What do mathematicians mean by proof? A comparative-judgement study of students' and mathematicians' views

Ben Davies, Lara Alcock, and Ian Jones

Contact:
Ben Davies
Loughborough University
Loughborough
LE11 3NT
Tel.: +44-1509-22-8212
Email: b.m.j.davies@lboro.ac.uk

Abstract

We present a study in which mathematicians and undergraduate students were asked to explain in writing what mathematicians mean by proof. The 175 responses were evaluated using comparative judgement: mathematicians compared pairs of student responses and their judgements were used to construct a scaled rank order. The judgements allowed us to compare the quality of student and mathematician responses and, via further qualitative analysis, to identify which features of responses the judges collectively valued. We establish the reliability of comparative judgement in this context and provide evidence for the validity of this approach to investigating beliefs about proof. Substantively, our findings reveal that despite the variety of views found in the literature, mathematicians broadly agree on what people should say when asked what mathematicians mean by proof.

Keywords: Comparative judgement, Beliefs, Proof, Reliability, Validity, Mathematicians

What do mathematicians mean by proof? A comparative-judgement study of students' and mathematicians' views

Authors redacted

Contact information redacted

Abstract

We present a study in which mathematicians and undergraduate students were asked to explain in writing what mathematicians mean by proof. The 175 responses were evaluated using comparative judgement: mathematicians compared pairs of student responses and their judgements were used to construct a scaled rank order. The judgements allowed us to compare the quality of student and mathematician responses and, via further qualitative analysis, to identify which features of responses the judges collectively valued. We establish the reliability of comparative judgement in this context and provide evidence for the validity of this approach to investigating beliefs about proof. Substantively, our findings reveal that despite the variety of views found in the literature, mathematicians broadly agree on what people should say when asked what mathematicians mean by proof.

Keywords: Comparative judgement, Beliefs, Proof, Reliability, Validity, Mathematicians

1. Introduction

Proof is central to mathematics education and is a well-known stumbling block for many students (Moore, 1994; Weber, 2001; Alcock and Weber, 2005). It has been argued that one reason for this is that students' beliefs about proof often do not align with those accepted by the community of practising mathematicians (Dawkins and Weber, 2017; Harel and Sowder, 1998; Healy and Hoyles, 2000; McLeod, 1992). However, despite the growing literature on proof, we have limited evidence documenting students' beliefs (Stylianou et al., 2015). Relevant research does exist: Healy and Hoyles (2000) for instance, asked a large number of secondary school students for 'written descriptions of the purposes of proof' (p. 404). Their coding, however, was based on a simplified version of the taxonomy of de Villiers (1990) and captured only purposes related to truth, explanation or discovery. Questionnaire-based studies also offer only coarse-grained analyses: Stylianou, Blanton and Rotou (2015), for instance, administered a closed-form test of proof conceptions comprising five prejudged multiple-choice items. To understand the role of proof conceptions in learning, Stylianou et al. compared these test scores to self-efficacy beliefs and attainment on more standard proof comprehension tasks. Both Healy and Hoyles and Stylianou et al. provide important initial insights into students' conceptions of proof and their role in the complex process of learning. However, both approaches necessarily miss the variety in beliefs about proof; we lack fine-grained analyses of what individuals say when offered the opportunity to express their views freely. Moreover, studies to date provide no systematic empirical evidence on mathematicians' beliefs about proof. This weakens investigations we might wish to undertake in which student and expert views are compared.

We address these gaps in the literature by investigating beliefs about proof as expressed by undergraduate and research-active mathematicians. Our study asked students and mathematicians to 'explain what mathematicians mean by proof in 40 words or less'. We evaluated the responses to the task using a technique based on comparative judgement which we describe later in the manuscript. First, we summarise the literature on beliefs and mathematical proof.

2. Theoretical Background

2.1. Defining mathematical proof

Understanding beliefs about proof is not straightforward, in part because a definition of proof is elusive. Various definitions have been proposed, from the strictly formalistic to more socially oriented, context-dependent definitions such as those offered by Aberdein (2009) and Dawkins and Weber (2017). Consequently, emphases in research vary. Aberdein, for instance, focused on proof as a form of argument, intended as nothing more than communication between mathematicians. Dawkins and Weber also noted the communicative role of proofs, while emphasising their role in establishing knowledge in the field by turning mathematical assertions into theorems. Others have emphasised the role of proof in ascertaining and persuading (Harel and Sowder, 1998), in systematising mathematics (de Villiers, 1990), and as the fundamental mechanism for mathematical progress (Fawcett, 1938; de Villiers, 1990).

Researchers in mathematics education have also operationalised the notion in different ways for different research contexts. In a survey of the various definitions of proof present in the education literature, Balacheff (2008) observed that a multitude of epistemologies motivate the various definitions and that we 'need an organisation of our work at an international level, beyond our idiosyncratic views or possible tendency to accept ready-made ideas' (p. 16). Both Reid and Knipping (2010) and Hanna and de Villiers (2012) responded to the need identified by Balacheff, providing comprehensive reviews of progress in the field. Both reviews serve as resources for future researchers to situate their work within the field, but do not resolve the problem of diversity identified by Balacheff.

