Assessment of Mathematics Proof Construction: Preservice Mathematics Teacher Training and Upper Secondary School Students

Mohamad Waluyo*, Sutama, Nining Setyaningsih

Author Affiliations
Universitas Muhammadiyah Surakarta, Indonesia

Author Emails
**mw192@ums.ac.id*

Abstract. Constructing valid mathematical proofs is a fundamental yet persistent challenge for students. This study investigates the developmental trajectory of proof construction abilities among Indonesian students, spanning from upper secondary school (Grades 10–11) to undergraduate mathematics levels. Adopting a descriptive cross-sectional design, data were collected from 479 participants using open-ended proof construction tasks and a survey on proof perception. The results reveal a pervasive dominance of empirical proof schemes, where students rely on specific numerical examples rather than generalized logical arguments. Notably, formal proof construction was virtually non-existent at the secondary level, highlighting a significant pedagogical gap. While the frequency of formal proof attempts increased among university students, valid execution remained limited, indicating a disconnect between students' strategic knowledge and their normative understanding of "correct" proof. Furthermore, findings show a critical misalignment between perception and performance; students often recognized formal algebraic arguments as superior yet reverted to empirical methods in practice. The study concludes that the transition from inductive to deductive reasoning does not occur naturally and requires explicit instructional interventions to bridge the cognitive rupture between secondary calculation-based mathematics and tertiary formal reasoning.

Keywords: mathematical proof, proof construction, empirical proof scheme, secondary-tertiary transition, proof assessment, mathematics education.

INTRODUCTION

Mathematical proof is central to the discipline of mathematics, serving not merely as a tool for verification but as a primary means of communication and explanation within the mathematical community (Hanna, 2000; Knuth, 2002; Yackel & Cobb, 1996). A proof must be logically presented to effectively convey ideas and convince readers of an argument's validity. Consequently, the standard for acceptance is rigorous; even a single logical error within the argumentative chain can render a proof invalid. Despite its critical role, writing a well-structured and acceptable proof remains a persistent challenge for students ranging from secondary education to undergraduate levels (e.g., Lew & Zazkis, 2019; Sari et al., 2018; Stavrou, 2014; Weber, 2001).

To construct a proof successfully, students require not only relevant content knowledge but also the capacity to organize a logical and well-structured argument (Moore, 1994). However, the structure of a valid argument is not monolithic; it varies significantly depending on the proving method employed. A proof may require a direct approach (Leron, 1985), the use of mathematical induction (Stylianides et al., 2016), or a proof by cases (Aricha-Metzer &

Zaslavsky, 2019; Nardi & Knuth, 2017). This diversity in proof structures adds a layer of complexity for students, who often struggle to identify the appropriate framework for a given problem.

Due to the complexity and importance of this domain, mathematical proof has attracted considerable research interest. However, a substantial portion of the existing literature focuses on proof comprehension (Mejía-Ramos et al., 2012), the classification of student errors, or the teaching of specific proof techniques. There is comparatively less research dedicated to developing a robust assessment framework specifically for proof construction that tracks development across educational levels. While some studies examine university students' proof writing (Selden & Selden, 2003a, 2003b), few investigate the developmental trajectory of this skill as students transition from secondary school to university.

To contextualize the developmental trajectory examined in this study, it is necessary to understand the Indonesian educational landscape. Secondary education in Indonesia spans grades 7 through 12, divided into lower secondary (grades 7–9) and upper secondary (grades 10–12). The upper secondary level typically distinguishes between vocational schools and general senior high schools. Within the general track, students specialize in either Natural Sciences or Social Sciences. This study focuses specifically on students from the Natural Sciences stream, as their curriculum places a heavier emphasis on mathematics and science instruction compared to other tracks. Consequently, these students represent the cohort most likely to pursue university degrees in mathematics and related fields, making them the ideal population for analyzing the transition of proof construction skills from school to university.

Addressing the research gap within this specific context, the present study aims to develop an assessment instrument for mathematical proof writing and to analyze the progression of this skill. Furthermore, this study investigates potential factors influencing proof construction, including gender differences and students' perceptions of what constitutes a "correct" proof—an area where beliefs often diverge from formal mathematical standards (Harel & Sowder, 2007).

