Validity objections to comparative judgement

Ian Jones

Loughborough University, Department of Mathematics Education, UK; I.Jones@lboro.ac.uk

Comparative judgement approaches to assigning grades to students' work have received interest from mathematics education researchers over recent decades. These approaches involve assessors deciding which of two presented pieces of work is 'better', and the decisions are then converted into scores. Several objections have been raised to using comparative judgement for summative assessment and in this theoretical paper I respond to objections that such approaches are 'not valid'. These include objections that there is a lack of evidence supporting validity, and that researchers assert comparative judgement is 'intrinsically valid' in ways that are incomplete and inconsistent. I argue that most validity objections are addressed by published evidence, and that the validity of applying comparative judgement to mathematics education assessments is a special case.

Keywords: Comparative judgement; alternative assessment; summative assessment; validity.

Comparative judgement.

Comparative judgement approaches to generating scores or grades from students' exam scripts or test responses have been gaining attention in education over recent decades (Bartholomew & M. Jones, 2022; I. Jones & Davies, 2023). This has been notably so in mathematics education, where the motivation to consider comparative judgement instead of rubric-based scoring is that the latter tends to not to produce reliable outcomes for open-ended, relatively unstructured tests that require longer answers from students (Newton, 1996). Here, 'reliable' refers to the extent to which a student would have obtained a different score or grade had a different assessor marked their work. Traditionally, mathematics assessment designers have shied away from open-ended test questions, and instead embraced short, objective items in order to maximise reliability (Swan & Burkhardt, 2012). This comes at the cost of validity if we value mathematics students performing longer-form mathematical activities such as problem solving, sustained reasoning and explaining their understanding (NCTM Research Committee, 2013).

Comparative judgement has been demonstrated to produce highly reliable outcomes for open-ended, relatively unstructured mathematics assessments across a range of mathematical topics and student age ranges (I. Jones & Davies, 2023). This is achieved by applying Thurstone's (1927) well-established Law of Comparative judgement which states that human beings are poor at making absolute judgements (e.g. determining the temperature in a room in Celsius) and good at making relative judgements (e.g. determining if it is colder in the room or outside). Thurstone's law can be applied to educational assessment by harnessing the collective judgement of a group of subject experts (Pollitt, 2012). In practice, an online platform presents pairs of students' written mathematical work ('responses') to expert assessors via an internet browser (Wheadon et al., 2020). Interested readers can try judging pairs of responses to the test prompt "What is an equation?" here: http://tinyurl.com/190523equation.

Once many judgements have been collected from several assessors the binary decision data is converted into a unique score for each response. Due to space constraints I do not detail the mathematics or other statistical details here, but see Pollitt (2012) and I. Jones and Davies (2023) for technical details.

Comparative judgement has been applied to and tested in a range of summative assessment contexts. In mathematics education these contexts have included undergraduate modules covering topics such as calculus (e.g. I. Jones & Alcock, 2014), statistics (e.g. Bisson et al., 2016) and proof comprehension (e.g. Davies, Alcock & I. Jones. 2020). At secondary school level, contexts have included fractions (I. Jones et al., 2013), statistics (e.g. Marshall et al., 2020) and problem solving (e.g. I. Jones & Inglis, 2015). Comparative judgement approaches have also been used to assess conceptual understanding across a range of curriculum topics at both secondary level (e.g. I. Jones & Karadeniz, 2016) and primary level (e.g. Hunter & I. Jones, 2018). Across these studies and others the outcomes of using comparative judgement have been found to be robust in terms of reliability and, as discussed below, in terms of validity (Bartholomew & M. Jones, 2022; I. Jones & Davies, 2023).

The validity objections.

Most of the objections that have I encountered have not been published in the scholarly literature but have arisen in other sources such as social media, blog posts, conversations at conferences or seminars, and reviewers' comments on submitted articles. I focus here on what I perceive to have been the most common, and I strive to present them coherently and fairly.

