What counts as a “good” argument in school?—how teachers grade students’ mathematical arguments

Michael Meyer¹ · Susanne Schnell²

Published online: 5 September 2020
© The Author(s) 2020

Abstract

As argumentation is an activity at the heart of mathematics, (not only) German school curricula request students to construct mathematical arguments, which get evaluated by teachers. However, it remains unclear which criteria teachers use to decide on a specific grade in a summative assessment setting. In this paper, we draw on two sources for these criteria: First, we present theoretically derived dimensions along which arguments can be assessed. Second, a qualitative interview study with 16 teachers from German secondary schools provides insights in their criteria developed in practice. Based on the detailed presentation of the case of one teacher, the paper then illustrates how criteria developed in practice take a variety of different aspects into account and also correspond with the theoretically identified dimensions. The findings are discussed in terms of implications for the teaching and learning about mathematical argumentation in school and university: An emphasis on more pedagogical criteria in high school offers one explanation to the perceived gap between school and university level mathematics.

Keywords Arguments · Argumentation · Summative assessment · Teachers’ practice · Toulmin

1 Introduction

Proving and argumentation are important activities in mathematics and thus also in mathematics education. As different studies show (e.g., Balacheff, 2010; Kempen & Biehler, 2015), dealing with proofs at a university level is a huge challenge for many students, even though

---

✉ Michael Meyer
michael.meyer@uni-koeln.de

Susanne Schnell
schnell@math.uni-frankfurt.de

¹ Institute of Mathematics Education, University of Cologne, Herbert-Lewin-Str. 10, 50931 Cologne, Germany

² Institute of Mathematics and Computer Science Education, Goethe University Frankfurt, Robert-Mayer-Str. 6-8, 60325 Frankfurt, Germany

they learn about mathematical arguments and proofs in school. While a lot of research is put into identifying challenges and scaffolding of mathematical argumentation in school, one aspect of practice has received less attention: the assessment of learners' arguments by teachers. As we will elaborate, the evaluation of arguments is complex. Especially criteria for grading are highly dependent on the way of teaching.

The intention of the present study is to address criteria for evaluating mathematical arguments first from a theoretical point of view, before identifying criteria developed in practice by teachers. Combining the two perspectives aims to provide comprehensive insights and to further the discussion on teaching argumentation and proof.

2 Proofs and argumentation in mathematics

Perceiving mathematics as an activity of exploring, conjecturing, and justifying puts one aspect at its core: arguments and proof (cf. Hanna, 1989; Meyer, 2010). Mostly, the word “proof” is reserved for an argument consisting of valid deductions, which guarantee the truth of an assertion by given premises (Jeannotte & Kieran, 2017). We use the term “argument” to refer to logical relations within a specific content (see below). “Argumentation” relates to the process of giving or developing arguments for an assertion, which also includes the broader context of how and why an argument is presented (cf. Schwarzkopf, 2000). Certainly, within the argumentation, an argument can be wrong and is thus not suitable to infer the truth of the assertion.

Despite the importance of proofs in mathematics and its education, the question “Why does something count as a proof?” is still not completely answered (Davis & Hersh, 1981). One approach attempts to understand proofs as derivations (e.g., Auslander, 2008; Meyer, 2010): Every step of the proof is inferred deductively from theorems, which are assumed to be familiar to the audience. Following this approach includes accepting the tacit assumption that every proof can be traced back to the basics of a theory (e.g., the axioms). Dreyfus (2002) illustrates that this is problematic in the practice of writing and publishing proofs—also due to the extension of this procedure.

Hanna and Jahnke (1996, p. 884) also state that there are no clear and commonly accepted criteria for the validity of proofs. Overall, the criteria for accepting a proof as a proof depend on different factors: “standards of proof vary over time and even among mathematicians at a given time” (Auslander, 2008, p. 62).

When mathematical proofs are discussed in the school curriculum, formal rigor is often not the core of consideration. Krauthausen (2001) states that students in school conduct proofs “before they are named” and refers to generic, pre-formal proofs (cf. Hanna, 1989; Kempen & Biehler, 2015). Even less formal but at least equally important is the mathematical argumentation in school. The different aspects of the nature and the discussion on proofs as explained here—and more importantly the way they transform into argumentation in the reality of mathematics education—show the unique challenge teachers are faced with in the classroom: They have to assess the quality of arguments (in the sense of whether one argument is deserving of a better grade than another) based on transparent criteria, even though there is no common agreement on these criteria even in the experts' community. Similar to a “good” proof, the definition of a “good argument” is highly subjective and depends on many different factors. Therefore, the question arises which criteria we use to call an argument a good one. This question can be answered from a theoretical perspective as well as from the perspective of teachers' developed practice of assessing arguments.

3 Dimensions for assessing arguments

In this section, we present three theoretically derived dimensions along which arguments can be evaluated. As a starting point served Kiel, Meyer, and Mueller-Hill (2015), who mentioned “dimensions for *explaining*” (p. 3, translated by the authors), which help understanding characteristics and affordances of mathematical explanations. Argumentation is closely related with explanation: “In the context of argumentation, premises are offered as proof of a conclusion or a claim, often to persuade someone or settle an issue that is subject to doubt or disputation. (...) In the context of explanation, the explananda (facts to be explained) are explained by a coherent set of explanans (facts that explain). The usual purpose of explanation is not necessarily to convince someone but rather to help someone understand why the explananda are the case” (Bex & Walton, 2016, p. 57). Kiel et al. (2015) summarized the characteristics of explaining as “explaining why something is the way it is” (p. 3 translated by the authors), called the “structural dimension,” “explaining so that the addressee understands it” (ibid, p. 3, translated by the authors), called the “recipient-orientation,” as well as the “content dimension” (what is the explanation about?) and the “language dimension” (how is the explanation worded?).

