DEVELOPING FURTHER SUPPORT FOR IN-SERVICE TEACHERS' IMPLEMENTATION OF A REASONING-AND-PROVING ACTIVITY AND THEIR IDENTIFICATION OF STUDENTS' LEVEL OF MATHEMATICAL ARGUMENTATION

Cornelia Brodahl, Niclas Larson, Unni Wathne and **Kirsten Bjørkestøl**
Department of Mathematical Sciences, University of Agder, Norway

Abstract

This is the third in a series of papers focusing reasoning-and-proving. Participants were in-service teachers enrolled in a continuing university education programme in teaching mathematics for grades 5–10. Data were collected from a course assignment in 2018 and 2019, where the in-service teachers reported about their students' work with a reasoning-and-proving task. Their reports included an identification of the levels the students' written argumentation reached, based on Balacheff's taxonomy of proofs. The course assignment's instructions were expanded for the 2019-cohort. Comparing in-service teachers' proof level identifications to the researchers' by statistical analyses, indicated an improvement of the general quality from 2018 to 2019. A higher consensus in 2019 included identifying generic arguments and an understanding that there might be examples falling outside of the taxonomy levels. Qualitative content analysis of the two cohorts' justifications of their identifications revealed an improved understanding of what is considered generic argumentation. The results encourage and contribute to further developments of the concept.

Keywords: Balacheff's four levels of proofs, identification of student arguments, mathematical reasoning-and-proving, written imaginary dialogues

Introduction

There is a shortage of research regarding how to prepare pre- or in-service teachers in engaging their students in proving activities in primary and lower secondary classroom (Stylianides, 2016). The aim for this paper is to add knowledge to this gap. This is the third paper in a series of papers focusing on studies on teaching of mathematical reasoning-and-proving. The subjects in the papers are in-service teachers (ISTs) from across Norway enrolled in the first term of Year 1 of a continuing university mathematics education for grades 5–10 (students of age 10 to 15) in the national strategy "Competence for Quality". ISTs in the online programme prepared and implemented a reasoning-and-proving task to their students. The mathematical problem was presented for the students as a dialogue, where two fictitious pupils already had taken the initial steps in discussing the task. Pairs of students were to continue the dialogue and put their arguments in writing. ISTs approached students' mathematical thinking processes in their imaginary dialogues (Wille, 2017) and analysed students' written arguments based on Balacheff's (1988) taxonomy of proofs, which classifies four hierarchical levels of thinking: naïve empiricism,

crucial experiment, generic example, thought experiment.

Each study in this series explored the combination of a mathematical task, the method of imaginary dialogues, and a taxonomy for analysis. The first study (Brodahl & Wathne, 2018), drawn on 2016-participants, investigated perceptions of first experiences with the complex combination as a whole, the second (Wathne & Brodahl, 2019), drawn on 2017-participants, the usefulness of the particular method and taxonomy, as perceived by the teachers. The current study of participants from 2018 and 2019 is independent from the previous two studies. It focused on the quality of ISTs' proof level identifications and the impact of expanding instructions for the 2019-cohort with a preparation brief on possible pitfalls in proof level identifications.

When we assessed the project reports from 2018, we noticed shortcomings in the proof level identifications made by ISTs. These observations resulted in a development of the task instructions for the course in 2019. We briefed on potential pitfalls and added a short video to support ISTs in their work. When assessing the reports from 2019, our impression was that the quality of the ISTs' proof level identifications had improved. However, that notion raised an important issue: How could we know that there really was any improvement of quality of the identifications? That issue led us to the first research question of this paper.

On average, 90.5% of ISTs from the 2018 and 2019 cohorts perceived Balacheff's taxonomy for classifying levels of proofs to be useful; whereas 54.1% perceived challenges in identifying proof levels in their students' written dialogues, and 59.5% found explaining their identifications to be challenging. Because it is important to distinguish invalid and valid proof argumentation in primary and lower secondary school (Ministry of Education and Research, 2020; Stylianides, 2016), these ISTs' perceptions indicated the understanding of what counts as a valid proof might need to be improved. Thus, as teacher educators, we concluded that we needed to know how to better facilitate ISTs in identifying the third level in Balacheff's taxonomy, the generic example. This insight could, in turn, help teacher educators in improving their instructions when adapting the concept 'a task, a method and a taxonomy' to develop ISTs' proficiency of teaching reasoning-and-proving. This issue led us to the second research question of this paper.

Subsequently, we developed the following research questions:

1. How did the quality of in-service teachers' proof level identifications of their students' work with a reasoning-and-proving task change from 2018 to 2019?

