Comparative judgment: An overview

Eva Hartell[1] and Jeffrey Buckley[1, 2]

[1]*KTH Royal Institute of Technology, Stockholm, Sweden*

[2]*Athlone Institute of Technology, Westmeath, Ireland*

Abstract

There is a growing demand for the use of digital tools in assessment. Few approaches show innovative benefits beyond being logistical aids. Comparative judgment (CJ) has the potential to enhance educational practices by providing a mechanism for reliable assessment, supporting formative feedback, and by supporting critical discourse on evidence of learning. This chapter provides an overview of CJ as it has been used in educational assessment and describes how it can be facilitated by digitalisation by providing illustrative examples of research studies, mainly undertaken for formative purposes. Specifically, this chapter provides an introduction to CJ and a description of its theoretical roots, presents possible approaches and agendas for the use of CJ ranging from being a pedagogical tool in a classroom to being a mediator for continuing professional development, and discusses implications for practice and future research needs. Ultimately, it is envisaged that this chapter will act as a source of inspiration for educational stakeholders who wish to use CJ to add value to their practice.

Key words: Comparative judgment; formative assessment; feedback; continuing professional development.

Introduction

The potential for the digitalisation of assessment is immense. However, most digital assessment tools rely on traditional means of assessing student achievement instead of adding value to student learning and/or reducing teacher workload. Comparative judgment (CJ) on the other hand has been proven to be a valid, reliable and efficient method of assessing open-ended tasks in a variety of subject areas (Bartholomew & Yoshikawa-Ruesch, 2018; Jones et al., 2015; Pollitt, 2012b, 2012a; Seery et al., 2012; The Royal Society, 2016) and offers significant formative opportunity (Bartholomew et al., 2019; Seery et al., 2019). In contrast to traditional criterion referenced assessment, the CJ process is premised on pairwise comparisons. Cohorts of assessors are individually presented with pairs of portfolios[1] of student work from which they must select the "better" of the two. By doing this, the question an assessor needs to ask themselves when evaluating student work is changed from asking what mark should be given relative to a criterion to which of the two portfolios in front of them provides more evidence of capability or learning. Based on research to date, which will be discussed in this chapter, it appears that this second question can be answered more reliably. In practice, the pairwise decisions made within the CJ process have generally been holistic, based on assessor expertise and prior experience, however it is feasible for external criteria to be provided to guide these judgements (e.g., Bartholomew, Strimel, & Jackson, 2018; Mortier et al., 2015). After a number of rounds of judgments, the result of this process is a rank order of the included portfolios, with relative distances (parameter values) between portfolios based on the assessors' judgements (Figure 1). Thus, the validity of the process is directly tied to the cohort of assessors (Lesterhuis, 2018). Critically, there are no absolute indicators of quality such as grades inherent within the rank. The highest-ranking portfolio may not necessarily be of high quality, and the lowest ranking portfolio may not necessarily be of low quality, they are just ranked as best and worst relative to all portfolios, which were included in the CJ session. The transposal of the rank to, for example, percentages or grades can be achieved after the CJ process through a variety of methods if desired based on the agenda of the assessment.

[1] The term "portfolio" will be used broadly throughout this chapter to describe all manners of student work (e.g., design outputs, essays, etc.) which could be included for assessment through CJ.

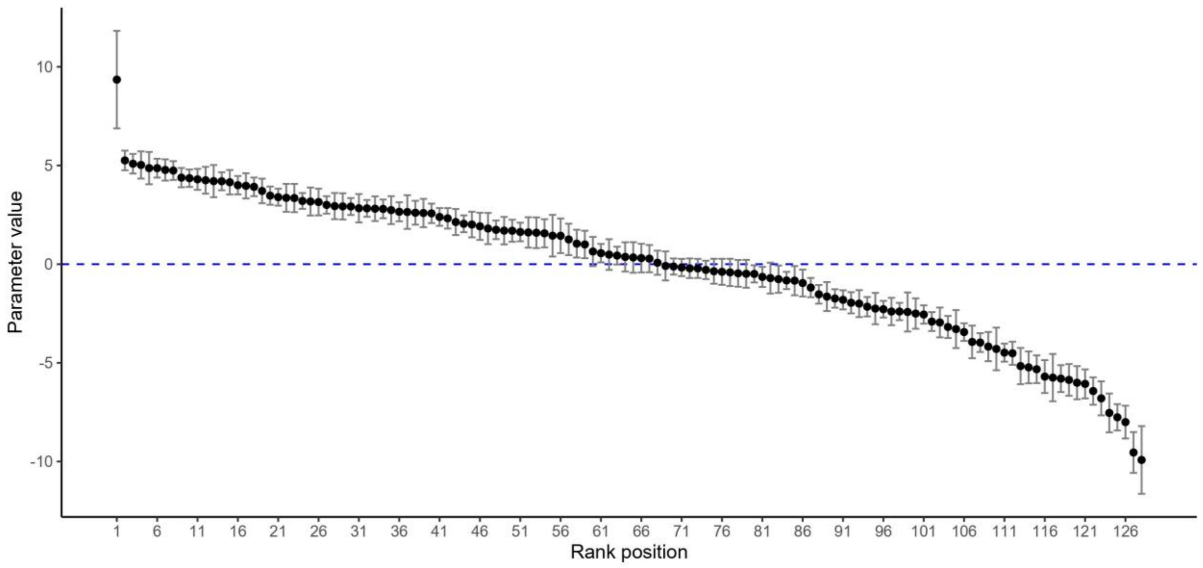


Figure 1. Output of a CJ session based on data from Seery and Buckley (2016). Each data point represents one portfolio. The x-axis describes rank position and the y-axis describes relative performance in terms of parameter values (z-scores). Error bars represent standard error.

An individual assessor can undertake this process of pairwise comparisons with the work of their own students independent of a digital solution. For example, a teacher could pick two random essays from their pile of student work, compare them and identify one as better, then repeat the process iteratively until a rank of quality from the pile is determined. Digital tools which have been developed to support the CJ process such as No More Marking (2020) and RM Compare (2020) do this by providing detailed reports of outcomes, managing the selection of portfolios to present for comparison, and by facilitating collaborative practices amongst teachers and schools which can occur both nationally and internationally (see Bartholomew et al., 2019, for an example of an international cohort of judges). A direct result of this digital facilitation is that CJ is most commonly undertaken by several assessors who independently complete the series of pairwise comparisons, which results in the outputted rank. According to Pollitt (2012b), this holistic approach embedded in CJ with multiple judges rules out personal biases, leading to higher consistency in judgement amongst the assessors.

