Tarricone and Newhouse *International Journal of Educational Technology in Higher Education* (2016) 13:16
DOI 10.1186/s41239-016-0018-x
UOC Universitat Oberta de Catalunya

International Journal of Educational Technology in Higher Education
a SpringerOpen Journal

RESEARCH ARTICLE Open Access

Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability

Pina Tarricone and C. Paul Newhouse*

* Correspondence:
p.newhouse@ecu.edu.au
Centre for Schooling and Learning Technologies (CSaLT), School of Education, Edith Cowan University, 2 Bradford St, Mount Lawley, 6050 Western, Australia

Abstract

In this paper we argue that comparative judgement delivered by online technologies is a viable, valid and highly reliable alternative to traditional analytical marking. In the past, comparative judgement has been underused in educational assessment and measurement, particularly in large-scale testing, mainly due to the lack of supporting technologies to facilitate the large number of judgements and judges. We describe the foundations of comparative judgement and dispel many of the old issues regarding its use in regards to time, cost and training for large-scale assessment. Studies in the use of comparative judgement and online technologies for the assessment and measurement of practical performance conducted by Edith Cowan University provide a context for further promoting its use in educational testing.

Keywords: Comparative judgement, Paired comparison, Pairwise comparison, Online assessment, Creative performance, Relative judgements, Absolute judgements

Background

Assessment is influenced by numerous factors including the teacher's expertise in the design and application of a measurement instrument, such as a test, and their judgement of student performance. Inferences and judgements are made about a student's learning development based on the evidence provided by a student's performance; and judgement making uses evidence in comparison to a standard, or another performance (e.g. by another student). Performance evidence, its quality, and judgements made are crucial to reliable assessment and have formed the problem of our research. In this paper we argue for the increased use of comparative judgement in the assessment of performance, particularly where creative expression is an integral component of that performance. Initially the concept of comparative judgement is defined and explained, then a range of literature is used to provide a rationale for its application to assessment, the important role of digital technologies is explained, and finally our own research journey over the past decade in the Centre for Schooling and Learning Technologies (CSaLT) is used to illustrate the potential.

Comparative judgement

The fundamental principle of measurement involves comparison, whether between two objects or performances or between a performance and a theoretical standard (e.g. marking criteria). Comparison provides the basis upon which measurement instruments are developed. Measurement instruments, assessment tasks, are deliberately designed and developed to measure ability and to classify performance relative to another performance. Essential to measurement is the construction of the task and its categories, which are used to compare, classify and identify ability through performance in relation to a well-defined outcome space (Humphry & Heldsinger, 2007; Wilson, 2005). Therefore, the basis of measurement is comparison, based on a specific outcome space and developed criteria that identify what performance outcomes are being measured and the categories of the performance (Andrich, 1988; Rasch, 1961; Wilson, 2005).

Comparative judgement, comparative pairs, pairwise comparison or paired comparison are terms that are used to describe a measurement method, which involves making inferences based on specific task criteria about one student's performance compared with another's, in contrast to comparison with a theoretical standard in analytical methods. This means that through comparison, items, objects or performances are measured with reference to another item, object or performance (Andrich, 1988; Rasch, 1961; Thurstone, 1927). Arguably comparative judgement began when Fechner (1860) first introduced the psychophysical method which involved ranking objects of physical magnitude; this method was later investigated and extended by Thurstone (1927, 1954). Comparative judgement was originally used for the pairing of psychophysical stimuli, such as weight, height and for measuring values and attitudes (Andrich, 1988; Bramley et al. 1998; Fechner, 1860; Jones & Alcock, 2014; Jones et al. 2015; Thurstone, 1927, 1954). The original conception of comparative judgement was that repeated comparisons of pairs of items, objects or performance used for an immediate perception with minimal cognitive processing. A comparison for quality takes more time than a comparison for magnitude. An example of an early study applying comparison for quality was by Bradley and Terry (1952), who investigated the feasibility of using comparative judgement to determine the quality of pork meat. Inferences of quality are made based on the criteria and this informs the estimation of the scale location. The process and outcome of the comparison involves estimating or inferring the scale location of the performances/instances, not based on the origin being relative to an absolute zero (Bond & Fox, 2001; Humphry & Heldsinger, 2007).

