Educational Studies in Mathematics (2020) 105:181–197
https://doi.org/10.1007/s10649-020-09984-x

Comparative judgement, proof summaries and proof comprehension

Ben Davies[1] · Lara Alcock[2] · Ian Jones[2]

Published online: 1 October 2020

Abstract
Proof is central to mathematics and has drawn substantial attention from the mathematics education community. Yet, valid and reliable measures of proof comprehension remain rare. In this article, we present a study investigating proof comprehension via students' summaries of a given proof. These summaries were evaluated by expert judges making pairwise comparisons, which were used to generate a score for each summary. This approach, known as comparative judgement, has been demonstrated to generate reliable and valid scores when assessing other mathematical constructs. Our findings suggest that comparative judgement can produce valid and reliable assessments of the quality of student-produced proof summaries. We also explored which features of students' proof summaries were most valued by the expert judges, and found that high-scoring summaries referenced a proof's logical structure and the mechanism by which it reached a contradiction.

Keywords Proof comprehension · Comparative judgement · Reliability · Validity · Assessment

1 Introduction

Undergraduate mathematics is commonly assessed using tasks in which students construct proofs (Mejia-Ramos & Inglis, 2009; Weber, 2012). However, such assessments have been criticised as relying too heavily on recall and near-transfer tasks and thus as having limited validity as meaningful measures of comprehension (Cowen, 1991; Conradie & Frith, 2000). Consequently, there is growing interest in more directly assessing students' proof comprehension. For example, Mejia-Ramos, Lew, de la Torre, and Weber (2017) designed multiple-choice tests to assess comprehension of proofs drawn from real analysis and number theory. These tests were rigorously developed, and psychometric work demonstrated that they offer valid and reliable assessments.

✉ Ian Jones
I.Jones@lboro.ac.uk

[1] West Virginia University, Morgantown, WV, 26506, USA

[2] Loughborough University, Loughborough, UK

Such test development, however, requires time- and resource-intensive iterative work for every new proof. In the present article, we investigate an alternative approach to assessing proof comprehension, based on the Law of Comparative Judgement (Thurstone, 1927). This approach asks experts to make numerous pairwise comparisons of the "quality" of students' responses to a task, usually one that is simple to state and open-ended. These comparative judgements are then statistically modelled to produce a score for each response. Comparative judgement (CJ) has previously been used to assess mathematical competencies that are difficult, time- and resource-intensive to address using traditional methods, notably conceptual understanding (Bisson, Gilmore, Inglis, & Jones, 2016; Jones & Karadeniz, 2016; Hunter & Jones, 2018; Jones & Alcock, 2014) and problem solving (Jones et al., 2014; Jones & Inglis, 2015).

Following the methods developed by this previous research, we investigated the reliability and validity of a comparative judgement technique for assessing students' summaries of a given proof; we used summaries because summarising is a recognised aspect of proof comprehension (Mejia-Ramos, Lew, de la Torre, & Weber, 2017) but is difficult to assess via traditional methods. Specifically, we provided students with a proof that the open unit interval is uncountable, and asked for a 40-word summary. We then generated CJ scores for participants' responses. To investigate reliability, we used measures standard in CJ studies. To investigate validity, we first compared CJ scores with outcomes of Mejia-Ramos et al.'s (2017) multiple-choice test and with student achievement data; positive correlations indicate criterion validity. We also conducted a content analysis of the responses followed by regression analyses predicting the CJ scores based on these contents. Content validity was evaluated by considering the relationship between the content of high-scoring responses and earlier research on what mathematicians value in proofs. These validity analyses yielded acceptable results. So, in addition to our methodological contribution, we contribute to theoretical discussions on proof comprehension by offering empirically evidenced insight into what constitutes a "good" proof summary.

2 Comparative judgement theory and application

In this section, we review two areas of relevant literature. First, we describe the theory and methodological application of CJ, explaining its earlier use in mathematics education for assessing conceptual understanding and problem solving. We then review research on assessing proof comprehension.

2.1 Comparative judgement

Comparative judgement offers a general method for positioning a set of complex objects on a unidimensional scale. It has been used in psychological laboratory studies for almost a century (Thurstone, 1927) and is based on the observation that humans are relatively poor at making isolated absolute judgements of quantities or constructs such as weight or social attitudes, yet vastly more reliable at making pairwise comparisons of those quantities (Thurstone, 1927; 1954). CJ is applied by asking people to make pairwise decisions about which of two objects has the greatest "amount" of the construct of interest (Thurstone, 1927). The binary decision data from many such pairings are then fitted to a statistical model to produce a unique score for each object (Bradley & Terry, 1952).

CJ has been successfully applied in a range of educational assessment contexts (Tarricone & Newhouse, 2016). An early application in mathematics education was in assessing problem solving. Jones, Swan, and Pollitt (2014) compared CJ outcomes with rubric-based scoring of students' responses to innovative problem solving tasks (based on Bowland Mathematics materials; (Onion, 2011)). They reported a high correlation, suggesting good criterion validity (Newton & Shaw, 2014). Similarly, Jones and Inglis (2015) commissioned professional examination writers to construct a test of mathematical problem solving and found that CJ outcomes had good criterion validity as compared to outcomes from a scoring rubric generated iteratively from a sample of student responses. Similar methods have been employed to assess conceptual understanding in mathematics. CJ outcomes have been compared to independent measures for students' understanding of mathematical concepts including fractions (Jones & Wheadon, 2015), differential calculus (Bisson et al., 2016; Jones & Alcock, 2014; Bisson, Gilmore, Inglis, & Jones, 2019) and variables (Bisson et al., 2016; Jones, Bisson, Gilmore, & Inglis, 2019). Good criterion and divergent validity were reported in each case.