CJ has emerged quite recently in various educational subjects, such as Modern Languages, Design and Technology, Music, Mathematics, and Geography, and at various levels of education ranging from primary level to higher education in different parts of the world. It was first used by Pollitt and Murray (1993) in the assessment of a foreign language speaking assessment. This was followed by further use by Pollitt (2012b) with English writing scripts and was adopted for use in other subject areas such as Design and Technology education by Technology Education Research Unit (TERU) at Goldsmiths, University of

London (Kimbell, 2012; Kimbell et al., 2009). While it was first used primarily for summative purposes, as it was integrated in more educational settings such as technology education (Bartholomew, Nadelson, et al., 2018; Hartell, 2018; Kimbell et al., 2005, p. 1, 2007, 2009; Seery et al., 2012; Seery & Canty, 2017; Stables & Lawler, 2012), mathematics (Jones & Inglis, 2015) and geography (Whitehouse, 2013; Whitehouse & Pollitt, 2012) there was a shift towards using CJ formatively. Since the pioneering work of Pollitt, CJ has also been further refined as a tool for assessing student writing (Coertjens et al., 2017; Daal et al., 2019; Jones & Wheadon, 2015; Lesterhuis et al., 2018; Steedle & Ferrara, 2016).

In providing an overview of the use of CJ in education, the following sections of this chapter will discuss the underpinning theory of CJ, present examples of its use in education, synthesise criteria for the successful incorporation of CJ for educational assessment, and discuss future possibilities for the use of CJ in terms of research and practice. The chapter will not discuss technical information or suggest particular software to use; instead, it provides illustrative examples to inspire readers to try to embed it in their own context. Most of the examples provided in this chapter are taken from the STEM context, in particular technology and engineering, due to the relatively high volume of CJ studies in these areas. However, it is important to note that CJ as a process is transferable to any context, at least in which the student work being assessed was generated in response to an open-ended task, and it has and continues to be used in other subject areas such as English, Mathematics and Geography.

A primer on the underpinning theory of comparative judgement

The pairwise comparison methodology inherent to CJ was first adapted in the 1920’s by the American psychologist Luis Leon Thurstone in his quest to find reliable measures of people’s attitudes about the seriousness of various crimes (Thurstone, 1927). Thurstone argued that people found it difficult to describe how serious a crime is, especially in absolute terms. Instead, he asked them to compare two crimes and then judge which one was more serious. From this work he formulated the *Law of Comparative Judgement*, which essentially says that people are more reliable when comparing two stimuli, such as two crimes, than when giving an absolute value to a stimulus. However, it is arguable that the history of CJ predates Thurstone, as making such pairwise comparisons is something that humankind does across many aspects of day-to-day life. For example, when posed with multiple options of what to eat for dinner, what clothes to wear, what to watch on TV, or when choosing a perfume they prefer, people are engaging with a decision making process similar to that which underpins the use of CJ for educational assessment. Indeed, Laming (2003) built on Thurstone’s work and concluded that all assessment is a comparison of one thing to something else, arguing that absolute judgment is not possible. This was corroborated by Gill and Bramley (2013), who found that assessors made more accurate judgments when making relative rather than absolute assessments and that assessors felt more confident using the comparative approach than assessing texts absolutely by using scores or rubrics.

Central to CJ is the idea that two stimuli (such as portfolios of student work) must cause a reaction in the observer. These differences in reactions to the two stimuli are called *just-noticeable differences* (Thurstone, 1927). From these, the observer formulates a judgment about the relationship between the two stimuli, such as the seriousness of two crimes or the best perfume scent. To provide an example of this process, Figure 2 shows an example of two pieces of student work, in this case pictures of sunflowers, from an assessors perspective in the RM Compare CJ digital solution. As an assessor, if the assessment was about which piece of artwork shows greater evidence of capability, you would be tasked with reflecting on the just-noticeable differences you experience between the two in order to make your decision. Once you have done, you would select each option A or B as the winner, and be presented with a further two pieces of work to compare. Importantly, in the arts there are personal preferences whereas in art education there are certain concepts that needs to be taught and practiced. Therefore, judgements and comments relating to the student work in Figure 2 cannot just be personal opinions. Instead they should be tied to the context and circumstances in which these sunflowers are undertaken.

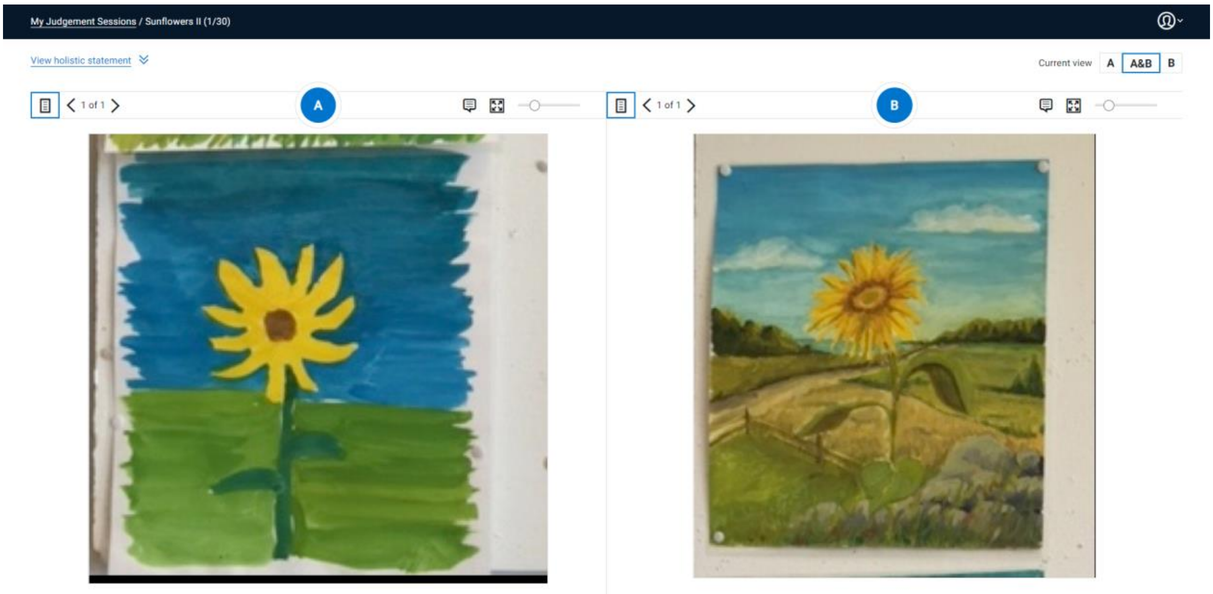


Figure 2. Example of two pieces of student work presented for a pairwise comparison in the RM Compare digital CJ solution.