Measurement of ability (i.e. latent trait) from performance relies upon comparison to infer thresholds and form an interval scale (Andrich, 1988; Andrich & Luo, 2003). Thresholds, classifications or intervals on the latent-trait continuum represent a range of abilities. But differences in ability are not necessarily equal, and therefore the thresholds or intervals on the continuum cannot be proportioned equally. The location of a student's ability on the continuum is inferred by the judgment of their performance on a task or tasks in a specific knowledge domain. A foundation of the latent-trait continuum is that learning is cumulative and prior knowledge becomes the basis for the development of new knowledge and understandings (Humphry & Heldsinger, 2007). Inference and subsequent probabilistic mapping of ability and growth on a continuum, or 'underlying growth continua', is fundamental to the Rasch (1980) latent-trait theory (Andrich, 1988). Andrich (1988) argued that "ability is a latent trait, whereas performances are the observable manifestations of ability … in making observations that reflect properties, the actual

properties are abstractions based on the patterns of observations" (p. 14). Based on Andrich's conception, the latent nature of ability means that comparisons cannot be deterministic, but are probabilistic.

Deterministic classifications involve comparisons that are determined with certainty (Andrich, 1988). However, comparisons cannot be deterministic if they are measuring ability because it is difficult to determine with certainty the classification of abilities as they are not necessarily fully observable. Therefore, the comparisons are probabilistic; meaning that the outcomes of the comparisons are based on frequency, or some likelihood and qualitative judgment. In addition, invariance of classification is essential to measurement in a two-way frame of reference, such as in comparative judgement (Andrich, 1988). The Guttman structure provides a two-way observational frame of reference where comparisons are made between two items, objects or performances in a process of classification (Andrich, 1988; Bond & Fox, 2001; Humphry & Heldsinger, 2007). The outcomes of such comparisons are labelled dichotomous as they involve two outcomes represented by the integers of 0 or 1. Student work samples are compared to each other in this two-way frame of reference to identify which has achieved the criteria of the assessment at a higher or lower level. Therefore, comparison is used to determine whether each work sample is better (1) or worse (0) than another work sample. Hence, the ordering or comparison process is not deterministic but is probabilistic (Humphry & Heldsinger, 2007).

Referring to Thurstone's work on comparative judgements, Pollitt (2012a), who we understand initially collaborated with Andrich in the 1970's, argued that comparisons can be made with any kind of object including educational performances. Essentially, assessors' holistic judgements provide a shared contribution to the development of a measurement interval scale. The scale represents the outcomes of the assessors' cumulative comparative judgments of the quality of the psychometric latent trait being measured (Bramley, 2007; Pollitt, 2012a).

Relative and absolute judgements

"There is no absolute judgment. All judgments are comparisons of one thing with another" (Laming, 2004, p. 9). Absolute and relative judgements are terms that have been used to describe analytical and comparative judgement assessment methods respectively. Absolute judgements, analytical traditional marking/scoring, refers to the allocation of marks/score points or a grade (Gill & Bramley, 2013; Pollitt, 2012a). It has been contended by researchers that relative, comparative judgements are highly reliable, more so than absolute, analytical judgements (see Jones & Alcock, 2012, 2014; Laming, 2004; Pollitt, 2012b). Jones and Alcock (2012) argued that "people are far more reliable when comparing one thing with another than when making absolute judgements" (p. 64). Pollitt has also consistently supported the argument that comparative judgements, specifically adaptive comparative judgements, can generate "extremely reliable scores, far higher than traditional marking" with little training (Pollitt, 2012a, p. 168). For example, a study investigating absolute and relative judgements and standard-setting by Gill and Bramley (2013) considered the accuracy of holistic comparative judgements in relation to the original analytical marks of examination scripts that were 'close together in overall mark' in a standard-setting exercise. The assessors made absolute (mark/grade of each script), relative (script quality compared to another script) and confidence judgements (how confident they were of their judgements). Gill and Bramley found that in the

standard setting/awarding process, the assessors made more accurate relative judgements than absolute judgements, and that the judgements that were described by the judges as 'very confident' had higher accuracy than other judgements.