Based on these objectives, this research addresses the following questions: (1) How does students' ability to construct proof develop across grade and year levels? (2) Is there a significant gender effect on proof-writing ability? (3) How do students' perceptions of "correct proof" influence the type of proof they construct?

METHODOLOGY

Research Design and Participants

This study employs a descriptive cross-sectional design to assess and compare students' proof construction abilities across different educational levels (Creswell, 2014). The participants were selected using a purposive sampling technique to represent a continuum of mathematical training in the Indonesian context.

The secondary school cohort consisted of 120 students from Grades 10 and 11 specializing in Natural Sciences (IPA). This group was selected because the Natural Science curriculum provides the most rigorous mathematics foundation at the secondary level, serving as the primary pipeline for STEM-related university programs. The undergraduate cohort was drawn from the Mathematics Education program at Muhammadiyah Surakarta University, comprising 141 first-year, 118 second-year, and 100 third-year (and above) students. These university students represent individuals who have explicitly chosen to pursue advanced mathematics, thus allowing for an analysis of how proof writing skills evolve from general education to specialized training.

Instrument Development

The data collection utilized a two-part instrument designed to assess both students' perceptions of proof and their actual performance in constructing proofs. Part 1: Students' Perceptions. Before attempting the construction tasks, students completed a multiple-choice survey regarding their beliefs about "good" proofs. They were presented with

four distinct types of arguments for a mathematical problem: empirical (examples), narrative (verbal explanation), formal (algebraic), and pictorial. Students were asked to identify which argument would receive the best mark from their teacher. This design aligns with the framework established by Healy and Hoyles (2000), who distinguished between arguments students choose for themselves versus arguments they believe earn high grades. Part 2: Proof Construction Tasks. The core assessment consisted of two open-ended questions requiring students to construct their own proofs. The tasks were designed to be accessible to secondary students while remaining challenging enough to reveal the depth of understanding in university students.

Question 1 (Number Theory): "What do you think if two odd numbers are added? Is the result always odd or even? Prove your answer!"

Question 2 (Algebraic Structure): "Prove the statement: The multiple of three consecutive numbers is always a multiple of six."

These tasks were selected because they require a transition from empirical verification to generalized algebraic reasoning, a key indicator of proof maturity (Stylianides, 2009).

DATA ANALYSIS

Student responses were analyzed quantitatively and qualitatively. Two trained raters evaluated the proofs based on two dimensions:

1. Type of Proof: Responses were categorized into four distinct forms: Empirical, Narrative, Formal (Algebraic), and Others. This categorization draws upon Harel and Sowder's (1998) theory of proof schemes, distinguishing between inductive (empirical) and deductive (analytical) reasoning.
2. Correctness of Proof: The quality of the argument was scored using an analytic rubric ranging from 0 to 4:
   - 0: Not related / No response.
   - 1: Basic response with irrelevant information.
   - 2: Relevant information provided but lacks logical reasoning.
   - 3: Partial proof (relevant reasoning but incomplete or unclear).
   - 4: Complete proof with clear, valid reasoning.

To ensure the reliability of the assessment, Inter-Rater Reliability (IRR) was calculated. Cohen's Kappa was used for the categorical variable (Type of Proof), while the Intraclass Correlation Coefficient (ICC) was employed for the ordinal variable (Correctness Score). The interpretation of these coefficients followed the guidelines by Landis and Koch (1977) to determine the strength of agreement.

RESULTS AND DISCUSSION

Reliability of the assessment

There are several methods used to assess measurement reliability and in this study, the type of reliability used is interrater reliability as referred in (Gliner et al., 2009) that "When observation is the method of collecting data, then reliability must be established among the judges' scores to maintain consistency". To compute the reliability of proof correctness that students write, intraclass correlation coefficient (ICC) with two-way mixed effects model where people effects are random and measures effects are fixed, is performed. Meanwhile for measuring the reliability of students' proof type, Kappa statistics is used because the data are nominal.

