WHEN PROOFS DO(N'T) HELP: THE ROLE OF ARGUMENTS IN VALIDATING UNIVERSAL STATEMENTS

Milena Damrau

LMU Munich

This experimental study investigates how first-year university students validate universal statements and how their success is influenced by statement characteristics and argument types. A total of 430 participants validated five statements that varied in familiarity and truth value and were randomly assigned to receive no arguments, empirical arguments, generic proofs, or ordinary proofs. Logistic regression analyses show that both familiarity and truth value significantly affect validation success. Reading empirical arguments positively affects validation success, whereas ordinary proofs do not. These findings suggest that students need support in using proofs to validate statements, and that example-based reasoning can help students develop an initial understanding of universal statements.

INTRODUCTION

Finding and validating mathematical statements—i.e., judging whether a statement is valid within a given theoretical framework—are core mathematical activities. Accordingly, curricula worldwide suggest that students should be able to make, investigate, evaluate (e.g., validate), and justify mathematical statements (e.g., Kultusministerkonferenz, 2012; National Council of Teachers of Mathematics, 2000). These activities are considered an integral part of mathematical argumentation (e.g., Kultusministerkonferenz, 2012), as mathematical statements can only be validly justified by mathematical proofs. Moreover, as investigating and proving statements constitute a major part of mathematics programs, respective competencies are particularly relevant at the beginning of university studies.

Given the prominent role of validating statements in mathematics, it is not surprising that there is a large body of research investigating learners' "estimation of the truth of a conjecture" (e.g., about 12% of the analyzed articles on proof and argumentation in Mejía Ramos & Inglis, 2009). Reported success rates in validating statements vary widely, ranging from approximately 30% to 100% (e.g., Barkai et al., 2002; Hoyles & Küchemann, 2002; Ko, 2011; Ko & Knuth, 2009; Riley, 2003; Zeybek Simsek, 2021). Reasons for such variation have been attributed partly to the characteristics of the statement, such as its truth value (e.g., Barkai et al., 2002; Zeybek Simsek, 2021), but also to students' success in constructing or using mathematical arguments. However, only a few studies have systematically investigated influencing factors, and the results are inconsistent. In particular, to my knowledge, no studies have systematically analyzed relations to proof-related activities yet. This paper, therefore, aims to

2026. In H. Krzywacki, M. Hähkiöniemi, E. Haataja, M.S. Hannula & P. Hästö (Eds.). *Proceedings of the 49th Conference of the International Group for the Psychology of Mathematics Education: Research Reports*, pp. 315-322. PME.

investigate factors influencing first-year university students' success in validating universal statements, in particular, statement characteristics, as well as reading and constructing mathematical arguments.

FACTORS INFLUENCING THE VALIDATION OF STATEMENTS

In addition to investigating students' or teachers' success in validating mathematical statements, i.e., estimating the truth value, some studies have examined the relationship to statement characteristics (e.g., true vs false, universal vs existential) and/or the type of arguments or argumentation process that leads students or teachers to reach a respective conclusion.

Characteristics of the statement

As already mentioned in the introduction, the research on factors influencing the success of validating mathematical statements is, so far, inconclusive. Overall, learners appear to struggle particularly when validating false statements (e.g., Ko, 2011; Zeybek Simsek, 2021). For instance, Ko (2011) conducted semi-structured interviews with eight secondary mathematics education majors from a US university. While most participants (approximately 83%) correctly validated six statements (four of which were true and two of which were false, from different content areas), one of the false statements was correctly validated by only half of them (the other one by 87.5%). A false statement that has been used in several studies is that "if the perimeter of a rectangle increases, its area also increases". The percentage of preservice teachers who *incorrectly* validated this statement as true has been reported to be comparatively high (between 57 and 72%, e.g., Riley, 2003; Zeybek Simsek, 2021). 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 are predominantly asked to prove true (universal) statements and only rarely to disprove false ones (e.g., Buchbinder & Zaslavsky, 2007; Ko, 2011). However, other studies have found no noteworthy differences in the validation of true and false statements (Ko & Knuth, 2009). More research is needed to better understand how a statement's truth value influences students' success in validating the statement.