Of course, researchers are interested not just in validity but also in content: assessment research has a tradition of investigating how experts make decisions when assessing student work (e.g., Crisp, 2017). Accordingly, most published studies on CJ in mathematics education include analyses of which features of student responses expert judges value. Judges have been asked to complete surveys (e.g., Jones et al., 2014) and interviews (e.g., Jones & Alcock, 2014) about their decisions. And, as for the present study, researchers have conducted content analyses of students' responses. For example, Jones and Karadeniz (2016) applied an existing coding scheme to lower secondary students' written explanations of a range of mathematical concepts. They reported that expert judges preferred responses that contained numbers and graphics but had no preference for responses that contained letters (variables), references to real-world contexts, or connections between different mathematical topics.

Informed by all of this work, we report a study in which CJ was applied to proof summaries in order to assess undergraduates' comprehension of given proofs, and in which the preferences of expert judges were investigated using content analysis.

2.2 Assessing proof comprehension

Research into students' understanding of proof has traditionally focused on proof construction (e.g., Weber, 2001; Hoyles & Healy, 2007); until recently, less attention was paid to comprehension of written proofs (Mejia-Ramos & Inglis, 2009). Early interest in how students read proofs tended to focus on validation, defined as "the process in which an individual determines whether a proof is correct and actually proves the particular theorem it claims to prove" (Yang & Lin, 2008, p. 60). For instance, Selden and Selden (2003) and Alcock and Weber (2005) observed that some students did not pay attention to a purported proof's global structure, focusing instead on verifying calculations or specific implications. Weber (2010) and Ko and Knuth (2013) found that mathematics majors did not adequately identify localised logical flaws. These largely interview-based findings were corroborated using eye-movement analyses by Inglis and Alcock (2012), who found that undergraduates focused less than mathematicians on the words in purported proofs. Inglis and Alcock also reported that students were less inclined to switch their attention back and forth in a manner consistent with identifying implicit links between statements or attending to a global structure.

Such findings prompted Mejia-Ramos and Inglis (2009) to argue that students might not fully comprehend the proofs they are expected to read and that there was a need for research focused on reading comprehension. They noted that a barrier to progress was a lack of methods for assessing students' proof comprehension. There was work to build on: dissatisfaction with standard examinations had prompted Conradie and Frith (2000) to discuss proof comprehension tests, and Yang and Lin (2008) had developed tests to assess students' comprehension of geometry proofs. Mejia-Ramos, Fuller, Weber, Rhoads, and Samkoff (2012) combined this work with an extensive literature review and semi-structured interviews with mathematicians to construct a seven-part model of proof comprehension (see Table 1).

Mejia-Ramos et al.'s model has since been used by various authors to design proof comprehension tests (e.g., Hodds, Alcock, & Inglis, 2014; Roy, Alcock, & Inglis, 2017). But only the model's originators have undertaken rigorous test design, producing multiple-choice comprehension tests for three specific, commonly studied proofs in number theory and real analysis (Mejia-Ramos et al., 2017). They constructed these using a mixed-methods approach comprising interviews with students and mathematicians as well as large-scale test administration and refinement to evaluate and confirm internal reliability. The internal consistencies of the three resulting tests were found to be acceptable, all Cronbach's α's $>$ .7. To the best of our knowledge, these are the first psychometrically validated proof comprehension instruments.

These tests were, however, highly time- and resource-intensive to produce. Moreover, multiple-choice tests work better for some aspects of proof comprehension than for others. It is not obvious, in particular, how to use them to test "summarising via high-level ideas" (aspect 4). Mejia-Ramos et al. (2012) noted that "objectively grading the quality of a proof summary might be difficult" (p. 11), and proposed that tests could provide several possible summaries and "ask the student to choose which summary captures the main idea of the proof". However, according to Amman et al. (2020), designing summary-based items proves challenging for many proofs. Happily, the features of summarising that create difficulties for traditional assessment are precisely those that make it a perfect task for comparative

Table 1 A summary of the assessment model presented by Mejia-Ramos (2012, p. 15)

Aspect	Assessment evidence
Local	
1. Meaning of terms and statements	Understanding of key terms and statements in the proof
2. Logical status of statements and proof framework	Knowledge of the logical status of statements in the proof and the logical relationship between these statements and the statement being proven
3. Justification of claims	Comprehension of how each assertion in the proof follows from previous statements in the proof and other proven or assumed statements
Holistic	
4. Summarising via high-level ideas	Grasp of the main idea of the proof and its overarching approach
5. Identifying the modular structure	Comprehension of the proof in terms of its main components/modules and the logical relationship between them
6. Transferring the general ideas or methods to another context	Ability to adapt the ideas and procedures of the proof to solve other proving tasks
7. Illustrating with examples	Understanding of the proof in terms of its relationship to specific examples

judgement. “Provide a summary” is a simple, open-ended task. Varied and unexpected responses are unproblematic and indeed advantageous: CJ works best when responses are diverse. And we need not specify quality criteria or know in advance what experts will value (Bisson et al., 2016). We simply ask judges to make pairwise comparisons then use these to generate scores. Of course, CJ is less overtly transparent than rubric-based scoring. So, as in earlier reported studies, its reliability and validity must be carefully considered. But, if these turn out to be acceptable, we have a strong basis from which to ascertain and report on what is collectively valued and rewarded by judges.

3 Methods

3.1 Materials

To collect both proof summaries and data relevant to validity, we constructed a task booklet containing three sections. First was the theorem and proof about the uncountability of the unit interval, as used by Mejia-Ramos et al. (2017) and shown in Fig. 1. Second was the 12-item multiple-choice comprehension test for this proof developed by Mejia-Ramos et al.