TABLE 1. Interrater reliability score

	Type of proof (kappa)	The correctness of proof (ICC)	N
Question 1	0.883	0.914 (95% CI: 0.895-0.930)	380
Question 2	0.932	0.930 (95% CI: 0.915-0.943)	380

Cohen's κ was run to determine if there was agreement between two raters judgement on whether 380 students' proof writing are formal, empirical or narrative type. There was a good agreement between the two rater judgements for two questions, $\kappa_1 = .883$ and $\kappa_2 = .932$, $p < .0005$. Similarly, A high degree of reliability was found between students' correctness of proof measurements. The average measure ICC was .914 and .930 for question 1 and question

2 respectively with a 95% confidence interval from 0.895 to 0.930 ($F(379,379)= 11.636$, $p<0.001$) and 0.915 to 0.943 ($F(379,379)= 14.277$, $p<.001$).

Students' Proof Construction Patterns

There are two mode of assessment from students writing of mathematical proof namely type or form of proof and the validity or correctness of the writing. Overall the data from both question 1 and question 2, the dominan form of proof is empirical proof, with more than a half of total samples. The second biggest is formal form, then followed by narrative form and other form as the lowest.

TABLE 2. Distribution of students' answer by types of proof

type of proof	Question 1		Question 2	
	N	%	N	%
Formal (algebraic)	144	28,3	83	16,3
Narrative	58	11,4	23	4,5
Empirical	292	57,4	374	73,5
others	15	2,9	29	5,7

The second measurement is the score of the correctness. This variable is an ordinal variabel starts from 0 to 4. Generally, for all type of proof and both question, scores '3' get the highest percentage among other score except for formal proof form on the second question. In the formal proof for the second question, more students get score '2' (41%) rather than '3' (19%) or 4 (0%). The difficulties to answer the second question in a formal way is the main reason in this case.

TABLE 3. Crossectional data of type of proof and correctness

Scores	Empirical		Narrative		Formal (algebraic)	
	N	%	N	%	N	%
Question 1						
0	0	0	1	2	2	1
1	19	7	10	17	37	26
2	85	29	20	34	25	17
3	164	56	25	43	50	35
4	24	8	2	3	30	21
total	292	59	58	12	144	29
Question 2						
0	1	0	0	0	2	2
1	46	12	7	30	31	37
2	108	29	6	26	34	41
3	178	48	8	35	16	19
4	41	11	2	9	0	0
total	374	78	23	5	83	17

Development of writing proof across the year level

The ratio for each grade is not equal. Students grade 10 and 11 only a half than each year of undergradute students. The highest number of the empirical form of proof is performed by first year students accounted for 97 and 125 for question 1 and 2 respectively. Meanwhile no formal proof is shown by students in secondary school level for both grade and question. The absent on formal proof is contrast with their performance in empirical proof form which is shown a high proportion (higher than 70%). For more detailed information is listed in the table 4.

TABLE 4. Distribution of proof type toward grades

	empirical		narrative		formal		no proof		total
	N	%	N	%	N	%	N	%	
Question 1									
10 grade	63	91	4	6	0	0	2	3	69
11 grade	46	77	3	5	0	0	11	18	60
1st year	97	69	24	17	20	14	0	0	141
2nd year	43	36	15	13	59	50	1	1	118
3rd year	43	36	12	10	65	54	1	1	121
Question 2									
10 grade	58	84	5	7	0	0	6	9	69
11 grade	42	71	2	3	0	0	15	25	59
1st year	125	89	4	3	7	5	5	4	141
2nd year	72	61	9	8	37	31	0	0	118
3rd year	77	64	3	2	39	32	2	2	121

Developmental Trajectory

The cross-sectional data highlights a distinct developmental trend. As illustrated in the data, the reliance on empirical proofs is highest among Grade 10 students (91% for Q1) and gradually declines as students progress through university. Conversely, the production of formal proofs—which is virtually non-existent at the secondary level (0%)—begins to emerge in the first year of university and peaks in the second and third years. However, even at the undergraduate level, the transition is not absolute; a significant portion of third-year students still reverted to empirical or narrative methods when faced with simpler number theory problems.