Regarding arguments and argumentation, we subsumed the language dimension in the recipient-orientation as the understandability of an argument also depends on the way it is worded. Therefore, the three dimensions of mathematics arguments we are looking at are structure, content, and recipient-orientation.

Structural dimension For this dimension, Kiel et al. (2015) draw on the model by Hempel and Oppenheim (1948) which dissects explanation into explanandum (the description of a phenomenon that is to be explained) and explanans (class of sentences, which are used in order to explain the phenomenon).

In terms of the structure of arguments, a variety of different models can be found (e.g., Buth, 1996; Klein, 1980). In the framework for the present study, we draw on the model by Toulmin as it is widely accepted in mathematics education research (e.g., Inglis, Mejia-Ramos, & Simpson, 2007; Nardi, Biza, & Zacharides, 2012; Schwarzkopf, 2000). This can be attributed to (a) the kind of logic on which Toulmin’s model focuses, as its use is not necessarily in need of formal logical answers and (b) the ability of this perspective to also grasp the social aspects of argumentations. Therefore, it can cover a variety of instances occurring within mathematical learning contexts.

According to Toulmin’s model, arguments consist of different functional elements, namely datum, conclusion, rule, backing, and rebuttal (cf. Fig. 1).

A “datum” is composed of undisputable statements (Toulmin, 1969, p. 88). Based on this, a “conclusion” (ibid.) can be inferred, which might have been a doubtful statement before. The “rule” (ibid.) shows the connection between datum and conclusion and is necessary to legitimize the conclusion. If the rule’s validity is questioned, then the arguer could be forced to assure it by a so-called “backing” (ibid, 93ff), which can, for instance, give further details about the context of the rule. The “rebuttal” (ibid.) states exceptions in which the argument does not hold true.

A variety of questions can be addressed to evaluate if an argument is acceptable or not (cf. Meyer & Prediger, 2009): Is a claim given? Are there any exceptions given (understood as rebuttals)? Are there any reasons given? And are they relevant and effective for the claim? Are there evidences given? Are these evidences sufficient, credible and accurate? Etc. These

Fig. 1 Structural components of an argument according to Toulmin (1969)

questions assess the quality of an argument along its structure and functional elements as outlined above. Therefore, we will speak of the *structural dimension* of arguments. It describes the existence of functional elements and thus also the complexity of the argument (e.g., are there one or more steps of reasoning?) as well as the rigor of the logical derivation (e.g., can *this* rule be used to derive *this* conclusion from *this* data?).

An important specification of the structural dimension is a pragmatic one: Normally, an argument is presented with a certain intention, such as to fulfill a specific need for justification or to infer a specific conclusion requested by a given task (cf. Schwarzkopf, 2000). In contrast, the goal for a mathematical explanation is the compensation for a knowledge deficit (Kiel et al., 2015, p. 5). Therefore, assessing an argument can take into account if or to what degree a pragmatic goal was reached.

Content dimension According to Kiel et al. (2015), the content dimension of a mathematical explanation looks at which knowledge deficit is to be compensated and what type of phenomenon is to be explained (e.g., a mathematical rule) by drawing on a specific subject matter.

The same questions can be asked for the structural components of arguments (i.e., what is the (mathematical) content of the datum and conclusion; to which contents do the rule and backing refer).

The content dimension includes the validity (Kopperschmidt, 2000, p. 62–64) and accuracy (cf. Sampson & Clark, 2008) of the mathematical content (e.g., is the used rule mathematically correct?). Furthermore, the content dimension also includes the completeness of the argument in the specific mathematical context (e.g., are all reasons given which are necessary from a mathematical point of view?).

Recipient-oriented dimension As elaborated above in the context of proofs, the validity of an argument depends on whether or not the recipient accepts it. Klein (1980) perceives the beginning of an argumentation as group interaction process as bringing something up that is questionable for the group. “To achieve [a successful argumentation], the social group develops an answer that is accepted by everyone due to rational reasoning” (Schwarzkopf, 2000, p. 157). The people who indicate the need for a justification (in school contexts often the teachers) of a statement are the ones to indicate if the presented argument is sufficient to fulfill this need (Schwarzkopf, 2000) (e.g., is the argument up to my personal expectations of how to justify the statement?). The criteria of the different groups of evaluators can be different. Jahnke and Ufer (2015) emphasize how norms are developed in discourse (cf. socio-mathematical norms, Voigt, 1995; Yackel & Cobb, 1996) for evaluating proofs. These norms can even differ in different contexts. Therefore, they speak of a process of enculturation for getting familiar with them.

Similarly, Kiel et al. (2015) highlight the role of the beliefs and expectations of the recipient for (not) accepting an explanation. Furthermore, the accessibility of the explanation (and also an argument) depends not only on the pre-existing knowledge and preference of the recipient but also on the way it was worded: “An explanation, which poses a problem for the recipient already on the language level, cannot be understandable” (Kiel et al., 2015, p. 6, translated by the authors). As Healy and Hoyles (2000) showed in an empirical study, a more formal language prompts students to be more likely to accept a given argument even though it is not correct. Furthermore, the recipient has to decide if the argument (its functional elements and their interplay) is presented in a way so that he or she is able to trace the logical structure (e.g., am I able to fully understand the presented argument? If not, am I able to fill in the gaps?).

In summary, the identified three dimensions along which an argument can be assessed are the following:

- Structure; including existence and complexity of functional elements, rigor and pragmatic aspects
- Content; including validity and completeness
- Recipient-orientation; including presentation and traceability

While these dimensions define the possible ways to evaluate an argument in general, in school, it is the teachers’ practice which defines the criteria on which an argument is assessed. Thus, we take a look at teachers evaluating given arguments.

