Assessment in Education: Principles, Policy & Practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

Critiquing the rationales for using comparative judgement: a call for clarity

Kate Tremain Kelly, Mary Richardson & Talia Isaacs

To cite this article: Kate Tremain Kelly, Mary Richardson & Talia Isaacs (2022): Critiquing the rationales for using comparative judgement: a call for clarity, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2022.2147901

To link to this article: https://doi.org/10.1080/0969594X.2022.2147901

Published online: 18 Nov 2022.

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
https://doi.org/10.1080/0969594X.2022.2147901

Routledge
Taylor & Francis Group

Critiquing the rationales for using comparative judgement: a call for clarity

Kate Tremain Kelly [a,b], Mary Richardson [a] and Talia Isaacs [a]

[a]IOE - Curriculum, Pedagogy & Assessment, UCL Institute of Education, University College London, UK; [b]Research and Development, AQA, Manchester, UK

ABSTRACT

Comparative judgment is gaining popularity as an assessment tool, including for high-stakes testing purposes, despite relatively little research on the use of the technique. Advocates claim two main rationales for its use: that comparative judgment is valid because humans are better at comparative than absolute judgment, and because it distils the aggregate view of expert judges. We explore these contentions. We argue that the psychological underpinnings used to justify the method are superficially treated in the literature. We conceptualise and critique the notion that comparative judgment is 'intrinsically valid' due to its use of expert judges. We conclude that the rationales as presented by the comparative judgment literature are incomplete and inconsistent. We recommend that future work should clarify its position regarding the psychological underpinnings of comparative judgment, and if necessary present a more compelling case; for example, by integrating the comparative judgment literature with evidence from other fields.

ARTICLE HISTORY
Received 17 August 2021
Accepted 8 November 2022

KEYWORDS
Comparative judgment; pairwise comparisons; validity; scoring; human judgements

Introduction

Broadly put, a comparative judgment is one in which two or more stimuli are judged in relation to each other on the basis of some criterion. In contrast, absolute judgment is where a stimulus is judged in isolation. In absolute judgment, absolute quality is decided: in comparative judgment, only relative quality is determined. In an educational context, most UK examination boards use absolute judgements for their large-scale assessments – such as the GCSE qualification – to evaluate student performance and award marks or grades. However, methods which use comparative judgements are generating interest (e.g. Curcin et al., 2019). Typically, these methods involve judging performances by comparing two items and deciding which of them best meets the agreed criteria. The process is repeated a large number of times, with a pool of judges, so that the items can be placed on a scale. This scale is created using a statistical model that provides more useful information than if the items had simply been placed in rank order (Pollitt, 2012a).[1]

Fundamental to the case for using comparative judgment are two claims. First, that humans are better at making comparative than absolute judgements (Bramley, 2005;

CONTACT Kate Tremain Kelly katherine.kelly.15@ucl.ac.uk UCL Institute of Education, 20 Bedford Way, Bloomsbury, London WC1H 0AL, UK

Jones & Inglis, 2015). Second, and dependent in part upon the first, is that judgements made by experts using comparative judgment necessarily have high levels of validity. As Pollitt (2009) suggests that validity is inherent to the scoring process in comparative judgment, we refer to this interpretation of validity henceforth as the 'intrinsic validity' argument.

In this paper, we interrogate these claims as they are laid out in the comparative judgment literature. Pertaining to the first, we highlight that the use of comparative judgment is justified by its psychological underpinnings, but these are treated superficially in the literature. We then suggest that the intrinsic validity argument is not explicitly situated within a particular validity framework, and represents an attempt to create a new definition of validity specific to comparative judgment. We explain and then critique the intrinsic validity argument, to better understand its requirements and implications. Consistent with this conceptualisation of validity, we do not explicitly espouse a particular validity framework in this article, but argue below that intrinsic validity is insufficient alone for some of the suggested applications of comparative judgment.

To provide some context, we first outline the way comparative judgment is used in educational assessment, and then explain why it is being considered as an alternative to absolute judgment.

Comparative judgment in educational assessment

Pollitt (2012b) lays claim to the introduction of comparative judgment to educational assessment in England, and through his advocacy comparative judgment has become the subject of a nascent but growing international research literature (Bramley, 2007).

Comparative judgment has been used to assess a wide range of attributes across a range of subjects, such as problem-solving in mathematics (Jones & Inglis, 2015; Jones et al., 2015), proficiency in English as a foreign language (Pollitt & Murray, 1993), academic writing (Van Daal et al., 2019), quality of electronic portfolios in design and technology (Kimbell, 2012), and essays in both English literature (Steedle & Ferrara, 2016) and geography (Whitehouse & Pollitt, 2012).

Most often, comparative judgment has been used to generate estimates of performance quality intended to replace marks. It has also been used for moderating teacher-assessed work (Wheadon & Christodoulou, 2016), aligning standards across different tests (Black & Bramley, 2008; Bramley, 2005; Humphry & McGrane, 2015; Suto & Novaković, 2012), evaluating the quality of mathematics problem-solving questions (Holmes et al., 2017), and investigating comparability of standards across examination boards (Ofqual, 2015; Pollitt & Elliott, 2003) and of performance standards over time (Jones et al., 2016).