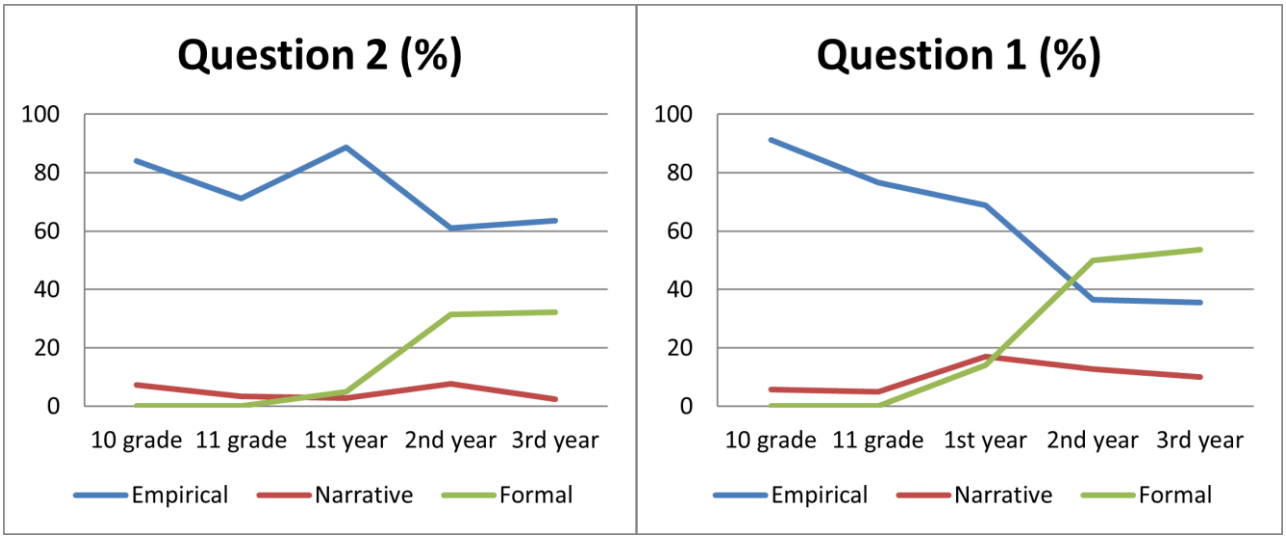

FIGURE 1. Trend of proof type among the grades

As the important form of proof in the mathematics world, the score formal form of proof from question 2 is shown below. The bar chart explain that the students from second and third year get score ‘3’ the most whereas the first student only get score ‘1’.

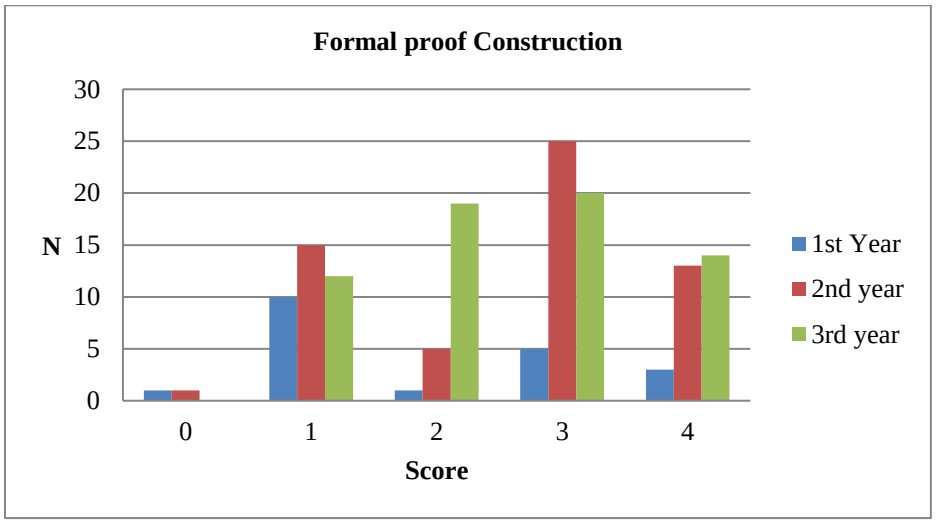

FIGURE 2. Formal proof score for each grade

Gender Differences

The statistical analysis shows that there is association between gender and type of proof used by students. It is shown by Chi-Square value by 9.465 and Likelihood Ratio value by 8.467 (p<0.05). The Association between two variables is quite low indicated by the value of Phi and Cramer's V by 0.136. The likelihood of female students will write an empirical, narrative, and formal are 56%, 12%, and 30% respectively. While male students shows 61%, 10% and 23% for the same category with female students.

