International Journal of Technology and Design Education (2022) 32:1515–1529
https://doi.org/10.1007/s10798-021-09654-w

Examining the reliability of Adaptive Comparative Judgement (ACJ) as an assessment tool in educational settings

Richard Kimbell[1]

Accepted: 29 January 2021 / Published online: 23 February 2021

Abstract
Conventional approaches to assessment involve teachers and examiners judging the quality of learners work by reference to lists of criteria or other 'outcome' statements. This paper explores a quite different method of assessment using 'Adaptive Comparative Judgement' (ACJ) that was developed within a research project at Goldsmiths University of London between 2004 and 2010. The method was developed into a tool that enabled judges to distinguish better/worse performances not by allocating numbers through mark schemes, but rather by direct, holistic, judgement. The tool was successfully deployed through a series of national and international research and development exercises. But game-changing innovations are never flaw-less first time out (Golley, Jet: Frank Whittle and the Invention of the Jet Engine, Datum Publishing, Liphook Hampshire, 2009; Dyson, Against the odds: an autobiography, Texere Publishing, Knutsford Cheshire, 2001) and a series of careful investigations resulted in a problem being identified within the workings of ACJ (Bramley, Investigating the reliability of Adaptive Comparative Judgment, Cambridge Assessment Research Report, UK, Cambridge, 2015). The issue was with the 'adaptive' component of the algorithm that, under certain conditions, appeared to exaggerate the reliability statistic. The problem was 'worked' by the software company running ACJ and a solution found. This paper reports the whole sequence of events—from the original innovation, through deployment, the emergent problem, and the resulting solution that was published at an international conference (Rangel Smith and Lynch in: PATT36 International Conference. Research & Practice in Technology Education: Perspectives on Human Capacity and Development, 2018) and subsequently deployed within a modified ACJ algorithm.

Keywords Adaptive comparative judgement (ACJ) · Assessment · Reliability · Holistic judgement

I am indebted to Tom Bramley of Cambridge Assessment for his helpful conversations and generosity with providing sources and to Ruth Wright formerly Head of Research at the Engineering Council UK for her valuable review of the manuscript.

✉ Richard Kimbell
r.kimbell@gold.ac.uk

[1] Goldsmiths University of London, London, UK

Introduction

This paper concerns a story of innovation in assessment. Historically, the assessment of learners' performance was undertaken by *ranking* candidates rather than by *marking*, which only emerged in the 18thC as the industrial revolution expanded the number of candidates for examination. For the last 200 years, with numbers-based marking as an overwhelming methodology, 'true-score-theory' has dominated educational assessment. This paper examines a radical departure from this norm. Specifically it concerns the creation of Adaptive Comparative Judgement (ACJ), an assessment method that was developed within a research project at Goldsmiths University of London between 2004 and 2010.

The project in question was funded by the Qualifications and Curriculum Authority (QCA UK) who were interested to develop digital approaches to assessment for General Certificate of Secondary Education (GCSE; 16+) examinations in England & Wales. The interest of QCA was two-fold; first to create digital approaches to learner-portfolio-building in 'performance' subjects like design & technology (d&t), and second to develop digital approaches to the assessment of those portfolios. The original project was entitled 'e-scape' (e-solutions for creative assessment in performance environments) and work progressed through three phases. Phase 1 was dominated by the concerns of web-portfolio-building. Phase 2 involved a small schools trial of the resulting portfolio approach and a prototype version of ACJ was developed for web-based assessment of the portfolios. This was then fully explored in phase 3, which involved a trial in 17 schools in four regions of England & Wales and produced 357 d&t portfolios. And it was in this phase 3 that a fully developed version of ACJ was first employed for the assessments. (Kimbell et al. 2009; Kimbell and Stables 2007).

The story of the project was fully articulated in a Special Issue of the International Journal of Technology and Design Education in 2012 (Williams and Kimbell [eds] 2012). In that Special Issue, not only was the e-scape project and the development of ACJ itself reported, but additionally it included a number of developments from it, including the use of ACJ in other settings. Seery and Canty (in Ireland) explored its use as a peer-review tool in Higher Education; McLaren (in Scotland) explored its use in primary schools as a feedback and learning tool; Davies applied it to the assessment of science enquiry skills; and Williams (West Australia) to the assessment of engineering performance. The value of ACJ as a formative assessment and learning tool (from primary school through to Higher Education) has subsequently been explored and reported by several authors, most extensively by Scott Bartholomew now at Brigham-Young University USA (see for example Bartholomew et al. 2019).