One might ask, however, whether this apparent lack of consensus on the meaning of proof is important in understanding the behaviour of mathematicians. Weber and Czocher (2019) investigated this question, asking nearly 100 mathematicians in an online study to evaluate five proofs with varied characteristics. Of the five, two were described as prototypical textbook proofs, one was exclusively visual, one was computer-assisted and one based on empirical evidence alone. Consistent with the theoretical diversity discussed above, the authors reported that mathematicians were divided on both the

computer-assisted and visual proof, deemed valid by 62% and 39% of participants, respectively. However, the other three proofs yielded near-total agreement ($> 98\%$). The prototypical proofs were accepted by mathematicians. The empirical argument was unanimously rejected. These findings paint a picture of a centralised consensus with ambiguity and diversity at its periphery, suggesting that mathematicians likely agree on validity for the majority of the proofs that they 'typically encounter' (p. 12).

Based on such observations of the mathematical community, Czocher and Weber (2019) adopted a different approach to defining proof. The authors argued that criteria-based accounts of proof are necessarily doomed to failure and that proof can be more profitably be viewed as a *cluster category* (Lakoff, 1987). Lakoff's notion of a cluster category is an extension of Wittgenstein's family resemblance (Wittgenstein, 1953), built on the premise that an object with more properties consistent with membership is more likely to belong than one with less. By extension, a *cluster category* is then a collection of properties that an object can satisfy, 'counting toward' membership of the given category. This probabilistic (as opposed to deterministic) structure denies the binary notion of belonging inherent in criteria-based accounts.

Czocher and Weber (2019) defined a cluster account of proof as a convincing, perspicuous, a priori, transparent 'justification that has been sanctioned by the mathematical community' (p. 20). This account acknowledges the diversity in the literature by embedding various criteria-based definitions as properties indicating membership of the category 'proof', without dictating that any given property be satisfied by every member. Further, as noted by the authors, this definition is consistent with the findings that mathematicians agree on the validity of many but not all purported proofs, and that some proofs are more 'typical' than others.

We agree with the approach of Czocher and Weber (2019), but note that this is a theoretical resolution to the problem of defining proof; it does not solve the problem of how to research how people do, can or should think about it. Of course, this problem is no worse than that encountered in any area of social science in which a human concept must be operationalised. But it demands careful attention, and preferably methods that allow us to assess the quality of knowledge and understanding while respecting diversity in the detail. Comparative judgement, as used here, can do exactly that. It permits a

meaningful evaluation of responses to open-ended questions without specifying criteria for what should or should not be included. To date, its use in mathematics education has been commonly directed at assessing conceptual understanding (Bisson et al., 2016; Jones and Karadeniz, 2016) and problem solving (Jones and Inglis, 2015). In the wider psychological literature, it has been applied to the investigation of beliefs (e.g. Thurstone, 1928, 1954). Here, we apply comparative judgement to investigating beliefs about mathematical proof.

2.2. Evaluating students' beliefs about proof

Epistemological beliefs are considered important in understanding students' engagement with educational opportunities. They have a long history in education research, with much work traceable to Perry (1968), who mapped the developmental trajectory of Harvard undergraduates from simple 'dualistic' thinkers to believers in complex, dynamic and tentative knowledge. In recent decades, the study of beliefs in mathematics education has grown into its own subfield of educational research. In her review, Muis (2004) summarised this body of work into five categories: beliefs, effects of beliefs, development of beliefs, changing beliefs, and beliefs as domain-specific vs. domain-general entities. Depaepe et al. (2016) added a further category focused on teachers' mathematical epistemologies, reflecting the research published in the intervening years between reviews.

Across this research, there is a general view that some beliefs are more naive and some more sophisticated. Naive epistemological beliefs often characterise knowledge in terms of absolute truths that are handed down by authorities; more sophisticated beliefs characterise it as evolving in social contexts (Perry, 1968). Naive mathematical beliefs often characterise mathematics as a set of procedures to be memorised; more sophisticated beliefs characterise it as a web of logically connected information (Crawford et al., 1998). Naive beliefs about proof are often empiricist; more sophisticated beliefs recognise the value of deductive arguments (Harel and Sowder, 1998; Muis, 2004).

Methodologically, interest in beliefs in relation to other constructs generates a need for reliable measures. The result of this has been a natural shift from more detailed, qualitative studies to the development of quantitative scales: Depaepe et al.

(2016) observed that recent research evaluating students' beliefs about mathematics has been dominated by large-scale questionnaires (Healy and Hoyles, 2000; Nasser and Birenbaum, 2005; Schommer-Aikins, 2004; Stylianou et al., 2015). Implementations vary between the extremes: Healy and Hoyles (2000) requested open-ended descriptions of 'proof and its purposes' then categorised these based on a pre-defined taxonomy; Stylianou et al. (2015) asked students both to identify the purposes of given proofs and to select responses that matched their beliefs about proof and themselves as learners thereof; Muis (2008) used inventories reflecting students' epistemic profiles and learning strategies. Results from these studies broadly reflect what we would expect: where data are related to students' performance on other mathematical tasks, those assessed as having more sophisticated beliefs typically performed better (Stylianou et al., 2015; Healy and Hoyles, 2000).