Theorem: The open interval $(0,1)$ is uncountable.

Proof: The interval $(0,1)$ includes the subset $\left\{\frac{1}{2^k} : k \in \mathbb{N}\right\}$, which is infinite. Thus, $(0,1)$ is infinite. Suppose $(0,1)$ is denumerable. Then, there is a function $f : \mathbb{N} \to (0,1)$ that is one-to-one and onto $(0,1)$. Now, we write the images of f, for each $n \in \mathbb{N}$, in their decimal form:

$$\begin{aligned}
f(1) &= 0.a_{11}a_{12}a_{13}a_{14}a_{15}\ldots\\
f(2) &= 0.a_{21}a_{22}a_{23}a_{24}a_{25}\ldots\\
f(3) &= 0.a_{31}a_{32}a_{33}a_{34}a_{35}\ldots\\
f(4) &= 0.a_{41}a_{42}a_{43}a_{44}a_{45}\ldots\\
&\vdots\\
f(n) &= 0.a_{n1}a_{n2}a_{n3}a_{n4}a_{n5}\ldots\\
&\vdots
\end{aligned}$$

Since some elements of $(0,1)$ have two different decimal representations (one with an infinite string of 9’s and another one with an infinite string of 0’s), we do not use representations that contain an infinite string of 9’s. That is, for all $n \in \mathbb{N}$ we represent $f(n) = 0.a_{n1}a_{n2}a_{n3}a_{n4}a_{n5}\ldots$ in such a way that there is no k such that for all $i > k$, $a_{ni} = 9$.

Now let b be the number $b = 0.b_1b_2b_3b_4b_5\ldots$, where $b_i = 5$ if $a_{ii} \neq 5$ and $b_i = 3$ if $a_{ii} = 5$. Because of the way b has been constructed, we know that $b \in (0,1)$ and that b has a unique decimal representation. However, for each natural number n, b differs from $f(n)$ in the nth decimal place. Thus $b \neq f(n)$ for any $n \in \mathbb{N}$, which means b does not belong to the range of f. Thus, f is not onto $(0,1)$. This contradicts our assumptions. Therefore, $(0,1)$ is not denumerable. □

Fig. 1 Uncountability theorem and proof

Question 1 What inferences in the proof are drawn from the assertion that f is one-to-one? Please select **the best** option.

○ This assertion is not used later in the proof.
○ This assertion is used to conclude that b differs from $f(n)$ in the nth decimal place.
○ This assertion is used to conclude that the decimal representations of $f(n)$ are unique.
○ This assertion is used to conclude that f is not onto $(0,1)$.

Question 2 How do we know that $b \neq f(n)$ for any natural number n? Please select **the best** option.

○ We know $b \neq f(n)$ because every decimal place of $f(n)$ is different from the corresponding decimal place of b. That is, if $f(n) = 0.a_{n1}a_{n2}a_{n3}a_{n4}a_{n5}...$, then for all $i \in \mathbb{N}, b_i \neq a_{ni}$.
○ We know $b \neq f(n)$ because the decimal representation of b only contains 3's and 5's, and there is no $n \in \mathbb{N}$ such that the decimal representation of $\frac{1}{2^n}$ only contains 3's and 5's.
○ We know $b \neq f(n)$ for any natural number n because f is one-to-one, which means that different elements of $\mathbb{N}$ have different images under f.
○ We know $b \neq f(n)$ because b has a unique decimal representation and that representation is different than the decimal representation of $f(n)$ for any natural number n.

Fig. 2 Example items from the comprehension test of Mejia-Ramos et al. (2017), evaluating students' reading comprehension of the proof in Fig. 1

Summarise the proof, given on the previous page, **in 40 words or less**.

Note: You are not being asked to reproduce the proof. The best responses will be those that succinctly communicate the most important aspects/ideas in the proof.

Write your summary in the box below:

Fig. 3 Summary task

$(0,1)$ is infinite. Suppose its denumerable. there exists a one-to-one function (f) onto (0,1). For any image of f, $f(n) = 0.a_{n_1}a_{n_2}, a_{n_3}\ldots$, we can construct a number $b \in (0,1)$ where $b \neq f(n)$, therefore f isn't onto (0,1), which contradicts our assumption.

Score: 2.78 Codes assigned: 1, 4, 5, 6, 7, 8, 10, 13, 14, 15.

Assume that $f : \mathbb{N} \to (0,1)$. There is no representation for strings of 9s and 0s. There is a number $b = 0.b_1b_2b_3\ldots b_n$, $b_i = 5$ iff $a_{ii} \neq 5$, $b_i = 3$ iff $a_{ii} = 5$. The n^{th} term of b differs from the n^{th} term of $f(n)$. $\therefore b \neq f(n)$, b is not in the range of the function. f is not onto (0,1) #.

Score: 2.59 Codes assigned: 1, 4, 5, 9, 10, 11, 12, 14

Interval (0,1) is infinite as contains infinite subset $\{\frac{1}{2k} : k \in \mathbb{N}\}$. Suppose (0,1) is denumerable. Some elements of (0,1) have 2 different decimal representations so we don't use one's with infinite 9's. no k s.t. $\forall i < k, a_{n_i} = 9$. Let $b = 0.b_1b_2b_3b_4b_5\ldots$ where $b_i = 5$ if $a_{ii} \neq 5$ and $b_i = 3$ if $a_{ii} = 5$ so $b \in (0,1)$ and has unique decimal representation. $\therefore \forall n \in \mathbb{N}, b \neq f(n)$ so f is not onto (0,1) #. (0,1) is not denumerable.

Score: 2.54 Codes assigned: 1, 2, 3, 4, 9, 10, 11, 12, 13, 14, 15.

