"Is It Valid or Not?": Pre-Service Teachers Judge the Validity of Mathematical Statements and Student Arguments

Zulfiye Zeybek Simsek [1]*

[1] Department of Mathematics and Science Education, Tokat Gaziosmanpasa University, Tokat, TURKEY
* Corresponding author: zulfiye.zeybek@gop.edu.tr

Abstract:
There is a wide recognition that reasoning abstractly, constructing arguments, or critiquing arguments should be an important educational goal in the mathematical experiences of all students in the standards for school mathematics. Seeing these standards as an essential element for developing deep mathematical understanding; however, call for a strong knowledge of proof for teachers. Thus, the purpose of this study is to investigate how pre-service middle school teachers (PSMTs) decide whether a presented mathematical statement is true or false and how they verify student arguments presented for these statements. 50 PSMTs participated in the study. Individual interviews were conducted with 7 PSMTs to further delve into the verification processes of the PSMTs. The results of the study demonstrated that meeting the expectations of the current standards is not an easy feat by documenting that most of the PSMTs struggled with evaluating mathematical tasks and constructing arguments.
Keywords: counterexamples, empirical arguments, geometry, verifications

INTRODUCTION

Standards for school mathematics have increasingly focused on the importance of student reasoning abstractly, constructing viable arguments, critiquing others' reasoning, and attending to precision across their K-12 experience (Ministry of National Education [MEB], 2018; National Governors Association Center/Council of Chief State School Officers [NGA/CCSSO], 2010; National Council of Teachers of Mathematics [NCTM], 2000; Stylianides, Bieda, & Morselli, 2016). While there has been a strong emphasis in various policy documents for the inclusion of constructing and critiquing mathematical arguments in all grades, these documents are generally thin in describing how to teach proofs in this vision and what this requires for teachers. In reality, the place of proof in mathematics classrooms is far from that vision (Buchbinder & McCrone, 2020). Stylianides et al. (2017) used the phrase classroom-based interventions in the area of proof to describe interventions designed to improve understanding or use of proof as captured by these standards at any grade level. Stylianides et al. (2017) stated, "the number of such studies is small and acutely disproportionate to the number of studies that have documented problems of classroom practice [in the area of proof] for which solutions are sorely needed" (p. 253).

Seeing these standards as an essential element for developing deep mathematical understanding and making it a crucial element of students' mathematical experiences obviously call for a strong mathematical knowledge for teachers (Buchbinder & McCrone, 2020; Lesseig, 2016; Mata-Pereira & da Ponte, 2017). Teachers are expected to decide what conjectures proposed by students or textbooks are worth pursuing, to judge whether students have the requisite background such as key definitions or

theorems to produce a valid argument for the conjectures, and to check what implicit or explicit warrants support the argument (Dawkins & Weber, 2017; Mata-Pereira & da Ponte, 2017; Stylianides, 2007).

Yet, research indicates that teachers do not have experience supporting mathematical argument in their classroom and, in fact, are not sure what mathematical proof is, or should look like in a classroom setting (Dawkins & Weber, 2017; Mata-Pereira & da Ponte, 2017; Stylianides, Bieda, & Morselli, 2016). All these suggest that teachers should have more experience with evaluating mathematical grounds on which claims or arguments could be accepted or rejected in classrooms.

The purpose of this study is to investigate how pre-service middle school teachers (PSMTs) decide whether a presented mathematical claim is true or false and then how they verify student arguments presented for these claims. It is hypothesized that using statements that are not always hold true can better illuminate the ability of pre-service teachers' verification of mathematical claims. Furthermore, it is also hypothseized that using arguments that use invalid modes of reasoning can better illuminate the parts of pre-service teachers' conceptions of what constitutes a proof. More specifically, this study is guided by the following two research questions:

1. How do pre-service middle school teachers (PSMTs) verify given statements that are not always true?
2. How do PSMTs judge the validity of given arguments?

Clarification of the Terms

Ellis et al. (2012) mentioned that the development of new knowledge passes through several stages of which the construction of a mathematical argument is considered as the last stage. Earlier stages of this complex processes of developing new knowledge typically includes exploration of particular cases, generation (or refinement) of conjectures, and then attempts to develop arguments that may translate into a proof (Ellis, Bieda, & Knuth, 2012; Stylianides, 2008; Zazkis, Liljedahl, & Chernoff, 2008). This conceptualization is thought to be useful for comprehending PSMTs' ability to verify mathematical statements and then to evaluate students' arguments. Thus, the definition of these concepts will be addressed in this section so that the definitions could be used to shed light onto the participants' verification and evaluation processes.

Conjecturing involves reasoning about mathematical relationships to develop statements that are tentatively thought to be true but are not known to be true (Lannin, Ellis, & Elliott, 2011, p. 13). Stylianides (2008) defined conjecture as a reasoned hypothesis about a general mathematical relation based on incomplete evidence (p. 11). Stylianides (2008) described that the term 'reasoned' was used to emphasize the non-arbitrary character while the term 'hypothesis' was used to indicate a level of doubt in the definition. Similarly, Harel and Sowder (2007) defined conjecture as an observation made by a person who has doubts about its truth. These elements of doubt and non-randomness; therefore, are essential components of conjecturing (Lannin, Ellis, & Elliott, 2011). Similarly, Canadas and colleagues highlighted the non-arbitrary character of conjecturing by arguing that conjecturing involves the following seven stages: (1) observing cases, (2) organizing cases, (3) searching for and predicting patterns by imagining that such patterns might apply to the next unknown case, (4) formulating a conjecture about all possible cases based on empirical facts, (5) validating the conjecture for a specific case through some independent method, (6) generalizing the conjecture, and (7) justifying the generalization (2007, p. 63).

Although Canadas and colleagues (2007) conceptualized justifying as the last step of conjecturing, justifications could be constructed separately from conjecturing processes. Justifying, in a general sense, is a coordinated collection of reasons that an individual provides for believing that a mathematical statement is true (Czocher & Weber, 2020, p. 51). Thus, justifying includes any attempts to use mathematics to convince oneself or others, regardless of whether the argument is complete or would be

accepted as a proof. Indeed As Czocher and Weber (2020) stated, all proofs are justifications but not all justifications could be counted as proofs. Then the question remains: what properties a justification must possess to qualify as a proof (Cirillo, Kosko, Newton, Staples, & Weber, 2015; Czocher &Weber, 2020).

Although there is not a consensus among mathematics educators and mathematicians as to what a mathematical proof should look like (Czocher & Weber, 2020), Stylianides (2007) proposed a definition of a proof as follows: "Proof is a mathematical argument, a connected sequence of assertions for or against a mathematical claim, with the following characteristics:

1. It uses statements accepted by the classroom community (set of accepted statements) that are true and available without further justification;
2. It employs forms of reasoning (modes of argumentation) that are valid and known to, or within the conceptual reach of, the classroom community; and
3. It is communicated with forms of expression (modes of argument representation) that are appropriate and known to, or within the conceptual reach of, the classroom community." (Stylianides, 2007, p. 291)

Thus, justifications should encapsulate several characteristics to be considered as a mathematical proof as follows: the arguments accepted as proofs use true statements, valid forms of reasoning, and appropriate forms of expression, whereby the terms "true", "valid" and "appropriate" should be conceptualized as part of classroom community. This conceptualization of proof is thought to be helpful to evaluate how the PSMTs verify the statements that are not always true as well as student arguments, which will be addressed next.