Large-scale questionnaire-based work thus provides useful insights into how students view the world of mathematics and into how these views relate to performance. However, pre-defining responses to a question limits the scope of research for accessing new ideas. This is disadvantageous where simple categories do not do justice to the complex reality of the situation, as is the case with beliefs about proof. Research has established that professional mathematicians do not always think in terms of deductive inference: they also make sophisticated use of empirical evidence, visual information and authoritative sources (Weber et al., 2014). Undergraduate students who offer empirical evidence for a claim often know that this form of evidence is not adequate for proof (Weber, 2010). Such nuances merit attention if we are to establish ways to engage students with authentic mathematical activity and to assess their performance. However, researchers who have explicitly asked students to express beliefs on mathematical topics (Knuth, 2002; Zaslavsky and Shir, 2005) have to date lacked a systematic way to compare responses. Through our comparative judgment-based approach, we capture the richness of individuals' beliefs about proof with an open-ended task while using a systematic tool to evaluate and score responses.

2.3. Comparative judgement

Comparative judgement addresses a specific need in assessment and thereby in educational research: it offers a systematic way to quantify responses to tasks for which success criteria are hard to define and rubric-based scoring is likely to fall short (Jones et al., 2019). It works by asking judges to make pairwise comparisons of responses to a task. For each comparison, the judges simply decide which response is better, meaning that they can make global judgements without having to decide that any specific content should always be present or always be equally valued. The judgements are then fitted to the Bradley-Terry Model to generate a scaled rank order in which each response receives a parameter estimate reflecting its quality (Bramley, 2007). Scores are therefore 'grounded in the collective expertise of the judges' (Bisson et al., 2016, p. 143).

In mathematics education, comparative judgement has been used to assess constructs including conceptual understanding (Jones and Karadeniz, 2016; Jones et al., 2019; Bisson et al., 2016), problem solving (Jones and Inglis, 2015) and general reasoning in primary school (Hunter and Jones, 2018). In all cases, the construct being assessed is considered important but resists rubric-based assessment. In all cases, the researchers reported evidence for satisfactory reliability and validity of their comparative judgement-based analyses.

Our work differs from these studies in that beliefs about proof are not the same as conceptual understanding or problem solving. However, assessing beliefs presents similar methodological challenges: beliefs, too, are considered important but resist criterion-based assessment. Comparative judgement is therefore a suitable approach because we do not need to force respondents to choose between pre-defined responses, so we do not constrain their ability to respond authentically. We do not need to classify their responses according to pre-defined categories, so we do not lose nuances of meaning. We do not need to decide *a priori* what constitutes a 'correct' or 'good' or 'sophisticated' view of proof, so we do not need to align ourselves with any specific view in the existing literature. We do not need our judges to agree to a set of pre-defined criteria, so we respect their individual professional expertise and collective diversity, while also capitalising on shared understanding of typical cases.

3. Methods

3.1. Task, participants and procedures.

Our research used a task in which participants were asked to 'explain what mathematicians mean by proof in 40 words or less'. Respondents were 130 undergraduate students and 45 research-active mathematicians. Student respondents were all enrolled in the same introductory real analysis course at a UK university. Their responses were collected during a lecture in week eight, when they were given 10 minutes to complete the task[1]. All attendees completed the task; research participation was made voluntary by providing an opt-out option. No module credit was associated with the task. Mathematician respondents were invited to participate by email or in person after academic presentations at UK universities. An e-version of the 'explain' task was produced using onlinesurveys.com to facilitate remote recruitment. All responses to the 'explain' task were typeset in an identical format to remove the potential influence of handwriting.

3.2. Judging data

The typeset task responses were judged twice, each time by an independent group of judges. One group of judges were experts and the other were non-experts.

Expert judges. The 29 expert judges were research-active mathematicians. They were asked to complete between 20 and 100 judgements each. The minimum was given to encourage judges not to perform a trivial number of judgments and the software was set to allow no more than 100 judgements per judge. The expert judges completed a total of 1941 judgements, with each completing between 11 and 100 judgements (median 86). Each response received between 20 and 27 expert judgements (median 22), and the median time spent on each judgement was 10.6 seconds. The expert judges were not compensated for their time.

[1]The task appeared third in a three-task booklet also containing a proof and two associated comprehension tasks. Students had 40 minutes to complete the booklet, of which 10 minutes were nominally allocated to the 'explain' task.