Familiarity with statements has also been considered a relevant factor influencing understanding of proof in general (e.g., Hanna, 1989; Stylianides, 2007). Although I am not aware of any systematic research, I would expect that learners are more successful at validating familiar statements than unfamiliar ones, because they have usually already applied and understood them. But this needs empirical confirmation.

In summary, the success rates of students' and teachers' validation of mathematical statements differ substantially and seem to depend on characteristics of the statement, such as its actual truth value, and the familiarity with the statement and the respective framing theory. However, more research is needed to identify specific relations.

Relations to learners' usage of mathematical arguments

In addition to statement characteristics, previous research has examined how learners approach the validation of mathematical statements and what types of arguments they construct, for instance, empirical arguments (based on examples), generic proofs (based on specific observations that can be generalized), or ordinary proofs (deductive arguments typically constructed by mathematicians). Many studies report that learners often use empirical arguments (e.g., Barkai et al., 2002; Healy & Hoyles, 2000) or 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" (Ko, 2011, p. 481) and only few learners use other strategies such as deductive arguments (e.g., Barkai et al., 2002; Kempen, 2019; Ko, 2011).

Students' usage of empirical arguments has been critically discussed in the mathematics education literature. While some researchers argue that it *may* indicate insufficient understanding of generality (e.g., Conner, 2022), others emphasize that experimentation is not only common but also potentially useful for exploring conjectures and gaining an initial understanding (e.g., de Villiers, 2010). Indeed, empirical arguments are frequently used and often successfully during the initial exploration phase (e.g., Buchbinder & Zaslavsky, 2007; Ko, 2011). Also, findings reported by Alcock and Inglis (2008) show that even mathematicians use examples when evaluating and proving statements, albeit at different stages, highlighting the usefulness of skills involving experimentation for exploring statements.

In summary, empirical arguments are used frequently. Students sometimes attempt to identify patterns or structures, while deductive arguments are comparatively rare (e.g., Ko, 2011). What remains open is whether and how the construction of different types of arguments affects students' validation success and/or the degree of their conviction. For instance, do students who construct empirical arguments gain absolute or mainly relative conviction (Weber & Mejía Ramos, 2015) in the truth of the statement? Also, the extent to which *reading* different types of arguments (such as empirical ones or generic or ordinary proofs) influences the success in validating mathematical statements has not yet been systematically investigated. Since proofs are considered the central method for determining the validity of a statement, reading and validating such arguments should have a positive effect.

RESEARCH INTEREST AND THE CURRENT STUDY

This contribution examines the influence of statement characteristics (truth value and familiarity) and of reading and constructing different types of arguments (empirical arguments, generic proofs, and ordinary proofs) on first-year university students' success in validating mathematical universal statements. The focus in this study is on universal statements, mainly as the overall project also investigates students' understanding of the generality of statements (see Damrau, 2023). The following

questions were investigated: (RQ1) How successfully do students validate universal statements, and how does their success depend on the familiarity and truth value of a statement? (RQ2) To what extent does reading different types of arguments influence students' success in validating universal statements? (RQ3) How does the construction of different types of arguments relate to students' validation of statements?

Methodology

An experimental online questionnaire study was conducted in November 2020 during two first-year mathematics lectures at a large German university, with 430 preservice mathematics teachers and mathematics majors. The participants were randomly assigned to one of four groups: no arguments (Group A), empirical arguments (Group B), generic proofs (Group C), or ordinary proofs (Group D).

Participants worked through five universal statements in sequence: two familiar statements from school geometry (sum of interior angles in a triangle and Pythagorean theorem), two unfamiliar from elementary number theory (sum of two odd numbers is even and product of two odd numbers is odd), and one false statement (sum of three consecutive numbers is divisible by 6). They read each statement and the corresponding provided argument (if applicable; non-general *proof-like* arguments were used for the false statement) and were then asked to evaluate the validity of the statement. Response options allowed them to express both absolute and relative conviction in their judgments (Weber & Mejía Ramos, 2015), as well as state that they do not know. As participants in Group A received no arguments, they were instead asked to justify why the statements were true or false. These justifications were coded according to a scheme based on Harel and Sowder (1998). The main categories consisted of external convictions (i.e., authority-based, reference to a rule, or *pseudo-arguments*, see also Kempen, 2019), empirical arguments, counterexamples, analytical arguments (e.g., deductive or transformative), and unclear (if no decision could be made). In theory, counterexamples can be seen as empirical arguments; however, because they *prove* the falsity of a statement, in contrast to other empirical arguments, it seemed oversimplifying to code them as empirical. Because neither of the other main categories introduced by Harel & Sowder seemed to be appropriate either, counterexamples were treated as another main category. The full coding scheme and all items can be found in Damrau (2023).