METHODS

Participants

Participants of the study were 50 pre-service middle school teachers (PSMTs) who are certified to teach mathematics in grades 5 through 8. The participants were juniors at a public university in Turkey when the data was collected. They completed several mathematics courses and two consecutive mathematics methods courses prior to the study. The PSMTs enrolled in a mathematics education course, which was taught by the author of this study in the spring semester of 2019. However, it should be noted here that the data was not collected as a part of the course. Instead, the data was collected by the end of the course and all participants were informed that the participation to the study was voluntary. 50 PSMTs volunteered to complete a questionnaire by the end of the semester. Among these 50 PSMTs, semi-structured individual interviews were conducted with 7 PSMTs. All these 7 PSMTs volunteered to further participate to the study.

Tasks

The tasks used in the study were only true for some cases but not for all. Several researchers have called for increased emphasis on such tasks for instructional purposes (Ball, Hoyles, Jahnke, & Movshovitz-Hadar, 2002; Brown, 2014; Stylianides & Stylianides, 2009). For instance, Harel and Sowder (2007) argued that using example-based reasoning should be cautioned due to its tentative nature; therefore, employing patterns that do not always hold true could be helpful to get students recognize the limitation of using empirical-based arguments. Employing patterns that would not always hold true could also be used to justify the norm that students should only employ inferential techniques that are valid (or explicitly justify why a technique is valid in this situation). Stylianides and Stylianides (2009) and Brown (2014) found some success with employing such patterns to teach students the dangers of empirical induction, non-generalizable deduction, and diagrams. Three tasks that were only hold true for some cases were used to investigate the PSMTs' ability to verify the tasks (see **Table 1**).

Table 1. Tasks and arguments used in the study

|	
Task 1: Students have been working on an area and perimeter task. One student—Nermin— proudly proclaims that she discovers a new math conjecture: "whenever the perimeter of a rectangle increases, its area also increases". Do you think her claim is true or false? If so, how would you prove it?	Task 1 a: You think that it is a good opportunity to engage students with proofs and ask them to prove whether Nermin's claim is true or false. Nermin provides an argument as follows:
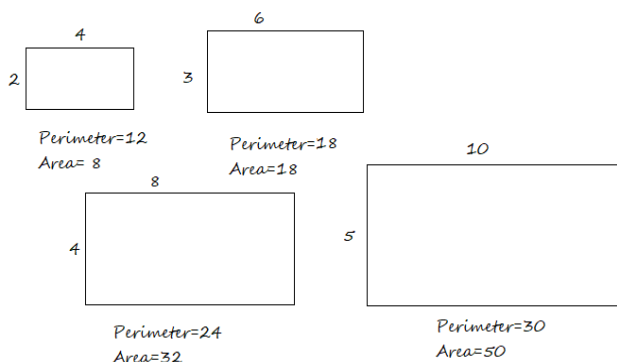
Would you accept her argument as a proof? Why? Why not?

|	
Task 2: Ali claims that "if the vertices of a quadrilateral are on the consecutive sides of a rectangle, then the area of the quadrilateral inside is always half of the area of the rectangle". Do you think Ali's claim is true or false? If so, how would you prove it?	Task 2a: Ali provides an argument as follows: If you take four points on the sides of a rectangle as follows, there are eight congruent triangles formed. Since four of these triangles are inside of the inner quadrilateral, the area of the quadrilateral is half of the area of the rectangle.
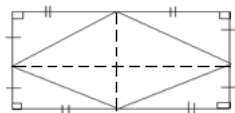
Would you accept his argument as a proof? Why? Why not?

Task 3: "At least one of the diagonals cuts the area of a quadrilateral in half" Do you think that this claim is true or false? How would you prove it? (adapted from: Ball, Hoyles, Jahnke & Movshovitz-Hadar, 2002).	Task 3a: Leyla: "Diagonals cut a square in two congruent triangles so that the areas will be the same. If we fold a square along its diagonal like this, we can see that the areas of the triangles are the same. We can do the same for a rectangle, parallelogram, and a rhombus. So, it is true."
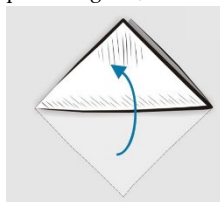
Would you accept her argument as a proof? Why? Why not?

The tasks were in the conceptual reach of the PSMTs, yet they still provided a productive struggle since the tasks only held true for some cases and required a justification. Three student arguments for the tasks were also used to investigate the PSMTs' processes of evaluating mathematical arguments. The presented student arguments were designed as empirical arguments that purported to justify the statements for a subset of the classes covered; therefore, fell short of being accepted as mathematical proofs (see Harel & Sowder, 2007 for details). Studies have shown that students at all levels have difficulty recognizing universally and existentially quantified statements (especially when the quantifier is implicit) and struggle to understand that a universally quantified statement must be proved for all elements in the domain, and fail to recognize the limitation of relying on supportive examples for proving universal statements (Buchbinder & Zaslavsky, 2019). Therefore, using empirical arguments is thought to be useful for investigating how the participants evaluate student arguments.

Data Collection Process

The participants were administered a questionnaire which consisted of eight open ended questions with several sub questions during 75 minutes by the end of the spring semester of 2019. The questionnaire consisted of two parts. The first part contained the tasks that the PSMTs were asked to verify and then to provide a justification while the second the part included hypothetical student arguments. The PSMTs were instructed to complete the first part and then to move to the second part. Three of the tasks were analyzed in this study (see **Table 1** for the tasks employed in the study). The PSMTs were informed that their responses would not be graded and would only be used for educational purposes to ensure that they reflected their own thoughts comfortably in their responses. 7 PSMTs volunteered to further participate to the study. The PSMTs were interviewed individually among 30-45 minutes and were asked to elaborate more upon their responses to the three questionnaire questions during the individual interviews. The individual interviews were recorded by a video camera. The video camera was positioned in such a way that the participants' gestures, written responses, and drawings were captured. The individual interviews took place in an office where only the interviewer and the interviewee were present. All the papers that the participants used during the interviews were collected by the interviewer for data analysis.

Data Analysis

The data analysis started with transcribing the individual interviews and reviewing the PSMTs' responses to the three questions. A constant comparative method (Glaser & Strauss, 1967) was employed to construct a coding scheme as follows: (1) the author independently reviewed all the responses and created an initial coding scheme depending upon the related literature and the related definitions mentioned previously; (2) two graduate students, who are familiar with the literature related to reasoning-and-proving, and the author compared the descriptions of the codes in the preliminary coding scheme with the sample of responses to see whether the features of the responses captured by the codes or indicated any mismatches with the codes that could lead to the generation of new codes or adjustment of existing codes. After finalizing the coding scheme as displayed in **Table 2** by collaborating with the coders, coding of the data process started and occurred in two steps.