4 Teachers’ assessment of students’ arguments

Following Shulman’s (1986) model of teacher knowledge, the assessment of students’ arguments can be categorized as pedagogical content knowledge if we emphasize the necessity of knowing about students’ specific argumentation-related cognition or typical (mis-)conceptions. Summative assessment by formally grading written tests is still the most common activity in German mathematics classrooms to determine and communicate whether a specific learning goal was reached or not. The process of grading students’ written products is necessarily a process of interpreting: Teachers use their experience and knowledge to generate hypotheses about possible meanings implicated by the statements of students. Thus, the processes of grading involved in the assessment of (written) students’ answers are highly complex, multifaceted, and dependent on the individual examiner (Crisp, 2008; Philipp, 2018).

Philipp (2018) describes an idealized diagnostic process (in a formative assessment setting) identified in interviews with six in-service teachers as consisting of different steps. Among those steps are the following: first solving a task themselves, identifying obstacles and pre-requisites, and decomposing the solutions into elements, which can be analyzed step by step. Overall, she describes the teachers’ approaches as systematical while drawing on a variety of different resources (such as pre-existing content knowledge), which differ from teacher to teacher.

Furthermore, Nardi et al. (2012) identify a range of different influences when teachers evaluate students’ written responses. “Our point is relatively simple: teachers’ acceptance, skepticism or rejection of students’ mathematical utterances – as expressed in their evaluation of these utterances (...) – does not have exclusively mathematical (epistemological) grounding. Their grounding is broader and includes a variety of other influences, most notably of a

pedagogical, curricular, professional and personal nature” (Nardi et al., 2012, p. 161). While from a purely mathematical point of view, the content and structural dimension are probably the most important, the significance of the other dimension increases in educational settings in which students learn to argue mathematically (cf. Schwarzkopf, 2000).

Overall, it is an open question which criteria teachers use when assessing arguments and how their criteria relate to the dimensions of arguments we elaborated before.

Thus, the present study aims to research the following questions:

1. Which *criteria* do teachers use to evaluate the elements of the students’ arguments?
2. How do teachers’ individual criteria (implicitly) relate to the identified dimensions?

5 Methodology

In the following section, we describe the design of the interview study, which aims to reconstruct teachers’ criteria for evaluating students’ arguments. Although a good evaluation is in need of mathematical and logical knowledge, one should keep in mind that this study is not about investigating the validity and reliability of teachers’ grading against the backdrop of some kind of external expert evaluation (as in, e.g., Jones & Inglis, 2015). Rather, the study aims to explore the teachers’ individual criteria to gain insights into this practice.

For data gathering, semi-structured interviews (between 60 and 120 min) with 16 teachers from German secondary schools (grade 5 to 13; teaching experience of 1 to 25 years) were conducted. In the beginning of the interview, the teachers had to solve a probability task (Fig. 2) themselves.

Next, they were confronted with fictional and real student answers, which contained calculations, diagrams, and/or symbolic considerations. One argument is shown in Fig. 3.

The arguments were chosen or constructed in terms of the presence and/or elaboration of different functional elements, drawing on the Toulmin model (cf. Fig. 4 for an example). Overall, 12 arguments were presented, of which three were complete and correct—one being more informal, one more symbolic, and one focusing on an iconic presentation. In the other arguments, rules and/or backings were missing or wrong (cf. Fig. 3).

The analysis of the students’ arguments during the interviews consisted of three phases: First, teachers had to give each a specific grade between 1 (best) and 6 (worst)—which is the normal grading range in the German school system—for each argument according to their individual criteria. Next, the teachers were asked to focus on the argumentation and argumentative competence shown by the students and accordingly alter their grades if necessary. Lastly, we asked the teachers to improve some of the students’ arguments, which we will not discuss here in detail.

All 16 interviews were videotaped. Eight interviews were fully transcribed, as they were especially rich, covering all types of secondary schools and different amounts of teaching

Two dice are thrown and the sum is written down. The student Basti wins when the two dice show the sum 1, 2, 3, 10, 11 or 12. Student Derya wins when they show 4, 5, 6, 7, 8 or 9. Basti complains: “I hardly have a chance to win!”
Is that correct? Justify your answer.

Fig. 2 Probability task for eliciting argumentations (cf. Meyer & Schnell, 2016)

Tobi’s argument:

I calculated the probabilities for each event. We have $6^2 = 36$ possible results, therefore the calculation is:

$$\frac{4}{36} \cdot 2 + 2 \cdot \frac{2}{36} + \frac{3}{36} < \frac{3}{36} + \frac{4}{36} \cdot 2 + \frac{5}{36} \cdot 2 + \frac{6}{36}$$

$$\frac{2}{36} + \frac{4}{36} + \frac{3}{36} < \frac{3}{36} + \frac{3}{36} + \frac{10}{36} + \frac{6}{36}$$

$$\frac{9}{36} < \frac{27}{36}$$

This calculation shows that the probability of 27/36 for Derya is higher than 9/36 for Basti. Therefore, Derya will win more games, if they play many times against each other. If they play only a few times, Basti can also win sometimes. The reason is that this is a game of chance and Basti can be lucky once in a while of course. Overall, the complaint is justified.

Fig. 3 Incomplete argument of fictional student Tobi; the argument does not make the probability distribution explicit (i.e., explaining the numerator and the factor before the fractions in the calculation) and thus lacks the explication of the first two data and the corresponding rules (see Fig. 4)

experience (between 1 and 15 years). For the other interviews, partial transcripts were created of the phases in which teachers justified or revised their decisions on the grades.