This paper is not intended as a further exploration of ACJ applications within curriculum or assessment. Rather, it arises because of a technical challenge to ACJ that was raised by Tom Bramley of Cambridge Assessment in 2015. The original comparative judgement algorithm of ACJ had been developed by a team of people including Alastair Pollitt, and Karim Derrick who had both contributed papers to the 2012 IJTDE Special Issue. Bramley ran simulations with the method and was convinced that the reliability levels of ACJ assessment sessions reported in the literature were inflated by the adaptivity of the algorithm (Bramley 2015). In this paper, I present the evolution of that dispute; the flaws as identified by Bramley and the remedial solution as subsequently developed by the software team running ACJ. The question that prompted this paper exists in two parts. Is the Bramley criticism accepted as real by the ACJ software team, and (if it is) does the solution

developed by the ACJ team satisfy Bramley that the ACJ algorithm, in its modified form, now produces trustworthy reliability data?

How assessment works

The conventional approach to assessment, eg in current GCSE examinations, is to set questions or challenges that require 'answers' or other outcomes that are then measured against what is deemed (by the examining authority) to be an ideal answer or outcome. This ideal answer or outcome is specified through sets of criteria that are used to decide how thoroughly/accurately the learner has responded to the questions or challenges. The criteria are typically associated with numerical scores, and by adding up learners' scores against each of the criteria, a final score is arrived at reflecting the learners overall level or ability in that examination or assessment (Gipps 1994; Finn 2015; Marshall 2017).

The process of identifying grade-related criteria, or performance-indicators, or Statements-of-Attainment, expanded hugely through the 1980s and 1990s (Kimbell 1997, ch 2). Elongated lists of criteria (in every subject) provided many more qualities for teachers to look for in learners' performance. By the same token however, that detailed process of examination took far longer than assessments had ever previously taken. The Task Group on Assessment and Testing (TGAT) for the UK National Curriculum warned the Department of Education and Science (DES) that '… there is a potentially serious problem because of the size of the burden that could be placed on teachers' (DES 1988, para 119). And this raised a related problem. The Chair of TGAT, Professor Paul Black, first identified the associated problem of the *reliability* that could be expected of teachers' assessments when using such a vast number of detailed Statements.

> It seemed absurd to me that SEAC could countenance reporting separately on such a large number, but it was not until the Examination Groups pointed out that they could not possibly do this at GCSE level with any respectable degree of accuracy that the absurdity was accepted. (Black 1993, p. 60).

There is a long history of challenges to this atomised view of assessment, and particularly in the professional milieu of classrooms and the behaviour of learners and teachers, as Schon (1983) pointed out.

> In his day-to-day practice he (sic) makes innumerable judgements of quality for which he cannot state adequate criteria, and he displays skills for which he cannot state the rules and procedures. (Schon 1983, pp. 49–50).

This notion of 'tacit' knowing was presented many years earlier by Polanyi (1958) in his discussion of connoisseurship. He uses the concept to discuss the work of any skilled professional group; wine-blenders, tea-tasters or medical diagnosticians, who (he argues) develop tacit understanding of their professional practice. 'We can know more than we can tell' (Polanyi 1966 p4), being able to judge or act skillfully without being able to articulate exactly what it is that he or she knows. (See also Eisner 1981; Angoff 1974).

Wiliam (1998a, b) takes Polanyi's idea of connoisseurship explicitly into the world of school-based assessment. His critique of criterion-based assessment leads him to postulate the existence of an altogether different view of the processes involved in making judgements; *construct*-based assessment.

> I have argued elsewhere that most summative assessments are interpreted not with respect to criteria (which are ambiguous) nor with respect to norms (since precisely-defined norm groups rarely exist), but rather by reference to a shared construct of quality that exists in some well defined community of practice (Lave & Wenger, 1991). For this reason, I have termed these assessments 'construct-referenced' assessments (Wiliam 1998a, p. 8).

And the same year …

> To the extent that the examiners agree, they agree not because they derive similar meanings from the regulation, *(ie they are not criterion-driven)* but because they already have in their minds a notion of the required standard. The consistency of such assessments depend on what Polanyi (1958) called connoisseurship … (Wiliam 1998b, p. 6) (*my insert in italics).*

The final word on this should be Polanyi's where he points out that connoisseurship, like skill, can be communicated only by example, not by precept. (Polanyi 1958 p. 56). The problem with precepts (general rules.. or we might say 'criteria') is that their meaning inevitably migrates according the context in which they arise. Polanyi is making the point that *examples* not only carry a general rule (a criterion) but they also do it in context, making its meaning clear.