Advocates for the method vary in their expectations for the role of comparative judgment in high-stakes national assessments. Some, such as Pollitt, and Tarricone and Newhouse, have suggested that comparative judgment methods should replace marking and grading activities (Pollitt, 2009; Pollitt & Crisp, 2004; Pollitt & Elliott, 2003; Pollitt et al., 2004; Tarricone & Newhouse, 2016). Proponents of CJ have also suggested that comparative judgment could play a role in quality assurance. For example, the comparability of equivalent qualifications from different examination boards could be

investigated using the methods outlined in Pollitt and Elliott (2003) or Ofqual (2015), as could the comparability of performance standards over time (Jones et al., 2016). Similarly, Wheadon and Christodoulou's (2016) work on teacher moderation within schools could be extended to assuring the quality of marking and grading.

Arguments for using comparative judgment in educational assessment

One motivation for using comparative judgment in educational assessment is that the rank orders it produces tend to have high levels of reliability. In an overview of 16 comparative judgment exercises conducted between 1998 and 2015, the Scale Separation Reliability (SSR) indices – equivalent to Cronbach's alpha (although see, Verhavert et al., 2017) – ranged from .73 to .99 (Steedle & Ferrara, 2016). This is higher than is typically obtained through marking using a scoring rubric, at least for essay questions (Pollitt, 2012a; Steedle & Ferrara, 2016).

Comparisons of SSR with other reliability indices is not straightforward; Verhavert et al. (2017) query the way it is conceptualised within comparative judgment, and some research suggests that the measure is artificially inflated, particularly for adaptive comparative methods (Bramley & Vitello, 2019; Crompvoets et al., 2019). Nevertheless, there are plausible reasons for the reliability to be high. For example, in contrast to conventional marking, where each response is likely to be viewed only once or twice, scripts are viewed multiple times (Benton & Gallacher, 2018; Pollitt & Elliott, 2003). It is probably not possible for individuals to accurately distinguish between performances within a small range of marks (Baird & Dhillon, 2005), although the accuracy may vary by subject and by judge (see, e.g. Gill & Bramley, 2013; Van Daal, 2020; Van Daal et al., 2017). By aggregating the views of a number of judges, as occurs in comparative judgment, Benton and Elliott (2016) argue that accurate distinctions can be obtained.

Eliciting relative judgements also has the effect of cancelling out differences in examiner severity and leniency (Bramley, 2007). With absolute judgment, severity or leniency in judges risks an evaluation of performance that is too high or too low.[2] With comparative judgment, it matters only that judges are consistent with one another in terms of the rank order of scripts (Pollitt, 2012a).

A second motivation derives largely from the work of experimental psychologist Donald Laming, and his argument that humans are better at drawing comparisons than at making absolute estimates of value. In fact, Laming took this a step further, stating that 'there is no absolute judgment. All judgments are comparisons of one thing with another' (Laming, 2004a, p. 9). If we accept, for the meantime, that this statement is true, we must modify our definition of comparative judgment. Accounting for Laming's argument, a comparative judgment is more properly understood as a judgment in which the comparator is explicitly specified and presented, rather than a judgment in which a remembered stimulus, an internal benchmark, a platonic ideal or some other mental representation acts as the comparator (Raikes et al., 2008; Suto & Novaković, 2012). For simplicity, this latter scenario will be referred to as absolute judgment throughout this paper.

The first motivation is based on the reliability of comparative judgment, and has been discussed in depth elsewhere (e.g. Bramley, 2015; Pollitt, 2015). The second motivation contributes to the novel perspective of validity proposed by advocates for comparative

judgment, and has been central to the rationale for using comparative judgment. Neither the argument that humans are better at comparative judgment nor the intrinsic validity argument (i.e. that judgements made by experts using comparative judgment are necessarily valid) have been extensively considered in the literature. This paper is an opportunity to explore these ideas in some more detail and we present them as two, discrete claims. We turn first to the claim that humans are better at comparative than at absolute judgements.

Claim 1: humans are better at comparative judgment: the foundation of the rationale for using comparative judgment

Psychological arguments underpin the rationale for using comparative judgment. Those who advocate for comparative judgment do so in part because these judgements are supposedly easier to make accurately. Comparative judgment researchers have asserted that 'there is plenty of evidence from psychology that humans are much better at making comparative judgments than absolute judgments' (Bramley, 2005, p. 204), and that 'the underlying theoretical basis is a well-established psychological principle' (Jones & Inglis, 2015, p. 341). Yet the evidence has not been presented to support these claims. These statements are rarely supported by citations other than those of Thurstone's or Laming's work, and it is rare for this evidence to be critically evaluated.