2. What justifications do in-service teachers suggest when they incorrectly identify the proof level "the generic example" in students' reasoning and proving?

The paper describes two trails associated with the two research questions. The first research question in this observational study was treated as well by a statistical approach to examine if there existed a change in the quality of the identifications, as by a qualitative approach to describe some of these changes. The second research question was explored by a qualitative content analysis.

Theoretical considerations

This study investigated how ISTs identified their students' mathematical arguments used in a reasoning-and-proving process. The definition of "proof" and conceptualisation of the meaning of proof in school mathematics that guided the current study, come from A. Stylianides (2007): "*Proof* is a *mathematical argument*, a connected sequence of assertions for and against a mathematical claim" (p. 291). The argumentation uses statements accepted by the classroom community and forms of reasoning and communication known by or within the conceptual reach of the classroom community.

The study followed G. Stylianides (2008) using the term "reasoning-and-proving" to encompass the breadth of the activity associated with identifying patterns, making conjectures, providing proofs, and providing non-proof arguments (p. 9). His analytical framework (p. 10) captures activities involved in reasoning-and-proving. Four of them are identified in Balacheff's (1988) taxonomy of proofs and used to categorise different levels of mathematical arguments: "naïve empiricism", "crucial experiment", "generic example" and "thought experiment" (p. 218).

The terms naïve empiricism and crucial experiment are acknowledged as special kinds of empirical arguments for or against a mathematical claim, as example-based reasoning, but not as valid proofs. Due to the taxonomy, only level 4 counts as a rigorous proof, although even level 3 requires general argumentation similar to a formal proof. A generic example is in terms of A. Stylianides' (2007) proof definition, a valid mode of proving a conjecture, using a particular case seen as representative of the general case, while thought experiment constitutes formally established modes of mathematical proof.

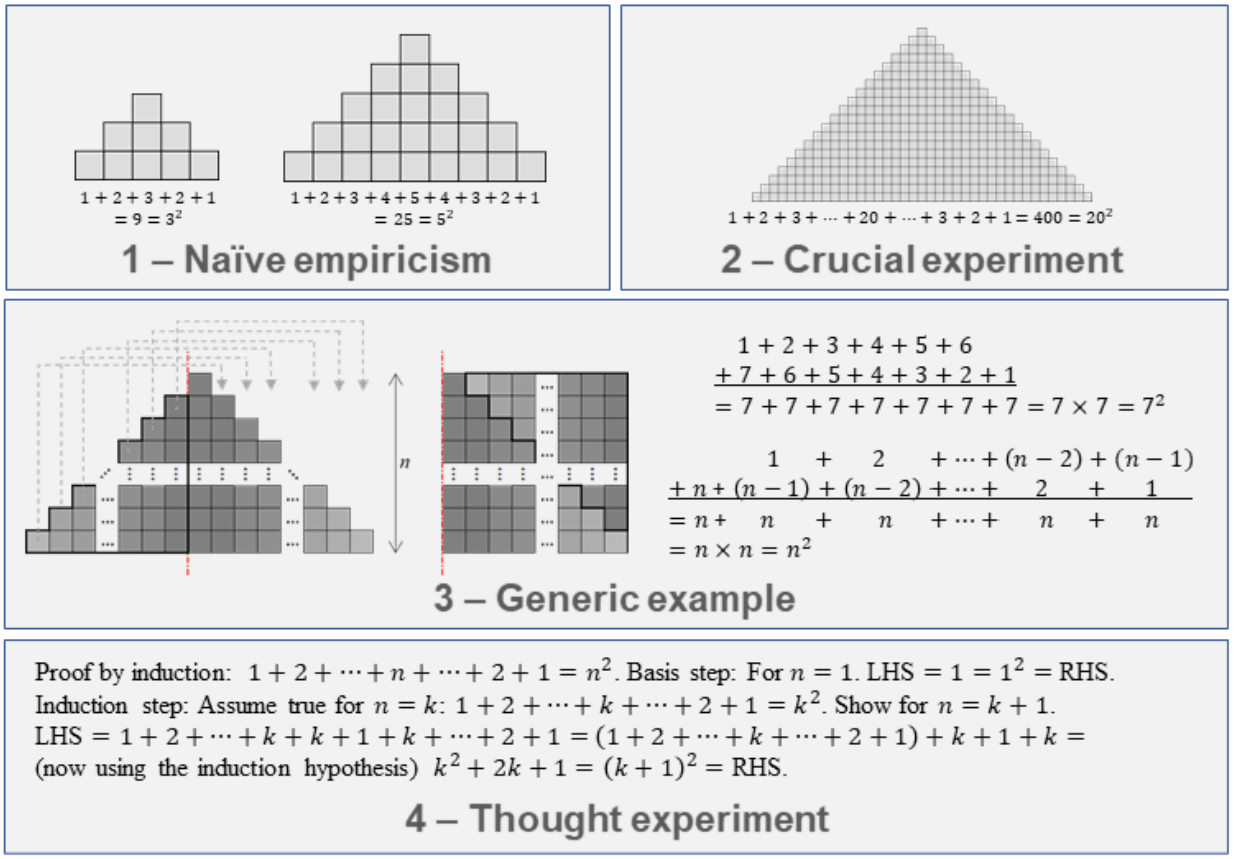
The current study drew on ISTs' application of Balacheff's taxonomy of mathematical proofs to identify how far the students' argumentation reached on their way to a valid proof of a mathematical conjecture. Using the same conjecture as in the course assignment, Figure 1 illustrates examples of elements to be considered. The conjecture was to argue that there are n^2 blocks in an n step up-and-down staircase.

