conceptions and Consequences of Mathematical Argumentation, Justification, and Proof* (pp. 1–10). Springer. https://doi.org/10.1007/978-3-030-80008-6_1

Steele, J. R., & Ambady, N. (2006). "Math is Hard!" The effect of gender priming on women's attitudes. *Journal of Experimental Social Psychology, 42*(4), 428–436. https://doi.org/10.1016/j.jesp.2005.06.003

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education, 38*(3), 289–321. Retrieved from https://www.jstor.org/stable/30034869

Stylianides, A. J. (2016). *Proving in the elementary mathematics classroom.* Oxford University Press.

Stylianides, A. J., & Al-Murani, T. (2010). Can a proof and a counterexample coexist? Students' conceptions about the relationship between proof and refutation. *Research in Mathematics Education, 12*(1), 21–36. https://doi.org/10.1080/14794800903569774

Stylianides, A. J., & Stylianides, G. J. (2009). Proof constructions and evaluations. *Educational Studies in Mathematics, 72*(2), 237–253. https://doi.org/10.1007/s10649-009-9191-3

Stylianou, D. A., Blanton, M. L., & Rotou, O. (2015). Undergraduate students' understanding of proof: Relationships between proof conceptions, beliefs, and classroom experiences with learning proof. *International Journal of Research in Undergraduate Mathematics Education, 1*(1), 91–134. https://doi.org/10.1007/s40753-015-0003-0

Stylianou, D. A., Chae, N., & Blanton, M. (2006). Students' proof schemes: A closer look at what characterizes students' proof conceptions. In S. Alatorre, J. L. Cortina, M. Sáiz, & A. Méndez (Eds.), *Proceedings of the 28th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 54–60). Universidad Pedagógica Nacional.

Tabach, M., Barkai, R., Tsamir, P., Tirosh, D., Dreyfus, T., & Levenson, E. (2010). Verbal justification-is it a proof? Secondary school teachers' perceptions. *International Journal of Science and Mathematics Education, 8*(6), 1071–1090. https://doi.org/10.1007/s10763-010-9230-7

Tabach, M., Levenson, E., Barkai, R., Tirosh, D., Tsamir, P., & Dreyfus, T. (2010). Secondary school teachers' awareness of numerical examples as proof. *Research in Mathematics Education, 12,* 117–131. https://doi.org/10.1080/14794802.2010.496973

Tabach, M., Levenson, E., Barkai, R., Tsamir, P., Tirosh, D., & Dreyfus, T. (2011). Secondary teachers' knowledge of elementary number theory proofs: The case of general-cover proofs. *Journal of Mathematics Teacher Education, 14*(6), 465–481. https://doi.org/10.1007/s10857-011-9185-9

Tall, D. (1989). The nature of mathematical proof. *Mathematics Teaching, 127,* 28–32. Retrieved from https://homepages.warwick.ac.uk/staff/David.Tall/pdfs/dot1989a-nature-proof-mt.pdf

Tall, D., Yevdokimov, O., Koichu, B., Whiteley, W., Kondratieva, M., & Cheng, Y.-H. (2012). Cognitive development of proof. In G. Hanna & M. de Villiers (Eds.), *Proof and Proving*

in Mathematics Education (Vol. 15, pp. 13–49). Springer. https://doi.org/10.1007/978-94-007-2129-6_2

Tanswell, F. (2015). A problem with the dependence of informal proofs on formal proofs. *Philosophia Mathematica, 23*(3), 295–310. https://doi.org/10.1093/philmat/nkv008

Thiede, K.W., Griffin, T. D.,Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 85–106). Routledge/Taylor & Francis Group.

Thomson, K., & Oppenheimer, D. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making, 11,* 99–113. https://doi.org/10.1037/t49856-000

Thurston, W. P. (1994). On proof and progress in mathematics. *Bulletin of the American Mathematical Society, 30,* 161–177. https://doi.org/10.48550/arXiv.math/9404236

Tieben, N. (2019). Brückenkursteilnahme und Studienabbruch in Ingenieurwissenschaftlichen Studiengängen. *Zeitschrift für Erziehungswissenschaft, 22*(5), 1175–1202. https://doi.org/10.1007/s11618-019-00906-z

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Toulmin, S. E. (2003). *The Uses of Argument* (Updated ed.). Cambridge University Press.

Ufer, S., Heinze, A., Kuntze, S., & Rudolph-Albert, F. (2009). Beweisen und Begründen im Mathematikunterricht. *Journal für Mathematik-Didaktik, 30*(1), 30–54. https://doi.org/10.1007/BF03339072

Ufer, S., Heinze, A., & Reiss, K. (2008). Individual predictors of geometrical proof competence. In O. Figueras & A. Sepúlveda (Eds.), *Proceedings of the Joint Meeting of the 32nd Conference of the International Group for the Psychology of Mathematics Education, and the XX North American Chapter* (Vol. 4, pp. 361–368). PME.