Table 2. Coding framework

Verifying the Statements	Decision Support	Evaluating Student Arguments	Reasons for Decisions
Correct	Valid Counterexample	Proof	Valid/Mathematical
	Invalid Counterexample		Appropriate for Student Level
	No /Unfinished Counterexample		Other Reasons
In Correct	Incorrect Inference	Not Proof	Not General
	Empirical Argument		Invalid/Not Mathematical
	No/Unfinished Argument		

In the first step, the PSMTs' responses to the tasks were coded in two categories as *Correct*—if the PSMTs were able to realize that the tasks were not true for all cases—and *Incorrect*—if the PSMTs thought that the tasks were true for all cases. Later, how the PSMTs' attempted to support their decisions was coded from a mathematical point of view. If the PSMTs correctly identified that the tasks were not true (Correct Category), then they were expected to provide an example that succeed in refuting the statement, which was coded as *Valid Counterexample*. If the PSMTs provided an example that failed to refute the statement, then their responses were coded as *Invalid Counterexample*. *No /Unfinished Counterexample* category, on the other hand, included all the responses that had no counterexample constructed or unfinished attempts to provide a valid counterexample.

If the PSMTs thought that the tasks were true for all cases (Incorrect Category), they were then expected to provide an argument to support their decisions. Given that the presented tasks included the statements that were only true for some cases but not for all, the PSMTs were expected to construct a justification rather than a mathematical proof. Analyzing their constructed justifications, if the PSMTs provided an invalid general argument that used a sequence of assertions that refer to all cases in the

domain of the statement but one or more of these assumptions used in the argument were built upon an incorrect mathematical inference, their responses were coded as *Incorrect Inference*. Incorrect Inference arguments, therefore, fail to meet the criterion of employing true sets of statements—the first criterion of the definition implemented in the study. The term "inference" was used instead of "conjecture" intentionally for two reasons: (1) The definitions of conjecture mentioned previously highlighted that conjecturing involves reasoning about mathematical relationships by observing, organizing cases, and then formulating the relationship that thought to hold true for all possible cases. Therefore, conjecturing involves non-arbitrary hypothesis. Since conjecturing was not one of the purposes of the study, the PSMTs' processes of observing, organizing cases and then formulating relationships were not evident in the study. Instead, it rather seemed like the PSMTs were formulating an invalid mathematical relationship based on their insights or previous knowledge given that there were no signs of investigating and/or organizing different cases in their responses. (2) The definition of conjecturing included an element of doubt in its nature. However, the PSMTs in this study did not mention any signs of doubt in their arguments. Rather, they seemed very confident in their inferences so that they did not attempt to further justify them. In addition to Incorrect Inference category, if the PSMTs provided an argument that purported to show the truth of the mathematical statement by validating the statement in a proper subset of all possible cases covered by the statement, their responses were coded as *Empirical Argument*. Thus, Empirical Arguments fail to meet the criterion of modes of argumentation by employing an invalid form of reasoning. *No/Unfinished Argument* included all the responses that were incomplete or no response at all. All the irrelevant arguments that were constructed to justify the tasks were also coded in No/Unfinished Argument category.

In the second step of the analysis process, the PSMTs' responses to the presented student arguments were coded in two categories as Proof—if the PSMTs thought that the student arguments could be classified as mathematical proofs—and Not Proof—if the PSMTs thought that the student arguments could not be classified as mathematical proofs. In the Proof category, the PSMTs responses were coded in one of the following three categories: Valid/ Mathematical, Appropriate for Student Level, and Other Reasons. The responses that considered the modes of reasoning used in the student arguments as valid and/or mathematical were coded as Valid/Mathematical. If the responses highlighted that the employed modes of reasoning or modes of representation was appropriate for middle grade standards and students, then these responses were coded in Appropriate for Student Levels. All the other responses that did not mention employed modes of reasoning as valid or connected the argument to students' levels of thinking were coded in Other Reasons category. In Not Proof category, the PSMTs responses were either coded as Not General or Invalid/Not Mathematical. Not General category included all the responses that highlighted the fact that the arguments did not guarantee the truth of the assertion for all cases in the domain of the statements. Invalid/Not Mathematical category, on the other hand, included all the responses that mentioned the limitation of the arguments as employing an invalid method for proving.

Given that the PSMTs were asked to evaluate the student arguments presented and to state their reasons — not restricted to provide only one reason—for their evaluations, the PSMTs sometimes provided more than one reason. In that case, the first reason that was stated by the PSMTs was accepted as their primary reason and were coded. Focusing only on the PSMTs' primary reasons while evaluating the presented student arguments as described above consisted of two reasons as follows: (1) reflecting the PSMTs' primary reasons since they thought to be important and should be more elaborated and (2) ensuring that the paper being more concise by displaying important piece of the data instead of all the data collected. Two graduate students coded a random sample of 20% of the PSMTs' responses. The coders reached an agreement on 85% of these codes, and all disagreements were resolved through discussion. In the results section next, the PSMTs' responses that belonged to each category of the coding scheme will be displayed and described.

RESULTS

This section will be organized around the two research questions. First, the results related to the PSMTs' ability to verify given the statements to be true or false will be presented. Later, the results about in what ways PSMTs judge the validity of presented student arguments will be shared.

Verifying Mathematical Tasks and Justifying Decisions

The results of the PSMTs verification of mathematical statements and then providing justifications to support their decisions are displayed cumulatively in **Table 3**. As can be seen in **Table 3** for Task 1, most of the participants (36 PSMTs) failed to recognize that the task was not true for all cases. Only 14 PSMTs were able to recognize that the task would not always hold true and coded in correct category. 13 PSMTs, who recognized that the task was not always true, were also able to provide a valid counterexample that refuted the task while 1 PSMT failed to provide a valid counterexample.

**Table 3.** Results of verifying and justifying the statements that are not always true

	Correct			Incorrect		
	Valid Counterexample	Invalid Counterexample	No / Unfinished Counterexample	Incorrect Inference	Empirical Argument	No/Unfinished Argument
Task1	13	1	-	29	7	-
Task2	14	2	14	4	14	2
Task3	39	3	-	1	6	1

29 PSMTs believed that to increase the perimeter of a rectangle, at least one side should be increased in length while the other side should be kept the same (or increased as well). Thus, they argued that the area of the rectangle should increase as a result. These responses were coded as Incorrect Inference for Task 1. A sample of these responses is displayed in **Figure 1**. The PSMT argued that increasing the perimeter of a rectangle by a certain amount—k—requires increasing one of the sides of the rectangle by k/2 while keeping the other side the same.
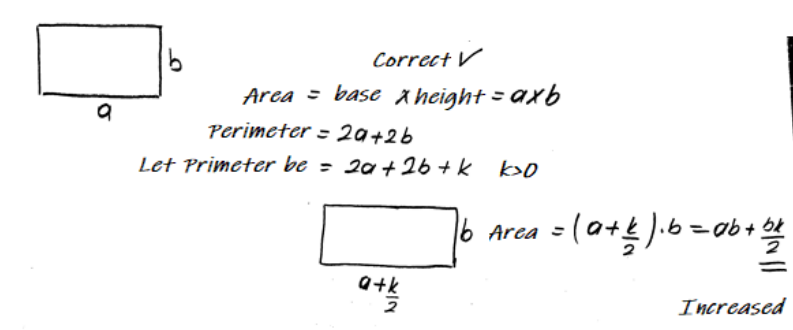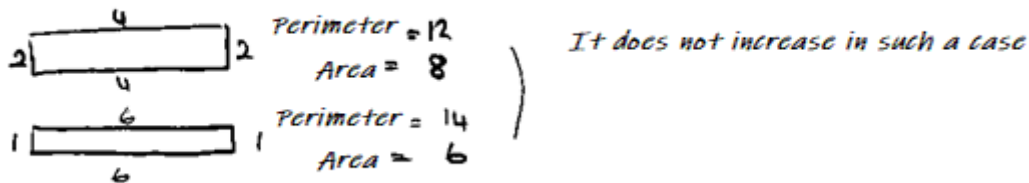