Three of the ideas outlined above—the uncertainty associated with generic criteria; the connoisseurship of expert judges; and the importance of judging through examples (Polanyi)—contributed to the awareness that Goldsmiths (QCA) 'e-scape' project might be an ideal vehicle within which to develop a quite new approach to assessment. An additional contributing factor was that, since the assessments were to be web-based, a computer-managed approach to the assessments would be essential. The issue (in 2005) resolved itself into the question of how teachers' professional judgements of quality could be reconciled within a computer-based assessment approach.

Judgement by comparison

The comparative judgement method was first articulated by Louis Thurstone in a series of articles concerned with the measurement of the psychological perceptions of physical stimuli such as tones and loudness, and of psychological variables such as values and attitudes (Thurstone 1927, 1959). He proposed several methods for constructing scales using his 'Law of comparative judgement', which describes how the comparative judgement of the quality of two objects depends on the magnitude of their difference in 'quality'. From these ideas, in the late 1990s, Alastair Pollitt, then Director, Research & Evaluation Division at Cambridge Assessment, investigated the potential for using the method of comparative judgement as a mechanism to reduce subjective bias within educational assessments.

In 2004, Pollitt had explored the difficulty of making judgements by reference to generic criteria or 'grade descriptions' in school-based assessment.

> When we try to judge a performance against grade descriptors we are imagining or remembering other performances and comparing the new performance to them. But these imagined performances are unlikely to be truly representative of performances of that standard, and very likely to vary in the minds of different judges. (Pollitt 2004, p. 5).

Pollitt goes on to recommend an approach that involves the direct comparison of one piece with another, and for a very good statistical reason. Imagine we are comparing two pieces of work. Which is more thorough.. A or B? As Pollitt explains..

> … when a judge compares two performances (using their own personal 'standard' or internalized criteria) the judge's standard cancels out. …(Pollitt 2004, p. 6).

An easy/lenient judge might think them both thorough, but one is more-so. Or a strict judge might think them both not-thorough, but one will still be more so than the other. In either event, and despite their different personal standards, they will both identify the same more-thorough piece. As Pollitt noted, their personal standards cancel out, and as Polanyi noted, *direct comparison* facilitates judgement far better than abstract 'precepts'.

Creating Adaptive Comparative Judgement

In 2006, as the Goldsmiths e-scape project began to address the challenge of making judgements of the emerging learner web-portfolios, they contacted Pollitt to see whether there might be any useful interaction with his ideas of comparative judgement. A trial was established with twenty paper-based portfolios of known quality (identified in a previous project at Goldsmiths) that could therefore be placed in a rank order. These twenty were judged by a new team of six researchers using a manual (spreadsheet-based) version of Pollitt's comparative judgement approach. The emerging rank correlated well with the original rank (Spearman's correlation co-efficient = 0.89). (Kimbell et al. 2007, pp. 57–58). But the process of 6 judges, each making 40 paired judgements of 20 portfolios created a serious logistical challenge.

> We had 6 judges and 20 pieces of work and we all sat at a big round table. Judge 1 was looking at script 15 and 5; judge 2 was looking at 17 and 2, and so on. Soon more than one judge needed the same piece of work and just had to wait till the other judge was finished with it. And by then the scripts had all got jumbled up in the middle of the table and did not come easily to hand. If this was a problem with just 20 scripts, imagine the challenge of a 100 or 1,000 or 50,000 scripts. The distribution process alone makes the process of repeated comparative judgments (by different judges) quite unmanageable. But the situation changes dramatically when all the portfolios are in a website. There, *every piece of work* is available *all the time*, *simultaneously,* and for *any of the judges*. (Kimbell et al. 2007, p. 58).

Pollitt's comparative judgement approach had formerly been constrained by these logistic difficulties and was used purely as a research tool. This trial experience made it obvious that for the e-scape project to work, two linked web-based systems needed to work together. The web-portfolio part of the system needed to speak directly to the comparative pairs assessment part of the system (at that time called the 'pairs engine'). The combined e-scape software systems would then make it possible—for the first time—for comparative judgement to become a front-line assessment tool. (Kimbell in Williams and Kimbell [eds] 2012, pp. 135–155).