The analysis was conducted under the interpretative paradigm. We followed a hermeneutic perspective, which has been elaborated for mathematics education by the working group of Bauersfeld (e.g., Voigt, 1995). The interpretation contained four steps, also drawing on Mayring (2000) in steps 2 and 3: First, we scrutinized the data for sections, in which the teachers explained their reasoning behind deciding for (or sometimes against) a certain grade for each student solution. Next, these sequences were analyzed in terms of which elements of the given arguments specifically they were focusing on (e.g., calculation, verbal statement, and diagram). This step was added due to all teachers basing their assessment on certain parts of the students’ argument as Philipp (2018) described to be a common approach. Understanding what the teachers are evaluating helps to support the interpretation of their reasoning and application of criteria.

The third step was to identify *which individual criteria* teachers seemed to use to justify their grades and condense these into categories. The results were contrasted and compared within and across the interviews in order to identify emerging categories. To reach a theoretical saturation, the analysis was conducted until no more codes for different criteria were found to explain the decisions of the teachers. To avoid artifacts specific to one individual teacher, criteria were only included in the presented list when they were used at least twice by two

Fig. 4 Structural representation of Tobi’s argument (Fig. 3), which consists of three steps; the fully framed boxes are explicitly given functional elements in the fictional student’s solution, and the dotted ones stay implicit

different teachers. As the study has an exploratory nature, overall frequencies for each criterion were not measured (cf. Saunders et al., 2018). The categories were validated consensually and independently from each other by the working groups of the authors. Lastly, we compared the *teachers' individual criteria* with the *theoretically identified dimensions* and carefully discussed relations between the two perspectives. All data presented here were translated from German and marginally enhanced in readability. In order to facilitate the reading, we will mark the identified elements on which the teachers focus (i.e., components of the student argument which are evaluated) by underlining them and the identified *criteria* (i.e., the aspects along which the quality is assessed) by printing them in italics.

6 Results

In the following section, we begin with presenting a case especially rich in terms of different criteria, i.e., Mr. Schwarz, in order to give insights into a grading process and the set (and interplay) of individual criteria teachers use (research question 1). These individual criteria are then contrasted with the theoretically identified dimensions (research question 2). In the second subsection, we add an overview of all observed criteria in the set of the 16 interviews.

6.1 Elements and individual criteria: the case of Mr. Schwarz

Mr. Schwarz (all names are pseudonyms) has been a teacher for 6 years and is employed at a German secondary school for higher achieving students. He has no trouble solving the probability task (Fig. 2). Presented with the twelve student arguments, he analyzes them one after the other while frequently comparing and sometimes altering his initial decisions on a grade in the process. The first argument he analyzes is Tobi's (Fig. 4). Interestingly, the elements and criteria Mr. Schwarz uses in order to evaluate all student solutions are already nearly completely verbalized during this first analysis. Thus, this scene is representative of how this teacher grades all arguments. To facilitate the reading, we section the statement and skip the interviewer's comments signaling her active listening. Overall, Mr. Schwarz's analysis process of this first student's solution takes roughly 15 min between initially reading and finally deciding on a grade.

Mr. Schwarz (*after reading Tobi's solution*):

- 1 What I notice here, for Tobi's solution, is that the probability concept is pretty consolidated.
- 2 His considerations are correct. The calculation is correct; the algebra is correct.
- 3 The question is, what is his starting point. Does he start at 36 possible outcomes and – (...) it's not clear to me what he thought here.
- 4 (*Mr. Schwarz sketches a tree diagram and takes some notes of the number of combinations for each possible sum of the two dice. Afterwards, he looks at Tobi's solution again*) (...) Okay, I understand the solution now. It wasn't clear to me that he combined the paths in the tree diagram because they exist more than once in the solution.
- 5 For me, the visualization was the easier step to complete this draft.
- 6 The question was "I hardly have a chance to win! Is that correct?" and then "Justify your answer." Justify. The "justify" is the important task for me. For me, it is an operator which I know from [my practice in] school.

- 7 “Is that correct” was answered. “Was the complaint correct,” yes, he answered that. That’d be okay.
- 8 And now it’s about the way of the justification. He calculated both probabilities here—(*reads out Tobi’s argument*) “the probability [for Derya’s win] is higher than [Basti’s] therefore Derya will win more games”—correct—and as justification he has this calculated estimation.
- 9 I am asking myself if this justification is completely traceable, gapless. Therefore, I would first look at how he starts. He says there are 36 possible outcomes, that is okay.
- 10 But what I just found hard is the next step: how does he arrive at this [calculation] here? (...)
- 11 His conceptual understanding of probabilities is excellent, but has nothing to do with the task. But it makes me happy.
- 12 So, I am missing the step how he gets from this initial consideration of 36 possible outcomes to this [calculation], this inequation. I would expect that in a justification.
- 13 The interpretation of the results is correct later on. (...)
- 14 Well, I have to think about what is important to me: with the operator ‘justify’, this [missing] step is important.
- 15 Thus, the good calculation is great, but it’s not useful when the justification is not complete. As I said, I am happy about his understanding of probability, I am happy about the confident use of fractions; it is always nice to see that.
- 16 Now, I would not decide for a [grade] 1. Even though it is really good and presented really well.
- 17 I also would not give it a 4, 5, or 6. Because there is only one part of the justification missing, this step though is—as I just realized, explainable, derivable.
- 18 I am really just missing the explanation how he arrives at this inequation. I would deduct that and thus give him a [grade] 2.