Non-expert judges. The 10 non-expert judges comprised eight post-graduate students and two working professionals deemed expert in the English language. Importantly, the non-experts had not studied mathematics beyond the age of 16 and so were unable to make judgments based on specialist mathematical knowledge. We expected the non-experts to be unable to replicate the outcomes of the expert judges, thereby enabling us to check that the experts did not base their judgements solely on surface features such as quality of prose (Jones and Alcock, 2014). The non-expert judges were asked to complete 175 judgments each. In practice, each performed between 172 and 175 judgements resulting in 1740 judgements. Each response received between 20 and 23 non-expert judgments (median 21), and the median time per judgement was 14.9 seconds. The non-expert judges were compensated for their time based on a pre-defined rate of 20 seconds per judgement.

For each group of judges, the binary decision data were fitted to the Bradley-Terry Model (Firth, 2005) to generate a parameter estimate of the perceived quality of each response. The parameter estimates were then used to construct scaled rank orders of responses. The experts' scale had a mean of 0.07 and standard deviation of 1.48, and the non-experts' scale had a mean of 0.08 and standard deviation of 0.93.

4. Analysis and results

4.1. Reliability

To investigate whether the expert judgement-based parameter estimates formed a meaningful measure of response quality, we first examined the internal consistency and the inter-rater reliability of the scaled rank orders.

Internal consistency was measured by calculating the Scale Separation Reliability (SSR), which is often considered analogous to Cronbach's α (Pollitt, 2012). Inter-rater reliability was measured using a split-half comparison technique described in Bisson et al. (2016). Briefly, the judges were randomly divided into two groups, a new scale constructed for the judgements of each group, and the Pearson Product-Moment correlation coefficient was calculated for the two sets of parameter estimates. The

process was repeated 100 times and we report the median correlation coefficient here. We also considered judge and response misfit (Pollitt, 2012) as secondary measures of reliability.

For the expert judges' parameter estimates, internal consistency and inter-rater reliability were both acceptable, $SSR = .83$ and $r = .68$. No judge and only five of the 175 responses (3%) had misfit scores more than two standard deviations above their respective means. For the non-expert judges' parameter estimates, internal consistency was acceptable, $SSR = .66$, but inter-rater reliability was low, $r = .39$. No judge and only three responses (2%) had misfit scores more than two standard deviations above their respective means.

We interpret these findings as evidence that comparative judgement conducted by experts produced a reliable scale of quality for responses to the task, whereas that conducted by the non-experts did not. We further interpret this to mean that despite the diversity of views of proofs across the literature, mathematicians do broadly agree on what they would like people to say about proof.

4.2. Validity

To investigate whether the expert judgement-based parameter estimates formed a measure of response quality that was not only internally meaningful but valid, we used three methods: comparison of expert and non-expert reliability, comparison of mathematician and student response rankings, and a detailed content analysis of the responses.

First, the reliability analysis reported above provides evidence not just that the experts were consistent with one another, but also that their judgements were based on the mathematical content of the responses rather than surface features such as quality of prose. In contrast, the non-experts, who by definition made their judgements based on features other than mathematical content, were substantially less consistent with one another. Importantly, the non-experts were not consistent with the experts. The correlation between the two scales was modest, $r = .54$, as shown in Figure 1, and was lower than the inter-rater reliability of the experts, $r = .68$; this difference was significant, $Z = 2.06$, $p = .04$. In summary, the outcomes of the non-expert judging

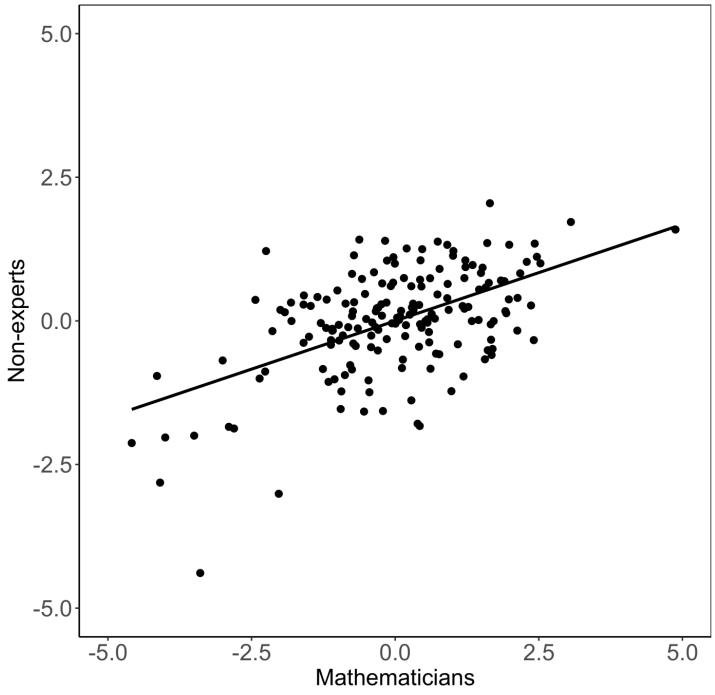

Figure 1: Scatter plot comparing the experts' and non-experts' parameter estimates.

support evidence for the divergent validity of the expert outcomes. In the remainder of the article we analyse only the experts' parameter estimates.