Fig. 4 The three responses awarded the highest CJ scores. Codes assigned are explained in Section 4.3

The interval (0,1) is infinite. If (0,1) is denumerable then $f : \mathbb{N} \to (0,1)$ is one-to-one and onto (0,1). Some elements have two different representations so we ignore all that have an infinitely occurring 9. B does not contain 2 numbers which = each other. When it gets to the nth decimal place b has a different number to n.

Score: 0.38 Codes assigned: 1, 2, 3, 4, 8, 10, 11, 12

If we represent $f(n)$ as $0.a_{n1}a_{n2}a_{n3}\ldots$ and choose a $b \in (0,1)$ s.t. $b_i = 5$ in all places other than where $a_{ii} = 5$, we then form a number that is not represented by $f(n)$ but is in the interval so it cannot be countable.

Score: 0.37 Codes assigned: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15

(0,1) contains an infinite subset which means (0,1) is infinite. Through contradiction, there isn't a one-to-one function $f(n)$, mapping onto (0,1), with a unique decimal representation, making (0,1) non-denumerable.

Score: 0.36 Codes assigned: 1, 5, 8, 9, 10, 12, 15

Fig. 5 The three responses awarded the middle CJ scores. Codes assigned are explained in Section 4.3

There are infinitely many value between 0 and 1. This uses proof by contradiction.

Score: -2.58 Codes assigned: 1, 4

The proof suggests and shows that seeing as there are a number of subsets in the interval (0,1) that have infinite values, then the interval (0,1) is uncountable.

Score: -3.67 Codes assigned: 1, 2, 3

Proves that for the interval (0,1) there is an infinite set of numbers between (0,1) that is countable and can be represented as a decimal expansion.

Score: -3.68 Codes assigned: 1

Fig. 6 The three responses awarded the lowest CJ scores (of those meaningfully addressing the task). Codes assigned are explained in Section 4.3

(2017). Figure 2 shows two example items from this test[1]. Third was the summary task prompt shown in Fig. 3, with space for students to write their responses. We included a 40-word limit to increase the distinctiveness of responses, expecting that without an upper bound, some would come close to replicating the original proof. Readers wishing to orient themselves to the nature of the student summaries might wish to inspect Figs. 4, 5 and 6 before reading on.

3.2 Participants

Participants were enrolled in a compulsory introductory course on real analysis for first- and second-year undergraduates, covering fundamental concepts related to sequences, series and $\epsilon - N$ definitions. Completing the task booklet was compulsory but did not contribute to formal assessment. Only responses from students who consented to their data being used in the research were included in the analysis. Of the 161 students in attendance when the task booklet was administered, 18 (11.2%) declined to consent, leaving 143 responses for analysis.

[1] We do not present the full test at the request of its developers, who wish to protect its integrity for future assessment and research. The full version can be requested at pcrg.gse.rutgers.edu/.

3.3 Procedure

Data collection took place in a week-eight lecture, with content directly related to the task booklet presented in the preceding lecture. Participants were allowed 30 min to complete the task booklet. They were advised to spend 20 min on the multiple-choice test and ten on the summary task. General feedback on overall performance was given to the course lecturer, based on the anonymised data. We also collected final course scores for each student, which we treated as a general measure of performance in proof-based mathematics. Final course scores ranged from 33 to 97%, with a mean of 56% ($SD = 13.5\%$); these were based on a weighted aggregate of students' coursework and final examination scores. Consent for analysis of final course scores was given by 134 of the 143 participants. Data from the nine students not consenting remained in all analyses not involving final course scores. All anonymised data are available at https://doi.org/10.17028/rd.lboro.8940149.

3.4 Comparative judgement

Participants' anonymised proof summaries were scanned and uploaded to nomoremarking.com, an online CJ engine that is free to use for researchers and educators. The responses were then comparatively judged by 11 experts. The experts were recruited using contacts from previous studies: seven were mathematics PhD students at the same English university as the undergraduates; three were mathematics PhD students from a second English university; one was a recent mathematics PhD graduate from a third English university. No background data were collected on the judges (e.g., research interests, teaching experience) because previous work on CJ has repeatedly confirmed that so long as judges are mathematically competent, outcomes are reliable (Bisson et al., 2016; Jones & Alcock, 2014). However, we note that people specialise early in the UK system and a PhD is a research degree (there is no course-based phase), so all PhD students are active in mathematics research.

Judges worked remotely in their own time. They were instructed to read the uncountability theorem and proof (Fig. 1) before judging and to keep it to hand throughout the process. Pairs of responses were displayed side by side on a screen, with judges clicking "left" or "right" to indicate which they judged better. The response pairs were different for each judge because they were selected randomly using a non-adaptive algorithm (Bramley & Vitello, 2019), with the constraints that all responses received approximately the same number of judgements and each response was seen by as many judges as possible. Each judge performed 143 pairwise comparisons, resulting in a total of 1573 judgements. Based on an informal pilot study and previous experience, judges were paid according to an expected average of 20 seconds per judgement. The actual median time per judgement was 21.6 s.

4 Data analysis and results

4.1 Reliability and example responses

To begin the analysis, the binary decision data generated by the pairwise comparisons were downloaded from the nomoremarking.com website and fitted to the Bradley-Terry (1952) model to generate a parameter estimate of collectively perceived quality for each summary. These parameter estimates—henceforth referred to as CJ scores—had a mean of 0.06 and a standard deviation of 1.69.