Stylianides and Stylianides (2009) regarded the particular case, chosen as a generic example, to illustrate a prototype offering conclusive evidence for the truth of a mathematical generalisation. Because all cases are considered when a particular case is used as representative of the general case, such cases do not have empirical status (p. 315). However, there are subjective aspects "to the acceptance of a generic proof as a proof, even within the same community of observers" (Zaslavsky, 2018, p. 295) and little consensus in literature on what constitutes a generic example (Doğan, 2020; Rø & Arnesen, 2020). Reid and Vargas (2018) suggested that generic examples in written work of students can be considered as both generic and empirical, depending on the reader's reasoning and reconstruction, and what is accepted within a certain community. They claimed two criteria for valid generic arguments: the evidence of awareness of generality, and the mathematical evidence of reasoning (p. 17). Rø and Arnesen (2020) expanded these claims

for examples providing a proof, that respectively "the argument concludes with the general claim that was to be proved", and "the argument contains both a mathematical reasoning concerning the example [...], as well as an explicit lifting of this reasoning to the general case" (p. 3).

Figure 1

Glimpses of arguments for the conjecture "There are n^2 blocks in an n step up-and-down staircase built from blocks", considered categorised according to the levels of Balacheff's (1988) taxonomy of proofs



In this paper, we followed A. Stylianides (2016) and other researchers (e.g. Rowland, 1998; G. Stylianides, 2008) regarding the strength and importance of the generic example in education, to lay in its potential as a didactic tool in school at all levels, as this level of abstraction is in accessible reach. The generic example was seen to help students discover the general in the particular and having the power to convince, as well as explain, and serve both as a mean in itself and supplying a mean to bridge the gap between informal arguments and more formal proof (Rø & Arnesen, 2020).

Methods

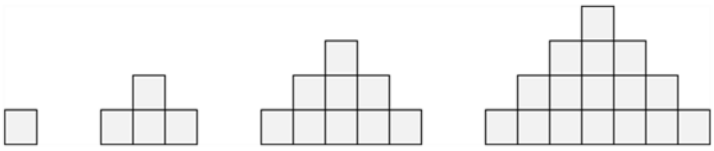
Our two research questions were partly explored by different methods. In this section, we present methods common for both research questions: the setting for this study, the sample, the analytic tool used to describe which proof levels were presents in the dialogues, and how this

analytic tool was employed to generate data. To facilitate the reading, analysis methods used in one trail of the investigation only, will be presented in the results section.

The data in this study came from a mandatory course assignment: IST should prepare and implement a reasoning-and-proving task to their students, with the mathematical problem presented as a dialogue between two fictitious pupils. The mathematical problem was about how to find the number of blocks in a specific pattern (Figure 2). The direct way is to add the number of blocks in each row or column, but the fictitious pupils in the initial dialogue "building up-and-down staircases" (http://bit.ly/buildingupanddownstaircases) suggest that a quicker way to find the number of blocks is by computing the square of the figure's number.

Figure 2

Up-and-down staircases built from blocks



The students' task was to continue the imaginary dialogue and prove that the square will always give the correct result. After the session, ISTs wrote a report, including three continuing dialogues produced by their students, followed by an identification of the levels of the mathematical argumentation in these dialogues, based on Balacheff's taxonomy of proofs. Our primary data consisted of the students' written argumentation, and ISTs' identifications of proof levels including justifications of their choices.