Heldsinger and Humphry (2013) identified some advantages of analytical marking including that the use of criterion marking rubrics provide specific information for assessors reducing subjectivity. Their review of the literature also suggested that there have been arguments highlighting concerns regarding the "assumption that using rubrics increases inter-rater reliability and validity, and the overall accuracy and quality of assessment" (p. 221). Finally, they concluded that assessors may mark leniently, harshly, to a different standard to others, and may value quality differently leading to a need for a substantial amount of time allocated to training and moderation. And Pollitt (2012a) argued that "a marker may award the same average mark, yet discriminate more finely amongst the objects, in effect being more generous to the better ones and more severe to the poorer—or vice versa" (p. 168). Therefore, one major downside of absolute, analytical marking is that ensuring reliability is challenging and costly to monitor. Whereas, relative judgements can be constantly monitored as they are made, and decisions can be made if additional judges or judgements are needed to increase reliability. Overall, there is agreement amongst researchers in the field with Jones and Alcock's (2012) argument "that people are far more reliable when comparing one thing with another than when making absolute judgements" (p. 64). Comparative judgement ensures that individual judge severity is overcome because the judgement process is focused on the 'relative merit' of each performance and not the individual judge's standard (Bramley, 2007, p. 265; Pollitt, 2012a). Also, if judges are making different quality discrimination decisions, the process of comparative judgment, after multiple rounds of judgements, can accurately rank performance quality (Pollitt, 2012a).

One of the main reasons for high reliability of relative comparative judgements compared to absolute analytical judgements is that, relative judgements require larger numbers of judges. Absolute analytical marking uses just one or two markers whereas comparative judgement requires more than two judges, e.g. ten judges (Pollitt, 2012a). For example, Pollitt (2012b) refers to his work with Richard Kimbell on the e-scape study of the assessment of e-portfolios as an example of the high reliability coefficients that can be achieved with digital assessment and the comparative judgement method. Pollitt posited that in Kimbell et al.'s (2009) study the high reliability coefficient of 0.96 generated by 28 judges assessing 352 e-portfolios with 3067 judgements was higher than any analytical marking system could achieve. Studies by Heldsinger and Humphry (2010), (2013) found that highly reliable comparative judgements of writing scripts by a large number of judges were made using calibrated exemplars as referents, as suggested by Thurstone (1928). They argued that comparative judgement, using calibrated exemplars, were successfully used as a highly reliable method to validate results from large-scale testing programs without the extensive assessor training and moderation processes that were required for absolute analytical judgements. In their 2010 study they found a high level of internal consistency of the teacher judgements and that the comparative judgements were highly correlated with the results of analytical marking from an Australian large-scale test for the same students, providing evidence of concurrent validity. These researchers were motivated by concerns about the accuracy of teacher judgements, the quality of information, including very little diagnostic information, provided by standardized testing. They found

that complex rubrics used for analytical marking required extensive training, whereas, comparative judgement required little training. So they found that the assessors' judgements of writing performance were highly internally consistent and highly reliable, considering that initially the teachers were concerned about making judgements across ability ranges with which they were not familiar. In a follow-up study Heldsinger and Humphry (2013) used calibrated writing exemplars as the referents for comparative judgements of scripts to assess student writing performance. The analysis showed a very high reliability and internal consistency.