The roots of comparative judgment go back to the early years of psychology, when psychophysicists used paired comparisons to explore how humans perceive physical stimuli (Bramley, 2007; Thurstone, 1931). Participants would make a series of comparative judgements on a set of stimuli, deciding which of each pair possessed more of the attribute under investigation (Bramley, 2007). Thurstone's main contribution to the method was to demonstrate how it could be used to order the stimuli along a scale – what Thurstone termed the 'psychological continuum' (Bramley, 2007; Thurstone, 1927a). The resulting Law of Comparative Judgment (Thurstone, 1927a) – a mathematical expression – allowed psychophysicists to produce measurement scales for psychological entities with no obvious physical correlates (Bramley, 2007). Thurstone himself used his law to evaluate properties such as excellence of handwriting and seriousness of crime (Bramley, 2007). However, as Pollitt (2012b) notes, the usefulness of Thurstone's innovation was limited because he rooted his analysis in the normal distribution, which is challenging to use computationally. This difficulty was alleviated in 1978 by Andrich, who noted that Thurstone's model could be reformulated as a type of Rasch model, which is more computationally manageable (Andrich, 1978).

By removing the requirement for a physical correlate, Thurstone's law made possible the measurement of a wide range of attributes. In fact, Thurstone argued that his approach could be applied to any attribute which could meaningfully be ranked on some basis (Thurstone, 1927a). Nevertheless, there are a number of reasons to be apprehensive about the application of his work to educational assessment. Firstly, Thurstone proposed five different cases of his Law of Comparative Judgment, in which additional statistical assumptions were made (Thurstone, 1927a). Adding assumptions simplifies the formula and improves its tractability, but one cannot place as much confidence in the outputs of the formula if those assumptions are not met. Case 5 contained the largest number of assumptions and is the simplest formula Thurstone

proposed. It is also the version used for comparative judgment (Andrich, 1978). Bramley notes that

> Thurstone seems to have had a rather ambivalent attitude to the Case 5 version of his law, saying "This is a simple observation equation which may be used for rather coarse scaling" … yet it seems to have been the one he used most often in practical applications! (Bramley, 2007, p. 253)

Pollitt (2012b) has argued that the assumptions made in Case 5 are plausible. Pollitt extended this argument, claiming that empirical evidence demonstrating the robustness of comparative judgment to missing data shows that

> Thurstone's original psychological underpinning for the method is unnecessary – whatever psychological reality underlies comparative judgment is general enough to apply in many contexts that do not meet the conditions that Thurstone assumed for his research. (Pollitt, 2012b, p. 160)

Regardless of one's personal stance on the value of theory, 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.

It also seems unclear whether Thurstone's work really can be applied to the stimuli typically used in assessment. While Thurstone used his Law to create measurement scales for purely psychological or subjective entities, he did so only with stimuli whose values could be evaluated relatively instantaneously, such as short statements and handwriting samples. He was also sceptical about its use for very heterogeneous stimuli (Thurstone, 1927b). In assessment, the stimuli presented for judgment are usually fairly long and complex, but short judgment times have often been recorded. Pollitt (2012a) reported mean judgment times of 3 minutes 15 seconds for the faster half of a sample assessing primary writing skills. This is significantly shorter than the estimate of 15 to 20 minutes suggested for marking the same scripts analytically (Pollitt, 2012a). McMahon and Jones (2015) reported one of the shortest times, with 33 seconds per judgment of experimental reports, although scripts in the same study took only 51 seconds to mark analytically. There is some evidence to suggest that fast judges are no less accurate than slower judges (Jones et al., 2015; Pollitt, 2012a), although the links between speed, accuracy and quality of judgment remain under-researched.

However, as Bramley (2007) noted, while it is possible to perceive two shorter stimuli simultaneously, it is not clear that this holds for longer, more complex, stimuli, as judges may need to recall the first stimuli in evaluating the second. He suggested this may lead to order effects in paired comparisons, where features of the first script interfere with the interpretation of the second, and vice versa,

> ironically, it is possible that the judges might resort to using an internal standard … when making their judgements – even though one of the main benefits of the method is that in theory it removes the need to do this. (Bramley, 2007, p. 273)

Applying comparative judgment to an assessment context goes beyond the work of Thurstone in terms of the complexity of the stimuli used and the statistical assumptions required. If Thurstone's Law is to be used in support of comparative judgment for assessment, his theory must be extended to cover the uses and applications that are required for assessment purposes.

Moving on to the work of Laming, while his statement that all judgments are comparisons is quoted regularly, it is a partial quotation, of which the other half is rarely acknowledged (although, see, Pollitt et al., 2004). Having stated that 'all judgments are comparisons of one thing with another', Laming added that 'these judgments are little better than ordinal' (Laming, 2004a, p. 68). In fact, Laming devoted a substantial portion of his book, from which that quotation is taken, to elucidating the many ways in which human judgment is flawed. Laming indisputably argued that humans are *better* at comparative judgments than at absolute judgments, but he did not argue that humans are *good* at either type of judgment. In the final chapter of the book, Laming concluded that when there is no objective basis for comparison, 'people are vulnerable to even the slightest and subtlest suggestions from others; indeed, they have no defence against extraneous suggestion' (Laming, 2004a, p. 268). Further, Laming argued that 'in most practical situations . . . there is some material evidence on which to base a judgment, but not so firm a basis for comparison but that some doubt remains' (Laming, 2004a, p. 268). This does not support the notion – outlined further below – that the collective expertise of judges is all that is needed for reliable and valid assessment, and the theoretical jump from Laming to comparative judgment in its modern incarnation has not been made within the comparative judgment literature. As Pollitt (2012a) has noted, Laming (2004b) actually used his research to argue that examinations should be limited to questions that could be marked objectively. Assessment literature exploring examiners' cognitive processes for marking, grading, and examiner training does suggest an important role for comparison in tasks such as marking, grading and standardising (e.g. Crisp, 2008, 2010a, 2010b; Elliott, 2017; Gill & Bramley, 2013). However, this work is still at an early stage, and does not seem to have been integrated fully with the literature on comparative judgment.