The requirement for a level of adaptivity to enhance the efficiency of the algorithm was later fully articulated by Pollitt in *The Method of Adaptive Comparative Judgement* (Pollitt 2012), and is discussed in detail later. Goldsmiths partnered with TAG Learning, a leading educational software developer, to create an assessment tool that utilised the ACJ method.

This assessment tool (the 'pairs engine') employed a specific adaptive algorithm to drive a scalable, paired-judgement process.

ACJ in project e-scape

There were many aspects of the e-scape project that were new, not least the challenge of creating on-line (web-based) digital portfolios of performance (including drawings, sound-files, video, photographs and text). The trick with e-scape was that these portfolios were created live (in real-time) from the studio/workshops where learners were undertaking the 7 h design task (see Williams and Kimbell [eds] 2012). Seventeen schools were involved in the e-scape trial in 4 regions of the country and at the end of the schools' work, there were 357 design & technology e-portfolios in a website, each representing 7 h of design studio work in response to a common task. (Kimbell et al. 2009).

The assessment judgements were made by a team of 28 judges who each judged approx 120 pairs of portfolios. In each case the software that used the ACJ method presented the judge with two portfolios that could be studied separately or together on a split screen. The principal role of the judge was to review them both and to decide which of the two portfolios (A or B) was the stronger. The 'pairs engine' software then presented another pair for comparison. The judges were very familiar with the work as most of them were the teachers in the trial schools, and others were from research groups [Australia/Ireland/Israel/USA] that were interested in—and had been closely following—the work involved in the e-scape project.

If one imagines the 350 portfolios distributed in a matrix, then that matrix is 350×350 and there are 122,500 units in the matrix. If every portfolio was to be compared with every other portfolio, then there were 61,075 potential pair combinations. In fact our 28 judges each did 120 comparisons or (in total) 3416 judgements, representing less than 6% of the possible matrix combinations. And yet it still generated a reliable outcome. Pollitt explains the rationale underlying this efficiency, and it turns on the adaptive mechanism for selecting the pairs of portfolios for comparison. As a simple illustration, imagine a sequence of paired comparisons; A beats B, B then beats C, and C then beats D. There is an extremely high probability that A would beat D so an adaptive algorithm uses this information to pick more useful (closer) pairings.

> The improvement in efficiency targeting generates is similar to that observed in computer-adaptive testing, where a student's 'ability' is re-estimated after every item, and the next item presented is chosen to match closely the new estimate. In adaptive testing, savings have been made of 50% or more in the number of items needed to reach the same level of accuracy as with conventional tests (Weiss 1982; Eggen and Straetmans 2000). In the CJ method there are no items but, if the next comparator is chosen as described above, similar gains in efficiency will be made. (Pollitt et al. 2009, para1.1).

It is important to understand the e-scape 'pairs engine' procedure, that operated through 'rounds' of judging. A round was completed when all the portfolios had been judged at least once against a selected other portfolio. In the first round the selection was random and the first four rounds of judging operated with a 'Swiss Tournament' procedure.

> The Swiss Tournament system comes from the world of tournament chess, where it is the most common way to arrange the pairings so that every player is fairly tested,

> and a winner is found even though no player is ever fully "knocked-out". Using the Swiss Tournament in this context, the first pairs presented to the markers were chosen at random, and the winner received one point. In following rounds the pairs were chosen from groups of 'players' with the same number of points. For example, at the start of round 3 some players had 2 points (having won 2 matches) others had 1 (won one and lost one) and some had zero (lost both). (Pollitt et al. 2009, para 2.2).

At the end of the four Swiss rounds, and with an approximated rank across 5 bands (0–4), the algorithm then moved into the full-scale Rasch estimation rounds and the process for choosing the pairs of portfolios changed.

> … the algorithm checks a script against all the other scripts in the system. It looks at how many times they were compared and how many times the current script won and lost. This data is then used to calculate the "ideal" parameter value for this script. A separate calculation is then made involving the number of wins and the current parameter value for each script. The difference between these two values is noted and a third calculation is made to generate an adjustment figure for the current script. (Pollitt et al. 2009 paras 2.2).

In choosing to compare two portfolios it is important how far apart they are on the putative scale of quality (a representation of differences in quality). Comparing pieces that are a long way apart (a very good piece and a very poor piece) makes it easy for the judge to decide which is better, but very little information is imparted to the system. On the other hand if they are close together, the judge may struggle to distinguish the better piece but the system gains much more information.