Several researchers have found that judgements based on quality, using specified holistic criteria, resulted in a more valid assessment of performance (Bramley, 2007; Heldsinger & Humphry, 2010; Pollitt, 2012a). According to Pollitt (2012a) the high reliability of comparative judgement is a "consequence of the constant focus on validity" (p. 168) and this "demands and checks that a sufficient consensus exists amongst the pool of judges involved" (p. 167). Comparative judgements are independent and additional judgements increase reliability (Bramley et al., 1998; Pollitt, 2012a). Thus, judge bias and misfitting scripts can be easily identified during the judgement process, and reliability can be improved by increasing the number of judges and/or the number of judgements (Bramley, 2007). Additional judgements of work at grade cut-offs can help to improve and consolidate the ranking of works and improve reliability (Pollitt, 2012a). However, Bramley (2007) has argued that it is 'implausible' that each comparison is independent as the judges would remember particular scripts. We believe that it may be the case if the comparative judgements were being conducted with a small number of hard copy scripts or performances but it would be unlikely with large numbers of scripts and judgements being made when using online software systems, because judges may not encounter individual scripts in subsequent comparisons.

Time, cost, training and large-scale testing

Researchers have identified several limitations of comparative judgement. One is the time that it takes to make the large number of judgements and the number of judges required to reach the high levels of reliability that can be achieved with comparative judgement (Bramley et al., 1998; Heldsinger & Humphry, 2013; Jones & Alcock, 2014). Heldsinger and Humphry (2013) explained that although comparative judgement was an efficient assessment method complex scripts required additional time to complete the number of judgements needed for reliability. Bramley (2007) used the term psychological validity to describe the additional cognitive requirements to make valid and reliable judgements of complex performances, such as scripts, because of the amount of time taken to make the judgement. Bramley's psychological validity may have some connection with what Thurstone (1927, 1954) described as the "discriminal process" (p. 48). According to Thurstone during comparative judgement the discriminal process encompasses the different reactions to each work sample, item or task as it is being compared to the next work sample, item or task. Thus the discriminal process leads to the comparative judgement and affects the time taken.

The issue of time in making comparative judgements was highlighted in a study by McMahon and Jones (2014), who investigated the use of comparative judgement in the assessment of a secondary school chemistry experiment. The researchers found that

the comparative judgement was highly reliable, consistent across judges and the judgements correlated with the analytical marking testing data. The study involved 154 test scripts and five teachers from a secondary school in Ireland. Altogether they made 1550 judgements using the 'No More Marking' online system. All of the scripts were also analytically marked by two of the five teachers. The researchers found that the comparative judgements totalled approximately 14 h (3 h per judge) and the analytical marking totalled approximately 3 h to mark all 154 tests. The two teachers who analytically marked the tests reported that they each took one and a half hours to mark the tests without providing feedback. The estimated reported marking time seems short for the number of scripts that were marked. Although the comparative judgements were more time consuming than the analytical marking, McMahon and Jones acknowledged that other studies, including our own, have shown otherwise (see Jones et al., 2015; Kimbell, 2012; Newhouse, 2011). Our review of the test instrument revealed that a possible reason for this time discrepancy, not identified by the authors, was that it was not conducive to making holistic judgements. Most of the questions required students to list responses rather than provide a descriptive response. These questions required the judge to assess each individual response rather than to make a holistic judgement of the script.

Another time and cost issue is assessor training. Several researchers have found that assessors needed little training in the use of comparative judgement systems and the process, including those in which judges had not had specific assessor training (see Heldsinger & Humphry, 2013; McMahon & Jones, 2014; Pollitt, 2012a). Training generally involves understanding the holistic performance criteria and assessors gaining an understanding of the quality of the evidence presented in a range of work samples (Kimbell et al., 2009). For example, in one study Heldsinger and Humphry (2010) found that training for comparative judgements took approximately half an hour, whereas, analytical marking rubric training needed a full day. In terms of judgement making, Heldsinger and Humphry (2013) found that with little training assessors using calibrated exemplars were able to efficiently judge 29 writing performances, one every three minutes, in one and half hours. Although they specifically did not refer to the lower cost of comparative judgement they did consistently argue the efficiency of the method, in terms of training and judgement making, in contrast to traditional, analytical marking. Pollitt (2012b) has consistently argued that comparative judgment can be conducted at a comparable or lower cost to analytical, traditional marking and that comparative judgement provides consistently high reliability. Jones and Alcock (2014) asserted that the larger number of judges in comparative judgement can be more costly than analytical marking, but this is offset by the greater reliability in the comparative judgment process.