Universität Bielefeld. (2020). *Studierende (Fallzählung) nach angestrebtem Abschluss und Fachsemester im Wintersemester 19/20.* Retrieved 2022-10-14, from https://www.uni-bielefeld.de/uni/profil/daten-zahlen/202008WS.pdf

Usiskin, Z. (1980). What should not be in the algebra and geometry curricula of average college-bound students? *Mathematics Teacher, 73,* 413–424. https://doi.org/10.5951/MT.73.6.0413

van der Linden, D., Frese, M., & Meijman, T. F. (2003). Mental fatigue and the control of cognitive processes: Effects on perseveration and planning. *Acta Psychologica, 113*(1), 45–65. https://doi.org/10.1016/s0001-6918(02)00150-6

Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics, 48*(1), 101–119. https://doi.org/10.1023/A:1015535614355

Weber, K. (2003). *Students' difficulties with proof.* Retrieved 2021-12-07, from https://www.maa.org/programs/faculty-and-departments/curriculum-department-guidelines-recommendations/teaching-and-learning/research-sampler-8-students-difficulties-with-proof

Weber, K. (2005). Problem-solving, proving, and learning: The relationship between problem-solving processes and learning opportunities in the activity of proof construction. *The Journal of Mathematical Behavior, 24*(3), 351–360. https://doi.org/10.1016/j.jmathb.2005.09.005

Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for Research in Mathematics Education, 39*(4), 431–459. Retrieved 2022-08-12, from https://www.jstor.org/stable/40539306

Weber, K. (2010). Mathematics majors' perceptions of conviction, validity, and proof. *Mathematical Thinking and Learning, 12*(4), 306–336. https://doi.org/10.1080/10986065.2010.495468

Weber, K. (2012). Mathematicians' perspectives on their pedagogical practice with respect to proof. *International Journal of Mathematical Education in Science and Technology, 43*(4), 463–482. https://doi.org/10.1080/0020739X.2011.622803

Weber, K. (2013). On the sophistication of naïve empirical reasoning: Factors influencing mathematicians' persuasion ratings of empirical arguments. *Research in Mathematics Education, 15*(2), 100–114. https://doi.org/10.1080/14794802.2013.797743

Weber, K. (2014). Proof as a cluster concept. In C. Nicol, S. Oesterle, P. Liljedahl, & D. Allan (Eds.), *Proceedings of the 38th Conference of the International Group for the Psychology of Mathematics Education and the 36th Conference of the North American Chapter of the Psychology of Mathematics Education* 36 (Vol. 5, pp. 353–360).

Weber, K., & Czocher, J. (2019). On mathematicians' disagreements on what constitutes a proof. *Research in Mathematics Education, 21*(3), 251–270. https://doi.org/10.1080/14794802.2019.1585936

Weber, K., Lew, K., & Mejía-Ramos, J. P. (2020). Using expectancy value theory to account for individuals' mathematical justifications. *Cognition and Instruction, 38*(1), 27–56. https://doi.org/10.1080/07370008.2019.1636796

Weber, K., & Mejia-Ramos, J. P. (2015). On relative and absolute conviction in mathematics. *For the Learning of Mathematics, 35*(2), 15–21. Retrieved from https://www.jstor.org/stable/44382751

Wilder, R. L. (1981). *Mathematics as a cultural system* (1st ed.). Pergamon Press.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wittmann, E. C. (2014). Operative Beweise in der Schul- und Elementarmathematik. *mathematica didacticta, 37,* 213–232. https://doi.org/10.18716/ojs/md/2014.1127

Wittmann, E. C., & Müller, G. (1988). Wann ist ein Beweis ein Beweis? In P. Bender (Ed.), *Mathematikdidaktik: Theorie und Praxis. Festschrift für Heinrich Winter* (pp. 237–257). Cornelsen.

Wußing, H. (2008). *6000 Jahre Mathematik. Eine kulturgeschichtliche Zeitreise—1. Von den Anfängen bis Leibniz und Newton.* Springer. https://doi.org/10.1007/978-3-540-77192-0

Wußing, H. (2009). *6000 Jahre Mathematik. Eine kulturgeschichtliche Zeitreise—2. Von Euler bis zur Gegenwart.* Springer. https://doi.org/10.1007/978-3-540-77314-6

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education, 27*(4), 458–477. https://doi.org/10.2307/749877

Yang, K.-L., & Lin, F.-L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics, 67*(1), 59–76. https://doi.org/10.1007/s10649-007-9080-6

Yee, S. P., Boyle, J. D., Ko, Y.-Y.W., & Bleiler-Baxter, S. K. (2018). Effects of constructing, critiquing, and revising arguments within university classrooms. *The Journal of Mathematical Behavior, 49,* 145–162. https://doi.org/10.1016/j.jmathb.2017.11.009

Young, A. G., Powers, A., Pilgrim, L., & Shtulman, A. (2018). Developing a Cognitive Reflection Test for school-age children. In C. Kalish, M. A. Rau, X. J. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* (pp. 1232–1237).

Zach, R. (2019). Hilbert's program. *The Stanford Encyclopedia of Philosophy.* Retrieved 2021-12-22, from https://plato.stanford.edu/archives/sum2019/entries/hilbert-program/

Zaslavsky, O., Nickerson, S. D., Stylianides, A. J., Kidron, I., & Winicki-Landman, G. (2012). The need for proof and proving: Mathematical and pedagogical perspectives. In G. Hanna & M. de Villiers (Eds.), *Proof and Proving in Mathematics Education: The 19th ICMI Study* (pp. 215–229). Springer. https://doi.org/10.1007/978-94-007-2129-6_9

Zeybek Simsek, Z. (2021). "Is it valid or not?": Pre-service teachers judge the validity of mathematical statements and student arguments. *European Journal of Science and Mathematics Education, 9*(2), 26–42. https://doi.org/10.30935/scimath/10772