Beyond the judgement of a piece of student work as better or worse, a second important aspect of CJ is the approach to selecting which two pieces of work to present to an assessor at a given time. This is generally managed by a digital solution. A central mechanism for this is the *Swiss tournament* approach. In this, the first round involves portfolios being paired randomly, with the results of each judgement being that one portfolio wins and the other loses. Following this, in a Swiss tournament pairing is conducted by selecting two portfolios with the same number of wins to be compared. Here, an important distinction must be made between CJ and adaptive comparative judgement (ACJ). Both are founded on the same principles, but in ACJ, the adaptive element relates to an algorithm (generally propriety so not disclosed in peer reviewed outputs) that presents two pieces of work to an assessor based on outcomes of the judgements made so far. In ACJ, after a number of rounds following the Swiss tournament method, the new adaptive algorithm is used to select the portfolios to present to assessors. In other words, the experience of the assessor is common in both CJ and ACJ, but ACJ was designed to be more efficient by reducing the number of judgements to be made by not presenting portfolios where the outcome if almost certain (Bramley & Wheadon, 2015; Pollitt, 2012b).

Reviewing the use of comparative judgment in education

Prior to exploring examples of the use of CJ methods in education, it is worth noting two often cited benefits of CJ over traditional methods of assessment. This first comes in response to the difficulties that emerge when trying to achieve high levels of reliability in the assessment of student work produced in response to an open ended task. In practice, many teachers start this process by roughly sorting student portfolios in order to get a sense of levels and indicators of quality. This process may be undertaken more or less tacitly and more or less systematically. Whilst teachers can often identify levels and indicators of quality in student work, how reliable their judgements are with respect to assigning grades relative to these indicators in comparison to the judgements other teachers would make needs to be considered. By having cohorts of teachers' act as assessors, the CJ process has repeatedly seen high levels of reliability, usually with cohort agreement levels being greater than 90% (see Bartholomew & Yoshikawa-Ruesch, 2018, for a summary). Indeed, a repeatedly purported strength of CJ is that it rules out personal standards and biases due to the involvement of several assessors. Where one assessor may hold certain views on what denotes capability, another may hold a slightly different construct of capability. Thinking back to the sunflowers presented in Figure 2, one person might suggest flower B as better evidence of capability due to the increased level of detail where another could appreciate the minimalist rendering of flower A. The final outcome (the relative rank order of portfolios) when using CJ is the result of all the judgments made by all the assessors involved, not just one assessor, as would usually be the case in traditional criterion referenced assessment. Thus, the outputted relative rank order of portfolios represents a shared consensus of the particular competence of all the assessors where individual biases are mitigated by the inclusion of multiple perspectives of quality.

Aside for reliability, a second suggested benefit of CJ relates to time. Some studies suggest CJ can take less time than traditional grading to assess student work (Newhouse, 2014; Steedle & Ferrara, 2016). However, evidence associated with time implications is mixed with others suggesting traditional grading takes less time (Bartholomew, Strimel, & Zhang, 2018) and others suggesting comparable time requirements for both approaches (Coertjens et al., 2017). In this regard, CJ is apparently not a silver bullet; it needs both practice and consideration in terms of the time it take to set up a judging sessions, assessment times are likely to be significantly impacted by other variables such as the complexity of the assessment rubric in traditional assessment, assessor expertise, and the nature of the student work, and the time commitments of all assessors needs to be considered (Buckley et al., 2020).

Illustrative example of a comparison between CJ and criterion referenced assessment

Even if the evidence is there, the complexity of embedding new models of assessment must be based on several factors, including compatibility with existing methods based on their relative advantages and utility. A group of researchers at Purdue University undertook an exploratory study to examine ACJ in comparison with traditional assessment methods in terms of validity, reliability and utility in the context of engineering education (Bartholomew, Strimel, & Jackson, 2018). In their study, a group of 16 undergraduate engineering students completed an engineering design challenge in response to an open-ended brief. Their work was then assessed by the course instructor using a traditional rubric, and a group of five independent experts (experienced in teaching engineering design) using ACJ. The assessors who used ACJ were asked to make holistic decisions based on their own professional opinion, but were aware of the criteria specified in the traditional rubric.

A very high level of reliability ($\alpha = .95$) was observed from the ACJ process and the researchers found strong a correlation ($\rho = -.79$, $p < .01$) between the grades awarded via the traditional rubric assessment method and the ACJ rank. Note that the correlation was negative as lower values in the rank indicate better performance, i.e., 1st place versus 2nd place. This was interpreted to suggest that ACJ is a valid, reliable and comparable tool to traditional assessment methods. Importantly, by using the correlation between ACJ and traditional assessment as evidence, this approach to interpreting ACJ being valid is based on the assumption that the traditional assessment method it is being compared to is also valid, and a limitation exists in this study in the researchers using the ACJ rank order in their analyses instead of the parameter values which denote relative performance between portfolios. While the parameter values and rank order will be very strongly correlated, as the rank is a direct product of the parameter values, by not reflecting relative performance the rank only offers a simplified understanding of the outcomes. Interestingly in this study, neither the results of ACJ nor the traditional assessment were significantly correlated with the actual performance of the student's design. Therefore, the authors noted that it might be time to question the current methods of assessing process instead of the actual performance of the final product/prototype.

Two further studies have compared CJ with traditional assessment practices. Jones and Inglis (2015), where secondary school level mathematics problems were the subject of assessment, found a strong correlation (*r* = .89) between CJ parameter values and predicted General Certificate of Secondary Education (GCSE) grades in the UK. They also found that when CJ was the intended mode of assessment, examination scriptwriters designed math assessments to be less structured and more problem-based than was typical of standard assessment papers. Coertjens et al. (2017) also conducted a study whereby they compared the use of CJ to traditional rubrics, however they looked more specifically at the time taken by assessors in each method. The participants in their study were in high school, and the assessments were conducted on their responses to a writing task. The researchers found that the time taken by assessors decreased as they made more pairwise comparisons or as they gained experience using the rubric, but added another important caveat to consider; it is important to see if conclusions from comparative studies such as these are transferable to other types of student work and when different rubrics are used.

Illustrative example focusing on unpacking learning intentions and criteria for success

Every student benefits when taught by teachers who are transparent about learning intentions and criteria for success. This practice will benefit all students, especially low achievers (Jönsson, 2010). An awareness of the criteria for success instils a sense of security among students which is beneficial to learning (Bandura, 1997). However, students' perception of learning intentions may not match teachers' expectations. Further, as professionals' teachers should strive to achieve consensus in their interpretations of learning and competency. Harrison (2009) stresses the need and importance for teachers to both plan and share assessment procedures with other professionals with Pettersson (2009) stressing this even more and warning that teachers are in risk of becoming misaligned with current regulations if they do not have access to professional discussions. CJ offers a potential solution to act as a mediator for these discussions by using pairwise comparisons as a stimulus to support teachers in articulating their thoughts. A clear example of how this can be achieved is demonstrable through a study conducted by Hartell and Skogh (2015).