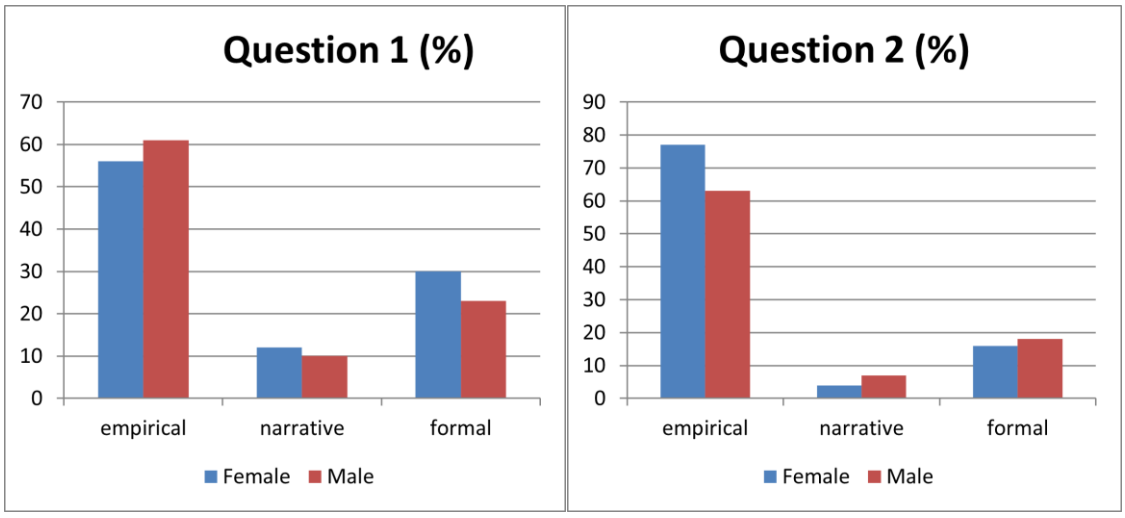

FIGURE 3. Gender differences in proof types

Students' assumption toward proof type

Before students writing their own proof they are asked about what kind of proof that they think will get the best mark. The resume of students' answer is described from the pie chart. Generally speaking, students of 10, 11 and first year students think that the narrative form of proof is the best way to write a proof by 41% and 49%. Contrary, the second and third year students think the best way to write proof is the formal one, accounted for 70% students who choose it. The rationale of this situation is that the 1st year students is starting their undergraduate mathematics class while the second year students have developed their understanding for a year previously. So that the way of thinking of first year students is more similar to secondary school students rather than the second year students or above.

Perception vs. Performance

There is a significant association between students' perceptions of "good proof" and the type of proof they constructed ($\chi^2 = 67.86$, $p < .001$). Interestingly, a misalignment exists: while 70% of second and third-year students correctly identified that a formal proof would receive the best mark, not all of them were able to produce one. Among students who actually *wrote* an empirical proof, a large portion (approx. 40%) actually believed that a narrative or formal proof was superior, yet they resorted to empirical methods presumably due to a lack of strategic knowledge to construct the formal argument.