Use of technology and comparative judgement

It has been noted by Heldsinger and Humphry (2013) that until recently there has been little use of comparative judgement methods for educational purposes, probably because it has been considered time consuming, tedious and laborious (see Bramley et al., 1998). The recent increase in the use of comparative judgment may be largely due to the availability and use of online software systems, using digitised student works, by small groups of researchers in the UK and in Western Australia. These online systems, such as the Adaptive Comparative Judgement System (ACJS) (Pollitt, 2012b) and the Pair-Wise Web Software

(Humphry et al. 2013–2015), are based on Thurstone's (1927) law of comparative judgement and have dealt with many of the issues, that in the past inhibited the use of comparative judgement in assessment (Jones & Alcock, 2012; Pollitt, 2012a). These online systems provide a platform for making comparative judgements of digitised works such as scripts, recorded performances and portfolios.

The recent use of these digital comparative judgment systems have been for standard setting (Bramley, 2007; Bramley et al., 1998; Gill & Bramley, 2013), moderation and marking of scripts and portfolios in areas such as writing (Heldsinger & Humphry, 2010, 2013; Pollitt, 2012b), mathematical problem solving (Jones et al., 2015), science (McMahon & Jones, 2014), design and visual arts (Newhouse & Tarricone, 2014), design and technology e-portfolio assessment (Kimbell et al., 2009), scientific and technology enquiry e-portfolio assessment (Davies et al. 2012), oral language assessment (Newhouse, 2011; Pollitt & Murray, 1996), and peer assessment (Jones & Alcock, 2012, 2014; Seery et al. 2012). These studies have provided evidence of the efficacy in using digital assessment systems for relative comparative judgement. Many researchers have also investigated the effectiveness of, and correlation between, absolute analytical marking and relative comparative judgements. For example, research by Jones, Swan and Pollitt (2015) investigated the use of comparative judgement, as an alternative to analytical criterion marking methods, to assess constructs such as mathematical problem solving, that can be somewhat difficult to capture and describe in analytical marking criteria. Their research involved two studies and two groups of mathematics education experts in the UK (Group 1 $N = 12$, Group 2 $N = 11$). The first study investigated the suitability of comparative judgements to assess the General Certificate of Secondary Education (GCSE) final examinations in the UK to determine whether comparative judgement could replicate the analytical marking using specific criteria. In the first study, Group 1 completed 151 judgements in 105 min, and Group 2 completed 150 judgements in 100 min, and showed that comparative judgement could be used to assess the GCSE. The second study used comparative judgement to assess mathematical problem solving skills, using 18 scripts from three tasks. The judgements were correlated with scripts' grades to determine the validity and reliability of comparative judgements as a measure to assess summative examinations and problem solving in mathematics. The groups of assessors, in both studies, produced ranked measurement scales with high reliability.

The CSaLT journey to comparative judgements

The Centre for Schooling and Learning Technologies (CSaLT) was formed in 2003 at Edith Cowan University, Perth Western Australia, with a mission to conduct research into, and promote the use of, digital technologies to improve schooling. The premise was that technologies could be used for teaching, learning and assessment (Newhouse, 2010). Increasingly, research in the Centre has found that the technologies were readily available in educational institutions but seriously underused especially in high-stakes assessment. In response to this situation, researchers in CSaLT decided to partner with the Western Australian curriculum and examination authority in testing the use of digital technologies in high stakes assessments for tertiary entrance. This decision resulted in three major studies from 2006 to 2014. Detailed descriptions of the methodologies and results have been previously reported (e.g., Newhouse, 2011; Newhouse & Tarricone, 2014).