Using CJ, Hartell and Skogh (2015) undertook a study in a Swedish primary school context with the purpose of understanding what teachers value as criteria for success. Hartell (2013) had previously found that teachers gathered their evidence of learning during classroom activities, which is aligned with previous research in technology education (Bjurulf, 2008; Kimbell, 2007). In their study, teachers were asked to assess students' work under authentic classroom conditions. Twenty-one year 5 pupils (average age ≈ 11) were tasked with designing and building a model of a robot friend which was capable of helping them in the home with particular actions. The pupils developed multimodal portfolios using iPads to capture evidence of their learning. This evidence consisted of voice recordings, sketches, videos, written text, mind maps and technical drawings, and was consolidated within the e-scape CJ software (Figure 3).

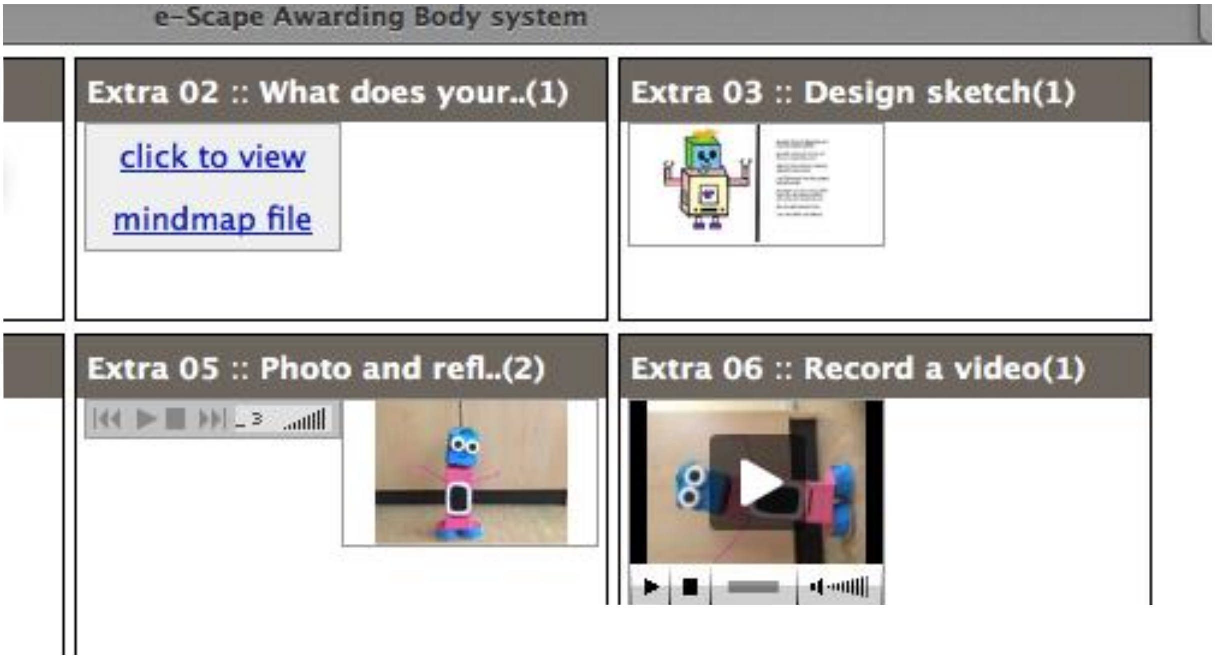
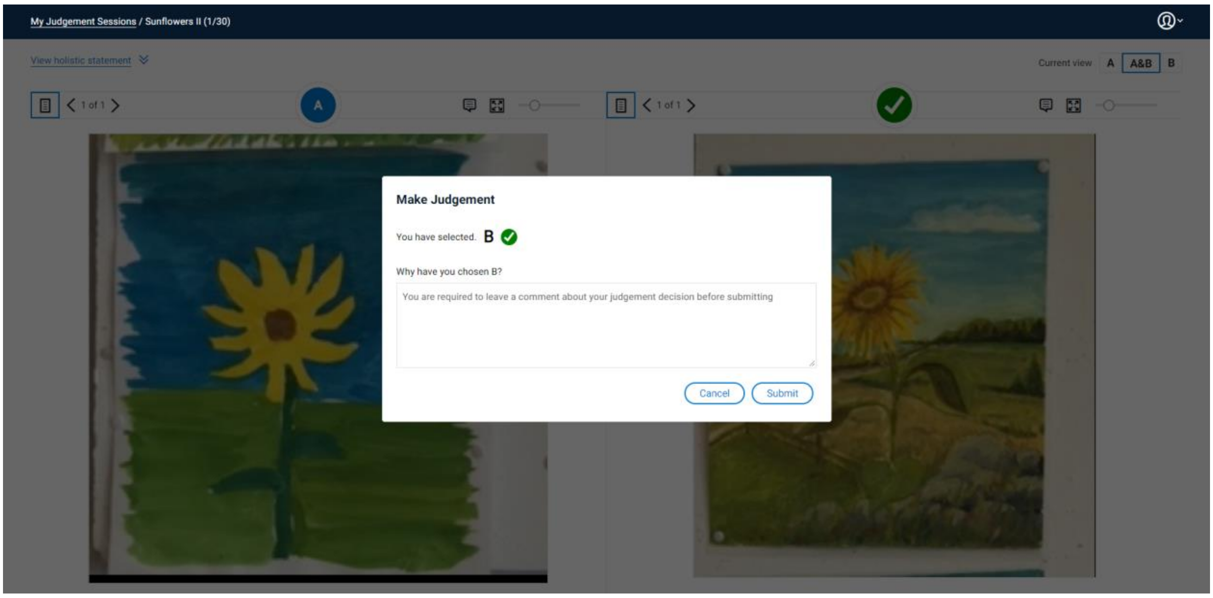

e-Scape Awarding Body system
Extra 02 :: What does your..(1)
click to view
mindmap file
Extra 03 :: Design sketch(1)
Extra 05 :: Photo and refl..(2)
Extra 06 :: Record a video(1)

Figure 3. Illustrative example of a student's online portfolio in e-scape.

Five teachers then assessed these portfolios using CJ, resulting in a rank order with very high reliability (α = .93). While they were judging the students' work they were asked to verbally provide reasons for the judgements they made through a think-aloud protocol. The analysis showed that these teachers all agreed on the importance of the narrative of the design process. They also questioned whether this had been communicated to all the students or whether some students had figured it out for themselves. The study also concluded that assessors value students' finishing their task, primarily to provide the narrative in the portfolios and in addition they wanted students to find value in finishing what they had set out to do, thereby emphasising the importance of providing sufficient time and instruction.