Second, we compared parameter estimates for mathematicians' and students' responses using a two-sample t-test. On average, mathematicians' responses received higher parameter estimates ($N = 45, M = 1.23, SD = 1.15$) than undergraduates' responses ($N = 130, M = -0.43, SD = 1.34$), as shown in Figure 2. This difference was significant, $t(88.75) = 7.95, p < .001$, with a large effect size of $d = 1.32$. We interpret these findings as evidence that comparative judgement provided a valid measure of quality for responses to our task: as we would expect, experts consistently judged responses from research mathematicians as better than responses from undergraduate mathematicians.

Third, we conducted a detailed content analysis of the responses; we report this in a separate section below.

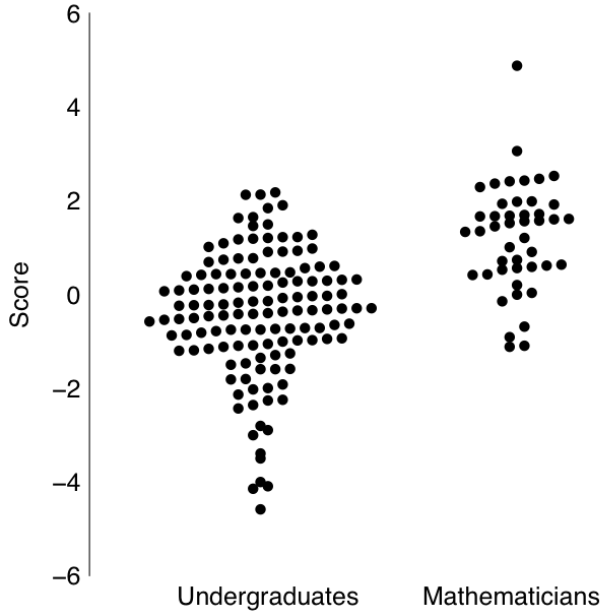

Figure 2: Comparison of experts' parameter estimates assigned to responses from undergraduate and research mathematicians.

4.3. Content analysis

The above analyses support the claim that comparative judgement yielded a reliable and valid scaled rank order of responses to the task. To further investigate validity we analysed the contents of the undergraduate and research mathematicians' responses, following the principles of thematic analysis set out by Braun and Clarke (2006). This facilitated a detailed analysis of what features of the responses were favoured in the experts' judging decisions, and a comparison of the differences between students' and mathematicians' responses.

To provide the reader with a sense of the types of responses provided by the two groups of participants, Table 1 shows the top three task responses from the research mathematicians and undergraduates, along with their overall ranked position within the parameter estimates. The three top responses all came from research mathematicians, although two undergraduate responses were included in the top ten. A full list of responses, ordered by parameter estimates, is available at [FigShare URL pending].

Rank	Response
	Research mathematician responses.
1st.	A logically coherent argument establishing the truth of an assertion from a known and agreed base.
2nd.	A comprehensive logical argument that a statement is true, based on clearly formulated assumptions and following generally accepted lines of reasoning and level of detail.
3rd.	A proof is a checkable record of reasoning establishing a fact from agreed, more basic assumptions.
	Undergraduate responses.
9th.	Proof is a logical argument in mathematics which uses previously proven theorems and ideas to build upon and generate new mathematics. It is there to show whether something is true or not.
10th.	A reasoning or logic that shows that a statement is the inevitable result of a set of assumptions.
11th.	Proof means to show something to be true by using things already proved to show the new thing your [sic] trying to prove to be true.

Table 1: The top three task responses from the research mathematicians and from the undergraduates, along with their overall ranked position within the parameter estimates.

4.3.1. Content analysis: coding

To analyse the content of all 175 responses, we developed a coding scheme via an iterative process of identifying common themes in subsets of the data. First, two researchers together examined a subset of 10 responses, generating an initial scheme of eight codes. Both researchers then independently applied this scheme to 10 further responses, noting possible additions to and mergers of the codes. Discrepancies were discussed and a new scheme with an additional two codes was generated. This process was repeated with a further set of 10 responses, leading to an 11-code scheme to be applied to the entire dataset.

A third researcher then joined the analysis team to replace a member who was no longer available. Both researchers analysed all 175 responses, noting points of uncertainty for further discussion and possible necessary amendments. In a penultimate analysis session, discrepancies were discussed, and a further four codes were added. Finally, both researchers re-examined all 175 scripts looking for evidence of these four

new codes. To evaluate inter-coder reliability, we examined pooled Cohen's Kappa; this yielded $\kappa = 0.79$, indicating acceptable inter-coder reliability. Given the volume of interaction throughout the coding process, this κ is likely an over-estimate of true inter-coder reliability. However, 0.79 is high enough to suggest that reliability is acceptable. All remaining disagreements were discussed case by case, until a consensus was found for each code as applied to each response. The resulting 15 codes, with frequencies of occurrence in the mathematicians' and students' responses, appear in Table 2.