In his complex analysis, Mr. Schwarz focuses first on elements of Tobi’s solution which he considers as *correct* (line 1–2), such as the probability concept, considerations, calculation, and algebra. Afterwards, he communicates difficulties in understanding the student’s thinking (line 3) which seems mostly related either to the products of natural numbers and fractions in the student’s solution or (possibly additionally) the different numerators in the inequation (line 4). Mr. Schwarz is able to overcome this problem by sketching a tree diagram himself. So far, his approach is in line with the “own solution approach” mentioned by Philipp (2018), where he clarifies the student’s solution by comparing it with his own, combined with the dissection of the task and the student’s solution. In this process, he seems to identify the missing step of how the student arrives at the calculation (i.e., determining the combinations for each sum). In line 5, he refers to this as “visualization,” which is an unusual choice of words in German. The reason might be that Mr. Schwarz sketched a tree diagram, which is indeed a more visual representation. In terms of the Toulmin model (cf. Fig. 4), this corresponds with the missing explication of the first datum (and possibly second) and rule.

In line 6, he goes back to the task itself, possibly as a backdrop for the following analysis. Declaring the “justify” in the task a known and important operator; he then decomposes the student’s solution in chunks for each of which he checks the quality. He evaluates four elements as *existing* and *correct*: the (conclusion as an) answer to the question “is that correct?” (line 7), the total number of outcomes as a datum (line 9), the calculation of win probabilities (line 8), and the interpretation of results (line 13). Furthermore, he mentions that he enjoys the “consolidated

probability concept” (line 1, 11, and 15) and “confident use of fractions” (line 15) but perceives them as *less important* as the focal point of the task is the justification (line 6 and 14). For this, he mentions the criteria *“gapless”* and *“traceable”* (line 9). He emphasizes repeatedly (line 10, 12 and 17) that he is expecting the student to explain also the step from the initial consideration of possible outcomes to the calculation but simultaneously calls this step *“easier”* (line 5), *“explainable [and] derivable”* (line 17). Lastly, in line 16, Mr. Schwarz mentions the overall *quality* (*“really good”*) and *presentation* before settling on the second-best grade.

Overall, the reasoning behind Mr. Schwarz’s assessment can be interpreted as follows: The student’s calculation of win probabilities is not completely comprehensible, as the *“explanation”* of the calculation and the underlying determination of combinations per sum (i.e., the explication of the datum) does not exist but is necessary in the justification task. Thus, the justification is incomplete. Nevertheless, as there is only one *“step”* (in the representation of the Toulmin’s model the first step of the argument, cf. Fig. 4) missing which has a low complexity, the student gets the second-best grade.

In regard to research question 1, summarizing this grading process in terms of elements and criteria, Mr. Schwarz first checks if the yes-no-question posed by the task is answered (element I, Table 1) and then begins a more in-depth diagnosis of the student’s justification (element II). Thus, he decomposes the task in two chunks and the student answer in several elements along these two components, which are presented in Table 1. They include the elements calculation and algebra (II-1) and interpretation of the results (II-2) as well as a conceptual element concerning the probability concept (II-3) underlying the student’s solution. For these elements, he uses a variety of criteria to evaluate them individually such as checking if the element is present or not (*existence*), if it is correct (*correctness*), necessary (*necessity*), etc. Interestingly, Mr. Schwarz goes into a deeper analysis of the missing element, which he calls “visualization” against the backdrop of his own consideration, evaluating its *complexity* and *derivability* to justify his decision for the grade.

Table 1 provides a summary of Mr. Schwarz’s individual grading profile along his individual criteria (columns) and his analyzed elements of the argument (rows). The inserted *“yes”* and *“no”* mean that the criterion is—in his perspective—(not) fulfilled for the element. The parenthesis indicates that Mr. Schwarz probably checks the criterion for the element but does not explicitly comment on it. *“Low,” “less,”* and *“good”* are the qualifications he uses for the corresponding criteria. Unspecified comments on quality such as *good* (line 1) were not explicitly included as criterion, as they linguistically may refer to either *existence* or *correctness*. The attribute *gapless* (line 9) was subsumed under the *existence* of all elements that were identified as *important*. The criterion *presentation* is not clarified and specifically applied to a certain element, but instead used as an overall additional evaluation aspect. The criterion *complexity* is considered explicitly only for the missing element visualization of combination. While all other criteria are mostly dichotomous as in *“existing”* or *“non-existing”* or *“correct”* or *“incorrect,”* the complexity of this missing step is evaluated as *“low,”* indicating the existing of either one or more discrete values such as *“medium”* and *“high”* or a more continuous perception of complexity. While Table 1 shows Mr. Schwarz’s profile for grading the solution of student Tobi, he uses the same criteria again for nearly every student throughout the interviews. In terms of element categories, he analyses most often aspects which belong to calculation while drawing on all the different criteria.

Concerning research question 2—comparing the evaluation of Mr. Schwarz with the theoretically elaborated dimensions—some observations can be made: The pragmatic goal (structural dimension) is to answer the posed yes-or-no-question and to justify this answer,

Table 1 Individual grading profile of Mr. Schwarz (filled in for grading Tobr’s solution)