A common objection is that there is a lack of evidence in support of the validity of comparative judgement for summative assessment. For example, van Daal et al. (2022) wrote, in the context of education generally, "only a limited number of studies dig into the validity of comparative judgement" (p.2). Others go further, suggesting there is something inherently challenging about validity when it comes to comparative judgment (e.g. Bokhove, 2019). A particular concern is the opacity of assessment decisions lacking detailed criteria, which it has been alleged means that comparative judgement "drains from the assessment any considerations of *what* [*sic.*] is being assessed & the whole question of validity worth having" (Davis, 2017).

Perhaps the most substantive objections to the claime validity of comparative judgement approaches, including for mathematics education, came from a theoretical paper by Kelly, Richardson and Isaacs (2022). Before I present and respond to the key objections presented in their paper, it is worth clarifying that the vast majority of literature on using comparative judgement for educational assessment is authored by advocates, of which I am one. This has resulted in a body of literature overtly committed to demonstrating its virtues with few dissenting voices. Kelly et al.'s paper is therefore a needed and refreshing counter to the growing pro-comparative judgement corpus, to which the current paper adds, and it would be healthy for the discipline if further sceptical researchers publish scholarly critiques.

Kelly et al. critiqued what they called the "intrinsic validity" (p.2022) rationale for using comparative judgement. The intrinsic validity rationale is that validity can be defined in terms of what experts collectively deem 'good' answers to be and therefore, advocates of comparative judgement argue, outcomes are inherently valid. Kelly et al. offer four specific critiques of the intrinsic validity

rationale: (i) there is no evidence relative judgements are superior to absolute judgements; (ii) researchers rarely define 'expert'; (iii) non-experts, specifically students comparatively judging peers' work, produce outcomes that correlate well with experts' outcomes; (iv) researchers sometimes remove poorly performing ('misfitting') experts thereby undermining their very definition of validity. They argue that "it is inconsistent to justify the use of a method based on its psychological underpinnings, and also to contend that the details of this theory are irrelevant provided the method works in practice" (p.679).

Finally, an objection related to validity is that comparative judgment can only produced norm-referenced grades (e.g. Dolan, 2021). Norm-referenced grading involves allocating grades statistically, such as awarding the top 10% of scores a grade A. This means that grade A for a 'weaker' cohort is not equivalent to grade A for a 'stronger' cohort. In contrast, criterion-referenced grading involves allocating grades against agreed standards. This means that we would expect fewer grade A's to be awarded to the 'weaker' cohort than the 'stronger' cohort (Lok, McNaught & Young, 2016). It has been asserted that comparative judgement can only be used for norm-referenced grading, presumably because each student's performance is judged relative to the other students' performances.

Responding to the validity objections.

It is the case that there could and should be research into the validity of comparative judgement for summative assessment, including across different mathematical topics, student age ranges, jurisdictions and so forth. However, most of the assertions of a lack of evidence seem not to acknowledge, let alone critique, the evidence that has been published over the past decade. Numerous studies have investigated the convergent, divergent and content validity of comparative judgement outcomes in mathematics education.

Convergent validity has been demonstrated by correlating or predicting outcomes with independent measures (standardised instruments, specially designed tests, achievement data, teacher estimates) across age ranges, learning contexts and mathematical topics (e.g. Davies et al., 2020; I. Jones et al., 2016; I. Jones et al. 2019; Marshall et al., 2020). For example, Bisson et al. (2019) conducted a randomised controlled trial using two outcome measures of students' conceptual understanding of calculus – one based on comparative judgement and the other based on a traditional standardised test – and compared the results.

Divergent validity has been investigated using independent measures we would not expect to correlate with or predict comparative judgement-based mathematics outcomes. For example, Bisson et al. (2019) found that secondary students' mathematics but not English GCSE grades predicted comparative judgement scores (GCSE refers to a terminal school qualification in parts of the United Kingdom). In another example, I. Jones et al. (2013) used comparative judgement to assess secondary students' explanations of how to put fractions in order. They found that conceptual but not procedural measures of mathematical knowledge predicted comparative judgement scores, thereby demonstrating the potential of comparative judgement to reliably assess conceptual understanding rather than procedural knowledge. Other studies have shown that non-expert judges (peers, lay

people) produce comparative judgement outcomes that diverge from those of experts (e.g. I. Jones & Alcock, 2014; I. Jones & Wheadon, 2015).