Figure 1. An argument that was coded as Incorrect Inference for Task 1

Although the numbers of the PSMTs were not as high, 7 PSMTs argued that the task held true and they provided an empirical argument to support their decision for Task 1. A sample of these responses is displayed in **Figure 2**. The PSMT drew a general conclusion based on a particular case—rectangles with the side lengths of 4 by 6 and 6 by 8.
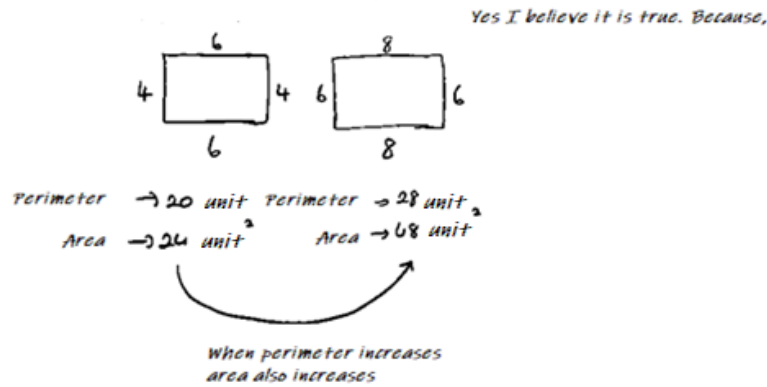
Figure 2. An argument that was coded as Empirical Argument for Task 1

13 PSMTs were not only able to recognize that the statement was not always true, but they were also able to provide a valid counterexample (see **Figure 3**).

Figure 3. An argument that was coded as Valid Counterexample for Task 1

For Task 2, while 30 PSMTs were able to recognize that the task was not true for all cases (Correct Category), 20 PSMTs failed to recognize that the task was only true for a subset of the classes covered by the statement (Incorrect Category). Out of these 30 PSMTs, who correctly evaluated the task, 14 PSMTs were able to provide a valid counterexample while 14 PSMTs provided no counterexample at all or failed to complete their counterexamples. For instance, the PSMT in **Figure 4** attempted to construct a general counterexample by selecting four random points on the side lengths of the rectangle and assigning different variables to the side lengths to signify the randomness. Later, he attempted to calculate the areas of the polygons to justify that the statement would not hold true. However, the PSMT failed to calculate the areas of the polygons formed inside of the rectangle correctly due to lengthy calculations. The question mark that he put by the end of his response may demonstrate that he got stuck by the lengthy calculation and failed to complete his counterexample.
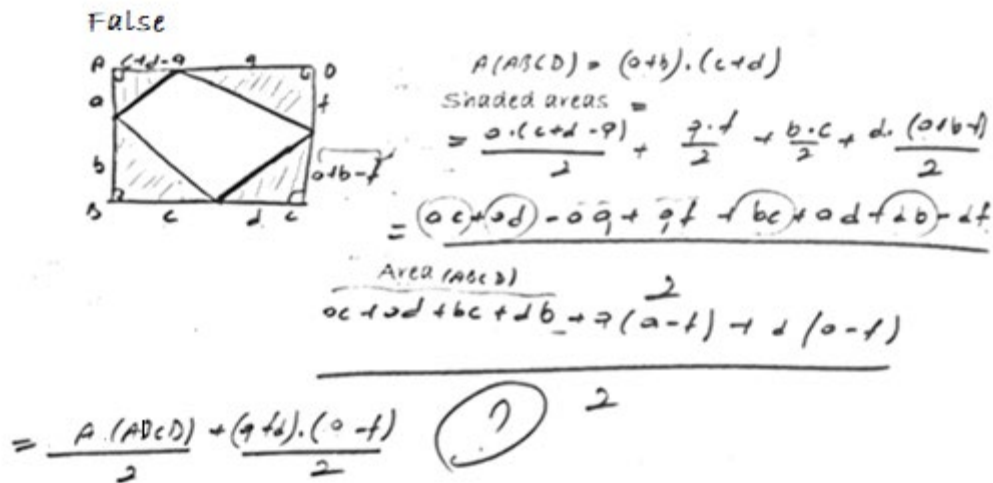
Figure 4. An argument that was coded as Unfinished Counterexample for Task 2

Only 4 PSMTs provided arguments that were coded as Incorrect Inference for Task 2. For instance, the PSMT in **Figure 5** made a logical flaw by arguing that in a right trapezoid, the area of the triangle formed by connecting two vertices with a vertex taken on the right side of the trapezoid is equal to the sum of the areas of the other two triangles. However, this assumption would only be true if the vertex taken on the right side of the trapezoid was the midpoint. Although the PSMT attempted to justify the task for all possible cases covered by the statement, his argument was built upon an incorrect inference and did not provide further justification for why it might be the case.
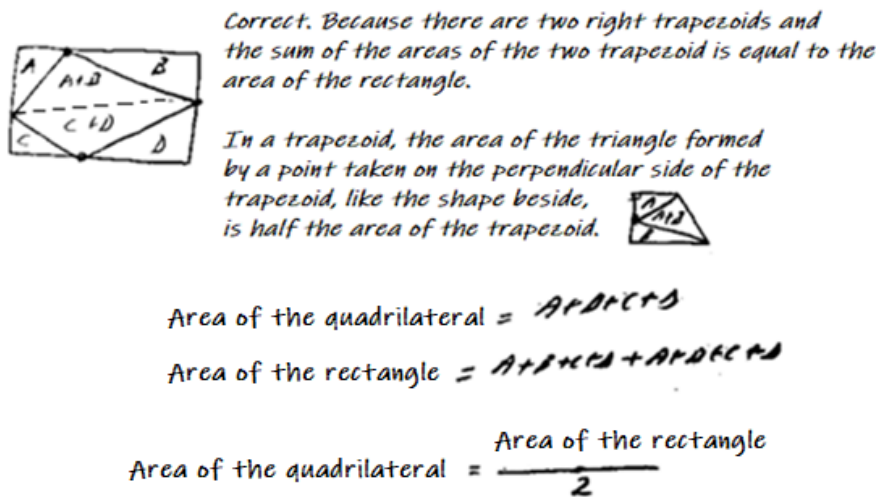