This example shows how CJ, by requiring assessors to articulate indicators of quality using pairwise comparisons as a stimulus, can be used to unpack what teachers' value as criteria for success and then create a basis for professional collective discussions. A more recent study had a similar aim but for undergraduate students in a teacher education degree programme. Buckley et al. (2020) conducted a study whereby the students completed an open-ended design task requiring them to design and make a flower which conveyed an emotion but which had no face, and an accompanying pictorial scene. No assessment criteria were provided, instead the students themselves acted as assessors in an ACJ session once they had all completed the task. The ACJ software solution used had the capacity for students to leave text based commentary after each judgment explaining the reason behind their decision (see Figure 4 for an example of this functionality). From this, it was possible for the criteria being used to make decisions, and thus the features of the students' portfolios, which denoted quality and evidence of learning, to be identified.



Make Judgement

You have selected. B ✓

Why have you chosen B?

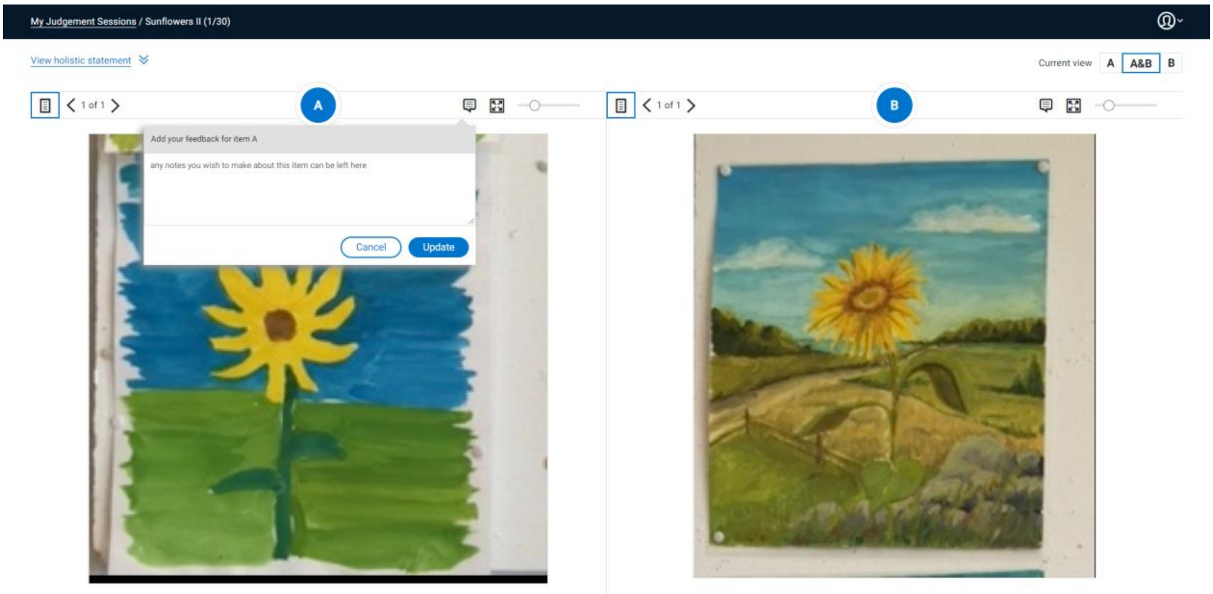
You are required to leave a comment about your judgement decision before submitting

Figure 4. Example of leaving of comment after making a pairwise comparison in RM Compare.

This in itself offers a further layer of understanding of students learning beyond what can be seen in the outputs from a task. From a pedagogical perspective such an activity could be used to stimulate discussion between teachers and/or students on what should evidence of learning look like within a task, or across a topic/unit of learning.

Illustrative examples of feedback facilitated by comparative judgment

Feedback is put forward as critical in education. Unfortunately, feedback is often misinterpreted as any kind of information provided to learners. There is a strong body of evidence showing that feedback can cater to student learning but can also hamper learning when focused on personal traits instead of process and effort, causing lower self-esteem instead of promoting learning (Wiliam, 2006). CJ does not provide feedback to learners per se. Knowing one's place in the ranking is not particularly helpful, especially at the end of learning. As noted, some software solutions enable writing comments while judging which can be fed back to learners as feedback (see Figure 5 for an example of this functionality). This section provides three examples of how CJ can be used as a feedback mechanism either through the provision of such comments, or through simply engaging with the act of making comparative judgements on the work of peers.



Add your feedback for item A

any notes you wish to make about this item can be left here

Figure 5. Example of leaving of comment on a piece of student work as feedback while making a pairwise comparison in RM Compare.

In 2015, a group of Belgian researchers examined CJ as an alternative method for peer assessment of competences in the context of argumentative writing (Mortier et al., 2015). They focused their research on students' attitudes towards feedback provided, which included their position on the rank relative to their classmates and to an expertly derived benchmark. Specifically, the students' comments on the perceived honesty, relevance and trustworthiness of the feedback as well as the importance of specific advice received from the CJ-based feedback. The researchers found that students did find the feedback to be reliable, relevant and honest. The students appreciated personalised tips on how to improve being included in the feedback report along with the quantitative measure showing how they had performed relative to peers and the benchmark. The authors concluded that "CJ-based feedback is a potential fruitful way to ameliorate students learning" (p. 79), and were hopeful that their work would encourage further investigation into the potential of CJ as a feedback tool.

The work of Mortier et al. (2015) did indeed inspire further research on CJ and feedback. A recent study by Seery et al. (2019) explored whether the act of making comparative judgements on peers work alone could act as a useful feedback mechanism. In their study, a group of 136 undergraduate students in a design and communication graphics module completed four consecutive graphics design tasks. Each task was followed by an ACJ session, where the students acted as the judges of the outputs produced in response to the task. In other words, the students completed a design task, assessed it through ACJ, and then begun the next design task, in a process that repeated for four design tasks. Performance in the first task was used as a benchmark, and students were grouped into quartiles based on this performance. The only form of feedback received by students was their exposure to the work of their peers by having to make pairwise comparisons through the ACJ process. The results of this study were that the poorest performing students (quartile 1) in the initial task saw a mean increase of 41% between the first and fourth tasks, the students in the second quartile saw a mean increase of 28% between the first and fourth tasks, the students in the third quartile saw a mean increase of 19% between the first and fourth tasks, and the top performing students initially (quartile 4) actually saw a mean decrease of 1% between these tasks. The researchers interpreted these results as poorer performing students initially having more room for improvement over time relative to initially higher performing students due to a potential ceiling effect, and through the ACJ process, poorer performing students were likely to be exposed to work of a higher standard and thus were receiving better quality feedback in comparison to students who were initially highest performing who were more likely to be comparing work of a poorer standard to their own.