The study drew on the project reports from all ISTs from the cohorts in 2018 and 2019, who gave their informed consent to participate. Excluding reports from ISTs' teaching classes for non-Norwegian speakers gave 37 participants (out of 56) from the 2018 cohort (aged 27 to 55), and 31 participants (out of 48) from the 2019 cohort (aged 28 to 54). Their classes ranged in 2018 from grade 4 to 11 (age 9 to 16), and in 2019 from grade 5 to 11.

After excluding reports where students' argumentations were not presented as dialogues, the remaining 186 dialogues and the corresponding identifications constituted our sample. In 2018, there were 33 dialogues from primary school, 65 from lower secondary, and 3 from first year at upper secondary level. The corresponding numbers from 2019 were 21, 61 and 3. All project reports were anonymised before analysis.

The reports should include which Balacheff levels ISTs identified in each dialogue, and a justification for their decision. In our analysis, we categorised these texts written by ISTs by marking which of Balacheff's levels we interpreted ISTs had identified. However, we found the four levels of Balacheff's taxonomy to fall short, for example when an IST reported "the students have started a reasoning towards level 3, but they are not close to finishing their proof". We found it unsatisfactory either to mark level 3, or to not at all expose that IST had identified some reasoning at this level. Hence, we developed an analytic tool, which also included precursors for

each of the four levels in Balacheff's taxonomy. In addition, there were dialogues falling outside of Balacheff's taxonomy, as no argumentation for the validity of the suggested formula existed. We chose to identify such dialogues as 'category 0.'

Thus, our analytic tool consisted of nine categories, 0, p1, 1, p2, 2, p3, 3, p4 and 4. The prefix p indicated the precursor for the numbered levels. Category 0 was used for argumentations falling outside of Balacheff's taxonomy and constituted the only category that cannot be combined with any other category. Arguments categorised as p1 comprise students giving a single example with few blocks and showing that n^2 gives the right number. Argumentation categorised as category 1 (identical to Balacheff's level 1 – naïve empiricism) meant the students showed that n^2 gives the correct number of blocks for a few examples. Examples of p2 comprise students showing that n^2 works for several examples, but they did not try an example where the number of blocks was distinctly higher than in the other examples. Category 2 (crucial experiment) meant the students showed that n^2 will give the correct number for a staircase with many blocks. That is, the students became convinced of the validity of the formula, by applying it on a more advanced example. Arguments in category p3 comprise students for specific examples claiming the staircase can be rearranged to a square, but not arguing for why this is always possible. An argumentation where the students, in addition, showed that every staircase can be rearranged as a square, was identified as category 3 (generic example). Arguments in category p4 comprise starting an algebraic proof, but not finishing. Category 4 (thought experiment) would usually be a complete proof. In this task, that would be likely to be a proof by induction. Overall, one dialogue might contain arguments on more than one category, however a level and its precursor exclude each other.

The 186 dialogues were scrutinized by three researchers, first individually, and then together to agree on which categories ISTs had identified. In addition, these three researchers made their own analysis on which categories were present in these dialogues, independently of what categories ISTs had identified. The identifications made by ISTs and by the researchers were then compared for each dialogue. This was made to investigate how the disagreement changed from 2018 to 2019.

Results

We will present the results in two subsections, one for each research question. In each subsection, we will explain the analysis methods used to answer the corresponding research question only.

The Quality of ISTs' Identifications of Categories

In this subsection, we present results regarding how well ISTs succeeded in their identifications of Balacheff's levels in students' argumentation. The quality was measured through comparisons between ISTs' and the researchers' identifications, first on the whole dataset,

then on dialogues where ISTs identified any generic argument (category p3 or 3). We chose to focus on these categories, because reasoning-and-proving at a general level is important for students' learning (Rowland, 1998; Rø & Arnesen, 2020; A. Stylianides, 2016; G. Stylianides, 2008), as well as a core element in the syllabus for compulsory school in Norway (Ministry of Education and Research, 2020). The notion of 'disagreement' between ISTs and the researchers, was central in our comparisons and connected to the notion of 'quality' of ISTs' identifications. Disagreement described differences in the identifications of categories due to our analytic tool. We used the researchers' agreed identification of categories as a unanimous conclusion and treated it as the correct solution. Thus, low disagreement between the IST's and the researchers' identifications on a dialogue was equivalent to high quality of that IST's identifications. At the end of this subsection, we also explore ISTs' use, or rather lack of use, of category 0, that is the category for dialogues falling outside of Balacheff's taxonomy.

Comparison of all dialogues. We constructed an algorithm to measure the disagreement between ISTs and the researchers for each dialogue in the cohorts 2018 and 2019. The algorithm has three different types of disagreement, named A, B and C, graded from minor (A) to major (C). The algorithm does not distinguish whether it was the IST or the researchers who identified a category, only that the other had not. Thus, in the following explanation of types of disagreement, IST and the researchers can always be swapped.