Figure 5. An argument that was coded as Incorrect Inference for Task 2

The numbers of empirical arguments constructed for Task 2 is higher than the other tasks. The PSMTs, who constructed empirical arguments for Task 2, only considered choosing the midpoints of the sides of a rectangle as opposed to considering any random points. As can be seen in **Figure 6**, the PSMT picked the midpoints of a rectangle and calculated the area of the rectangle and the quadrilateral formed by the midpoints to justify that the task was true.
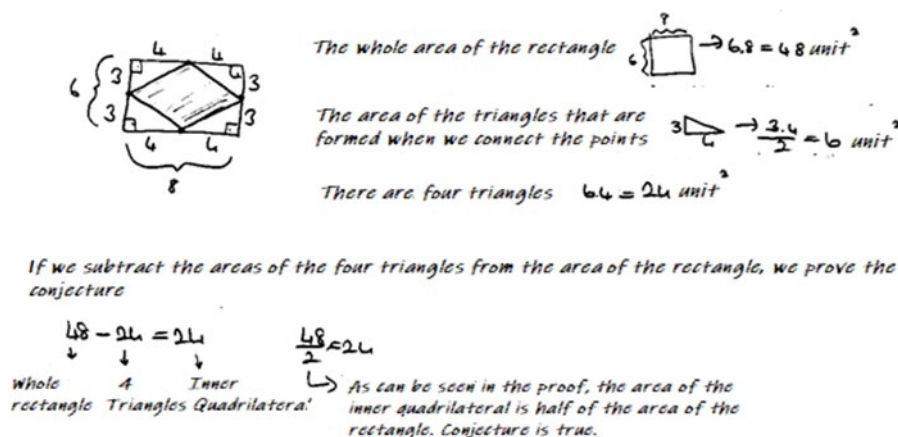
Figure 6. An argument that was coded as Empirical Argument for Task 2

2 PSMTs believed that the statement was true; however, they could not complete their arguments to justify their decisions. In the excerpt below, One PSMT—Mustafa[1]—attempted to construct an argument to justify the statement; however, he failed to complete his argument.

> Mustafa: I know that this statement is true. But, I could not prove it. Well, if I picked the midpoints, I could do it easily. Because, I could show that the triangles were the same. But, I did not know how to do it otherwise. Like, if I did not pick the midpoints, I could not do it. Let's choose arbitrary points (Labelling the side lengths and angles in **Figure 7**). But these do not intersect perpendicularly (Referring to the diagonals of the inner quadrilateral). I am trying to show that the triangles are congruent. There are eight triangles in total and four of them formed the quadrilateral. But I do not know how to do that. I know they are the same but do not know how to show it.
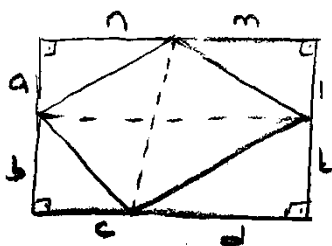

Figure 7. Mustafa's argument that was coded as Unfinished Argument for Task 2

Mustafa believed that the statement was true. Seeing the statement held true for a specific case—connecting the midpoints of the sides of the rectangle—convinced him that the statement would hold true for all cases. When asked to justify the statement, he indeed attempted to construct an argument for a more general case. However, he failed to complete his argument since he did not know how he could show that the triangles had the same area. Although he could not proceed with how to justify that the triangles were congruent, he still was convinced that the statement held true. Therefore, Mustafa's response was coded as an unfinished argument since his attempt to construct an argument to justify that the statement was true was not completed. The responses in **Figure 4** and in **Figure 7** could be interpreted similarly since both responses attempted to investigate the case in which arbitrary points were selected as opposed to the mid points of the rectangle. However, the purposes of constructing

---

[1] All names used in the study are pseudonyms

these examples differed since one was constructed to refute the statement (see **Figure 4**) while the other one was constructed to justify the statement (see **Figure 7**).

For Task 3, on the other hand, most of the PSMTs (42 PSMTs) were able to evaluate the task correctly as opposed to the other two tasks. Out of these 42 PSMTs, 39 of them were also able to construct a valid counterexample while 3 PSMTs failed to provide a valid counterexample that refuted the statement. The majority of the PSMTs constructed a trapezoid as a counterexample for Task 3. As can be seen in **Figure 8**, the PSMT not only refuted the task by providing a valid counterexample, but the PSMT also demonstrated why the example contradicted to the statement by showing that the areas of the triangles formed by the two diagonals were not the same. Thus, it was coded as a valid counterexample for the task.
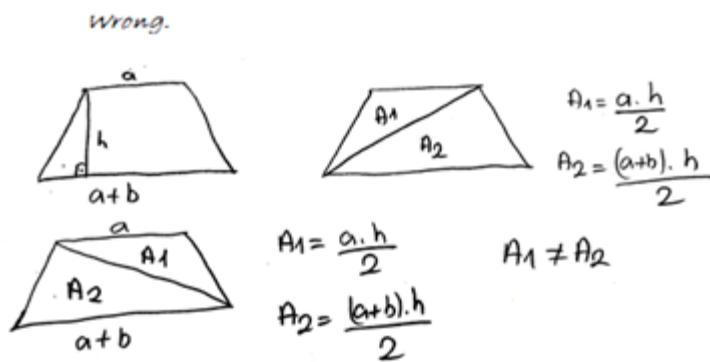
Figure 8. An argument that was coded as Valid Counterexample for Task 3

Although many of the PSMTs successfully constructed valid counterexamples, 3 PSMTs failed to do so. As can be seen in **Figure 9**, the PSMT provided four examples, one of which was a kite since she used same notations on the adjacent sides to show that they were congruent. Then, she circled the kite and labelled the areas of the triangles formed by one of the diagonals as A and B to show that they were not equal. Although the PSMT evaluated the task correctly, she provided an invalid counterexample since the example did not contradict the statement. The PSMT only focused on one diagonal and ignored the other one, which indeed cut the area of the kite in equal halves.
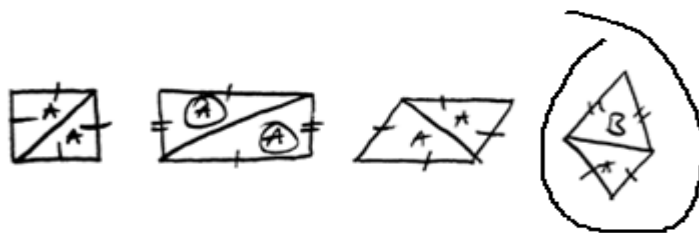
Figure 9. An argument that was coded as Invalid Counterexample for Task 3

When looking at the results cumulatively, it was seen that most of the participants struggled to evaluate the presented tasks correctly and were coded in Incorrect Category. Among the PSMTs who failed to evaluate the presented tasks correctly, most of them attempted to justify the statement for all cases covered by the domain of the tasks; however, their arguments built upon a logical flaw—an incorrect inference drawn from particular conditions. Although the number of the arguments that were coded as empirical arguments were not as high, those types of invalid ways of justifications still existed among the participants. Thus, these results could be interpreted that the PSMTs attempted to construct general arguments to justify the statements for the domain of the statements more than they constructed arguments that purported to show the truth of the statement by validating it in a proper subset of all possible cases covered by the statement. Yet, the PSMTs struggled with employing true sets of accepted

statements in their arguments. Among the PSMTs who correctly recognize that the statements were partially true for some specific conditions, most of them were also able to construct a valid counterexample. However, some of the participants failed to construct a valid counterexample or complete a counterexample at all. How did the PSMTs evaluate the presented empirical arguments will be discussed next.