To investigate whether the judges were collectively consistent in their judgements, and therefore whether CJ scores were meaningful on their own terms, we calculated two standard reliability measures. The first, Scale Separation Reliability (SSR), is analogous to Cronbach's α in interpretation (Pollitt, 2012) and is a measure of internal consistency. This yielded $SSR = .86$, which is considered acceptable. The second is a measure of inter-rater reliability, estimated using a split-half comparison method introduced to the CJ literature by Bisson, Gilmore, Inglis, and Jones (2016). Post-judging, the judges were split randomly into two groups and CJ scores recalculated for each group. Inter-rater reliability was estimated by computing the Pearson Product-Moment correlation coefficient between the two groups. We repeated this procedure 100 times and took the median correlation coefficient as our estimate of inter-rater reliability. This was again acceptable, $r_{\text{median}} = .73$.

Taken together, these statistics demonstrate that the judging process produced internally consistent outcomes: the judges were collectively consistent in their judgements of which summaries were better and which were poorer. This means that the CJ scores constitute a reliable measure of proof summary quality. To demonstrate the content of the summaries, Figs. 4, 5 and 6 each present three: those awarded the highest, middle and lowest CJ scores respectively. In Fig 6, we present the lowest-scoring responses that meaningfully addressed the task—seven were awarded lower scores but were fragmented or blank. As is discussed in a later section, a conjecture that students' unfamiliarity with the summary task may have led to a number of such responses.

4.2 Criterion validity

To investigate whether the judges were consistent not only with one another but also with empirically and theoretically reasonable criteria for proof summary quality, we next considered validity. We began with *criterion validity* (Newton & Shaw, 2014), investigating whether CJ scores were correlated with our two benchmark measures: the multiple-choice test scores and final course scores. We would expect students who are more successful on these other proof-related performance measures to produce better proof summaries, so we expect positive correlations. We calculated Pearson Product-Moment correlations after first processing the multiple-choice test scores.

Performance on the 12-item multiple-choice test was significantly above chance but nevertheless low, $M = 4.19$, $SD = 2.16$, $t(133) = 6.35$, $p < .001$. This indicates that the test was difficult for this sample of students. Unfortunately, internal consistency was also low, Cronbach's $\alpha = .53$, in contrast to values reported by Mejia-Ramos et al. (2017) when the test was administered in the USA (all α's $> .7$). Consequently, we investigated its factor structure using principal component analysis (KMO $= 0.62$ with Bartlett's test of sphericity at $p < .001$). Two components were extracted, accounting for 29% of the variance, with seven items loaded on to Component 1, two items loaded on to Component 2, and two items loaded on to neither component at $> .4$. The internal consistency of the subset of seven items that loaded on to Component 1, $\alpha_{\text{subset}} = .61$, was higher than that for all twelve items. Consequently, we based our analysis on the subset of seven items that loaded onto Component 1, although repeating it with all twelve items made no substantive difference to the results.

The correlation between the CJ scores and the seven-item multiple-choice test scores was significant but modest, $r = .28$, $p < .001$ (a scatter plot is shown in Fig. 7). This correlation is lower than those reported in other studies where scores from standardised instruments were correlated with the parameter estimates of open-ended tests assessed using CJ. Bisson et al. (2016), for instance, measured conceptual understanding of three topics in secondary and tertiary mathematics—p-values in statistics, letters in algebra, and

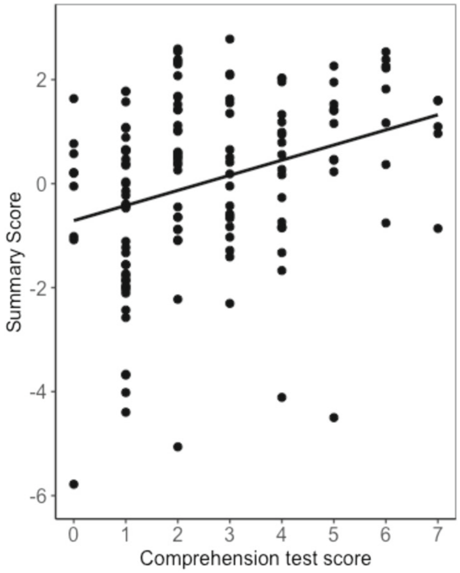

Fig. 7 Scatter plot of CJ-based parameter estimates against scores from the 7-item multiple-choice test

derivatives in calculus—and reported correlation coefficients between .35 and .56. To further understand our result, we correlated the seven-item multiple-choice test scores with the final course scores. The correlation was moderate and significant, $r = .55$, $p < .001$. This suggests that, despite the modest internal consistency of the multiple-choice test, it was meaningful in terms of assessing students' performance in proof-based mathematics.

The correlation coefficient between the CJ scores and the final course scores was also significant but modest, $r = .23$, $p < .001$. This means that, while low, the two correlations involving our CJ scores were significant and in the expected directions. These analyses thus offer modest support for the criterion validity of the CJ-based assessment, while perhaps indicating that overlapping but different skills are required for summarising a single proof, immediate overall comprehension of a single proof, and performance in course examinations. We offer further interpretation of these statistical results in the Discussion.

4.3 Content validity and what judges valued

To investigate whether the judges were consistent with theoretically reasonable criteria for proof summary quality, we next considered *content validity*. We would expect good summaries to have contents consistent with theoretical analyses and empirical outcomes from earlier research on proof comprehension. But we do not assume that theoretical analyses necessarily capture what expert judges actually value in real-world judgements, and we do not believe that earlier empirical outcomes predict how judges might weigh different factors when making those judgements. So we first coded the contents of the summaries according to the principles of thematic analysis (Braun & Clarke, 2006); this involved no value judgements about which contents were "good". We then conducted multiple regression analyses to identify which content in proof summaries predicted high CJ scores. This allowed us to investigate empirically what the expert judges valued (Hunter & Jones, 2018; Jones & Karadeniz, 2016).