Content validity has been demonstrated using various techniques including expert review (e.g. I. Jones & Inglis, 2015), thematic analysis and related coding methods (e.g. Davies et al. 2021), and interviewing or surveying judges (e.g. Hunter & I. Jones, 2018; I. Jones et al., 2014; Marshall et al., 2020). An increasingly common method is to qualitatively code students' responses using existing frameworks (e.g. I. Jones & Karadeniz, 2016) or grounded approaches (e.g. Davies et al., 2020), and then use multiple regression techniques to identify the features of high-scoring responses. For example, I. Jones and Karadeniz (2016) applied a published coding scheme (Hunsader et al., 2014) and found that quantity written, use of numbers, and use of graphics predicted comparative judgement scores, but other features such as use of letters or use of 'real-world' examples did not. Similarly, Davies et al. (2020) developed a grounded coded scheme for undergraduates' definitions of proof, and found that comparative judgement scores produced by research mathematician judges were consistent with typical characterisations of proof reported by philosophers of mathematics (Davies et al., 2020).

This body of evidence partly addresses Kelly et al.'s objection that researchers assume that comparative judgement is intrinsically validity. We have seen convergent, divergent and content validity has been investigated across a variety of contexts, and not merely assumed to be inherent, the case for summative assessment in mathematics education.

I now turn to Kelly et al.'s specific objections of the intrinsic validity rationale.

(i) Kelly et al. claimed thar researchers have not established that relative judgements are superior to absolute judgements for producing reliable outcomes. However this is not the case, at least for the case of secondary school peer assessment. Jones and Wheadon (2015) showed that for absolute judgement outcomes, inter-rater reliabilities were effectively zero (mean $r = –.02$), whereas the relative (comparative judgement) outcomes, reliabilities were high (mean $r = .86$).

(ii) It is fair to claim that researchers tend not to provide a precise or universal definition of 'expert'. It is also fair to counter that expertise is clearly defined within the context of many published studies, such as "mathematics PhD students" (Davies et al., 2020, p. 188). Experts are also sometimes contrasted against non-experts such as peers or novices (e.g. I. Jones & Alcock, 2014). Nevertheless, these operationalisations and contrasts tend to be buried in methods sections, and researchers could and should be more upfront and explicit about the term 'expert'.

(iii) A related point is Kelly et al.'s objection that non-experts' outcomes sometimes "correlated well" (p. 682) with the outcomes of experts. This has indeed been the case in some studies (e.g. I. Jones & Wheadon, 2015), and moreover has been used as an argument for peer judgements contributing to summative outcomes. However, there is an important caveat: studies typically use many times more peer judgements than expert judgements because the former are easier to obtained. Crucially, the number of judgements per student answer collected is correlated with reliability (see Verhavert et al., 2019, for a detailed explanation and demonstration), and this largely explains peers' robust outcomes. Moreover, when novices (e.g. participants who have received no mathematics education beyond secondary school and have not studied calculus) make comparative judgements of undergraduate

mathematicians' responses to an open-ended calculus question, correlations are lower (e.g. I. Jones & Alcock, 2014).

(iv) Kelly et al.'s objection that removing 'misfitting' experts undermines claims that validity is intrinsic to collective expert judgement has merit. Opinions as to whether or not misfitting experts should be identified and removed varies across comparative judgement researchers in my experience. I am of the view that removing misfitting experts is generally not necessary or desirable (see I. Jones & Davies, 2023) bar rare occasions when an expert appears not to have judged in good faith, or with adequate attention (e.g. they completed their judgements suspiciously quickly). Therefore, Kelly et al. make a good theoretical point and researchers should consider carefully the rationale and validity of removing misfitting expert judges.