Evaluating Student Arguments Provided for the Statements

The results regarding to the PSMTs evaluating student arguments are displayed cumulatively in **Table 4**.

Table 4. Results of evaluating student arguments

	Proof			Not Proof	
	Valid / Mathematical	Appropriate for Student Level	Other Reasons	Not General	Invalid / Not Mathematical
Task1	11	1	-	27	11
Task2	16	-	1	32	1
Task3	4	1	1	38	6

For Task 1, while 38 PSMTs evaluated the student argument as Not a Proof, 12 PSMTs believed that the argument could be classified as a Proof. For Task 2, more of the participants again argued that the presented student argument could not be considered as valid. 33 PSMTs evaluated the student argument as Not a Proof while 17 PSMTs classified the arguments as a Proof. Similarly, for Task 3, 44 PSMTs believed that the presented argument could not be classified as a Proof and 6 PSMTs argued that the presented argument could be considered as a Proof. Among these PSMTs who believed that the presented student arguments could not be considered as mathematical proofs, most of them argued that the arguments were Not General.

The PSMT—Tugce—, for instance, argued that the presented student work for Task 1 did not implement a valid proving method as the primary reason. Tugce stated: "To prove a statement, she (Referring to the hypothetical student in the task) should either use a direct proving method or should prove by induction. If a statement is wrong, then she should provide a counterexample. What Nermin did is not a proof since providing two examples that show that the statement is true does not fall into any of the proving methods. Her [Nermin's] argument should have shown that the statement was true for all rectangles." Tugce argued that the employed mode of reasoning in the presented argument was not valid since it did not fall into any of the valid proving methods that she mentioned in her response. Therefore, Tugce's response was coded as Invalid/Not Mathematical.

Dilara, on the other hand, argued that the student argument could not be considered as a proof since it only showed that the statement was true for a proper subset of all the cases covered by the statement by stating that the student only tried some numbers. She stated: " ….She only tried some numbers. The fact that these two examples showed that the statement was true does not mean that it would always hold true for all examples. She only could have proved a false statement with this method. Because when she found one wrong example, we could understand that this statement was not true." Dilara mentioned the generic aspect of the argument as questionable and argued that the presented argument failed to provide conclusive evidence to justify the statement for all examples.

It could be argued that both Tugce and Dilara used not being general and not implementing a valid way of proving in their responses. Tugce stated that "her argument (Referring to the hypothetical student in Task 1) should have shown that the statement was true for all rectangles" by the end of her response, which indeed addressed the limitation of the student argument that failed to provide conclusive evidence for the truth of the statement for all cases. Thus, she also questioned the generality of the presented student argument along with implemented proving methods in the argument. Similarly, Dilara stated that "She (Referring to the hypothetical student in Task 1) only could have proved a false

statement with this method" to describe the limitation of the employed method of reasoning. However, as described above in the data analysis section, the reasons first stated by the PSMTs were accepted as their primary choices and coded in the case of the PSMTs provided more than just one reason while evaluating the presented student arguments.

CONCLUSION AND DISCUSSIONS

The results of the study will be discussed under the light of current studies in this section. This section is organized around the two research questions that guided the study.

Verifying Presented Statements and Constructing a Justification

There has been a strong emphasis in various policy documents for the inclusion of constructing and critiquing mathematical arguments in all grades (MEB, 2018; NGA/CCSSO, 2010; Stylianides & Stylianides, 2017). However, verifying the truth or falsity of statements accurately is a complex process as individuals should have adequate understandings of mathematical concepts and be able to apply such knowledge flexibly (Buchbinder & McCrone, 2020). Before constructing an argument for a true statement or generating a counterexample for a false one, students and teachers need to be able to accurately decide the truth or falsity of a given proposition. Research investigating undergraduate students' and mathematics teachers' ability to evaluate a given proposition suggest that many of them have difficulty verifying the truth and falsehood of given statements due to their inadequate understanding of the mathematical content (Riley, 2003; Zeybek Simsek, 2020). For instance, Riley (2003) found that roughly 57% of 23 prospective secondary mathematics teachers believed that a false statement in geometry was true. The results of this study documented that the PSMTs struggled with deciding whether the presented three statements held true. The results also demonstrated that the PSMTs struggled with verifying Task 1 the most. When analyzed Task 1 separately, 36 PSMTs believed that the statement held true; while only 14 PSMTs verified the falsehood of the statement correctly. Given that teachers need to critically evaluate and determine what is entailed in student-generated conjectures (Stylianides, 2007), the results of this study demonstrated that it is not an easy feat for pre-service teachers.

Zeybek (2017) argued that refuting conjectures and justifying invalid claims is a complex process that goes beyond deductive proof and requires the development of rationality and a specific state of knowledge. Given that counterexamples have power to illustrate why a mathematical statement is false and to refute a mathematical statement only requires a single counterexample (Kinzel & Cavey, 2017), counterexamples play such a significant role in comprehending mathematics and axiomatic system of it. Yet, studies demonstrated that students and teachers struggled to provide a valid counterexample (Zaslavsky & Peled, 1996; Zeybek, 2017) The results of this study also demonstrated that the PSMTs struggled with constructing valid counterexamples (or counterexamples at all) to refute the statements. The possible sources of difficulty in generating such examples were presumed to include the following: incomplete knowledge, inability to process existing knowledge, misconceptions, and insufficient logical knowledge (Zaslavsky & Peled, 1996). The PSMTs, who struggled to construct a counterexample or constructed counterexamples that were coded as invalid, also demonstrated limited knowledge of the contents that underpinned the statements.

For instance, for Task 1, most of the PSMTs believed that there was a relationship between area and perimeter of a rectangle so that they did not even attempt to test the method or to generate examples. Further, 14 PSMTs struggled to provide an example that satisfied the condition for a counterexample for Task 2, which might be resulted from their limited subject matter knowledge. Although the PSMTs struggled with constructing counterexamples for Task 1 and Task 2, for Task 3, on the other hand, it was much easier for them to construct a valid counterexample. Zazkis et al. (2008) argued that the process of constructing counterexamples depends on the extent to which they are in accord with individuals' example spaces. In other words, the process of constructing counterexamples while refuting false claims should be conceptualized with individual's example spaces. Thus, the fact that the

PSMTs performed better at constructing counterexamples for Task 3 compared to other two tasks could then be interpreted as a result of the PSMTs' possible example spaces regarding to the underpinning concepts of the statements.

From a mathematical standpoint, the main difference between empirical arguments and proofs lies in the modes of argumentation (Stylianides, 2007, p. 291). Empirical arguments provide inconclusive evidence by verifying its truth only for a proper subset of all the cases covered by the generalization, whereas proofs provide conclusive evidence truth by treating appropriately all cases covered by the generalization. Stylianides and Stylianides (2009) highlighted the importance of realizing this limitation of empirical arguments as methods for validating mathematical generalizations. Yet, the results demonstrated that empirical arguments were pervasive among the participants. Students at all levels have difficulty recognizing universally and existentially quantified statements (especially when the quantifier is implicit) and struggle to understand that a universally quantified statement must be proved for all elements in the domain, and fail to recognize the limitation of relying on supportive examples for proving universal statements (Buchbinder & Zaslavsky, 2019). The PSMTs who constructed empirical arguments in this study indeed failed to recognize the fact that empirical arguments provide inconclusive evidence so that they could not be generalized for all cases covered by the statements.