All participants then completed a *Cognitive Reflection Test* (CRT, based on Frederick, 2005), which was used to control for inter-individual differences (for details, see Damrau, 2023). Lastly, participants answered demographic questions to collect, for instance, information about their final high school mathematics grade.

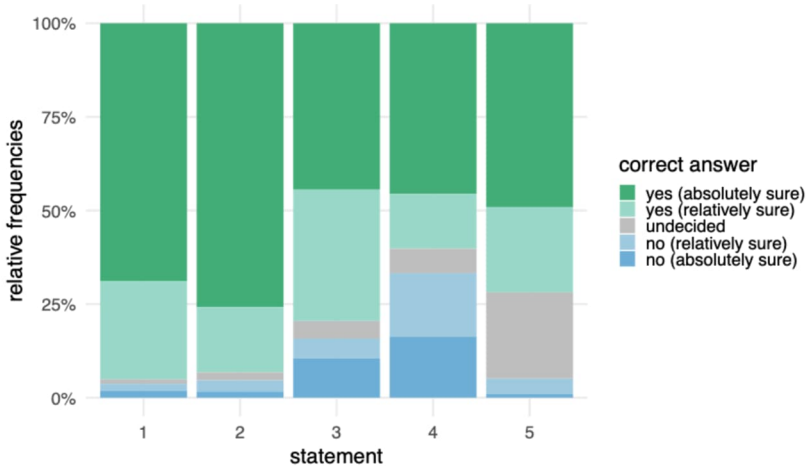
Data analysis included descriptive statistics and logistic regression analyses using cumulative link mixed models (CLMMs) for ordinal data (Christensen, 2022), using participants' success in validating the statements as the dependent variable, CRT performance as a control variable, the group and the statement characteristics (truth

value and familiarity) as fixed factors, and the person as random factor. Multiple testing was corrected for using the Holm's method.

Results

Success rates for the correct validation of statements ranged from approximately 45% to 75% with respect to absolute conviction (see Fig. 1). The CRT performance was predictive for participants' validation success ($B = .63, p < .001$).

Figure 1: Correct validation of statements. Statements 2 and 5 are familiar ones, statements 1 and 3 are unfamiliar ones, and statement 4 is false



Surprisingly, one of the familiar statements—the Pythagorean theorem (statement 5)—showed relatively low success rates, with only about 50% of participants expressing absolute certainty that it is valid. Regression analyses nevertheless showed that familiarity had a positive effect on validation success. As expected, the false statement was validated correctly less frequently than the true statements. Participants who received empirical arguments validated statements more successfully than those who received no arguments ($B = .44, p = .004$). Generic proofs appeared to have a similar, though slightly weaker, positive effect, which did not reach statistical significance after Holm's correction ($B = .27, p = .088$). Reading ordinary proofs had no positive effect on students' success in statement validation ($B = -.09, p = .537$).

Participants in Group A, who were asked to justify their estimation, predominantly used empirical (45%) or pseudo-arguments (such as restating the claim, 16%) for the unfamiliar statements; some used (mostly incomplete) deductive arguments (12% incomplete, 3% complete). For familiar statements, participants most often relied on external convictions, such as arguments based on authorities (e.g., referring to the school, 24%) or a rule they had learned (37%), again frequently giving pseudo-arguments (25%); very few produced incomplete deductive arguments (2%). Most participants who correctly rejected the false statement used counterexamples (51% of all participants); those who claimed the statement is true mostly used empirical (18% of all participants) or unclear arguments (20% of all participants).

The degree of expressed conviction (absolute vs. relative) appeared to be related to the type of constructed argument. Participants who constructed (incomplete) deductive arguments were almost all absolutely convinced of the validity of the statements, whereas the majority of participants who relied on empirical arguments expressed only relative conviction.