4.3.1 Coding of proof summary contents

We coded the contents of students' proof summaries using a scheme developed in three iterations. We first examined the original proof (Fig. 1) alongside ten student summaries that were not part of the main dataset. Most statements in the summaries could be directly mapped to a discrete part of the proof, so the first version of the coding scheme simply partitioned the proof into 11 parts, each with a code. In the second iteration, two researchers independently coded ten responses from the main dataset. Comparing these analyses led to a revised scheme, clarifying existing codes or dividing one code into several. Having established a final 15-code scheme, we turned to the remaining 123 responses. Each researcher coded 75 responses, with an intersection of 27 for estimating inter-coder reliability: we calculated a pooled Cohen's κ of .88, suggesting good agreement between coders. Inconsistencies were found to be either coder errors or unique cases not anticipated by our scheme. In the latter case, a decision was reached to maintain the clarity of each code, opting to assign no code to a statement for which we did not have an obvious code.

The final codes are shown in Table 2. Column 1 shows the code number, where these numbers are also used to show how the codes were assigned to specific responses in Figs. 4, 5 and 6. Column 2 of Table 2 explains each code, and column 3 shows the percentage of student proof summaries assigned each code. Columns 4 and 5 show the Spearman correlation (ρ) between each code and the CJ scores, and the significance (p) of these correlations. The remaining columns refer to the regression analyses; both correlations and regressions are discussed next.

4.3.2 Content validity and contents judges valued

If the CJ scores have good content validity, we would expect codes that correlate with high CJ scores to be consistent with earlier theoretical and empirical studies on proof comprehension. And this seems to be the case. Eight of the 15 content codes correlated significantly with CJ scores, and all eight (codes 4, 5, 8, 9, 10, 11, 12 and 14) can be viewed as relevant to the construction of objects used in the proof, and to their properties in relation to the proof's contradiction argument. For example, the highest correlating code ($\rho = .54$, $p < .001$) was 10 ("Addressed the constructed b from the given proof in any way"); the second highest ($\rho = .51$, $p < .001$) was 12 ("Explicitly stated that b is not in the range of f (or, that $b \neq f(n)$ for any n)"). Both seem highly pertinent to summarising via high-level ideas because they show a "grasp of the main idea of the proof and its overarching approach" (Mejia-Ramos, Fuller, Weber, Rhoads, & Samkoff, 2012, p. 15).

Conversely, the seven codes that did not correlate significantly with parameter estimates (1, 2, 3, 6, 7, 13 and 15) arguably capture information that is more obvious or readily inferable by a mathematically experienced reader. For example, code 1 ($\rho = .15$, $p = .101$) captured information about the interval in question ("Explicitly stated that the interval (0, 1) is infinite"); code 2 ($\rho = .06$, $p = .536$) elaborated this point ("Addressed a subset of (0, 1) (not necessarily explicitly naming $1/2^k : k \in \mathbb{N}$)"). These are accurate descriptions of the original proof's content, but they seem less essential to communicating the proof's main idea or overarching approach.

The correlations thus provide initial support for the content validity of the CJ-based assessment: codes found to correlate significantly with CJ scores were those that are more theoretically pertinent to summarising the proof via high-level ideas. We investigated further by finding the predictive weights associated with these codes. First, the eight codes that yielded significant correlations were used in a forced-entry regression predicting the

Table 2 Content analysis and subsequent regression analyses

Code	Description	Frequency	ρ	p	8-code model				4-code model			
					B	SE	β	p	B	SE	β	p
1	Explicitly stated that the interval $(0, 1)$ is infinite	46%	.15	.101								
2	Addressed a subset of $(0, 1)$ (not necessarily explicitly naming $1/2^k : k \in \mathbb{N}$)	48%	.06	.536								
3	Explicitly related the infinitude of $(0, 1)$ to an infinite subset	33%	.13	.159								
4	Appealed to proof by contradiction (need not have explicitly associated wording, just looking for evidence of the logical structure)	51%	.33	$< .001$*	0.70	0.26	0.19	.008*	0.68	0.28	0.19	.012*
5	Described the function f as a mapping $\mathbb{N} \to (0, 1)$	46%	.38	$< .001$*	0.85	0.26	0.23	.012*	1.01	0.26	0.28	$< .001$*
6	Described the function f as injective	29%	.21	.024								
7	Described the function f as surjective	19%	.19	.031								
8	Described the images of f using decimal representation (any reference to decimal representations of $f(n)$'s is sufficient, reference to decimal representations of other values is not)	29%	.37	$< .001$*	0.61	0.31	0.15	.052				

Table 2 (continued)

Code	Description	Frequency	ρ	p	8-code model				4-code model			
					B	SE	β	p	B	SE	β	p
9	Appealed to 0's or 9's in reference to the decimal representations (accept references to elements of the range of f or $(0, 1)$, also accept ambiguity)	37%	.39	< .001*	0.31	0.29	0.08	.293				
10	Addressed the constructed b from the given proof in any way	70%	.54	< .001*	0.93	0.39	0.23	.024*	1.62	0.29	0.40	< .001*
11	Constructed b explicitly (sufficient to describe b as a number differing from each $f(n)$ in the n^{th} entry)	41%	.38	< .001*	0.51	0.30	0.14	.086				
12	Explicitly stated that b is not in the range of f (or, that $b \neq f(n)$ for any n)	50%	.51	< .001*	0.30	0.34	0.08	.387				
13	Explicitly stated that the constructed b is in $(0, 1)$	25%	.23	.009								
14	Explicitly stated that f is not surjective as a result of the surmised argument	17%	.32	< .001*	0.71	0.34	0.15	.043*	0.77	0.35	0.16	.029*
15	Included the term 'denumerable' anywhere	42%	.22	.017								