> In statistics the 'information' contributed by a single judgement, is quantified in terms of the modeled probability that it will have one or other output: where $\boldsymbol{p}$ is the probability that the first portfolio would be judged to have more quality, $\boldsymbol{q}$ (equal to 1-p) is the probability that the second would be, and $\boldsymbol{I}$ is the amount of information the comparison adds to the analysis. This function is at a maximum when $\boldsymbol{p}$ and $\boldsymbol{q}$ are both equal to 0.5, it declines slowly at first but more rapidly as $\boldsymbol{p}$ rises beyond 0.7 or falls below 0.3. (Pollitt et al. 2009, paras1.1).

Based on earlier phases of the trial, 0.67 logits (a unit of probability) was used as the separation factor for these Rasch analysis rounds meaning that the odds of one of the portfolios winning were 2:1. The judges found this acceptable in terms of distinguishing the quality of the two portfolios, and the information gathered was still 90% of that from a statistically 'ideal' pairing. A number of other safeguards were also built into the 'pairs engine' algorithm, eg to balance the overall number of judgements for each portfolio, to balance the number of times a script has been seen by the same judge, and to prevent the same pair being shown to the same judge. The rounds of judging continued in this way with each new estimation round being triggered as before, when all the scripts have been involved in one more comparison. As each estimation round was completed, the parameter values for each script were reviewed as well as the summary data for the whole estimation process.

Theoretically the process could just go on and on, until the conditions are met that terminate the process. In the 2009 trial the conditions were met after 17 rounds. At that point Pollitt reported as follows:

> The final scale spread the portfolios out with a standard deviation of almost 3 units. The average measurement uncertainty for a portfolio was about 0.67 units,

> and the ratio of these two figures was 4.45. This means that the standard unit of the scale was almost 4.5 times as large as the uncertainty of measurement. This means the portfolios were measured with an uncertainty that is very small compared to the scale as a whole; this ratio is then converted into the traditional reliability statistic – a version of Cronbach's alpha or the KR 20 coefficient. The value obtained was 0.95, which is very high in GCSE terms. (Pollitt in Kimbell et al. 2009 p. 29).

The principal output from the pairs engine was a set of parameter scores on a scale representing the quality of each portfolio. Theoretically, if this were to be a GCSE assessment (QCA was interested to do that), then grades could be calculated along that scale, but readers should ignore these here as they were imposed through a separate process and were not calculated automatically within the algorithm (Fig. 1).

Many additional elements of data were available through the algorithm. Each portfolio had a 'portfolio-misfit' calculation that indicated which (if any) portfolios had caused judges to disagree; each portfolio score (parameter value) had a 'standard error' calculation which indicated the degree of confidence the system had about the accuracy of the parameter value (standard error reduces with more rounds); each judge had a 'misfit' calculation identifying their consensuality (judgement consistency coefficient) with the judge group as a whole; the judging interface allowed judges to enter notes about each portfolio that could be reviewed by the administrator; each judgement was timed and average times were available for each judge. These and many more features made this a very carefully monitored assessment process.

We should note a most important, and perhaps the most astonishing, features of the process. The judges were able to make reliable assessments of very complex multimedia portfolios of performance by just comparing pairs of portfolios and using holistic judgement. And moreover they could do it speedily. Our judges were mainly experienced teachers who were familiar with prevailing school-based assessment methods and in the post-judgement review they were clear about the contrast. All the judge comments are from Kimbell et al. (2009) pp. 69–72.

> Easier assessment; no need to calculate grades and points (RM).
> Speed of judging (VG).
> much, much faster … less scary (re individual marker impact on individual learner life chances)… get a whole view much more readily (RW).

They also commented on the holistic nature of the assessment.

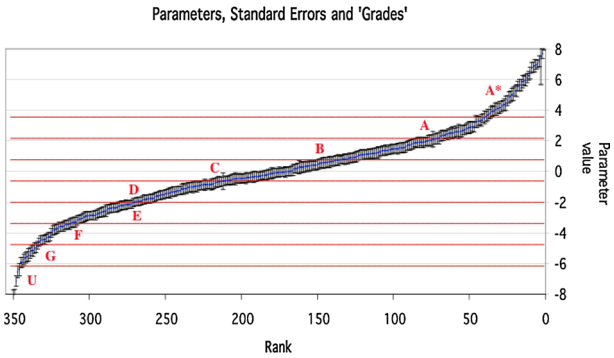

Fig. 1 The ranking output from the e-scape 'pairs engine'