The first study, from August 2006 to December 2007, was funded by a University collaborative research grant, and involved collaboration between CSaLT and the School Curriculum and Standards Authority (Newhouse, 2010). The researchers sought to investigate using digital technologies to support alternative forms of external assessment of student performance in tertiary entrance courses that included a major practical component (Newhouse, 2008). The focus of this study was the use of digital technologies to 'capture' and score performance on practical tasks in an engineering and an applied IT course. The study involved 13 small teacher-class case studies. During this study Professor Richard Kimbell from London University was a visiting scholar at the centre; he introduced the e-scape project, assessing complex practical performance using evidence collected in digital forms, and the use of comparative judgements for generating scores (Kimbell, 2012). As a result one of the outcomes of this study was the successful application of a comparative judgements method of scoring to the assessment of digital posters created by students under exam conditions. To facilitate this an online database tool was custom-designed using Filemaker Pro to allow judges to view pairs of student work and enter their judgements. On completion of all judgements, the researchers exported and analysed the data using RUMMcc software.

The first study was then used as a pilot for a more extensive three-year study investigating the feasibility of using digital representations of work for authentic and reliable performance assessment in foreign language, engineering, information technology, and physical education tertiary entrance courses (Newhouse, 2011). The study was funded by an Australian Research Council Linkage grant, with the local curriculum authority as the partner again, and also involved collaboration with researchers in the British e-scape project (Kimbell et al. 2007). For each course, digital forms of assessment were devised to improve on the standard form of assessment used at the time, with the evidence (e.g. student work, recording of performance) stored in digital files within online systems. Assessors accessed these files using online scoring tools to facilitate analytical marking and comparative judgement. By the second year of the study a commercial system had been developed by TAG Learning for the e-scape project (eventually known as ACJS), so this system was used for comparative judgements. It not only displayed the assessment evidence and captured assessor judgements, it also included the Rasch mathematical calculations to progressively provide scores and reliability statistics. As a result, assessors no longer had to make a set number of judgements before completing the exercise; completion was determined by reaching a required level of reliability.

Rasch measurement was used to analyse all scoring data to determine reliability coefficients and related statistics. The comparative judgement method, for all four courses, generated highly reliable sets of scores, with Cronbach's Alpha and Separation Index coefficients exceeding 0.9. For all four courses the comparative judgment method showed higher levels of reliability than analytical marking, but the comparative judgement scores and rankings were typically moderately to highly correlated with the scores from traditional analytical marking (Newhouse, 2011). Issues related to the time taken to complete comparative judgments when compared with analytical marking, were mainly related to file size, download time and the amount of text and pages of the digital evidence, rather than the time it took to make the comparative judgment. Overall, the study demonstrated the feasibility of using Rasch measurement with analytical

marking and comparative judgements for the digital assessment of performance as a reliable and valid alternative to paper-based high-stakes examinations. Ultimately the researchers found that the comparative judgment method of scoring was more reliable than analytical marking.

In the third study, which began in 2011 and was completed in 2014 (Newhouse & Tarricone, 2014) the researchers sought to investigate the efficacy of digitisation for online analytical marking and comparative judgement methods by building on Dillon and Brown's (2006) use of e-portfolios. This study involved three phases; the first two phases investigated the efficacy of digitisation of student practical work and the effectiveness of comparative judgement of portfolios for the purpose of summative assessment in the Visual Arts and Design tertiary entrance courses. Once again the ACJS was used to facilitate the comparative judgements. In the first phase the researchers explored the potential of representing practical work in the two courses, whereas the second phase concerned whether it was feasible for students to create and submit such representations. The third phase focused solely on the online scoring of Visual Arts digital portfolios for the purpose of moderation to ensure consistent standards. The researchers believed that this may be efficiently and effectively accomplished using the approach used in this study to the online scoring of digital portfolios.