Criteria: Elements:	Existence/ completeness	Correctness/ validity	Necessity	Importance	Derivability/ traceability	Presentation
I Answer to the claim—“Is the needed conclusion given?”	Yes	Yes	Yes			Overall: good
II Justification	Yes	Yes	Yes			
1 Calculation, algebra	Yes	(Yes)	(Yes)			
1a Number of possible outcomes	No		Yes		Yes	
1b Visualization/explanation of combinations	Yes	Yes	(Yes)	Less		
1c Use of fractions	Yes	Yes	(Yes)			
1d Calculation and comparison of win probabilities	Yes	Yes	(Yes)			
2 Interpretation of (calculated) results	Yes	Yes	(Yes)			
3 Probability concept	Yes	Yes	No	No		

which is the first thing Mr. Schwarz checks. The step he perceives as missing is the explication of the first argument (from the results of two dice to the probability distribution, cf. Fig. 4) and the explication of the rule in the second step of the argument (from the probability distribution to the comparison of the probabilities of the compound events). Thus, his individual criterion corresponds with the structural flaws which were hidden in the student's solution. However, looking at the rule and backing within the third step of the argument (interpreting probabilities in terms of frequencies on the backdrop of the empirical law of large numbers), Mr. Schwarz perceives these structural elements as unnecessary. The reason for this can be assumed to come from the content dimension: Mr. Schwarz regards the content dimension mainly in terms of different elements related to the calculation. Here, he evaluates the completeness and validity of the arguments. This might imply that he regards the content of the task mainly as the calculation of the theoretical probabilities, expecting the answer to be purely based on theoretical probabilities rather than their interpretation as frequencies for wins and losses in the game context, which was given in the task. Thus, the implicitly formulated expectations from the recipient dimension, which he checks against his own solution and understanding of each step, make the student's correct third rule obsolete.

6.2 Overview on the observed criteria

Similar to Mr. Schwarz, criteria used by the other teachers were identified and categorized by comparing, contrasting, and summarizing teachers' statements. The following table shows the richness of criteria the teachers made use of in the interviews. In addition to the criteria from Mr. Schwarz, new criteria such as presentation or concreteness were important for other teachers. In order to address research question 1, we will first present an overview on the criteria (Table 2) before discussing some observations.

As the exemplary statements show, the criteria are not evaluated in isolation but rather in combination. For example, *correctness* cannot only be checked concerning explicit results of the calculation but also concerning more general (logical) considerations. We also decided to include the criterion *language* separately (though it is certainly related to *presentation*), as some teachers clearly focused on the language used by the students in terms of different levels of common language (formal/informal terms, grammatical structure, etc.) or mathematical language.

Overall, some elements of the students' arguments were evaluated along different criteria and also reached opposing assessments: For instance, some teachers perceive the use of different representations (*presentation*) such as algebraic expressions and additional diagrams as positive in terms of the *comprehensibility*. Others evaluate this as unnecessary for the argument and thus as a decrease in the *efficiency* of the solution. Lastly, most teachers indicated at some point in the interview difficulties with grading as they did not conduct the lessons preceding the given task themselves. This highlights the subjectivity of grading, which is dependent on a wider context of individual teaching and learning practices (Nardi et al., 2012).

Regarding research question 2, the criteria identified in practice can be aligned with the theoretical dimensions.

The content dimension is most prominently covered by the criteria of *existence of specific functional elements* and *correctness*, which are applied to all sorts of elements focused on in the students' solution. The pragmatic dimension (as part of the structural dimension), which is the

Table 2 Teachers' criteria when grading students' arguments

Criterion	Decisive questions (hypothetically) asked by the teacher when applying this criterion	Exemplary statements from the interviews	Corresponding dimension(s)
Rigor, rationality, logic (including derivability, traceability)	Is the argument complete, coherent, well-structured, conclusive?	"I think he used his own structure. (...) he gave it a system. He put it down systematically and comprehensibly, (...) I think that's good" (Ms. Bayer)	Structure
Necessity	Is the element necessary in order to solve the task/make the argument complete?	"His justification is a bit implicit. You can see that he put some thought into it, but his reasoning behind his opinion is not completely conclusive" (Ms. Gerau)	Structure
Importance	Are all necessary functional elements given?	"This diagram is not necessary; one representation is enough" (Mr. Bicker)	Content, Structure
Existence of specific functional elements	Is a specific functional element present, e.g., a conclusion as an explicit answer to the question?	"Looking at the numbers is not the only important thing" (Ms. Hartmann)	Content, Structure
Correctness	Is a (functional) element of the argument correct?	"Of course, he is just missing - compared to what all the others did - the mathematical justification" (Ms. Klein)	Content, Structure
Concreteness	Is the argument (too) general? Is the argument concrete enough?	"His calculation is wrong but his conclusion is correct" (Ms. Nack) "It's so nicely described, not vague, but really concrete" (Ms. Ost)	Structure, Content
Comprehensibility	Is the argument/are its elements comprehensible?	"To me, it's not enough of a justification, because I find it too general" (Ms. Ringer) "I am missing some numbers to underpin his statement. (...) I always write on tests 'as you remember, your way of calculating must be comprehensible'. And I miss a bit of the comprehensibility here" (Ms. Nack)	Recipient-Oriented
Presentation (including efficiency, esthetic and creativity)	Is the argument creative, concise, well-illustrated, 'elegant'?	"Why should I write down so much when I succeed in such a short way" (Mr. Jonas) "Well, in [the student's] favor is that he solved it a lot more skillfully by using colors" (Ms. Klein)	Recipient-Oriented
Language	Does the argument contain formal language? Is the language fitting/ precise?	"Well, the argument is great, remarkably precisely worded" (Ms. Ost) "He is a bit better, maybe, as he also puts it down nicely by using the formal symbols for probabilities" (Mr. Bicker)	Recipient-Oriented

orientation towards a specific goal, can, in addition to the *existence of specific functional elements*, be associated with the *concreteness* (as in “is the concrete yes-no-answer addressed?” and “is a concrete justification given?”). The orientation towards a recipient is evaluated by the comprehensibility; here, the teachers tacitly assumed themselves (i.e., mathematical experts) as addressee and checked if they themselves understood the solution. In addition, the criteria *language* and *presentation* are crucial for the orientation towards the recipient. Finally, the structure and logic of the functional elements of the argument is addressed by the criteria *rigor*, *rationality*, and *logic* as well as the *existence of specific functional elements*. As a side remark about the latter criterion, the explication of the general rule seemed to play a subordinate role for the teachers. While Mr. Schwarz in the example above implicitly criticized it missing in Tobi’s solution, other teachers did not perceive this as a problem. If we consider mathematics at university level, the explication of rules to justify the steps in deductions is very important. This might be interesting concerning the discussion of a gap (e.g., Balacheff, 2010) between mathematics in school and in university concerning proofs: Teachers seem to attach little value to the explication of rules in arguments at school level.