DISCUSSION AND OUTLOOK

The main goal of this study was to gain experimental evidence for the effects of statement characteristics and the reading and construction of arguments on students' validation success. The results of this study experimentally support assumptions in the literature (e.g., Buchbinder & Zaslavsky, 2007; Hanna, 1989; Ko, 2011) that both familiarity and truth value influence students' success in validating mathematical statements. It should be noted that the statements differed not only in familiarity but also in content domain (geometry vs elementary number theory), which may have also played a role and should be investigated further. However, the different types of arguments constructed by participants in Group A indicate that familiarity with a statement indeed strongly influences students' argumentation processes, as expected. The unexpectedly low validation rate for the Pythagorean theorem may be explained by its formulation in natural language, which led some participants not to recognize it as the Pythagorean theorem. Success rates were lower for the false statement than for the true ones. Although this study considered only one false statement, this finding confirms prior research (e.g., Ko, 2011; Zeybek Simsek, 2021). Thus, I agree with other colleagues who argue that learners should also be given more opportunities to validate false statements, not only to improve their validation skills but also to gain a broader understanding of mathematical statements and proofs.

Reading empirical arguments (and to a lesser extent, generic proofs) had a positive effect on validation success. In line with prior research, one possible explanation is that considering examples helps learners better understand the statement and develop a stronger intuition about its validity (e.g., Buchbinder & Zaslavsky, 2007). Contrary to expectations, reading ordinary proofs did *not* improve validation performance, suggesting that students were unable to use proofs to determine whether a statement is true or false. They may lack sufficient proof comprehension or understanding of the verification function of proofs, which aligns with other research (e.g., Conradie & Frith, 2000). The relations between students' proof comprehension and their success in validating statements after reading corresponding proofs warrant further investigation in future studies.

Although few students constructed deductive arguments, those who did showed absolute conviction in the validity of the statements, whereas those who constructed empirical arguments predominantly showed only relative conviction. This highlights the importance of distinguishing between absolute and relative conviction when studying validation and proof competence, as suggested by Weber and Mejía Ramos (2015). Moreover, it shows that for students who *have* sufficient skills to actually *prove*

a universal statement (which, again, were not that many in this study), these proofs *do* help to successfully validate the statement with absolute conviction.

For instruction, the combined findings highlight two consequences: The potential of carefully using empirical arguments to support students in better understanding the statements under investigation *before* attempting further proof-related activities. And, as many researchers have highlighted, a clear need to improve students' proof comprehension and construction. The findings also suggest that generic proofs may be promising to help students move from an empirical to a more deductive understanding. Further research is needed to investigate how working with generic proofs can support students' learning.

References

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

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.

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).

Christensen, R. H. B. (2022). *ordinal: Regression Models for Ordinal Data* (Version R package version 2022.11-16) [Computer software]. https://cran.r-project.org/web/packages/ordinal/ordinal.pdf

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). *Understanding the generality of mathematical statements: An experimental study at the transition from school to university*. Springer. https://doi.org/10.1007/978-3-658-43763-3

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

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Hanna, G. (1989). More than formal proof. *For the Learning of Mathematics*, *9*(1), 20–23.

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). American Mathematical Society.

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

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

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

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.

Kultusministerkonferenz (Ed.). (2012). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife: Beschluss der Kultusministerkonferenz vom 18.10.2012*.

Mejía Ramos, J. P., & Inglis, M. (2009). What are the argumentative activities associated with proof? *Research in Mathematics Education*, *11*, 77–78. https://doi.org/10.1080/14794800902732258

National Council of Teachers of Mathematics (Ed.). (2000). *Principles and standards for school mathematics*. National Council of Teachers of Mathematics.

Riley, K. J. (2003). *An investigation of prospective secondary mathematics teachers' conceptions of proof and refutations* [Doctoral dissertation, Montana State University]. https://scholarworks.montana.edu/xmlui/bitstream/handle/1/8385/31762103856330.pdf;sequence=1

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, *38*(3), 289–321.

Weber, K., & Mejía Ramos, J. P. (2015). On relative and absolute conviction in mathematics. *For the Learning of Mathematics*, *35*(2), 15–21.

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