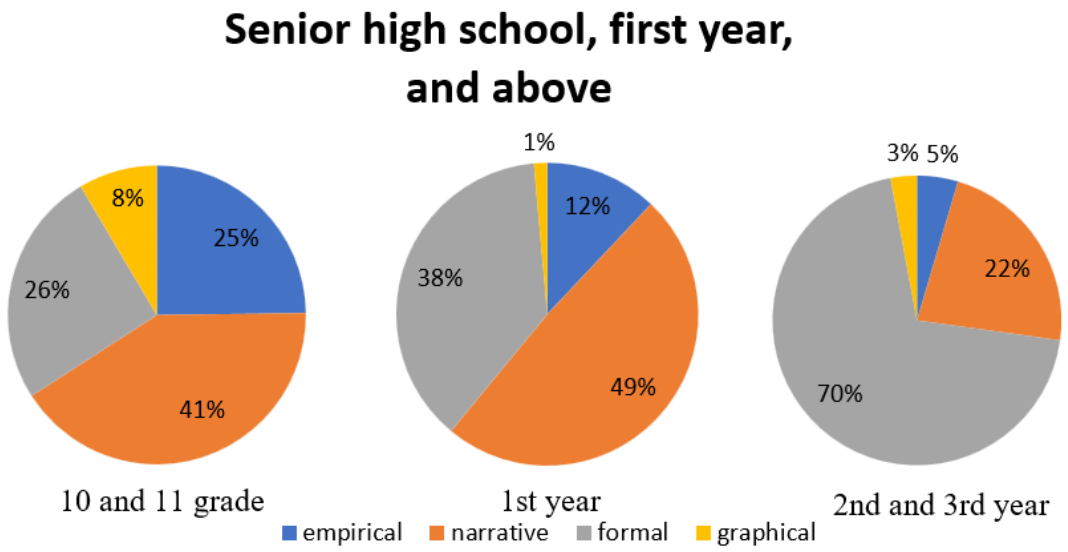

FIGURE 4. Proportion of proof type on each grade

The findings of this study provide critical insights into the developmental landscape of mathematical proof construction in Indonesia. The most pervasive finding is the persistent dominance of empirical proof schemes—the tendency to validate mathematical statements using specific numerical examples rather than generalized logical arguments. This aligns with the framework established by Harel and Sowder (1998), who described the "empirical proof scheme" as a cognitive stage where conviction is derived from the verification of cases. In our study, even undergraduate students, who are expected to operate within the "analytical proof scheme," frequently reverted to empirical verification. This corroborates the findings of Weber (2010) and Stavrou (2014), who noted that students often view example-checking not just as exploration, but as sufficient evidence for proof.

A key contribution of this study is the identification of the developmental gap between secondary school and university. The total absence of formal proof construction among Grade 10 and 11 students suggests a systemic issue in secondary mathematics instruction, which appears to prioritize algorithmic calculation over deductive reasoning. This supports Moore's (1994) conclusion that the transition to formal proof involves a massive cognitive leap (a "cognitive rupture") that students are ill-prepared for. Unlike the gradual development observed in some international

contexts, the Indonesian data suggests a "shock" transition: students are introduced to formal proof rigor only upon entering university, resulting in the high failure rate of valid formal proof construction among first-year students.

Furthermore, this study illuminates the complex relationship between competence and conception. Consistent with Healy and Hoyles (2000), we found a disconnect between what students *value* and what they *produce*. Students demonstrated a clear preference for formal algebraic arguments when asked what would earn the best grade, yet they frequently constructed empirical arguments themselves. This indicates that the primary deficit is not necessarily a lack of understanding of what a proof *should look like* (normative understanding), but rather a lack of strategic knowledge (Weber, 2001) to instantiate that understanding. They know algebra is the "gold standard," but they lack the procedural fluency to translate their informal ideas into formal algebraic syntax.

Regarding gender, while a statistically significant association was found between gender and proof type, the effect size was weak (Cramer's V = 0.136). This finding adds nuance to the debate on gender in mathematics. Unlike Recio and Godino (2001) who found distinct institutional obstacles, our data suggests that the struggle with proof construction is largely a cognitive and pedagogical issue shared by both genders, rather than a gender-specific deficit. Female students showed a slightly higher propensity for narrative and formal structures compared to males, but both groups were overwhelmingly dominated by empirical approaches.

In summary, the ability to construct formal proofs does not develop naturally with age but requires explicit pedagogical intervention. The "natural" tendency of students is to remain in the empirical stage, and without specific training in the *structure* of arguments—not just the *content* of mathematics—students struggle to cross the divide from inductive to deductive reasoning.

Implication and Conclusion

The assessment of the proof construction and become a limitation in this study is an analysis only based on students' document on writing proof. Inter rater reliability is used to evaluate the measurement reliability. This study shows that the most form of proof written by students is empirical one. Moreover the secondary students have not developed their ability on constructing a formal proof. This may happen because of the lack of opportunity that teacher give to students to think and reasoning some mathematics phenomenon. The students is tend to think to use empirical case or example and think inductively rather than think deductively. This ideas is supported by the few number of students who use the narrative form. The narrative form can be a bridge where students can try before they are able to write in a formal way. Even for undergraduate level, the ability to construct the formal proof is still low. From 144 students who write formal form, only 21% can get score 4 which means they can write a complete formal proof and 35% get score 3 which means they are able to write proof with some reasoning is missing.