Finally, it is not the case as asserted by some (e.g. Dolan, 2021) that comparative judgment can only be used to produce norm-referenced grades. There are several methods are available for criterion-referencing comparative judgment scores when producing grade boundaries. For example, Marshall et al. (2020) included exemplar grade-boundary scripts in the judging pot of students' responses. The boundary scripts' scores were then used as cut-scores to assign grades. In fact comparing the standards of grades across different cohorts is a particular strength of comparative judgement approaches. For example, I. Jones et al. (2016) used comparative judgement to investigate changes over five decades of a terminal qualification (A-level Mathematics) in England. Conversely, albeit in the context of creative writing, Heldsinger and Humphry (2013) used comparative judgement to identify grade boundary scripts, which were then disseminated to teachers in order to establish consistent writing standards in Australia.

Discussion.

I have presented objections to the validity of using comparative judgement to produce summative scores or grades. I have argued that in the main these objections are addressed by the published research evidence. That is not to say more validity evidence would be unwelcome. For example, recent developments have included eye-tracking equipment to investigate how experts make judgement decisions of argumentative writing (Gijsen et al., 2021), and such methods should be applied in the context of mathematics education. In addition, while criterion-referencing methods have been applied to comparative judgement outcomes (e.g. Marshall et al., 2020), these methods have not been widely published or, to the best of my knowledge, systematically investigated.

I agree with Kelly et al.'s conclusion that there is a need for a "comprehensive, systematic review of the evidence to explore to what extent the rationales for using comparative judgment have empirical support" (p. 684). Critical review is essential to scholarly progress and, based on the theoretical and detailed objections of Kelly et al., comparative judgement researchers should consider giving greater thought to and being clearer about their use of the term 'expert'. They should also reflect carefully on the implications for validity, including Kelly et al.'s critique of what they call intrinsic validity, when considering removing misfitting expert judges.

Final comment.

Objections to the validity of comparative judgement tend not to distinguish between different subjects. There is a motivation to use comparative judgement that is specific to *mathematics* education which I set out in the introduction to this paper: comparative judgement can reduce exam and test designers' dependence on short, objective test questions, and allow greater use of open-ended, relatively unstructured formats, without the reduced reliability associated with rubric-based scoring. This is not the case for other subject areas. For example, researchers have investigated using comparative judgement to assess students' creative writing (e.g. Wheadon et al., 2020). Here the motivation, unlike for assessing mathematics, is not to enable different types of test questions, but to improve the reliability of assessing existing test questions (Pollitt, 2012).

To conclude, Kelly et al.'s "intrinsic validity" could be amended for the case of mathematics education to include the theoretical argument that comparative judgement enables the inclusion of more open-ended questions in exams and tests, and therefore can increase their validity with no loss of reliability.

References

Bartholomew, S. R., & Jones, M. D. (2022). A systematized review of research with adaptive comparative judgment (ACJ) in higher education. *International Journal of Technology and Design Education, 32*, 1159–1190. https://doi.org/10.1007/s10798-020-09642-6

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education, 2*(2), 141–164. https://doi.org/10.1007/s40753-016-0024-3

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, 22*(3), 284–303. https://doi.org/10.1080/14794802.2019.1692060

Bokhove, C. [@cbokhove]. (2019, June 27). *The time is absolutely not ripe for Comparative Judgement at a national scale in year 6 sats. Validity, *total* time, accountability, logistics are challenges. Maybe just as additional moderation. I feel such desires mainly based on dislike of current system, not eval of CJ.* Twitter. https://twitter.com/cbokhove/status/1144137828621680640?s=43&t=ohKI7Uu4ddr7VZQXzHd2sA

Davis, A. J. [@ded6ajd]. (2017, April 8). *Comparative judgment drains from the assessment any considerations of *what* is being assessed & the whole question of validity worth having.* Twitter. https://twitter.com/ded6ajd/status/850639513399545856?s=43&t=ohKI7Uu4ddr7VZQXzHd2sA

Davies, B., Alcock, L., & Jones, I. (2020). Comparative judgement, proof summaries and proof comprehension. *Educational Studies in Mathematics, 105*(2), 181–197. https://doi.org/10.1007/s10649-020-09984-x