A similar investigation into the use of CJ as an assessment tool was undertaken by a group of researchers at Purdue University (Bartholomew et al., 2019). In their study, four class groups of middle school students (12-13 years old) in the US engaged with a learning activity requiring them to research, design and produce a travel brochure for a location of their choice in Southeast Asia. A total of 10 class periods were afforded for this. After five class periods, the research team divided the students into two groups. Two class groups became the control group who printed their draft brochures and engaged in a peer feedback session. The other two class groups became the experimental group who, at this midpoint in their assignment, engaged with an ACJ session. In addition to making pairwise comparisons, students in the experimental group using ACJ were also asked provide feedback on how the work could be improved in the comment sections for each portfolio they were judging, an act which itself has evidence indicating it is a valuable learning activity (McConlogue, 2015; van Popta et al., 2017). After receiving their feedback, all students were asked to continue and complete their assignments. Once all of the design tasks were completed, the researchers consolidated the work of the experimental and control groups and an ACJ session was performed with all portfolios with the students and their teacher acting as judges. The result of this final ACJ session was a highly reliable rank (α = .96). The researchers then compared the average position on the ACJ rank between the experimental group and control group and found a statistically significant difference indicating that the experimental group outperformed the control group at the end. While this study does have limitations such as not controlling for baseline competency and comparing rank position rather than parameter values, the results do suggest potential for ACJ as a feedback mechanism, possibly due to the added value students received from exposure to and having to make pairwise comparisons on the work of peers which is in addition to giving and receiving written or oral feedback.

In these examples, it is interesting to see the combination of CJ and the provision of written feedback, as beyond traditional feedback mechanisms this includes student exposure to a wide range of work, having to choose the better of two portfolios and then being forced to comment on why they make particular judgments. Benefits from this may be the result of the *worked example* effect (Sweller, 2006), wherein the students are exposed to other students' work and they can then base their own work on what to do and what not to do. This, in combination with providing and receiving peer feedback in an anonymous environment should be investigated further. Even though CJ has limitations (in particular technical and cost implications) there appears significant potential to support learning. As noted by Bartholomew et al. (2018, p. 381), "ACJ may be a potent tool for solidifying student perceptions of quality".

As a final comment on commentary feedback in these studies, it is important to remember that feedback can be shallow and not focused or aligned with the task or learning outcomes. The students in these examples were not trained in feedback or in the use of CJ, and therefore future related research should consider the quality of feedback provided so as to identify more clearly what benefits, if any, stem from the CJ process.

Assessors in comparative judgement as the most important component for successful implementation in the classroom

Undoubtedly, CJ offers a solution for very reliable assessment. It also enables remote assessment so the cohort of assessors is not geographically restricted and it is possible that engaging students as judges in CJ sessions can have positive educational implications. Implementing CJ, insofar as setting up and running a judging session, is also not difficult. The question becomes what makes a CJ session valuable, and therefore the "successful" implementation of CJ would be characterised differently depending on whether the agenda was summative or formative. Arguably, in either case the primary consideration is who is involved in making the judgements. If the goal is to expose students to the work of their peers as in Seery et al. (2019), then the students themselves can act as judges. If the goal is to describe quality or provide feedback from the perspective of experts as in Mortier et al. (2015), the design of the assessor cohort becomes more complicated.

By way of example, a Swedish–American team of researchers, Bartholomew et al. (2020), asked groups of judges from Sweden (n = 9), Ireland and the UK (n = 7), and the US (n = 5) to identify criteria for success by using a design similar to that used in Hartell and Skogh (2015). Judges were asked to compare and assess student work through ACJ and provide comments on why they chose one output over another. The judges were assessing 175 design portfolios (focussing on the design process) and 175 products (as a result of this design process) from 760 American secondary students' (ages 13–14). The students had worked in small groups on designing a travel-friendly pill dispenser. Each group of judges engaged with two ACJ sessions, one for the students' design portfolios and one for the products, meaning a total of six ACJ sessions were conducted. Each session had a high level of reliability (α > .95) indicating that within each group there was a high level of agreement. However, there were clearly differences between groups in what they valued as criteria for success.

Only one of the students' prototypes was in the top ten of the ranks from each cohort of judges (three prototypes were in two of the top ten ranks), and similarly only one of the design portfolios was in the top ten from each rank (two portfolios were in two of the top ten ranks). The judges' comments were analysed qualitatively to elicit valuable insights into cultural differences. Where the Americans and the Swedes focused on usability, size and design, the UK and Irish judges also declared innovation as important. The Swedes emphasised communication; the judges wanted to see whether the student could communicate the process, results, and conclusions, i.e. the narrative. Judges from the UK and Irish group focused on the process; how developed the portfolios were or how well they demonstrated progress in design. The comments from the US-based group revealed their focus on students fulfilling the task, following the criteria and constraints; that is, how well they had completed their portfolios. From one perspective, this study emphasises one of the strengths of CJ. People will hold different perspectives on evidence of learning and CJ offers the potential to design an assessment cohort consisting of people who hold such different perspectives. On the other hand, it needs to be considered what perspectives are valid. For example, narrative is emphasised in the Swedish secondary level curriculum for technology education. This was not commented on as much by the US, UK or Irish judges. If this had been a summative task for Swedish students, the resulting ranks from the US, and UK and Irish judges could be misaligned with curricular intent. If this had been a formative task for Swedish students, the value that could be gained from feedback coming from less-informed assessors needs to be taken into account. Therefore, and as noted earlier, while the actual implementation of CJ in a classroom is a concern of relatively basic IT competency which could be achieve through training and an associated cost if a digital solution is desired, the successful implementation with respect to a valid summative output or relevant formative feedback will be very much dependent on the cohort of assessors.

Discussion

Implications for practice

Research can never foresee what will happen in future practices; however, research may inform future educational practices to better meet learners needs. While there is not a very extensive body of published research on CJ, the evidence that does exist suggests significant positive educational potential. CJ provides a highly reliable form of assessment with formative opportunity. Beyond its immediate use in the classroom, there is potential to adopt CJ in continuing professional development (CPD) contexts. As demonstrated by Hartell and Skogh (2015), CJ can be used by cohorts of teachers as a mechanism to encourage discussion about what constitutes evidence of learning which would address the concerns raised by Pettersson (2009). In line with the validation studies of CJ conducted by Bartholomew et al. (2018), Jones and Inglis (2015), Lesterhuis, (2018) and Coertjens et al. (2017), CJ could be used in CPD to discuss the validity of traditional rubrics, or it could be used in the design of rubrics for national assessment. There is also potential to establish local or national clusters of schools whereby student work is submitted, for example as part of a national assessment, and teachers from that cluster act as judges. At a national level this would permit teachers to assess students work reliably and equitably while at the same time maintaining anonymity through a shared consensus if such concerns existed.