Disagreement A is disagreement inside one of Balacheff's levels (e.g. IST identified 1 and the researchers identified p1). In dialogue 1 in Table 2, there is one disagreement of type A.

Disagreement B is applied for disagreements between Balacheff's levels 1 and 2 (e.g. IST identified p2, and the researchers identified 1) as in dialogue 2 in Table 2. That gives two disagreements of type B, since each identified a category that the other did not. If IST identified categories 1 and 2, and the researchers identified 1, as in dialogue 3, then both agreed on category 1. The only disagreement is that Identification I included 2, which means one disagreement B.

Disagreement C is any other disagreement between categories (e.g. IST identified 3 and the researchers neither p3 nor 3). In dialogue 1, category 3 has no equivalent in Identification II, which gives one disagreement C. Notice that the categories at Balacheff's levels 1 and 2 also can generate disagreements of type C. If IST identified category 2, that means disagreement C if the researchers did not identify any of the categories at level 1 or 2. Only if the researchers identified p1 or 1, that IST's identification of category 2 would change to disagreement B.

It remained to decide what coefficients to assign the three types of disagreement. Our choice was to let A = 0.5, B = 0.7 and C = 1.5, based on that type C by far is the most serious disagreement, and that two disagreements of type B should be less serious than one of type C. For each dialogue, the total disagreement between IST and the researchers is obtained by adding all coefficients connected to the types of disagreements identified.

Table 1 demonstrates how to apply this algorithm. For example, in dialogue 1, there is one

disagreement of type A and one of type C, which means the measure of the disagreement is 0.5 + 1.5 = 2.0.

Table 1

Examples of measuring disagreements

Dialogue	Identification		Disagreement	
	I	II	ABC	Value
1	1, 3	p1	1A+1C	2.0
2	2	p1	2B	1.4
3	1, 2	1	1B	0.7
4	3	p3	1A	0.5
5	1, 2, 3	2	1B+1C	2.2

Note. Comparison of identifications for dialogues. Identifications consist of categories checked. Disagreement is calculated for A = 0.5, B = 0.7 and C = 1.5.

By using this algorithm, each of the 101 dialogues from 2018 and the 85 from 2019 got a measure of the disagreement between IST and the researchers. Our hypothesis, based on our first impression, was that the disagreement had decreased from 2018 to 2019. A comparison of the mean values of the disagreement suggested this to be true. The mean value decreased from 2.009 to 1.735. However, a t-test showed this change not to be significant ($p = 0.0975 > 0.05$). This meant, we could not surely conclude that the decrease was likely to be because of improvement in the ISTs' identifications.

To point at statistical significance would have been a strength in our claim that the quality of ISTs' identifications was higher in 2019. Despite the fact that the data indicated this endeavour failed, our impression was that there still existed an improvement, evidenced by the qualitative data; hence, we will now focus on these dialogues identified as category p3 or 3 by ISTs. Since reasoning at a general level is of special importance (cf. Rowland, 1998; Rø & Arnesen, 2020; A. Stylianides, 2016; G. Stylianides, 2008), our impression might be connected to identifications at this level.

Comparisons of dialogues identified as category p3 or 3. When this study was conducted, one core element in the forthcoming mathematics syllabus for grades 1–10 was "reasoning and argumentation" (Ministry of Education and Research, 2020). Arguments at the highest level in Balacheff's taxonomy (thought experiment) require an algebraic approach and may not be expected as common from students in compulsory school. Hence, it was plausible that ISTs and their students aimed for arguments at level 3 (generic example). Thus, we found it relevant to explore project reports where IST had identified categories p3 or 3 in their students' written dialogues. This could increase our knowledge of if the students actually reached p3 or 3, or if the identification made by IST was incorrect, and in that case what justification they provided. That would support us in further developments of the task instructions for future cohorts.

There were 37 dialogues in the 2018 cohort, where ISTs identified either p3 or 3. In 2019, 29 dialogues were identified as p3 or 3. The researchers identified category p3 in 42 of these 66 dialogues, and they did not identify any of these dialogues to fully reach category 3.

We divided the 66 dialogues into three groups, based on type of agreement between IST and the researchers. The first group consisted of dialogues of total disagreement, where IST had identified either p3 or 3, while the researchers had not. The second group consisted of dialogues of total agreement, where both IST and the researchers had identified p3. The third group consisted of "partial agreement", where IST had identified 3 and the researchers p3. The distribution between these groups is shown in Table 2.