In addition, no clear connection has been made between the works of the two authors. Laming's two most frequently cited outputs (Laming, 2004a, 2004b) do not cite Thurstone. Laming did cite Thurstone in his 1984 paper, but then only to disagree with a mathematical convention that he originated (Laming, 1984). Thurstone, working decades earlier, clearly did not cite Laming. While there are clear theoretical overlaps between the two authors' work, no attempt has been made to draw them together in the comparative judgment literature.

Claim 2: comparative judgments are necessarily valid: the 'intrinsic validity' argument

Conceptualising the intrinsic validity argument

Stemming from the argument that humans are naturally better at making comparative judgments is the argument that judgments made comparatively are *de facto* valid. Pollitt (2009) summarises the argument thus:

> What could be more valid than judging that one piece of work is more creative than another? Or more effective? And if many judges agree that the same one is better, isn't that the best evidence for validity we could ask for? (Pollitt, 2009, p. 2)

Put another way, comparative judgment is proposed as intrinsically valid because the eventual outcomes represent the aggregate view of the judges: distillations of the

collective expertise of the relevant set of experts (Bisson et al., 2020). We submit that this argument, as it is presented in the literature, represents an attempt to create a new definition of validity specific to comparative judgment. Here we term this perspective the 'intrinsic validity' argument (see, e.g. Pollitt, 2009). This perspective is not explicitly situated within a particular validity framework, although proponents draw on other types of validity in their research (e.g. concurrent validity, as in Steedle & Ferrara, 2016), and reference construct validity in their rationales for using comparative judgment (see below).

Key to the intrinsic validity argument is the removal of the scoring rubric, which is no longer required when judgments are relative, and which is thought to impose a particular view of the construct onto markers: this is the view taken by those who developed the assessment (Jones & Inglis, 2015). The claim is that comparative judgment allows performances to be judged directly against the construct (Pollitt & Crisp, 2004), rather than indirectly against the construct as represented through the artifice of a scoring rubric. Without a rubric, it is thought that judges are freed to make holistic (Van Daal et al., 2019), *and* more valid (Pollitt, 2012b) judgments. For example, an English teacher could use their own experience and expertise to judge the quality of student writing, rather than trying to follow someone else's attempt to articulate what good looks like.

This is considered a particular advantage when assessing constructs that do not readily lend themselves to quantification. For example, Bisson et al. (2016) justified their use of comparative judgment by stating that:

> . . . conceptual understanding is an important but nebulous construct which experts can recognise examples of, but which is difficult to specify comprehensively and accurately in scoring rubrics . . . rubrics attempt to capture the letter of a concept but risk losing the spirit. (Bisson et al., 2016, p. 143)

Comparative judgment, then, is thought to allow for reliable assessment of nebulous concepts, and to do so validly through the avoidance of 'reductionist' scoring rubrics. It is also suggested (see, Van Daal et al., 2019) that examiners may judge work holistically anyway, adjusting scores such that the final score is in line with their overall impression of the work. Crisp (2010a) reported some evidence of this approach: using a think-aloud method, examiners marking geography papers reported re-evaluating their original decisions in light of the total mark achieved by the student and the level of quality that was felt to represent.

Removing the requirement for a scoring rubric is also thought to improve validity by increasing the freedom of those who create the assessments (Jones & Inglis, 2015; Pollitt, 2009), as they do not have to consider how to ensure the answers will be marked reliably and validly. Comparative judgment may also facilitate reliable scoring for tasks that elicit a wide range of responses. When it is difficult to predict possible responses, it is difficult to write rubrics which adequately account for all responses (Bisson et al., 2016). Pollitt et al. (2004) do note that creating a rubric can be useful in focusing question writers' minds on the types of performances they wish to elicit. They suggest that it would probably be necessary to create a generic mark scheme expressing the qualities desired to facilitate the development of a common view of the construct.

Critiquing the intrinsic validity argument about comparative judgment

The argument that comparative judgment is intrinsically valid because the outcomes represent the aggregate view of experts is undermined by the vagueness with which an expert is defined. In fact, it is rarely defined at all. One of the most explicit definitions comes from Bisson et al. (2016), who state that expert refers to researchers for whom the assessed concept is relevant for their discipline. Bramley (2007) suggests further that the judges must have some shared conception of what makes one script better than another. This conception must also be related to the assessment intentions. Indeed, Whitehouse and Pollitt (2012) found that, without training, judges must have experience of teaching at the level being assessed in order to judge consistently (e.g. primary/elementary phase, or secondary for national tests and qualifications).