Davies, B., Alcock, L., & Jones, I. (2021). What do mathematicians mean by proof? A comparative-judgement study of students' and mathematicians' views. *The Journal of Mathematical Behavior, 61*, 100824. https://doi.org/10.1016/j.jmathb.2020.100824

Dolan, T. [@timdolan]. (2021, February 2). *Maths Teachers: Has anyone considered using comparative judgement as part of Y11 or Y13 assessment? I wonder whether it's worth exploring, for data to help with valid inferences about mathematical ability. Would only be norm-referenced data, which may be of limited value?* Twitter. https://twitter.com/timdolan/status/1356532458771087360?s=43&t=ohKI7Uu4ddr7VZQXzHd2sA

Gijsen, M., Van Daal, T., Lesterhuis, M., Gijbels, D., & De Maeyer, S. (2021). The complexity of comparative judgments in assessing argumentative writing: An eye tracking study. *Frontiers in Education, 314*. https://doi.org/10.3389/feduc.2020.582800

Heldsinger, S. A., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: An empirical study. *Educational Research, 55*(3), 219–235. https://doi.org/10.1080/00131881.2013.825159

Hunter, J., & Jones, I. (2018). Free-response tasks in primary mathematics: A window on students' thinking. *Proceedings of the 41st Annual Conference of the Mathematics Education Research Group of Australasia, 41*, 400–407.

Hunsader, P. D., Thompson, D. R., Zorin, B., Mohn, A. L., Zakrzewski, J., Karadeniz, I., Fisher, E. & MacDonald, G. (2014). Assessments accompanying published textbooks: the extent to which mathematical processes are evident. *ZDM, 46*, 797-813. https://doi.org/10.1007/s11858-014-0570-6

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787. https://doi.org/10.1080/03075079.2013.821974

Jones, I., & Davies, B. (2023). Comparative judgement in education research. *International Journal of Research & Method in Education*. https://doi.org/10.1080/1743727X.2023.2242273

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics, 89*(3), 337–355. https://doi.org/10.1007/s10649-015-9607-1

Jones, I., Inglis, M., Gilmore, C., & Hodgen, J. (2013). Measuring conceptual understanding: The case of fractions. In A. M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 113–120). IGPME.

Jones, I., & Karadeniz, I. (2016). An alternative approach to assessing achievement. In C. Csikos, A. Rausch, & J. Szitanyi (Eds.), *The 40th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 51–58). IGPME.

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. https://doi.org/10.1007/s10763-013-9497-6

Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation, 47*, 93–101. https://doi.org/10.1016/j.stueduc.2015.09.004

Jones, I., Wheadon, C., Humphries, S., & Inglis, M. (2016). Fifty years of A-level mathematics: Have standards changed? *British Educational Research Journal, 42*(4), 543–560. https://doi.org/10.1002/berj.3224

Kelly, K. T., Richardson, M., & Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: A call for clarity. *Assessment in Education: Principles, Policy & Practice, 29*(6), 674–688. https://doi.org/10.1080/0969594X.2022.2147901

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. Assessment & Evaluation in Higher Education, 41(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136

Marshall, N., Shaw, K., Hunter, J., & Jones, I. (2020). Assessment by comparative judgement: An application to secondary statistics and English in New Zealand. *New Zealand Journal of Educational Studies, 55*(1), 49–71. https://doi.org/10.1007/s40841-020-00163-3

NCTM Research Committee. (2013). New Assessments for New Standards: The Potential Transformation of Mathematics Education and Its Research Implications. Journal for Research in Mathematics Education, 44, 340–352. https://doi.org/10.5951/jresematheduc.44.2.0340

Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal, 22*, 405–420. https://doi.org/10.1080/0141192960220403

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

Swan, M., & Burkhardt, H. (2012). Designing assessment of performance in mathematics. *Educational Designer, 2*(5), 1–41. https://isdde.org/wp-content/uploads/2018/05/isdde09_burkhardt_swan.pdf

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273–286. https://doi.org/10.1037/h0070288

Verhavert, S., Bouwer, R., Donche, V., & Maeyer, S. D. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2020). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice*, 27(1), 46–64. https://doi.org/10.1080/0969594X.2019.1700212