Code	Description	Experts	UGs	χ^2	p
Argumentation	Reference to an 'argument', 'chain of reasoning' or 'derivation'.	80%	21%	50.90	<.001*
Object	Naming the object to be proved, e.g. 'theorem', 'statement', 'result'.	80%	82%	0.05	.820
Certainty	Reference to 'truth' or 'correctness'.	44%	76%	15.44	<.001*
Established knowledge	Reference to 'agreed assumptions' or 'shared knowledge'.	38%	29%	1.13	.287
Conviction	Reference to the readers' increased conviction in the statement.	22%	2%	22.39	<.001*
Conditions	Reference to the domain of applicability for a statement.	20%	25%	0.40	.529
Explanation	Reference to 'how' or 'why' the statement is true.	16%	23%	1.13	.287
Verification	Reference to 'confirms', 'validates', 'checks', 'justifies', or 'shows if...'.	16%	9%	1.38	.240
Axiom	Use of the term 'axiom'.	13%	8%	0.90	.341
Deconstruction	Reference to 'breaking down' the theorem into familiar truths.	7%	5%	0.10	.749
Discovery	Reference to proving something 'not already known'.	7%	9%	0.28	.596
Incontrovertability	Reference to 'undoubted', 'cannot be argued with'.	7%	13%	1.36	.244
Empiricism	Reference to empirical evidence.	4%	8%	1.36	.376
Falsification	Reference to disproving a statement.	2%	27%	12.48	<.001*
Generality	Reference to 'all cases'.	0%	18%	9.17	.002*

Table 2: 'Explain' task content analysis coding scheme. Each code was assigned to each response at most once. Experts = research-active mathematicians, UGs = undergraduate students. Codes ordered by frequency in expert responses. Significance determined based on alpha-level adjustment using the Holm-Bonferroni method.

4.3.2. Comparing students' and mathematicians' responses

To compare the responses given by students and mathematicians, independent chi-squared tests were run for each of the 15 codes; results are also shown in Table 2. For five codes there were significant differences between students' and mathematicians' responses. Mathematicians were significantly more likely to refer to argumentation and conviction, and students were significantly more likely to refer to certainty, falsification and generality. We discuss students' emphasis on falsification and generality first as we believe these to be artefacts of the immediate educational environment from which they were recruited. We then discuss the differences we find more epistemologically interesting.

The students were all from an introductory real analysis course with two features that might have promoted the student emphasis on *falsification* and *generality*. First, 'true or false' tasks were a common feature of formative and summative assessment, likely promoting a connection between proof and falsification. Second, emphasis on quantifiers was a common feature of lectured information, likely leading students to refer to generality in their explanations of proof. The emphasis on quantifiers and generality might be similar in other real analysis courses, although the UK context meant that this was a first-year, first-semester course, so it might be less explicit where real analysis is taught later. The extensive use of 'true or false' tasks is almost certainly unusual. We thus hesitate to suggest that one should expect similar findings in different contexts.

The remaining differences are consistent with the broad sweep of research on epistemological beliefs and their development. The notion that proofs provide *certainty*, more common in students' responses, is consistent with an idea of mathematics as the business of truth, and proof as the business of demonstrating that truth. Certainty can also be viewed as consistent with the day-to-day experience of students via the 'definition-theorem-proof' structure of much undergraduate mathematics (Moore, 1994), where proofs are often presented as bearing authority (Harel and Sowder, 1998). It is worth noting, however, that although certainty was more common in students' responses, it was also the third most frequently applied code for mathematicians' responses. Thinking

of proof in terms of certainty does not necessarily indicate a lack of sophistication.

The notion that proofs involve *argumentation* and provide *conviction*, more common in mathematicians' responses, suggest a socially constructed view of mathematics in which proofs are written for an audience. This is consistent with some characterisations put forth by philosophers, mathematicians and educators. Our frequency analysis provides evidence that of the various views of proof discussed in the literature, argumentation is most important in the views of working mathematicians: 80% of mathematicians' responses included it.

4.3.3. Features most valued by mathematician judges

We shift now from comparing responses to identifying the content most rewarded by the mathematician judges. We first examined the Spearman's rank-order correlations between response codes and parameter estimates; see Table 3. Only *argumentation* yielded a significant relationship with parameter estimates. This is consistent with the chi-squared analyses, confirming argumentation as the most important aspect of proof to the mathematicians in their judgements as well as in their stated views.

Code	r	p
Argumentation	0.48	$< .001$*
Object	0.14	.059
Certainty	0.00	.981
Established knowledge	0.21	.005
Conviction	0.14	.073
Conditions	0.00	.959
Explanation	-0.11	.141
Verification	0.10	.178
Axiom	0.21	.006
Deconstruction	-0.02	.812
Discovery	0.12	.123
Incontrovertibility	0.17	.095
Empiricism	-0.10	.207
Falsification	-0.02	.788
Generality	-0.12	.115

Table 3: Spearman rank-order correlation coefficients showing relationships between parameter estimates of proof conception quality and individual codes. Significance determined based on alpha-level adjustment using the Holm-Bonferroni method.

.

We then conducted regression analyses predicting parameter estimates based on the coding. Our first regression model used all 15 codes and yielded four significant predictors; see Table 4. Our second model used only those four codes; again, see Table 4.