Perhaps the broadest definition of who is qualified to make such judgments comes from Pollitt, who posits that 'any interested party could, in principle, be invited to make judgments'. The argument runs that 'parents, politicians and journalists' could 'try the system and so gain a better understanding of its strengths'. Pollitt goes on to say that 'the judgments they make might not be included as real assessment data – *if* they turn out to be "misfits"'[3] (Pollitt, 2012a, p. 297, emphases added). This suggests an expert is defined as anyone whose judgments cohere with the pool of judges as a whole.HT In practice, the choice of judges reflects a variety of perspectives on expertise – from subject matter expertise (e.g. Jones & Alcock, 2014) to assessment skill (e.g. Whitehouse, 2012) or experience teaching at the level being assessed (e.g. McMahon & Jones, 2015). While any of these could reasonably be considered relevant expertise, what is lacking is a clear theoretical framework linking the choice of expertise with the needs of the comparative judgment process.

However, the definition of expert in this assessment context is further complicated by studies suggesting that training is unnecessary for comparative judgment exercises to be conducted well. Jones and Wheadon (2015) found that untrained secondary school students (aged 12–15) produced outcomes which correlated well with the outcomes produced by expert judges. Heldsinger and Humphry (2010) similarly found that untrained assessors produced highly reliable outcomes that correlated well with results from a cognate test. Bisson et al. (2016) found that judges who received no guidance performed equivalently to judges who received guidance, and both showed good correlations with external measures of quality. In all cases cited, untrained assessors produced parameter estimates of high reliability, and which correlated well with scores produced by more conventional methods (Bisson et al., 2016; Heldsinger & Humphry, 2010; Jones & Wheadon, 2015). Even when judges in a comparative judgment exercise are provided with reference materials to guide their judgments, it seems they do not always refer to them (Pollitt, 2012a).

Another challenge to the intrinsic validity argument is that the aggregate approach and valuing of diversity in expert views (Van Daal et al., 2019) is in tension with the idea of removing misfitting judges (Pollitt, 2012a). If someone is considered an expert, and the validity of comparative judgment rests on the collective expertise of a community of practice, the decision to remove them and discard their judgments can only be justifiable on the assumption that those judgments do not reflect their expertise. For example, it may be that they were judging erratically in order to complete the exercise quickly.

However, this is not the usual argument proposed in such cases, because as Van Daal et al. (2019) note, misfitting judges are generally supposed to diverge in their conceptualisation of the construct (e.g. Bramley, 2007; Roose et al., 2019; Whitehouse & Pollitt, 2012). It is thus unclear whether or not this divergence is welcomed.

There is a paucity of discussion as to how many judges should be included. If the validity of the exercise lies in the distillation of collective expertise in a subject area, then presumably that validity must be ensured by involving a sufficiently large number of judges. Studies vary as to the number of judges included, from four or five (McMahon & Jones, 2015; Pollitt & Crisp, 2004), to more than 30 (Humphry & McGrane, 2015; Jones & Wheadon, 2015; Pollitt, 2012a). Bramley notes that comparability studies typically include between 10 and 20 judges, adding that Thurstone's studies usually included over 200 judges (Bramley, 2007).

Overall, there is a general dearth of advice and evidence regarding sampling in comparative judgment exercises. Bisson et al. (2016) suggest a general guideline of at least 10 judges in each exercise, although this is based on experience of running exercises rather than any kind of systematic trial. One of the most comprehensive investigations was conducted by Benton and Elliott (2016), which indicated between 10 and 20 judges would ensure an acceptable level of reliability. Suggestions also vary as to the minimum number of judgments that must be obtained to ensure an acceptable level of reliability – recommendations range from five to 25 judgments per stimulus (Pollitt, 2012b; Pollitt & Crisp, 2004; Steedle & Ferrara, 2016; Wheadon, 2015). To date, there is no evidence to support a minimum number of judges and judgments required to ensure an acceptable level of validity.

Thus far, we have critiqued the intrinsic validity argument using its own criteria: using the expertise of judges as the basis for validity requires a clear and common framework for who can judge and with what justification. Whether the intrinsic validity argument is in fact considered to be a distinct approach to validity, or whether it sits within existing theoretical frameworks, is also unclear. In short, we consider that the intrinsic validity argument has not yet been fully developed.

Moreover, for some purposes this perspective on validity is too limited. In England the results of national qualifications such as the GCSEs are used for multiple purposes, and thus evaluated against multiple criteria (Baird et al., 2000). If, as some have suggested (e.g. Pollitt, 2009), comparative judgment should be used for marking and grading in these and similar contexts, we need to broaden the perspectives and understanding of validity.