Columns 1-5 show code numbers, code contents, frequency of assignation to responses, Spearman correlations (ρ) between content code and CJ scores, and correlation significance (p). The remaining columns show 8-code and 4-code regression analyses, predicting CJ scores using codes identified as significantly correlated with CJ scores in the initial analysis

CJ scores; the resulting model, $F(8, 114) = 14.39, p < .001$, explained 50% of the variance[2]. Next, we generated a regression including only the four codes identified as significant predictors in the first model. The resulting model explained 43% of the variance, $F(4, 118) = 22.56, p < .001$ (see Table 2 for both analyses). This second analysis identified codes 5 and 10 as the most important predictors of CJ scores. These two codes refer to the two major mathematical objects used in the proof: the function f, which is to be the subject of the contradiction, and the number b used to demonstrate the contradiction. The other two codes in the final model, 4 and 14, capture the logical structure of the proof, appealing to proof by contradiction and explicating that contradiction.

Again, these results support the content validity of the CJ-based approach to assessment. And because the CJ scores show acceptable reliability, criterion validity and content validity, they also provide an empirical basis from which to make claims about quality in proof summaries. Principally, they reveal that judges rewarded summaries capturing the logical structure of the proof. In this, the judges appear to have acted in alignment not only with the summarising aspect of Mejia-Ramos et al.'s (2017) proof comprehension assessment model—as shown in Table 1—but also with the other aspects "logical status of statements and proof framework", "justification of claims" and "identifying the modular structure". The judges also rewarded a focus on constructed objects which, with hindsight, is perhaps not surprising for this proof, which relies on an inventive construction to reach its conclusion.

5 Discussion

This paper reports a study using comparative judgement (CJ) to score undergraduates' proof summaries, thus applying a novel method to a difficult-to-assess aspect of proof comprehension. Proof summaries provided by 143 students were comparatively judged by experts. The resulting CJ scores were reliable in the sense that the judges were collectively consistent in their judgements of which summaries were better. The CJ scores showed acceptable criterion validity in that there were significant, positive correlations with scores from a standardised proof comprehension test and with overall course scores. These correlations were, however, lower than those reported in studies using similar research designs for other mathematical constructs. This is a point we return to below. The CJ scores showed good content validity in that of 15 codes capturing the contents of the proof summaries, the eight found to be significantly correlated with CJ scores were pertinent to the main idea of the proof and its overarching approach; those not significantly correlated captured content that was more peripheral or readily inferable. Moreover, the four codes found via regression analyses to best predict the CJ scores were those that captured the logical structure of the proof—in this case, proof by contradiction—and the construction of objects used in the argument. This is in line with what we would expect given previous theoretical and empirical work on mathematicians' views of proof.

We believe, therefore, that this CJ approach performed satisfactorily in generating a quantitative measure of proof summary quality. However, the low correlations between CJ scores and multiple-choice and final course scores raise issues about the measurement of

[2] We also ran forced-entry regressions with all 15 codes, and with 12 codes significantly correlated with parameter estimates before Bonferroni correction ($\alpha = .05$). All led to very similar conclusions with respectively explained variance, 53% and 52%.

complex constructs in educational contexts. First, the multiple-choice test itself did not achieve acceptable internal consistency. A similar problem hampered Bisson, Gilmore, Inglis, and Jones (2019), who applied CJ-based assessment in differential calculus: there, too, a standardised instrument performed sub-optimally outside the context in which it was developed. If the substantial work done in generating standardised instruments needs to be repeated in each context, that further motivates a need for alternative low-resource assessment methods such as CJ. However, such instrument failure hinders the validation of those alternative methods.

Second, the low correlations could be part of a meaningful pattern in which the assessments simply assess different things. Unlike CJ scores, multiple-choice test scores did correlate moderately with final course scores. This could reflect the fact that a multiple-choice test is more similar to typical university mathematics assessments (Iannone & Simpson, 2011) than a summarising task which is likely to be unfamiliar to most students. Such a “test format effect” (Becker & Johnston, 1999) might have had more detrimental impact here than in earlier CJ-based mathematics assessments due to the shortness of the summarising task: participants were restricted to 40 words, where the typical response size in analogous studies is one page long. Because longer assessments tend to be more reliable and more valid (Newton & Shaw, 2014), a longer task—perhaps summarising a series of proofs—might increase measured criterion validity. On the other hand, it could be that summarising is an important skill but one that overlaps only partially with more standard assessments. A summarising task is a production task, whereas a multiple-choice test is a recognition task; perhaps scores on a summarising task would correlate more highly with scores from other production tasks such as proof construction. A mathematics exam is a production task, but one in which performance is influenced by many extra-mathematical variables such as commitment to regular study. Perhaps, then, more comparisons are needed, and it would benefit the field to conduct studies able to disentangle relationships between various proof-related skills.

For now, we conclude with emphasis on the positive. Students’ proof comprehension has attracted attention in mathematics education but has traditionally been under-researched due to a paucity of methods and tools for assessing comprehension. Recent progress has been enabled by the development of standardised instruments (Mejia-Ramos et al., 2017), but that development work is time- and resource-intensive and must be replicated for every proof. The CJ-based approach explored here offers a potentially efficient and inexpensive method for assessing comprehension of any proof. Further research is required in order to establish design principles for optimising validity in CJ-based approaches, to investigate their application to other aspects of proof comprehension, and to understand the relationships between proof comprehension and other proof-related skills. But, as one of a suite of methods, we believe that the CJ approach has much to offer, both to researchers and to lecturers wishing to understand students’ learning of proof.

Acknowledgements We would like to thank Pablo Mejia-Ramos for his contribution to the content analysis presented in this work.