In all three phases of the study the researchers demonstrated that the digitised student work could be efficiently and effectively assessed using online marking systems. For both courses, there was low consistency between the assessors for analytical marking, although the combined scores using a Rasch polytomous model exhibited good reliability. For both there was a high level of internal consistency for the results of the comparative judgements. For all three phases and for both Visual Arts and Design, the comparative judgement method generated highly reliable sets of scores, with Cronbach's Alpha and Separation Index coefficients for the first two phases exceeding 0.9. However, in the third phase a reliability coefficient of only 0.88 was achieved, with the most likely explanation being a lack of a consensus understanding of the criteria among some of the Visual Arts teachers who were inexperienced with marking these submissions.

Conclusion

In this paper we have used a range of literature and CSaLT's eight-year research journey to argue that the authenticity, validity and reliability of high-stakes summative assessment may be enhanced through the use of the comparative judgement method of scoring using digital representation of evidence and online tools to facilitate scoring and feedback. The changing requirements of tertiary courses necessitate that students be proficient in using a variety of digital technologies to support their learning. The potential for these technologies to be used for high-stakes assessment is fast becoming a reality, with the added capacity to use comparative judgement for scoring and Rasch modeling to analyse the data. The time is now to embrace digital technologies for high-stakes assessment and consider comparative judgement for scoring as a viable and reliable alternative to analytical marking, particularly for creative performances that rely on highly subjective judgements.

Authors' information
Paul is an Associate Professor at Edith Cowan University in Western Australia where he is the director of the Centre for Schooling and Learning Technologies (CSaLT) in the School of Education. His aim is to improve the opportunities for children to develop their potential through engaging and relevant schooling.
Pina's PhD in educational/cognitive psychology won the 2007 Edith Cowan University Research Medal. Psychology Press published her thesis research as a book titled The Taxonomy of Metacognition. She has a M Ed in Interactive Multimedia. In 2014 she completed a M Ed in Educational Measurement with Honours from the University of Western Australia. Her interests include educational psychology constructs, psychometrics and the use of technologies for assessment.

Acknowledgements
The study discussed in this paper was the work of a research team led by Paul Newhouse and included researchers Jeremy Pagram, Lisa Paris, Mark Hackling, Martin Cooper, Pina Tarricone, Alistair Campbell, Alun Price and many research assistants. The work of everyone in this team, particularly Martin Cooper and Pina Tarricone, and the teachers and students involved, contributed to the research outcomes presented in this paper.

Received: 17 July 2015 Accepted: 16 November 2015
Published online: 07 April 2016