Code	15-code model				4-code model			
	B	SE	β	p	B	SE	β	p
Argumentation	1.47	0.21	7.07	$< .001$*	1.52	0.19	7.83	$< .001$
Object	0.64	0.25	2.53	.013*	0.66	0.24	2.78	.006
Certainty	0.18	0.23	0.78	.436				
Established knowledge	0.57	0.23	2.47	.015*	0.78	0.20	3.89	$< .001$
Conviction	0.43	0.41	1.06	.292				
Conditions	0.20	0.23	0.88	.380				
Explanation	-0.35	0.24	-1.48	.140				
Verification	0.39	0.31	1.24	.218				
Axiom	0.49	0.35	1.40	.164				
Deconstruction	0.07	0.41	0.18	.856				
Discovery	0.49	0.35	1.41	.162				
Incontrovertibility	0.63	0.30	2.11	.036*	0.59	0.28	2.02	.045
Empiricism	-0.39	0.36	-1.07	.285				
Falsification	0.00	0.25	0.01	.993				
Generality	-0.05	0.29	-0.19	.850				

Table 4: Forced-entry multiple regression model predicting parameter estimates with coded analysis of proof conceptions. For the 15-code model $R^2 = .38, p < .001$, and for the four-code model, using significant predictors from the 15-code model, $R^2 = .33, p < .001$.

Again, *argumentation* was identified by the regression analyses as most closely associated with high value in the mathematicians' judgements. The relatively high value associated to responses coded as referring to the *object* being proven (the theorem, claim or similar) probably reflects clarity of responses: more than 80% of all responses featured a reference to such an object, so those that did not perhaps suffered from a lack of clarity or specificity. The remaining codes associated with high value captured the views that proofs build on *established knowledge* and should be *incontrovertible*. Both

are consistent with the values and norms of mathematics suggested by Dawkins and Weber (2017). Although these codes appeared less frequently than argumentation in mathematicians' responses to the 'explain' task, mathematicians deemed them important when presented. Again this provides empirical evidence on how mathematicians view their craft.

5. Discussion

We view our work as making two original and rigorously generated contributions: first, an empirical contribution to the theoretical debate around the meaning of proof; second, a methodological contribution regarding a novel application of comparative judgment to the domain of mathematical beliefs.

Empirically, we documented responses of students and mathematicians to a task asking them to explain what mathematicians mean by proof. We provided evidence that mathematicians commonly think of proof in terms of argumentation: this was reflected both in the number of mathematicians who included reference to argumentation in their own responses and in the statistical modelling identifying the priorities of the mathematician judges. Also valued by mathematicians were characterisations of proof in terms of certainty (from the frequency analysis) and in terms of building on established knowledge and generating incontrovertible arguments (from the statistical modelling). These are all themes that appear in the extant literature: what we add here is empirical evidence on their relative importance to mathematicians in describing their craft.

Methodologically, we established that comparative judgment can be applied successfully in the realm of beliefs, where reliable and valid instruments are difficult to generate and where other approaches necessarily involve coarse-grained analyses or lack the capacity for systematic quantitative comparison. Using comparative judgement with expert judges, we were able to generate a reliable scaled rank order of responses to an open-ended task without using a pre-determined definition of proof. Regarding validity, we presented three distinct pieces of evidence suggesting that our parameter estimates function as a meaningful measure of response quality. First, we found relatively poor performance of non-expert judges, suggesting that the original judgements were based

on mathematical expertise. Second, mathematicians' responses ranked higher than those from undergraduates, showing that mathematical expertise was systematically related to parameter estimates. Third, this difference was the result of content-based differences among the responses and was consistent with the literature on students' and mathematicians' experiences with proof.

Together, these findings have implications for research on both comparative judgment and beliefs. To the comparative judgment literature, we add another context for which the method appears to have utility. To the beliefs and conceptions literature, we offer a new tool for quantifying and understanding explicitly-stated beliefs on potentially wide-ranging topics.

References

Aberdein, A. (2009). Mathematics and argumentation. *Foundations of Science*, 14(1):1–8.

Alcock, L. and Weber, K. (2005). Proof validation in real analysis: Inferring and checking warrants. *Journal of Mathematical Behavior*, 24(2):125–134.

Balacheff, N. (2008). The role of the researcher's epistemology in mathematics education: An essay on the case of proof. *ZDM*, 40(3):501–512.

Bisson, M. J., Gilmore, C., Inglis, M., and Jones, I. (2016). Measuring Conceptual Understanding Using Comparative Judgement. *International Journal of Research in Undergraduate Mathematics Education*, 2(2):141–164.

Bramley, T. (2007). Paired Comparison Methods. In Newton, P., Baird, J., Goldstein, H., Patrick, H., and Tymms, P., editors, *Techniques for monitoring the comparability of examination standards*, (pp. 246–296) London, UK: Qualifications and Curriculum Authority.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Crawford, K., Gordon, S., Nicholas, J., and Prosser, M. (1998). University mathematics students' conceptions of mathematics. *Studies in Higher Education*, 23(1):87–94.