References

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

Amman, K., Athey, B., Guajardo, L., Alsen, J., Orr-Woods, C., & Mejia-Ramos, P. (2020). Designing proof comprehension tests in real analysis. In *Poster presented at the 23rd annual conference of the special interest group of the mathematical association of america on research in undergraduate mathematics education, Boston, MA*.

Becker, W., & Johnston, C. (1999). The relationship between multiple choice and essay response questions in assessing economics understanding. *Economic Record*, *75*(4), 348–357.

Bisson, M. J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate MathematicsEducation*, *2*(2), 141–164.

Bisson, M. J., Gilmore, C., Inglis, M., & Jones, I. (2019). Teaching using contextualised and decontextualised representations: Examining the case of differential calculus through a comparative judgement technique. *Research in Mathematics Education*. https://doi.org/10.1080/14794802.2019.1692060.

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles Policy and Practice*, *26*(1), 42–58.

Bradley, R., & Terry, M. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3), 324–345.

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

Crisp, V. (2017). The judgement processes involved in the moderation of teacher-assessed projects. *Oxford Review of Education*, *43*(1), 19–37.

Conradie, J., & Frith, J. (2000). Comprehension tests in mathematics. *Educational Studies in Mathematics*, *42*(3), 225–235.

Cowen, C. (1991). Teaching and testing mathematics reading. *The American Mathematical Monthly*, *98*(1), 50–53.

Davis, P., & Hersh, R. (1981). *The mathematical experience*. New York, NY: Viking Pengu in Inc.

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

Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In Lester, F. (Ed.) *Second Handbook of research on mathematics teaching and learning* (pp. 805–842). Greenwich, UK: Information Age Pub Inc.

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, *45*(1), 62–101.

Hoyles, C., & Healy, L. (2007). Curriculum change and geometrical reasoning. In P. Boero (Ed.) *Theorems in school* (pp. 81–115). Rotterdam, the Netherlands: Sense Publishers.

Hunter, J., & Jones, I. (2018). Free-response tasks in primary mathematics: A window on students' thinking. In J. Hunter, P. Perger, & L. Darragh (Eds.) *Making waves, opening spaces: Proceedings of the 41st annual conference of the Mathematics Education Research Group of Australasia* (pp. 400–407). MERGA: Auckland.

Iannone, P., & Simpson, A. (2011). The summative assessment diet: How we assess in mathematics degrees. *Teaching Mathematics and its Applications*, *30*(4), 186–196.

Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, *43*(4), 358–390.

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

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

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

Jones, I., & Karadeniz, I. (2016). An alternative approach to assessing achievement. In Csikos, C., Rausch, A., & Szitanyi, J. (Eds.) *Proceedings of the 2016 40th Conference of the International Group for the Psychology of Mathematics Education, Szeged, Hungary*. https://dspace.lboro.ac.uk/2134/21090.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, *13*(1), 151–177.

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, *47*(1), 93–101.

Ko, Y., & Knuth, E. (2009). Undergraduate mathematics majors' writing performance producing proofs and counterexamples about continuous functions. *Journal of Mathematical Behavior*, *28*(1), 68–77.

Ko, Y., & Knuth, E. (2013). Validating proofs and counterexamples across content domains: Practices of importance for mathematics majors. *Journal of Mathematical Behavior*, *32*(1), 20–35.

Küchemann, D., & Hoyles, C. (2006). Influences on students' mathematical reasoning and patterns in its development: Insights from a longitudinal study with particular reference to geometry. *International Journal of Science and Mathematics Education*, *4*(4), 581–608.

Mejia-Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics*, *79*(1), 3–18.

Mejia-Ramos, J. P., & Inglis, M. (2009). Argumentative and proving activities in mathematics education research. In F. Lin, F. Hsieh, G. Hanna, & M. de Villiers (Eds.) *Proceedings of the ICMI study 19 conference: Proof and proving in mathematics education* (Vol. 2, pp. 88–93). Taiwan: National Science Council. http://140.122.140.1/~icmi19/files/Volume_2.pdf.

Mejia-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, *19*(2), 130–146.

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

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. Cambridge, UK: Sage Publications.

Onion, A. (2011). Bowland maths–The CPD modules. *Mathematics Teaching*, *221*, 41–43.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, *19*(3), 281–300.

Rowland, T. (2001). Generic proofs in number theory. In Campbell, S., & Zazkis, R. (Eds.) *Learning and teaching number theory: Research in cognition and instruction* (pp. 157–184). Wesport, CT: Ablex.

Roy, S., Alcock, L., & Inglis, M. (2017). Multimedia resources designed to support learning from written proofs: An eye-movement study. *Educational Studies in Mathematics*, *96*(2), 249–266.

Selden, A., & Selden, J. (2003). Validations of proofs considered as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education*, *34*(1), 4–36.

Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics*, *29*(2), 123–151.

Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education*, *13*(2), 123–151.

Thurstone, L. (1927). A law of comparative judgement. *Psychology Review*, *34*(4), 273–286.

Thurstone, L. (1954). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–554.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy and Practice*, *26*(1), 59–74.

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. (2012). Mathematicians' perspectives on their pedagogical practice with respect to proof. *International Journal of Mathematical Education in Science and Technology*, *43*(4), 463–482.

Weber, K. (2015). Effective proof reading strategies for comprehending mathematical proofs. *International Journal of Research in Undergraduate Mathematics Education*, *1*(3), 289–314.

Weber, K., & Alcock, L. (2004). Semantic and syntactic proof productions. *Educational Studies in Mathematics*, *56*(2-3), 209–234.

Yang, K. L., & Lin, F. L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, *67*(1), 59–76.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.