In addition to the features of CJ discussed thus far, one more aspect which is particularly important for educational practice is the experience of the students and teachers who have engaged with it. In the study conducted by Hartell and Skogh (2015), teachers were interviewed about how they experienced assessing students' work through CJ. The teachers unanimously reported that they enjoyed the overall experience, especially the satisfaction of seeing the work of students other than their own. In his PhD thesis, Canty (2012) examined undergraduate students perceptions of using CJ over three years, finding a generally positive disposition towards CJ as it encouraged positive competition and the sharing of ideas. Finally, Seery et al. (2019) found a similar attitude from undergraduate students who saw positives in making pairwise comparisons as they could learn more from critically examining mistakes made by peers and as this caused self-reflection on limitations in their own work.

Future research

As it is relatively new, CJ is still quite unknown in general. It is however gaining interest across the world and has been put forward as a possible means of assessing students' learning in the future. In October 2016, the Royal Society invited a group of experts in educational assessment to discuss the future of assessment in science education, especially experimental science in years 11–18 of education. A selected group of international experts made presentations, and then a group discussion was held with the international experts who were invited as delegates. These presentations and discussions were summarised in the report from the event (The Royal Society, 2016), which suggested:

> Future research might look at how students should learn science and the skills this entails; the validity of teacher assessment (including the need to increase confidence in this by mapping and developing teacher assessment competences and the use of comparative judgment); and the integration of summative and formative assessment (p. 3).

The fact that they specifically named and demonstrated clear support for CJ does not come as a surprise to those of us who participated in the seminar. One of us (Hartell) was present at the meeting and can report that six of the seven groups of experts suggested that CJ was the future of assessment in experimental science education.

Importantly, CJ has utility in any area where the objects of assessment are responses to open-ended tasks. However, from a research perspective there are a number of unanswered questions which directly affect its education use. With respect to using CJ in a classroom, there is a need for further dedicated research into attitudes towards CJ from all involved stakeholders. Such research could reveal further important research questions, and positive dispositions towards CJ would be necessary for its broad uptake. Further, as CJ from a user perspective essentially has two elements, the pairwise comparisons and providing written commentary. As noted by Mortier et al. (2015) there has been insufficient research on whether argumentation through writing comments effects decision making in the comparison stage. This relates both to subject expert and student judges. While the act of providing feedback to peers can have positive benefits for students (McConlogue, 2015; van Popta et al., 2017), in situations where they could be acting as judges it may affect the resulting rank. There may not be the same agenda for expert assessors, but if their function is to provide a valid rank the circumstances which should be in place to enable this merit inquiry. A greater understanding of this would be needed to guide the valid use of CJ. Beyond research on the use of CJ in a classroom, there are many educational research agendas which could be supported by CJ such as:

- Understanding the “why” behind the judges’ choices for different tasks in different subject areas, which could aid in meaningful task design.
- Exploring CJ as a tool for building assessment literacy and self-efficacy in teachers and students.
- Using CJ as a method for investigating the formative effect of critically evaluating the work of peers based on the worked example effect (Sweller, 2006).

Perhaps what is most important to discussions concerning the future use of CJ and associated research is clarifying, from the perspectives of stakeholders, current assessment needs and identifying if and how CJ could help. Without the input of those would be benefit from the use of CJ, its development as a method may not serve its potential and intended purposes.

Concluding remarks

Perhaps the foremost value with CJ is its capacity to serve as a catalyst for discussion amongst stakeholders including teachers and students as well as curriculum designers and teacher educators. Similar to how wine connoisseurs taste and discuss wine in communities of practice, the potential of CJ to foster teachers’ assessment literacy and self-efficacy is immense and yet to be fully explored. CJ is a useful tool to unpack teachers’ assessment practices, to uncover epistemological values and constructs, and to explicate criteria for success in a much deeper way. Above all, CJ has great potential as a way to invite learners into the mystery of learning.

With all that said, it can be easy and tempting for people to get prematurely excited about new approaches or technologies especially when arguments have empirical support. In the case of CJ, while it is certainly promising and there are evidence based examples of some contexts in which it can be very useful (design tasks, essays and problem-based mathematics), there is a need for further research to establish a clearer remit for its potential use, taking subject area, school level, and learner expertise into account. While it could be used in the assessment of any open ended task, before widespread adoption it needs to be questioned what context specific added value is gained from using CJ over traditional assessment so as to make informed decisions around implementation. Further, even though there are multiple applications for CJ, appropriate use should always be kept in mind and this extends from task design to assessment. Learning outcomes must be designed and depending on these, teachers must choose appropriate tasks and exemplars both to ensure a meaningful learning experience and a useful assessment. For example, reflecting back on the sunflowers in Figure 2, an understanding of whether the task related to abstract art or if the task to was produce a still life painting is necessary in order to make judgements on capability and to provide appropriate feedback. It needs to be considered what type of evidence students should collect and present for assessment and who should act as judges or assessors. There are also ethical issues to be taken into account, for example who owns the copyright to the students' work, does the CJ software provider work in a GDPR safe environment, as a teacher is support needed to help interpret the data, and what is the data going to be used for. Even if the use of digital CJ is not possible in classrooms, it is hoped that teachers and CPD organisers will make use of pairwise comparisons as a pedagogical strategy to facilitate critical discussion which can support equity for students, and that this chapter provides a source of inspiration for further innovations linked to the use of CJ in education.

References

Bandura, A. (1997). *Self-efficacy. The exercise of control* (13th ed.). W. H. Freeman and Company.

Bartholomew, S., Nadelson, L., Goodridge, W., & Reeve, E. (2018). Adaptive comparative judgment as a tool for assessing open-ended design problems and model eliciting activities. *Educational Assessment*, *23*(2). https://doi.org/10.1080/10627197.2018.1444986

Bartholomew, S., Strimel, G., & Jackson, A. (2018). A comparison of traditional and adaptive comparative judgment assessment techniques for freshmen engineering design projects. *International Journal of Engineering Education*, *34*(1), 20–33.

Bartholomew, S., Strimel, G., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education*, *29*(2), 363–385. https://doi.org/10.1007/s10798-018-9442-7

Bartholomew, S., Strimel, G., & Zhang, L. (2018). Examining the potential of adaptive comparative judgment for elementary STEM design assessment. *The Journal of Technology Studies*, *44*(2), 58–75. https://doi.org/10.2307/26730731

Bartholomew, S., Yoshikawa, E., Hartell, E., & Strimel, G. (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

Bartholomew, S., & Yoshikawa-Ruesch, E. (2018). A systematic review of research around adaptive comparative judgement (ACJ) in K-16 education. In J. Wells (Ed.), *CTETE - Research Monograph Series* (Vol. 1, pp. 6–28). Council on Technology and Engineering Teacher Education.