7 Conclusion and discussion

As the literature on proof in mathematics (education) (e.g., Auslander, 2008; Inglis, Mejia-Ramos, Weber, & Alcock, 2013) shows, identifying commonly accepted benchmarks for evaluation of mathematical arguments (in school) is problematic. Therefore, this paper looked at two different sources for possible criteria: first, starting from a framework for explanations (Kiel et al., 2015), three theoretical dimensions were summarized (structure, content, recipient-orientation), which are perceived to be inherent in every argument. Second, an empirical interview study was conducted in order to identify in-service teachers’ criteria developed in practice for evaluating students’ arguments (cf. Table 2).

7.1 Teachers’ criteria for evaluating written arguments (research question 1)

The analysis of the empirical data shows how teachers use a variety of different criteria, to which they assign different values. Similar to the findings by Nardi et al. (2012), some of the criteria are more logical in nature while others are more pedagogical such as the demand for high comprehensibility by using different representations. Using different criteria can also lead to quite different assessments for the same argument. This observation can be explained by the teachers’ individual practices and preferences when teaching proof and argumentation (Jahnke & Ufer, 2015; Nardi et al., 2012; Philipp, 2018).

Following Schwarzkopf (2000), an argumentation in class is perceived as a process of uttering a need for justification, which is (not) satisfied by responses from students over time. Thus, criteria such as the presentation, comprehensibility, and language are especially important in an educational context, in which students learn how to develop, exchange, and evaluate mathematical arguments. A sole focus on criteria of rationality and logic therefore seems hardly useful.

The benefit of this project is presenting a list of possible criteria which are used in practice. Certainly, this is not exhaustive, but it is likely to be expanded or modified in the future and/or depending on the situation. Other criteria might also be in use and some of them might not be useful in every situation.

7.2 Practical criteria and theoretical dimensions (research question 2)

Contrasting theoretical dimensions and practical criteria shows how they can be aligned and thus together provide a detailed framework for assessing different aspects of mathematical arguments. Of the identified criteria, many involve the dimension structure and content at the same time. For example, the criterion “importance” can be described by the question “Are all necessary functional elements given?”. The aspect “necessary” can only be evaluated by looking at the content. The recipient-orientation showed no interplay with the others, as it has a different root in a more pedagogical and social setting (Klein, 1980; Schwarzkopf, 2000).

The overall fit between the two sources for criteria (theory and practice) can be interpreted in two ways: (a) in-service teachers develop criteria in their everyday practice, which focus similar values as theoretical considerations on the nature of mathematical arguments. Thus, these dimensions might naturally emerge from dealing with arguments in educational settings. (b) The theoretical dimensions are not purely hypothetical but can be used to categorize teachers’ individual criteria. Therefore, we believe they have also value for instance for teacher education to draw on different aspects (content, structure, recipient-orientation) when learning how to assess students’ mathematical arguments.

An open question, which cannot be addressed by this study, is a (suggested) prioritization among the dimensions and criteria. For instance, the role of the recipient-orientation and its respective criteria might be reducible over time, as students learn more and more to develop an argument and later to prove according to mathematical standards (e.g., notations and derivations). However, as the discussion on mathematical proofs shows, these standards are not shared among all mathematicians (Hanna & Jahnke, 1996).

Furthermore, a more specific operationalization of each criterion, especially for the ones with more room for interpretation such as comprehensibility or presentation, is necessary in order to give a more objective ground for assessment, especially across different classrooms and teachers. As none of the teachers drew on the national or federal curriculum as source of criteria, this might indicate the need for a critical revision of these standards.

The presented results have to be interpreted considering some limitations: The interviews were conducted in a laboratory setting with students unknown to teachers and thus might be limited in representing the everyday practice in a teachers’ own classroom completely. A few of the participants commented on how they perceived criteria for evaluation as context- and lesson-dependent, so that they had difficulties with applying their everyday practice. However, while we cannot assume to perceive naturalistic evaluation processes, we can gain insight into how teachers tackle the task to give a certain grade for a certain argument by drawing on knowledge, expectations, and values, which were most likely developed in their everyday practice.

Lastly, the study aims solely to explore the practice of grading students’ arguments in a summative assessment in order to rank their achievement; formative assessment with the intention of facilitating the students’ learning process plays an even more important role in teaching mathematics in classrooms. It remains to be analyzed if the presented criteria applied by teachers would be the same and especially if they would have the same importance in such a setting. However, the present study can be understood as an important step in the exploration and clarification of which criteria are and can be used when students are asked to develop and present arguments.

Taking these results seriously, one could summarize them pointedly: An argument can only be a good argument in a specific context, checked by a specific evaluator. This might highlight one of the perceived irritations when students advance from high school to university: While in

school, arguments are introduced and evaluated against more pedagogical criteria (which are probably not all made explicit), university emphasizes more structural criteria such as rigor and logic (among others the explication of the rule according to Toulmin, 1996).