Table 2

Comparison of identifications of p3 or 3

	2018		2019	
Dialogue	*n*	%	*n*	%
Disagreement	18	*49*	6	*21*
Agreement	8	*22*	14	*48*
Partial agreement	11	*30*	9	*31*
Total	37	*100*	29	*100*

Note. Frequency of dialogues identified as category p3 or 3 by IST.

The group of total disagreement decreased from 49% to 21%, between 2018 and 2019, while the group of total agreement increased from 22% to 48%. This finding indicated an improved quality of the identifications between 2018 and 2019. A χ^2-test indicated this change to be significant ($p < 0.05$).

This improvement remained, when we for the same 66 dialogues included all categories identified by ISTs and the researchers, respectively. When applying the algorithm for measuring disagreement, the mean value of the disagreements decreased from 2.281 in 2018, to 1.721 in 2019. A t-test showed this change to be significant ($p < 0.05$).

Identifications of category 0. We close the subsection dealing with research question 1, by highlighting the employment of category 0. In 2018, we discerned dialogues that did not deal with the task, which was to prove that the number of blocks for staircase n can be calculated by the 'formula' n^2. Such dialogues, or dialogues without any relevant argumentation for the validity of the formula, fall outside Balacheff's taxonomy, that is category 0 in our analytic tool. In 2018, no dialogue was identified as category 0 by the ISTs, while the researchers identified 32 dialogues as category 0. This difference indicated a lack of understanding of Balacheff's taxonomy by ISTs.

For the 2019 cohort, the task instructions were improved, stressing that a dialogue may fall outside the four Balacheff levels. This resulted in a modest improvement. The 2019 ISTs identified five dialogues as category 0, while the researchers identified 23.

ISTs' justifications of their identifications

In this subsection, we focus the 24 dialogues where ISTs identified either category p3 or 3, the researchers neither nor. We explore what justifications ISTs suggested when they incorrectly identified these categories. Qualitative content analysis was used to highlight underlying themes in their justifications, which allowed for the interpretation of textual data through finding and coding themes.

Two researchers independently analyzed and coded this subsample of 24 dialogues. Then, they organized themes and codes in a multifaceted codebook, in an iterative process using inductive and deductive approaches (Bryman, 2012). The process of having two researchers independently perform the coding of data, then compare and align, was used to improve the validity of coding. The outcome was five 'tags', which describe ISTs' justifications of category p3 or 3.

The first tag, 'incorrect employment of category', denoted that ISTs interpreted something as "generic example", although it was not. This suggests ISTs had not understood what a generic example is. One example was the IST that wrote "This group is at level 'generic example', where they with a couple of simple examples show the solution to be true. They have given a general formula, answer = steps · steps." Another example was "I mean that the dialogue and the proving is at level 2 'crucial experiment' since she uses higher numbers for testing, and that she actually approaches level 3 based on what she says regarding that it is valid for all figure numbers." In these two dialogues, ISTs used incorrect arguments for that the dialogues include generic examples.

The second tag was 'any figure used as warrant of p3 or 3'. Two examples were when IST wrote "Further, the students tried manipulatives, they drew a figure of the 'steps.' The students then applied the 'generic example'…" and when IST wrote "This is a generic example, since it has a figure in addition to the calculation." In these dialogues, the students had drawn just one or several pictures of a staircase, with no signs of rearranging the blocks into a square. This tag can be seen as a 'subtag' of the previous ('incorrect employment of category'). However, the researchers found this misunderstanding to be of such importance that it should be a tag of its own.

The third tag, 'conjecture is used as argument', can also be seen as a subtag of the first. One example was "Then he concludes that the formula is $x \cdot x$ and says he does not understand why. Here, I think he is searching for a generic justification." A second example was "The dialogue is due to Balacheff in 'generic example', since the students are aware of the relation and from this were able to construct a formula to solve the task generally." This meant no proof existed, but the construction or usage of the relation to be proved was interpreted as a generic example.

The fourth tag was 'wishful thinking'. This could be when ISTs used reasoning that was not in the written dialogue, for instance IST had heard the students discussing the point. The last

tag was ‘doubtful.’ This was used when ISTs clearly expressed that they did not know if this really was any of category p3 or 3.

Table 3 shows the frequencies of these tags in ISTs’ justifications. Note that more than one tag might be present in the same justification.

Table 3

Tags frequency dispersion

Tags	2018[a]	2019[b]
1. Incorrect employment of category	7	5
2. Any figure used as warrant of p3 or 3	4	0
3. Conjecture used as argument	8	2
4. Wishful thinking	3	0
5. Doubtful	3	3

Note. [a]*N*=18. [b]*N*=6. Dialogues identified as category p3 or 3 by IST, where the researchers disagreed.