The removal of scoring rubrics and trusting in expert judgment also removes transparency: it becomes challenging to explain the basis for awarding a particular mark or grade (Steedle & Ferrara, 2016). Bisson et al. acknowledge that the use of collective expertise can be seen as 'opaque and under-defined' but counter that 'this is a key strength: a given concept is defined by how it is perceived, understood and used by the relevant community of expert practitioners' (Bisson et al., 2016, p. 143). While this may be acceptable for low-stakes environments, it is less convincing when the stakes are higher. The expert reviewer defence is unlikely to reassure a young school-leaver who feels they have been unfairly denied their place at university.

The question of *who* is considered expert enough to undertake the judgments maintains its importance because the holistic nature of the judgments allows judges to differ –

within reason – as to what they value and how they weight different elements of a performance (Van Daal et al., 2019). The logic is that in aggregating the varied judgments, one comes to an overall view of the community of practice. However, this raises important questions about diversity of the judging pool. If it is acknowledged that judges may differ in what they value, it is not inconceivable that judges may differ as a function of their gender, age, ethnicity, culture, class, income, education and so on. Would the aggregate view of young, British-Asian men always be the same as the aggregate view of older, black women? For example, Bartholomew et al. (2020) found differences between what judges from the UK, Sweden, and the USA valued when assessing design portfolios, providing evidence to support divergence along demographic lines.

Such differences may not be detected through analyses of bias and misfit. McMahon and Jones (2015) found differences between students and teachers in what was valued in assessments of understanding of a chemistry experiment. Where students prioritised factual recall, teachers prioritised scientific explanation. Both groups produced highly reliable scales (reliability estimates of .893 for students, and .874 for teachers). Although, in this case, a clear argument as to whose consensus should take priority can be made, it demonstrates the existence of different, but equally reliable, sets of consensuses depending on who is asked. Assessments using absolute judgment are not exempt from this concern about different groups taking different perspectives (see, for example, Johnson, 2015). However, the scoring rubrics and marker monitoring provide an audit trail and a route for students to appeal against their outcomes. It is unclear how this could be achieved with comparative judgment. Pollitt (2012a) has argued that an audit trail is provided in comparative judgment through the record of each comparison made, and that student performances where marks are appealed can be sent for additional judgments. This does suggest a practical mechanism for dealing with appeals, but does not really address the principal issues of transparency and fairness. Adding more comparisons does not guarantee that the comparisons are appropriate.

The field of comparative judgment, while growing, is still relatively small, and at present we do not have compelling answers to fundamental questions such as: how many judges do we need? Who should judge? How many judgments should they make? And crucially, how can we be confident that they are basing their judgments on acceptable criteria? If judges are not using construct-relevant criteria, then the argument that comparative judgments are necessarily valid cannot be upheld. We recognise that there is unlikely to be a single answer to these questions for all the cases in which comparative judgment may be used. However, work to derive some general principles for good practice would be valuable.

Concluding remarks

Methods based on comparative judgment have a number of desirable features and present an appealing approach that could be implemented in different assessment contexts. However, comparative judgment advocates have not compiled a compelling case to support two of their central claims that,

(a) humans are better at comparative than absolute judgments, and

(b) comparative judgment is necessarily valid because it aggregates judgments made by experts in a naturalistic way.

What literature has been compiled does not wholly support the application of comparative judgment to assessment contexts. This is not to suggest that arguments made by comparative judgment advocates are inaccurate. Rather, the evidence – which likely exists in other literatures – has not been brought together to fully support those arguments.

We highlighted the inconsistency of justifying a method's use on the basis of its psychological underpinnings, while also contending that the details of these underpinnings are irrelevant if the method works in practice. For many applications, it may suffice for the method to work in practice – whatever 'work' means for any given application – and references to psychological underpinnings are unnecessary and unhelpful. This may even be appropriate for high-stakes testing applications, if the criteria and approach for validation is made sufficiently explicit (as should be the case for any method of judgment). However, if advocates wish to continue drawing on psychological arguments to support their use of comparative judgment, we propose that a more thorough exploration of these claims is needed. In particular, a theory of comparative judgment in its modern incarnation, which deviates in a number of significant ways from Thurstone's version, should be developed. We suggest that researchers should look to other fields, such as cognitive psychology, for the evidence required to build a case for its use in educational assessment.

We also unpacked the logic and implications of a validity argument centred on the use of expert judgments. We recognise that many of our procedural queries – how many judges to choose, how many judgments, what kind of training or shared understanding is required – are still being researched, and these will likely be addressed in time. However, we remain concerned that the implications of a judgment-based definition of validity should not be overlooked. We posit that the choice of judges and how they have internalised the construct is not necessarily objective, and the removal of a clear connection between score and rubric creates new challenges that have not been explored within the comparative judgment literature.

We are not arguing that comparative judgment is fundamentally flawed, but that its use requires further research to collate systematic evidence and a theoretical base to increase confidence in what could be a new way to assess high-stakes tests. Future work could conduct a comprehensive, systematic review of the evidence to explore to what extent the rationales for using comparative judgment have empirical support.