The explication of quality criteria for arguments also has a constructive component: Not only can teachers ask themselves “What else can I consider to evaluate arguments of my students in the future?” but also “How should I present my own argument to be easily understandable and function as a model for my students’ arguments?” (extending the suggestions in Meyer & Prediger, 2009). The presented study elucidated teachers’ reasons and practices for evaluating arguments and showed the richness of criteria which (can) be useful for answering the question “when is an argument an (good) argument?”

Acknowledgments We would like to thank the editor and the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Funding Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Auslander, J. (2008). On the roles of proof in mathematics. In B. Gold & R. A. Simons (Eds.), *Proofs and other dilemmas: Mathematics and philosophy* (pp. 61–77). Washington, DC: Mathematical Association of America.
- Balacheff, N. (2010). Bridging knowing and proving in mathematics. In G. Hanna, H. N. Jahnke, & H. Pulte (Eds.), *Explanation and proof in mathematics* (pp. 115–135). Heidelberg, Germany: Springer.
- Bex, F., & Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument & Computation*, 7, 55–68.
- Buth, M. (1996). *Einführung in die formale Logik unter der besonderen Fragestellung: Was ist die Wahrheit allein aufgrund der Form*. Frankfurt am Main, Germany: Lang.
- Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, 38(2), 247–264.
- Davis, P. J., & Hersh, R. (1981). *The mathematical experience*. Boston, NJ: Birkhaeuser.
- Dreyfus, T. (2002). Was gilt im Mathematikunterricht als Beweis? In W. Peschek (Ed.), *Beitraege zum Mathematikunterricht* (pp. 15–22). Hildesheim, Germany: Franzbecker.
- Hanna, G. (1989). More than a formal proof. *Learning Mathematics*, 9(2), 20–23.
- Hanna, G., & Jahnke, H. N. (1996). Proof and proving. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 877–908). Dordrecht, the Netherlands: Kluwer.
- Healy, L., & Hoyle, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, 31(4), 396–428.
- Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135–175.

- Inglis, M., Mejia-Ramos, J. P., & Simpson, A. (2007). Modelling mathematical argumentation: The importance of qualification. *Educational Studies in Mathematics*, 66(1), 3–21.
- Inglis, M., Mejia-Ramos, J. P., Weber, K., & Alcock, L. (2013). On mathematicians’ different standards when evaluating elementary proofs. *Topics in Cognitive Science*, 5, 270–282.
- Jahnke, H. N., & Ufer, S. (2015). Argumentieren und Beweisen. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme, & H.-G. Weigand (Eds.), *Handbuch der Mathematikdidaktik* (pp. 331–355). Berlin, Germany: Springer.
- Jeannotte, D., & Kieran, C. (2017). A conceptual model of mathematical reasoning for school mathematics. *Educational Studies in Mathematics*, 96(1), 1–12.
- Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, 89(3), 337–355.
- Kempen, L., & Biehler, R. (2015). Pre-service teachers’ perceptions of generic proofs in elementary number theory. In K. Krainer & N. Vondrová (Eds.), *Proceedings of the CERME 9* (pp. 135–141). Prague, Czech Republic: Charles University.
- Kiel, E., Meyer, M., & Mueller-Hill, E. (2015). Erklären – Was? Wie? Warum? *Praxis der Mathematik in der Schule*, 64(57), 2–9.
- Klein, W. (1980). Argumentation und Argument. *Zeitschrift fuer Literaturwissenschaften*, 38(39), 9–57.
- Kopperschmidt, J. (2000). *Argumentationstheorie*. Hamburg, Germany: Junius.
- Krauthausen, G. (2001). Wann faengt das Beweisen an? In W. Weiser & B. Wollring (Eds.), *Beitraege zur Didaktik der Mathematik fuer die Primarstufe* (pp. 99–113). Hamburg, Germany: Verlag Dr. Kovac.
- Mayring, P. (2000). Qualitative content analysis. *Forum Qualitative Social Research*, 1(2), Art. 20 <http://nbn-resolving.de/urn:nbn:de:0114-fqs0002204>. Accessed 16 Aug 2020.
- Meyer, M. (2010). Abduction – A logical view of processes of discovering and verifying knowledge in mathematics. *Educational Studies in Mathematics*, 74(2), 185–205.
- Meyer, M., & Prediger, S. (2009). Warum? Argumentieren, Begründen, Beweisen. *Praxis der Mathematik in der Schule*, 50(30), 1–7.
- Meyer, M., & Schnell, S. (2016). Was ist ein “gutes” Argument? In Institut fuer Mathematik und Informatik der Paedagogischen Hochschule Heidelberg (Ed.), *Beitraege zum Mathematikunterricht 2016* (pp. 667–670). Munster, Germany: WTM-Verlag.
- Nardi, E., Biza, I., & Zacharides, T. (2012). ‘Warrant’ revisited: Integrating mathematics teachers’ pedagogical and epistemological considerations into Toulmin’s model for argumentation. *Educational Studies in Mathematics*, 79(2), 157–173.
- Philipp, K. (2018). Diagnostic competences of mathematics teachers with a view to processes and knowledge resources. In T. Leuders, K. Philipp, & J. Leuders (Eds.), *Diagnostic competence of mathematics teachers* (pp. 109–128). Heidelberg: Springer.
- Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. *Science Education*, 92, 447–472.
- Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., ... Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, 52(4), 1893–1907. <https://doi.org/10.1007/s11135-017-0574-8>
- Schwarzkopf, R. (2000). Argumentation processes in mathematics education – Social regularities in argumentation processes. In GDM (Ed.), *Developments in mathematics education in Germany* (pp. 139–151). Potsdam, Germany: Franzbecker.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14.
- Toulmin, S. E. (1969). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- Voigt, J. (1995). Thematic patterns of interaction and sociomathematical norms. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning* (pp. 163–201). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 27(4), 458–477.