The most frequent tag was ‘incorrect employment of category’, where seven dialogues in 2018 and five dialogues in 2019 were categorised. The second most frequent was ‘conjecture used as argument’, with eight and two dialogues, respectively. For ‘any figure used as warrant for p3 or 3’ the number decreased from four dialogues in 2018 to zero in 2019. This indicated an improvement from 2018 to 2019 regarding ISTs incorrectly justifying category p3 or 3 with ‘conjecture used as argument’ and, in certain, ‘any figure used as warrant of p3 or 3’, while no clear improvement could be discerned for ‘incorrect employment of category’.

Limitations

Limitations of this study are mainly related to the small sample size and how to measure disagreement between IST and the researchers. When we scrutinised ‘the quality’ of ISTs’ identifications, we used our own identifications as a benchmark, highly aware there were no truly unambiguous ‘correct’ identification of the level of an argumentation (Zaslavsky, 2018), including our own. Although this is a potential concern, we presume that we as researchers can produce identifications that are ‘good enough’ to use as reference.

Even if one agrees on using the researchers’ identifications as references, another concern was how to measure the disagreement between ISTs and the researchers. It can be argued both for and against the algorithm employed. Even if we regarded the algorithm as appropriate, we do not claim it is fully developed. For example, the values of the coefficients A, B and C can be changed. Considering changes in the algorithm might be relevant in future, similar investigations. We will, however, not further discuss that issue in this paper.

Discussion and Implications

Our first research question was: 'How did the quality of in-service teachers' proof level identifications of their students' work with a reasoning-and-proving task change from 2018 to 2019?' The findings from our quantitative exploration of ISTs' identifications of proof levels briefly suggested there has been an improvement of the general quality from 2018 to 2019. However, this improvement was not significant for the entire sample of 186 dialogues and the corresponding identifications. Thus, we cannot present an unambiguous result that shows the general quality of ISTs' identifications did improve. We could, however, determine an improved quality in the subsample of 66 dialogues, where IST identified generic examples or incomplete generic examples. Our exploration showed their identifications to be of significantly higher quality in 2019 than in 2018. Because the assignment and the teaching model were unchanged, this improvement is likely to be related to the further development of the teacher educators' preparation brief for the 2019-course.

An interesting issue is ISTs identifying Balacheff levels in dialogues where no valid argumentation was present due to the taxonomy. Although a modest improvement of ISTs' use of category 0 appeared from 2018 to 2019, a large discrepancy to the researchers' remained. This indicates a potential for further improvement of future task instructions, so it becomes clear that not every argumentation is captured by the four Balacheff levels. Although we found the results from our first research question useful, results related to our second research question might be more gainful for future teaching.

This question was: 'What justifications do in-service teachers suggest when they incorrectly identify the proof level "the generic example" in students' reasoning and proving?' The most frequent tag was 'Incorrect employment of category', which shows there is a need to clarify what is required in argumentation to qualify as a generic example. It might, however, be rather difficult to grasp an abstract definition of the generic example and what constitutes a generic example (e.g. Doğan, 2020; Rø & Arnesen, 2020; Zaslavsky, 2018), which is a possible explanation of the rather small change of occurrence between 2018 and 2019.The tags 'any figure used as warrant of p3 or 3' and 'conjecture is used as argument' are more concrete concepts, although they, too, can be associated with a lack of understanding of the concept of generic example. These tags were clearly less frequent in 2019, which probably was a consequence of the improved instructions. One explanation could be that it was easier for ISTs to grasp instructions aiming to avoid errors spotted by these tags, than to interpret the abstract definition of a generic example, to avoid errors caught by 'Incorrect employment of category'. Instructions like 'just drawing some staircases is not enough to qualify as a generic example, it is necessary to show that the blocks can be rearranged to a square too' and 'just constructing or using the "formula" n^2 is not enough', are rather plausible.

These observations will support further development of instructions. For example, there is

still potential for improvement of the general description of the generic example, as well as what argumentation is required to fully qualify as a generic example. A better understanding of Balacheff's taxonomy will not just have the potential to improve the performance in the course assignment, it may also improve ISTs' own teaching about reasoning-and-proving in compulsory school.

Previous research has shown ISTs' knowledge of what counts as a valid proof to be insufficient (cf. Stylianides, 2016). Similar to the findings of Rø and Arnesen (2020), there were ISTs in our study using insufficient justifications for distinguishing empirical examples from generic examples, and complete generic proofs from incomplete. ISTs' insufficient knowledge about generic examples enforces teacher educators to highlight this level of reasoning-and-proving.