References

Andrich D (1988) Rasch models for measurement. Sage Publications, Newbury Park
Andrich D, Luo G (2003) Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. J Appl Meas 4(3):205–221
Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, N.J. ; London: L. Erlbaum
Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika 39(3/4):324–345
Bramley T (2007) Paired comparison methods. In: Newton P, Baird J-A, Goldstein H, Patrick H, Tymms P (eds) Techniques for monitoring the comparability of examination standards. QCA, London, pp 264–294
Bramley T, Bell JF, Pollitt A (1998) Assessing changes in standards over time using Thurstone paired comparisons. Education Res Perspectives 25(2):1–24
Davies D, Collier C, Howe A (2012) Assessing Scientific and Technological Enquiry Skills at Age 11 using the e-scape System. Int J Technology Design Education 22(2):247–263
Dillon SC, Brown AR (2006) The art of ePortfolios : insights from the creative arts experience. In: Ali J, Kaufman C (eds) Handbook of Research on ePortfolios. Idea Group Reference, Hershey, pp 420–433
Fechner, G. T. (1860). *Elements of psychophysics* (H. E. Adler, Trans. Holt, Rinehart & Winston Eds.)
Gill T, Bramley T (2013) How accurate are examiners' holistic judgements of script quality? *Assessment in Education: Principles*. Policy Practice 20(3):308–324
Heldsinger S, Humphry SM (2010) Using the method of pairwise comparison to obtain reliable teacher assessments. Australian Educational Res 37(2):1–19
Heldsinger S, Humphry SM (2013) Using calibrated exemplars in the teacher-assessment of writing: an empirical study. Educational Res 55(3):219–235
Humphry SM, Heldsinger S (2007) Using measurement principles in developing assessment processes. In: Educational Assessment and Measurement Series. University of Western Australia, Perth
Humphry SM, Wray WH, Wray FW (2013–2015) Pair-Wise Web Software. The University of Western Australia, Perth
Jones I, Alcock L (2012) Summative peer assessment of undergraduate calculus using adaptive comparative judgement. In: Iannone P, Simpson A (eds) Mapping university mathematics assessment practices. University of East Anglia, East Anglia, pp 63–74
Jones I, Alcock L (2014) Peer assessment without assessment criteria. Studies in Higher Education 39(10):1774–1787
Jones I, Swan M, Pollitt A (2015) Assessing mathmatical problem solving using comparative judgement. Int J Science Mathematics Education 13:151–177
Kimbell R (2012) Evolving project e-scape for national assessment. Int J Technology Design Education 22:135–155
Kimbell R, Wheeler T, Miller A, Pollitt A (2007) e-scape: e-solutions for creative assessment in portfolio environments. Technology Education Research Unit, Goldsmiths College, London
Kimbell R, Wheeler T, Stables K, Sheppard T, Martin F, Davies D, Whitehouse G (2009) e-scape portfolio assessment: Phase 3 report. Technology Education Research Unit: Goldsmith College, London
Laming D (2004) Human judgment: The eye of the beholder. Thomson Learning, London
McMahon, S., & Jones, I. (2014). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*. doi: 10.1080/0969594X.2014.978839
Newhouse CP (2008) Digital forms of performance assessment. In: Paper presented at the Australian Computers in Education. ACT, Canberra
Newhouse CP (2010) Aligning assessment with curriculum and pedagogy in applied information technology. Australian Educational Computing 24(2):4–11
Newhouse CP (2011) Comparative pairs marking supports authentic assessment of practical performance within constructivist learning environments. In: Cavanagh RF, Waugh RF (eds) Applications of Rasch Measurement in Learning Environments Research. Sense Publisher, Rotterdam, pp 141–180
Newhouse CP, Tarricone P (2014) Digitizing practical production work for high-stakes assessments. Canadian J Learning Technology 40(2):1–17
Pollitt A (2012a) Comparative judgement for assessment. Int J Design Education 22:157–170
Pollitt A (2012b) The method of Adaptive Comparative Judgement. Assessment Education 19(3):281–300

Pollitt A, Murray NL (1996) What raters really pay attention to. In: Milanovic M, Saville N (eds) Studies in language testing 3: Performance testing, cognition and assessment. Cambridge University Press, Cambridge, pp 74–89
Rasch G (1961) On general laws and the meaning of measurement in psychology. In: Neyman J (ed), Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, Berkeley, California, 1961 (Vol. IV: Contributions to biology and problems of medicine). University of California Press, Berkeley, California, pp. 321-333
Rasch G (1980) Some probabilistic models for the measurement of attainment and intelligence. MESA Press, Chicago
Seery N, Canty D, Phelan P (2012) The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. Int J Technology Design Education 22:205–226
Thurstone LL (1927) A law of comparative judgement. Psychol Rev 34:278–286
Thurstone LL (1928) Attitudes can be measured. American J Sociology 33(4):529–554
Thurstone LL (1954) The measurement of values. Psychol Rev 61(1):47–58
Wilson M (2005) Constructing Measures: An Item Response Modelling Approach. Lawrence Erlbaum Associates, Mahwah