Bjurulf, V. (2008). *Teknikämnets gestaltningar: En studie av lärares arbete med skolämnet teknik* [Doctoral thesis, Karlstad University]. http://kau.diva-portal.org/smash/record.jsf?pid=diva2%3A25379&dswid=6953

Bramley, T., & Wheadon, C. (2015). The reliability of Adaptive Comparative Judgment. *AEA-Europe Annual Conference, March*, 7–9.

Buckley, J., Canty, D., & Seery, N. (2020). An exploration into the criteria used in assessing design activities with adaptive comparative judgment in technology education. *Irish Educational Studies*. https://doi.org/10.1080/03323315.2020.1814838.

Canty, D. (2012). *The impact of holistic assessment using adaptive comparative judgement on student learning* [PhD Thesis]. University of Limerick.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., & De Maeyer, S. (2017). Judging texts with rubrics and comparative judgement: Taking into account reliability and time investment. *Pedagogische Studien*, *94*(4), 283–303.

Daal, T. van, Lesterhuis, M., Coertjens, L., Donche, V., & Maeyer, S. D. (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

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

Harrison, C. (2009). Assessment for learning – A formative approach to classroom practice. In A. Jones & M. de Vries (Eds.), *International Handbook of Research and Development in Technology Education* (pp. 449–459). Brill.

Hartell, E. (2013). Looking for a glimpse in the eye: A descriptive study of teachers' work with assessment in technology education. In I.-B. Skogh & M. de Vries (Eds.), *Technology Teachers as Researchers: Philosophical and Empirical Technology Education Studies in the Swedish TUFF Research School* (pp. 255–283). Sense.

Hartell, E. (2018). Teachers' self-efficacy in assessment in technology education. In M. de Vries (Ed.), *Handbook of Technology Education* (pp. 1–16). Springer International Publishing.

Hartell, E., & Skogh, I.-B. (2015). Criteria for success: A study of primary technology teachers' assessment of digital portfolios. *Australasian Journal of Technology Education*, *2*(1), 1–17. https://doi.org/10.15663/ajte.v2i1.27

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

Jönsson, A. (2010). *Lärande Bedömning*. Gleerups.

Kimbell, R. (2007). Assessment. In M. de Vries, R. Custer, J. Dakers, & G. Martin (Eds.), *Analyzing Best Practices in Technology Education* (pp. 247–258). Brill.

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

Kimbell, R., Martin, G., Wharfe, W., Wheeler, T., Perry, D., Miller, S., Shepard, T., Hall, P., & Potter, J. (2005). *E-scape portfolio assessment: Phase 1 report*. Goldsmiths, University of London. http://research.gold.ac.uk/1527/

Kimbell, R., Wheeler, T., Miller, S., & Pollitt, A. (2007). *E-scape portfolio assessment: Phase 2 report*. Goldsmiths, University of London. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606018/0107_RichardKimball_et_al_e-scape2report.pdf

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., Pollitt, A., & Whitehouse, G. (2009). *E-scape portfolio assessment: Phase 3 report*. Goldsmiths, University of London.

Laming, D. (2003). *Human judgment: The eye of the beholder*. Cengage Learning.

Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality: An assessors' perspective* [PhD Thesis]. University of Antwerp.

Lesterhuis, M., Van Daal, T., Van Gasse, R., Coertjens, L., Donche, V., & De Maeyer, S. (2018). When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality. *L1 Educational Studies in Language and Literature*.

McConlogue, T. (2015). Making judgements: Investigating the process of composing and receiving peer feedback. *Studies in Higher Education*, *40*(9), 1495–1506. https://doi.org/10.1080/03075079.2013.868878

Mortier, A., Lesterhuis, M., Vlerick, P., & De Maeyer, S. (2015). Comparative judgment within online assessment: Exploring students feedback reactions. In E. Ras & D. Joosten-ten Brinke (Eds.), *Computer Assisted Assessment. Research into E-Assessment* (pp. 69–79). Springer International Publishing.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy and Practice, 21*(2), 205–220. https://doi.org/10.1080/0969594X.2013.868341

No More Marking. (2020). *No More Marking*. https://www.nomoremarking.com/

Pettersson, A. (2009). Bedömning- varför, vad och varthän? In L. Lindström & V. Lindberg (Eds.), *Pedagogisk bedömning* (2nd ed., pp. 31–42). Stockholm universitets förlag.

Pollitt, A. (2012a). 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. (2012b). 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., & Murray, N. (1993). What raters really pay attention to. *Languages Testing Research Colloquium*, 1–7.

RM Compare. (2020). *RM Compare*. RM Compare. https://rmresults.com/digital-assessment-solutions/rmcompare

Seery, N., & Buckley, J. (2016). The validity and reliability of adaptive comparative judgements in the assessment of graphical capability. In J. Birchman (Ed.), *ASEE Engineering Design Graphics Division 71st Mid-Year Conference* (pp. 104–109). EDGD. https://edgd.asee.org/71st-midyear-meeting-proceedings/%0A

Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2019). Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels. *International Journal of Technology and Design Education*, *29*(4), 701–715. https://doi.org/10.1007/s10798-018-9468-x

Seery, N., & Canty, D. (2017). Assessment and learning: The proximal and distal effects of comparative judgment. In M. de Vries (Ed.), *Handbook of Technology Education* (pp. 1–14). Springer.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education, 22*(2), 205–226. https://doi.org/10.1007/s10798-011-9194-0

Stables, K., & Lawler, T. (2012). *Assessment in my palm: E-scape in Israel evaluation of phase 2*. Goldsmiths, University of London.

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

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction*, *16*(2), 165–169. https://doi.org/10.1016/j.learninstruc.2006.02.005

The Royal Society. (2016). *Assessing experimental science in 11—18 education: New research directions*.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, *34*(4), 273–286. https://doi.org/10.1037/h0070288

van Popta, E., Kral, M., Camp, G., Martens, R. L., & Simons, P. R.-J. (2017). Exploring the value of peer feedback in online learning for the provider. *Educational Research Review, 20*, 24–34. https://doi.org/10.1016/j.edurev.2016.10.003

Whitehouse, C. (2013). *Testing the validity of judgements about geography essays using the adaptive comparative judgement method*. Centre for Education Research and Policy.

Whitehouse, C., & Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment*. Centre for Education Research and Policy.

Wiliam, D. (2006). The half-second delay: What follows? *Pedagogy, Culture & Society*, *14*(1), 71–81. https://doi.org/10.1080/14681360500487470