Notes

1. Adaptive versions of comparative judgment are also available, in which pairings for comparison are selected adaptively in order to reduce the number of judgments while maintaining the level of reliability obtained (Pollitt, 2012a). The process is otherwise similar. As adaptivity increases the number of close and thus difficult judgments, it may affect the quality of the judgments. This distinction is important but does not affect the substance of the points we raise in the rest of the paper and so will not be explored further here.
2. For this reason, examining boards in England use statistical information alongside judgment to assure the fair treatment of students (see, e.g. Ofqual, 2011).

3. A judge is considered to be a misfit when their data deviates significantly from the model as a whole, indicating they may have interpreted the construct differently (Pollitt, 2012a).

Disclosure statement

The first author was an employee of the AQA awarding body until December 2021.

Notes on contributors

Kate Tremain Kelly is a PhD student at the UCL Institute of Education, University College London, UK. Her research interests are in comparative judgment and examiner experiences of marking and grading.

Mary Richardson is Professor of Educational Assessment at UCL Institute of Education where she leads that MA Education in Assessment. Her research interests include the ethics of assessment and testing, developing philosophical approaches to assessments, children's rights in education and more recently, the use of AI technology in assessment practice.

Talia Isaacs is an Associate Professor of Applied Linguistics and TESOL and Programme Leader of the MA TESOL In-Service at the UCL Institute of Education, University College London. She is a member of the TOEFL Committee of Examiners and an expert member of the European Association for Language Testing and Assessment (EALTA). Her research interests lie in the development and validation of scoring systems, particularly for second language speaking.

ORCID

Kate Tremain Kelly http://orcid.org/0000-0001-8939-7144
Mary Richardson http://orcid.org/0000-0003-0526-7479
Talia Isaacs http://orcid.org/0000-0003-4302-3379

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*(3), 449–460. https://doi.org/10.1177/2F014662167800200319

Baird, J.-A., Cresswell, M., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education*, *15*(2), 213–229. https://doi.org/10.1080/026715200402506

Baird, J.-A., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: Valid but inexact* (Report No. RPA 05 JB RP 077). Assessment and Qualifications Alliance.

Bartholomew, S. R., Ruesch, E. Y., Hartell, E., & Strimel, G. J. (2020). Identifying design values across countries through adaptive comparative judgment. *International Journal of Technology and Design Education*, *30*(2), 321–347. https://doi.org/10.1007/s10798-019-09506-8

Benton, T., & Elliott, G. (2016). The reliability of setting grade boundaries using comparative judgement. *Research Papers in Education*, *31*(3), 352–376. https://doi.org/10.1080/02671522.2015.1027723

Benton, T., & Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters: A Cambridge Assessment Publication*, *26*, 22–28. https://www.cambridgeassessment.org.uk/Images/514987-is-comparative-judgement-just-a-quick-form-of-multiple-marking-.pdf

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. (2020). 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

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, *23*(3), 357–373. https://doi.org/10.1080/02671520701755440

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement*, *6*(2), 202–223. https://www.researchgate.net/publication/7941436_A_rank-ordering_method_for_equating_tests_by_expert_judgment

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–294). Qualifications and Curriculum Authority.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgement*. Cambridge Assessment.

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 43–58. https://doi.org/10.1080/0969594X.2017.1418734

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, *38*(2), 247–264. https://doi.org/10.1080/03057640802063486

Crisp, V. (2010a). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education*, *36*(1), 1–21. https://doi.org/10.1080/03054980903454181

Crisp, V. (2010b). Judging the grade: Exploring the judgement processes involved in examination grading decisions. *Evaluation & Research in Education*, *23*(1), 19–35. https://doi.org/10.1080/09500790903572925

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2019). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, *45*(3), 316–338. https://doi.org/10.3102/1076998619890589

Curcin, M., Howard, E., Sully, K., & Black, B. (2019). *Improving awarding: 2018/2019 pilots*. Ofqual.

Elliott, V. (2017). What does a good one look like? Marking A-Level English scripts in relation to others. *English in Education*, *51*(1), 58–75. https://doi.org/10.1111/eie.12133

Gill, T., & Bramley, T. (2013). How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles, Policy & Practice*, *20*(3), 308–324. https://doi.org/10.1080/0969594X.2013.779229

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–19. https://doi.org/10.1007/BF03216919

Holmes, S. D., He, Q., & Meadows, M. (2017). An investigation of construct relevant and irrelevant features of mathematics problem-solving questions using comparative judgement and Kelly's Repertory Grid. *Research in Mathematics Education*, *19*(2), 112–129. https://doi.org/10.1080/14794802.2017.1334576

Humphry, S. M., & McGrane, J. A. (2015). Equating a large-scale writing assessment using pairwise comparisons of performances. *The Australian Educational Researcher*, *42*(4), 443–460. https://doi.org/10.1007/s13384-014-0168-6

Johnson, V. R. (2015). *Policy, practice and assessment: Revealing the relationship between the GCSE English assessment and educational reproduction* [Unpublished doctoral dissertation]. University of Manchester.

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., & 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., Swan, M., & Pollitt, A. (2015). 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

Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, *22*(2), 135–155. https://doi.org/10.1007/s10798-011-9190-4

Laming, D. (1984). The relativity of 'absolute' judgements. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 152–183. https://doi.org/10.1111/j.2044-8317.1984.tb00798.x

Laming, D. (2004a). *Human judgment: The eye of the beholder*. Thomson Learning.