Although empirical arguments for verifying the statements to be true failed to satisfy the generalization aspect of proofs, the arguments coded as incorrect inference satisfied the generalization aspect of mathematical proof. Yet, they were built upon an incorrect inference, so that they failed to implement true sets of statements. These types of arguments were common among the participants. This finding indeed demonstrated that the PSMTs who participated in this study struggled with employing true sets of statements in their arguments more than employing valid ways of reasoning. This shows that the PSMTs need not only an understanding of what counts as valid argument, but also an adequate knowledge of choosing accepted definitions, axioms, and facts. Various properties and postulates that underlie in an argument made in the proof are usually not spelled out, but rather are assumed to have been already learned and internalized by students (Schleppegrell, 2007; Weiss & Herbst, 2015). Therefore, it might not be surprising to see students have difficulties interpreting or using theorems on their own. Teachers should be able to evaluate the assumptions made during argument construction as well as to check what implicit or explicit warrants support the argument (Dawkins & Weber, 2017). To do so, teachers first should be cognizant about what they used in their arguments themselves. However, most of the PSMTs, in this study, failed to elaborate upon what principles are being used to derive new mathematical inferences and to warrant for the inferences used in their arguments. Given the difficulties that these PSMTs experience, this knowledge of proof entailments seems particularly critical.

Evaluating Presented Student Arguments

Researchers argued that student's poor argument constructions can be misleading indicators of what they think would meet the standard of proof, because they may be well aware of the limitations of their arguments but unable to produce better ones (e.g., Stylianides & Stylianides, 2009; Zeybek Simsek, 2020). Thus, Stylianides and Stylianides (2009) claimed that employing "construction-evaluation" activities together could better illuminate learners' understanding of proofs. Given that students do appear to be better at choosing correct proofs than constructing their own (e.g., Stylianides, Bieda, & Morselli, 2016; Zeybek Simsek, 2020), asking evaluating researcher generated arguments or constructing proofs separately, therefore, might draw different pictures about students' understanding of proofs. It is perhaps because generating a sequence of steps and conceptualizing someone else's proof demand different cognitive skills (Stylianides & Stylianides, 2009) and could not be captured by employing construction or evaluation activities separately. The results of this study showed that the PSMTs were more successful at evaluating arguments than verifying the falsehood of the statements and then justifying their decisions (see **Table 4** for details).

Given that the majority of the PSMTs were more successful at evaluating student arguments than verifying the truth of the statements and then constructing their own arguments, it could be argued that it was easier for them to recognize the limitation of the presented arguments. Among the PSMTs who successfully recognized the limitation of the arguments and classified them as not proofs, most of them argued that the presented arguments failed to provide conclusive evidence for the truth of the statement for all cases. Thus, the PSMTs questioned the generality aspect of the presented student arguments. Stylianides and Stylianides (2009) argued that recognizing the difference between a mathematical proof and empirical argument constitutes such an essential goal for mathematics teachers. The high number of the PSMTs, who seemed to recognize the limitation of empirical arguments as methods for validating mathematical statements and then correctly evaluated presented student arguments as not proofs, could, then, be seen as a hopeful picture since they will soon be expected to evaluate students' arguments in their classrooms (Stylianides, 2007).

Researchers argue that employing tasks that do not always hold true during instruction is important for developing an understanding of the role of mathematical proofs and gaining an appreciation for mathematical proofs (Ball, Hoyles, Jahnke, & Movshovitz-Hadar, 2002; Stylianides & Stylianides, 2009). For instance, Harel and Sowder (2007) argued that using example-based reasoning should be cautioned due to its tentative nature; therefore, employing patterns that do not always hold true could be helpful to get students recognize the limitation of using empirical-based arguments. The high number of the PSMTs, who seemed to recognize the limitation of empirical arguments as methods for validating mathematical statements might therefore be a result of employing tasks that do not always hold true. According to the results of the study, it could be argued that once the PSMTs recognized that the tasks would not always true, it was likely for them to argue that the presented student arguments would not constitute a valid way to prove. However, it should also be noted here that some participants, who believed that the tasks would always hold true, still evaluated the presented student arguments as not a proof. For instance, the majority of the participants failed to recognize that Task 1 would not always hold true. Yet, they still argued that the presented student argument for Task 1 would not constitute a proof (see **Table 3** and **Table 4** for details). Thus, it would be misleading to conclude that the PSMTs should verify the tasks correctly before evaluating presented student arguments properly. Employing the tasks that do not always hold true could be an essential implication of this study as will be addressed next.

Implications of the Study

There are clearly high pedagogical demands placed on teachers who strive to engage their students in proving at all grade levels as highlighted by current standards. Research show that creating and effectively managing these learning opportunities for students might be challenging and complicated (e.g., Stylianides, 2007). The results of this study demonstrated that the PSMTs struggle evaluating presented mathematical tasks as well as constructing arguments to justify their decisions regarding to the validity of the tasks. All these results suggest nothing but the need for pre-service teachers to gain more experiences with constructing and evaluating mathematical arguments. Possible ways to meet this suggestion of helping pre-service teachers to gain more experience with constructing or evaluating mathematical arguments will be explored next.

Research has demonstrated that various properties and postulates that underlie in an argument made in the proof are not spelled out, but rather are assumed to have been already learned and internalized by students (Schleppegrell, 2007; Weiss & Herbst, 2015). As a result, students have difficulties interpreting, or using theorems on their own (Zeybek Simsek, 2020). The number of the PSMTs who attempted to construct a general argument which failed to employ true sets of statements (Incorrect Inference Category) emphasize the need for spelling out underlying properties and postulates in textbooks or in classrooms. Although the arguments that the PSMTs constructed in this category (Incorrect Inference) captured adequately the generality of the tasks they aimed to justify, the arguments failed to capture the use of mathematical resources (e.g., relevant definitions, properties) that are known

or accessible to the PSMTs properly. Thus, the PSMTs' reliance on intuitive reasoning highlights the need for making mathematical resources more accessible to them.

Alcock (2004) argued that using examples would be such a useful approach to develop a 'guts feeling' regarding the validity of mathematical conjectures. However, it should be cautioned to use example-based reasoning (aka 'empirical proof scheme') due to its tentative nature and logical limitations in terms of generalization (Harel & Sowder, 2007). Researchers suggest that employing patterns that do not hold true for an infinite set could be helpful to get students recognize the limitation of using empirical arguments as a valid way of proving (Ball, Hoyles, Jahnke, & Movshovitz-Hadar, 2002; Stylianides & Stylianides, 2009). This study provided a support by showing that the PSMTs who recognized the tasks would not hold true for all cases also evaluated presented arguments as not proofs by highlighting this limitation of empirical arguments. This could be interpreted as that the statements that are not always true could be an essential instructional tool to help learners (i.e., pre-service teachers) begin to recognize the limitations of empirical arguments as methods for validating mathematical generalizations. Furthermore, students are often expected to prove results that seem obvious to them (Dawkins & Weber, 2017; Stylianides & Stylianides, 2009). Consequently, proof is likely to remain meaningless and purposeless in students' eyes. Thus, the element of uncertainty seems important to develop an appreciation of the need to prove. The statements that do not hold true for all cases such as the ones used in this study might therefore be a possible venue for highlighting an intellectual need to learn about more secure validation methods.