In addition, the gender is associated with the form of proof students write but it does not really show the gender differences in this area. Lastly, the students perception about the best form of the proof is shown a significant association with the proof form students written. Surely this understanding comes from how much the students read a proof from mathematics textbook or how their teacher explain them.

REFERENCES

1. Aricha-Metzer, I., & Zaslavsky, O. (2019). The nature of students' productive and non-productive example-use for proving. *The Journal of Mathematical Behavior*, *53*, 304–322. https://doi.org/10.1016/J.JMATHB.2017.09.002
2. Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th (ed.)). SAGE Publications.
3. Gliner, J. A., Morgan, G. A., & Leech, N. L. (2009). *Research Methods in Applied Settings*. Routledge. https://doi.org/10.4324/9780203843109

4. Hanna, G. (2000). Proof, explanation and exploration: An overview. *Educational Studies in Mathematics*, *44*(1), 5–23.
5. Harel, G., & Sowder, L. (1998). Students' Proof Schemes: results from exploratory studies. In A. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research on Collegiate Mathematics Education* (pp. 234–283). M.M.A. and A.M.S. http://www.sciepub.com/reference/178810
6. 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* (pp. 805–842). Information Age Publishing.
7. Healy, L., & Hoyles, C. (2000). A Study of Proof Conceptions in Algebra. *Journal for Research in Mathematics Education*, *31*(4), 396.
8. Knuth, E. J. (2002). Secondary School Mathematics Teachers ' Conceptions of Proof. *Journal for Research in Mathematics Education*, *33*(5), 379–405.
9. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310
10. Leron, U. (1985). A direct approach to indirect proofs. *Educational Studies in Mathematics*, *16*(3), 321–325. https://doi.org/10.1007/BF00776741
11. Lew, K., & Zazkis, D. (2019). Undergraduate mathematics students' at-home exploration of a prove-or-disprove task. *The Journal of Mathematical Behavior*, *54*, 100674. https://doi.org/10.1016/J.JMATHB.2018.09.003
12. 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. https://doi.org/10.1007/s10649-011-9349-7
13. Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics*, *27*(3), 249–266. https://doi.org/10.1007/BF01273731
14. Nardi, E., & Knuth, E. (2017). Changing classroom culture, curricula, and instruction for proof and proving: how amenable to scaling up, practicable for curricular integration, and capable of producing long-lasting effects are current interventions? *Educational Studies in Mathematics*, *96*(2), 267–274. https://doi.org/10.1007/s10649-017-9785-0
15. Recio, A. M., & Godino, J. D. (2001). Institutional and personal meanings of mathematical proof. *Educational Studies in Mathematics*, *48*(1), 83–99. https://doi.org/10.1023/A:1015553100103
16. Sari, C. K., Waluyo, M., Ainur, C. M., & Darmaningsih, E. N. (2018). Logical errors on proving theorem. *Journal of Physics: Conference Series*, *948*(1), 12059. https://doi.org/10.1088/1742-6596/948/1/012059
17. Selden, A., & Selden, J. (2003a). Validations of Proofs Considered as Texts: Can Undergraduates Tell Whether an Argument Proves a Theorem? *Journal for Research in Mathematics Education*, *34*(1), 4. https://doi.org/10.2307/30034698
18. Selden, A., & Selden, J. (2003b). Validations of proofs written as spoken. *The Journal of Mathematical Behavior*, *22*(1), 5–32. https://doi.org/10.1016/S0732-3123(03)00002-6
19. Stavrou, S. G. (2014). Common Errors and Misconceptions in Mathematical Proving by Education Undergraduates. *Issues in the Undergraduate Mathematics Preparation of School Teachers*, *1*. https://eric.ed.gov/?id=EJ1043043
20. Stylianides, G. J. (2009). Reasoning-and-proving in school mathematics textbooks. *Mathematical Thinking and Learning*, *11*(4), 258–288. https://doi.org/10.1080/10986060903253954
21. Stylianides, G. J., Sandefur, J., & Watson, A. (2016). Conditions for proving by mathematical induction to be explanatory. *The Journal of Mathematical Behavior*, *43*, 20–34. https://doi.org/10.1016/J.JMATHB.2016.04.002
22. 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
23. 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
24. 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