Laming, D. (2004b). Marking university examinations: Some lessons from psychophysics. *Psychology Learning and Teaching*, *3*(2), 89–96. https://doi.org/10.2304/plat.2003.3.2.89

McMahon, S., & Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 368–389. https://doi.org/10.1080/0969594X.2014.978839

Ofqual. (2011). *Maintaining standards in GCSEs and A levels in summer 2011*.

Ofqual. (2015). *A comparison of expected difficulty, actual difficulty and assessment of problem solving across GCSE Maths sample assessment materials*.

Pollitt, A. (2009, September 13-18). Abolishing marksism and rescuing validity [Paper presentation]. *35th Annual Conference of the International Association for Educational Assessment*, Brisbane, Australia. https://iaea.info/conference-proceedings/35th-annual-conference-2009/

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

Pollitt, A. (2012b). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x

Pollitt, A. (2015). *On 'reliability' bias in ACJ: Valid simulation of adaptive comparative judgement*. Cambridge Exam Research.

Pollitt, A., & Crisp, V. (2004, September 15-18). Could comparative judgements of script quality replace traditional marking and improve the validity of exam questions? [Paper presentation]. *British Educational Research Association Annual Conference*, Manchester, UK. https://www.cambridgeassessment.org.uk/images/109724-could-comparative-judgements-of-script-quality-replace-traditional-marking-and-improve-the-validity-of-exam-questions-.pdf

Pollitt, A., & Elliott, G. (2003). *Finding a proper role for human judgement in the examination system*. University of Cambridge Local Examinations Syndicate.

Pollitt, A., Elliott, G., & Ahmed, A. (2004, June). Let's stop marking exams [Paper presentation]. *International Association for Educational Assessment Conference*, Philadelphia, US: Cambridge Assessment. https://www.cambridgeassessment.org.uk/images/109719-let-s-stop-marking-exams.pdf.

Pollitt, A., & Murray, N. L. (1993). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 74–91). Cambridge University Press.

Raikes, N., Scorey, A., & Shiell, H. (2008, September 7-12). Grading examinations using expert judgements from a diverse pool of judges [Paper presentation]. *34th Annual Conference of the International Association for Educational Assessment*, Cambridge, UK; Cambridge Assessment. https://www.cambridgeassessment.org.uk/images/109766-grading-examinations-using-expert-judgements-from-a-diverse-pool-of-judges.pdf

Roose, I., Vantieghem, W., Van Damme, K., Lambert, P., Vanderlinde, R., & Van Avermaet, P. (2019). Measuring teachers' professional vision of inclusive classrooms through video-based comparative judgement. What does it mean to misfit? *International Journal of Educational Research*, *98*, 257–271. https://doi.org/10.1016/j.ijer.2019.09.004

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Suto, I., & Novaković, N. (2012). An exploration of the examination script features that most influence expert judgements in three methods of evaluating script quality. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 301–320. https://doi.org/10.1080/0969594X.2011.592971

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*(1), 16. https://doi.org/10.1186/s41239-016-0018-x

Thurstone, L. L. (1927a). Three psychophysical laws. *Psychological Review*, *34*(6), 424–432. https://psycnet.apa.org/doi/10.1037/h0073028

Thurstone, L. L. (1927b). Psychophysical Analysis. *The American Journal of Psychology*, *38*(3), 368–389. https://doi.org/10.2307/1415006

Thurstone, L. L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology*, *14*(3), 187–201. https://doi.org/10.1037/h0070025

Van Daal, T. (2020). *Making a choice is not easy?! Unravelling the task difficulty of comparative judgement to assess student work.* [Doctoral dissertation, University of Antwerp]. OSF. https://osf.io/7etq2/

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 & Practice*, *26*(1), 59–74. https://doi.org/10.1080/0969594X.2016.1253542

Van Daal, T., Lesterhuis, M., Coertjens, L., Van de Kamp, M.-T., Donche, V., & De Maeyer, S. (2017). The complexity of assessing student work using comparative judgment: The moderating role of decision accuracy. *Frontiers in Education*, *2*, 1–13. https://doi.org/10.3389/feduc.2017.00044

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2017). Scale Separation Reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*(6), 428–445. https://doi.org/10.1177/0146621617748321

Wheadon, C. (2015, September 22). *How many judgements do you need? No More Marking*. https://blog.nomoremarking.com/how-many-judgements-do-you-need-cf4822d3919f#.o3uxgi3hg

Wheadon, C., & Christodoulou, D. (2016, November 3-5). Improving moderation of teacher assessed work [Paper presentation]. *17th Annual Conference of the Association for Educational Assessment* – Europe, Limassol, Cyprus.

Whitehouse, C. (2012). *Testing the validity of judgements about geography essays using the Adaptive Comparative Judgement method* [Report No. CERP_RP_CW_24102012]. Centre for Education Research and Practice.

Whitehouse, C., & Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment* (Report No. CERP RP CW 2006201). Centre for Education Research and Practice.