REFERENCES

Alcock, L. (2004). Uses of example objects in proving. In M. J. Hoines & A. B. Fuglestad (Eds.), *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol.2, pp. 17-24). Bergen, Norway.

Ball, D. L., Hoyles, C., Jahnke, H. N., & Movshovitz-Hadar, N. (2002). The teaching of proof. In L. I. Tatsien (Ed.), *Proceedings of the International Congress of Mathematicians* (Vol. III, pp. 907-920). Beijing: Higher Education Press.

Brown, S. A. (2014). On skepticism and its role in the development of proof in the classroom. *Educational Studies in Mathematics, 86*(3), 311-335. https://doi-org.proxyiub.uits.iu.edu/10.1007/s10649-014-9544-4

Buchbinder, O., & McCrone, S. (2020), Preservice teachers learning to teach proof through classroom implementation: Successes and challenges. *The Journal of Mathematical Behavior, 58,*100779. https://doi.org/10.1016/j.jmathb.2020.100779

Buchbinder, O., & Zaslavsky, O. (2019). Students' understanding of the role of examples in proving: strengths and inconsistencies. *Journal of Mathematical Behavior, 53*, 129-147. https://doi.org/10.1016/j.jmathb.2018.06.010

Cañadas, M. C., Deulofeu, J., Figueiras, L., Reid, D., & Yevdokimov, O. (2007). The conjecturing process: perspectives in theory and implications in practice. *Journal of Teaching and Learning, 5*(1), 55-72. https://doi.org/10.22329/jtl.v5i1.82

Cirillo, M., Kosko, K. W., Newton, J., Staples, M., & Weber, K. (2015). Conceptions and consequences of what we call argumentation, justification, and proof. In T. G. Bartell, K. N. Bieda, R. T. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the thirty-seventh annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education: Critical responses to enduring challenges in mathematics education* (pp. 1343-1351). East Lansing: Michigan State University.

Czocher, J. A., & Weber, K. (2020). Proof as a cluster category. *Journal for Research in Mathematics Education, 51*(1), 50-74. https://doi.org/10.5951/jresematheduc.2019.0007

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

Ellis, A. B., Bieda, K., & Knuth, E. (2012). *Developing essential understanding of proof and proving for teaching mathematics in grades 9-12.* Reston, VA: National Council of Teachers of Mathematics.

Glaser, B. G., & Strauss, A. L. (1967). *Discovery of grounded theory*. Mill Valley, Ca.: Sociology Press.

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: A Project of the National Council of Teachers of Mathematics* (pp. 805-842). Reston, VA: National Council of Teachers of Mathematics.

Kinzel, M. T., & Cavey, L. O. (2017). The role of examples in teaching. *The Mathematics Teacher, 111*(2), 140-143. https://doi.org/10.5951/mathteacher.111.2.0140

Lannin, J. K., Ellis, A. B., & Elliott, R. (2011). *Developing essential understanding of mathematical reasoning for teaching mathematics in prekindergarten-grade 8*. National Council of Teachers of Mathematics.

Lesseig, K. (2016). Investigating mathematical knowledge for teaching proof in professional development. *International Journal of Research in Education and Science, 2*(2), 253-270. https://doi.org/10.21890/ijres.13913

Mata-Pereira, J., & da Ponte, J. P. (2017). Enhancing students' mathematical reasoning in the classroom: Teacher actions facilitating generalization and justification. *Educational Studies in Mathematics, 96*(2), 169-186. https://doi.org/10.1007/s10649-017-9773-4

MEB (2018). *Matematik dersi öğretim programı (ilkokul ve ortaokul 1, 2, 3, 4, 5, 6, 7 ve 8. sınıflar)* [Mathematics lesson curriculum (primary and secondary school 1st, 2nd, 3rd, 4th, 5th, 6th, 7th and 8th grades)]. https://ttkb.meb.gov.tr/

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

National Governors Association Center for Best Practices, & Council of Chief State School Officers (NGA & CCSSO) (2010). *Common core state standards mathematics*.

Riley, K. J. (2003). *An investigation of prospective secondary mathematics teachers' conceptions of proof and refutations* (Unpublished doctoral dissertation), Montana State University, Bozeman.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading and Writing Quarterly, 23*(2), 139-159. https://doi.org/10.1080/10573560601158461

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education, 38*, 289-321. https://doi.org/10.2307/30034869

Stylianides, A. J., & Stlyianides, G. J. (2009). Proof constructions and evaluations. *Educational Studies in Mathematics, 72*, 237-253. https://doi.org/10.1007/s10649-009-9191-3

Stylianides, A., Bieda, K., & Morselli, F. (2016). Proof and argumentation in mathematics education research. In A. Gutierez, G. Leder, & P. Boero (Eds.), *2nd handbook on the psychology of mathematics education* (pp. 315-351). Sense Publishers. https://doi.org/10.1007/978-94-6300-561-6_9

Stylianides, G. (2008). An analytic framework of reasoning-and-proving. *For the Learning of Mathematics, 28*(1), 9-16. http://www.jstor.org/stable/40248592

Stylianides, G. J., & Stylianides, A. J. (2017). Research-based interventions in the area of proof: The past, the present, and the future. *Educational Studies in Mathematics, 96*(2), 119-127. https://doi.org/10.1007/s10649-017-9782-3

Stylianides, G. J., Stylianides, A. J., & Weber, K. (2017). Research on the teaching and learning of proof: Taking stock and moving forward. In J. Cai (Ed.). *Compendium for research in mathematics education* (pp. 237-266). National Council of Teachers of Mathematics.

Weiss, M., & Herbst, P. (2015). The role of theory building in the teaching of secondary geometry. *Educational Studies in Mathematics, 89*(2), 205-229. https://doi.org/10.1007/s10649-015-9599-x

Zaslavsky, O., & Peled, I. (1996). Inhibiting factors in generating examples by mathematics teachers and student teachers: The case of binary operation. *Journal for Research in Mathematics Education, 27*(1), 67-78. https://doi.org/10.2307/749198

Zazkis, R., Liljedahl, P., & Chernoff, E. J. (2008). The role of examples in forming and refuting generalizations. *ZDM - The International Journal on Mathematics Education, 40*, 131-141. https://doi.org/10.1007/s11858-007-0065-9

Zeybek Simsek, Z. (2020). Constructing-evaluating-refining mathematical conjectures and proofs. *International Journal for Mathematics Teaching and Learning, 21*(2), 197-215.

Zeybek, Z. (2017). Pre-service elementary teachers' conceptions of counterexamples. *International Journal of Education in Mathematics, Science and Technology, 5*(4), 295-316. https://doi.org/10.18404/ijemst.70986