Students' stepping through the first two levels of proofs in Balacheff's taxonomy may be of importance for their learning process. However, ISTs need awareness that these steps do not constitute any valid form of proof. While encouraging students when they engage in this type of reasoning-and proving, focus also needs to be given to the difference between empirical and generic argumentation, when aiming to help students step further (Rø & Arnesen, 2020). Since algebraic proofs are beyond the reach of most primary school students, argumentation involving generic examples is of importance in students' learning process about proofs.

Conclusion

In this paper, we focus on how teacher educators can prepare pre- or in-service teachers to promote their students in proving activities in compulsory school. Our implications are derived from the findings of this study that adapted the concept 'a task, a method and a taxonomy' (Wathne & Brodahl, 2019). The major challenges concern teacher educators' instructions. These instructions should stress the overall awareness of the generality in proofs, and in particular emphasise how to support students in claiming generality in arguments, in order to prevent them from making incomplete generic examples in their writing. Further, the instructions have to guide ISTs in how to properly claim a given student argument to be either a generic or an incomplete generic example, in consideration of the criteria suggested by Reid and Vargas (2018), and Rø and Andersen (2020).

As to our series of studies, the teacher educators' preparations for the next course will utilise the experiences from this study, to enhance the concept and to achieve further progress in reasoning-and-proving for the future cohorts.

References

Balacheff, N. (1988). Aspects of proof in pupils' practice of school mathematics. In D. Pimm (Ed.), *Mathematics, teachers and children* (pp. 216–235). Hodder & Stoughton.

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

Brodahl, C., & Wathne, U. (2018). Imaginary dialogues: In-service teachers' steps towards mathematical argumentation in classroom discourse. *Journal of the International Society for Teacher Education, 22*(1), 30–42.

Doğan, M. F. (2020). Pre-service teachers' criteria for evaluating mathematical arguments that include generic examples. *International Journal of Contemporary Educational Research, 7*(1), 267–279. https://doi.org/10.33200/ijcer.721136

Ministry of Education and Research. (2020). *Curriculum for the common core subject of mathematics (MAT1-05)*. Norwegian Directorate for Education and Training.

Rø, K., & Arnesen, K. K. (2020). The opaque nature of generic examples: The structure of student teachers' arguments in multiplicative reasoning. *The Journal of Mathematical Behavior, 58*, 100755. https://doi.org/10.1016/j.jmathb.2019.100755

Rowland, T. (1998). Conviction, explanation and generic examples. In A. Olivier & K. Newstead (Eds.), *Proceedings of the 22nd Conference of the International Group for the Psychology of Mathematics Education: Vol. 4* (pp. 65–72). University of Stellenbosch.

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

Stylianides, A. J. (2016). *Proving in the elementary mathematics classroom*. Oxford University Press.

Stylianides, G. J. (2008). An analytic framework of reasoning-and-proving. *For the Learning of Mathematics, 28*(1), 9–16.

Stylianides, G. J., & Stylianides, A. J. (2009). Facilitating the transition from empirical arguments to proof. *Journal for Research in Mathematics Education, 40*(3), 314–352.

Wathne, U., & Brodahl, C. (2019). Engaging mathematical reasoning-and-proving: A task, a method, and a taxonomy. *Journal of the International Society for Teacher Education, 23*(1), 6–7.

Wille, A. M. (2017). Imaginary dialogues in mathematics education. *Journal für Mathematik-Didaktik, 38*(1), 29–55. 7 https://doi.org/10.1007/s13138-016-0111-7

Zaslavsky, O. (2018). Genericity, conviction, and conventions: Examples that prove and examples that don't prove. In A. J. Stylianides & G. Harel (Eds.), *Advances in mathematics education research on proof and proving: An international perspective* (pp. 283–298). Springer. https://doi.org/10.1007/978-3-319-70996-3_20

About the authors

Cornelia Brodahl is an associate professor in ICT and learning at the University of Agder with a master's degree in mathematics. Her research interests include multimedia in teaching and learning mathematics.

Niclas Larson is an associate professor in mathematics education at the University of

Agder. He holds a PhD in mathematics education. His research interest is focused on teaching and learning mathematics in higher education.

Unni Wathne, PhD, is an associate professor in mathematics education at the University of Agder. Her research interest is focused on teaching and learning mathematics in primary and secondary school.

Kirsten Bjørkestøl, PhD in statistics, is an associate professor in statistics at the University of Agder. She is interested in and does research on statistical methods, multivariate methods, and quantitative analyses in education.