Czocher, J. A. and Weber, K. (2019). Proof as a Cluster Category. *Journal for Research in Mathematics Education*. DOI: 10.1080/14794802.2019.1585936.

Dawkins, P. and Weber, K. (2017). Values and norms of proof for mathematicians and students. *Educational Studies in Mathematics*, 95(2):123–142.

de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras*, 24(1):17–24.

Depaepe, F., De Corte, E., and Verschaffel, L. (2016). Mathematical epistemological beliefs. In Greene, J., Sandoval, W., and Braten, I., editors, *Handbook of Epistemic Cognition*, (pp. 147–164.) New York, NY.: Routledge.

Fawcett, H. (1938). *The nature of proof. The national council of teachers of mathematics thirteenth yearbook*. Bureau of Publications of Teachers College, Columbia University, New York.

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, 12(1):1–12.

Hanna, G. and de Villiers, M. (2012). *Proof and Proving in Mathematics Education*. Springer Netherlands.

Harel, G. & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In Dubinsky, E., Schoenfeld, A., and Kaput, J. (Ed.), *Research in Collegiate Mathematics Education. III*, (pp. 234–283.) Washington, D.C.: American Mathematical Society.

Healy, L. and Hoyles, C. (2000). A study of proof concepts in algebra. *Journal for Research in Mathematics Education*, 31(4):396–428.

Hunter, J. and Jones, I. (2018). Free-Response Tasks in Primary Mathematics: A Window on Students' Thinking. In Hunter, J., Perger, P., and Darragh, L., editors, *Making waves, opening spaces: Proceedings of the 41st annual conference of the*

Mathematics Education Research Group of Australasia, (pp. 400–407), Auckland, New Zealand.

Jones, I. and Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, 39(10):1774–1787.

Jones, I., Bisson, M. J., Gilmore, C., and Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, 45(3):662–680.

Jones, I. and Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics*, 89(3):337–355.

Jones, I. and Karadeniz, I. (2016). An Alternative Approach To Assessing Achievement. In *Proceedings of the 2016 40th Conference of the International Group for the Psychology of Mathematics Education*, Szeged, Hungary.

Knuth, E. J. (2002). Teachers' conceptions of proof in the context of secondary school mathematics. *Journal of Mathematics Teacher Education*, 5(1):61–88.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. University of Chicago Press, Chicago, IL.

McLeod, D. B. (1992). Research on affect in mathematics education: A reconceptualization.

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics*, 27(3):249–266.

Muis, K. (2004). Personal Epistemology and Mathematics: A critical Review and Synthesis of Research. *Review of Educational Research*, 74(3):317–377.

Muis, K. (2008) Epistemic profiles and self-regulated learning: Examining relations in the context of mathematics problem solving. *Contemporary Educational Psychology*, 33(2): 177–208.

Nasser, F. and Birenbaum, M. (2005). Modeling mathematics achievement of Jewish and Arab eighth graders in Israel: The effects of learner-related variables. *Educational Research and Evaluation*, 11(3):277–302.

Perry, W. J. (1968). *Patterns of development in thought and values of students in a liberal arts college: A validation of a Scheme. Final report.* Cambridge, MA.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2):157–170.

Reid, D. and Knipping, C. (2010). *Proof in Mathematics*. Sense Publishers, Wolfville, Canada.

Schommer-Aikins, M. (2004). Explaining the Epistemological Belief System: Introducing the Embedded Systemic Model and Coordinated Research. *Educational Psychologist*, 39(1):19–29.

Stylianou, D. A., Blanton, M. L., and Rotou, O. (2015). Undergraduate Students' Understanding of Proof: Relationships Between Proof Conceptions, Beliefs, and Classroom Experiences with Learning Proof. *International Journal of Research in Undergraduate Mathematics Education*, 1(1):91–134.

Thurstone, L. (1928). Attitudes can be measured. *American Journal of Psychology*, 33(4):529–554.

Thurstone, L. (1954). The measurement of values. *Psychological Review*, 61(1):47–58.

Weber, K. (2001). Student difficulties in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics*, 48(1):101–119.

Weber, K. (2010). Mathematics Majors' Perceptions of Conviction, Validity, and Proof. *Mathematical Thinking and Learning*, 12(4):306–336.

Weber, K. and Czocher, J. (2019). On mathematicians' disagreements on what constitutes a proof. *Research in Mathematics Education*. DOI: 10.1080/14794802.2019.1585936

Weber, K., Inglis, M., and Mejia-Ramos, J. P. (2014). How mathematicians obtain conviction: Implications for mathematics instruction and research on epistemic cognition. *Educational Psychologist*, 49(1):36–58.

Wittgenstein, L. (1953). *Philosophical Investigations*. Macmillan Publishing Company, translated edition.

Zaslavsky, O. and Shir, K. (2005). Students ' Conceptions of a mathematical definition. *Journal for Research in Mathematics Education*, 36(4):317–346.