> It gives more appropriate results than atomised approaches which can lead to inaccurate overall assessment especially when the overall attainment is more than the sum of the parts. This often happens when the various elements of a designing process come together in a successful outcome that outstrips the quality of work in any (or all) the parts of the process. (DP).
> GCSE marking relies heavily on a tick box assessment of a pupil's work. It can be frustrating when confronted with an excellent piece of designing and making that does not meet the exam board's criteria. Too often the linear pattern of coursework requires the assessor to jump back and forth to find the marks that a student deserves. The e-scape judging is so simple in comparison. (AM).
> One of the major strengths of holistic judgements I see is its flexibility… in which you can give credit to students for what they have actually done rather than whether they are able to "tick the boxes" to match a set of assessment criteria. (DW).

But additionally, they had a view about the enhanced fairness of the approach, since with the e-scape 'pairs engine', each portfolio was seen by many judges—not just their teacher and (perhaps) a moderator as would normally be the case with portfolio assessments at GCSE.

> The judging system feels to be fair; it doesn't rely on only one person assessing a single piece of work. It removes virtually all risk of bias…. It feels safe knowing that even if you make a mistake in one judgement it won't significantly make a difference to the outcome or grade awarded to the student as other judges will also assess the same project. Also knowing that the system automatically checks the consistency of the assessor's judgements again reinforces the feeling of fairness that this process brings. (DW).

Interestingly, despite its origins within a summative assessment setting for GCSE, the use of ACJ in classrooms has increasingly been focussed on formative assessment for learning. This possibility arose when (in discussion with some of the e-scape trial teachers) we asked the question 'what if we ask the learners themselves to be the judges?' When learners themselves are asked to review two pieces of work and to identify which they think is better, and why, it inevitably leads to discussion about what the learners mean by 'good' and 'better', and the concrete examples of work make it easier for learners to crystalise and articulate their own constructs of quality – as Polanyi argued 60 years ago. McLaren, (primary classrooms in Scotland) and Seery & Canty (Higher Education in Ireland) pioneered this strand of pedagogic applications (both in Williams and Kimbell [eds] 2012), and subsequently Hartell and Skogh (2015) in Sweden, Bartholomew et al. (2018, 2019) in USA, and Williams and Newhouse (2013) in Australia have pursued it.

A problem emerges

The use of comparative judgement within a schools assessment context gained increasing attention during the years of the e-scape project (2004–10). Shortly after the publication of the e-scape phase 3 (2009) report, Cambridge Assessment released a two-page summary of '*Rank ordering and paired comparisons—the way Cambridge Assessment is using them in operational and experimental work'* (Bramley and Oates 2010). This acknowledged the e-scape work at Goldsmiths and outlined their interest in the wider use of the ACJ method in educational assessment 'We are actively exploring their applicability to more general

investigations of comparability and to mainstream qualifications and assessments'. In 2012 Pollitt described the ACJ method in full, also acknowledging the e-scape work and specifically the development of the 'pairs engine' algorithm that, for the first time, ".. turned ACJ from a mere concept into a practical assessment system." (Pollitt 2012).

Subsequently, in 2015, Bramley produced a Cambridge Assessment Research Report specifically … *Investigating the reliability of Adaptive Comparative Judgment* (Bramley 2015). None of the substance of the Bramley paper relates to pedagogic or school-related matters, but rather it concentrates on the technical issue of whether the ACJ method produces the level of reliability that Pollitt claimed. At the same time Digital Assess, (formerly TAG Learning) was exploring the same reliability issue with the International Baccalaureate (IB)—based on ranking essays. The approach to testing reliability was essentially to run two assessment sessions with different teams of judges but with the same work. The two resulting rank orders were then correlated to see how far they matched. This approach had been suggested by Bramley in his 2015 paper (p14/15).

The approach was used by Jones and Alcock (2014) and by Jones et al. (2015) for exploring the reliability of the ACJ method in mathematics assessment using the same 'pairs engine' algorithm from the e-scape project. The former produced correlations between the two ranks of 0.86 and the latter 0.87, with (in both cases) Scale Separation Reliability (SSR) around 0.9. In this mathematics setting, all appeared to be operating as intended. But in the IB study—two sets of judges ranking essays on the 'Theory of Knowledge'—there was a less ideal result. Whilst the SSR figures were good (0.95), the correlation was modest (0.64) suggesting that the two resulting ranks of the same work were somewhat different.

In combination, it was Tom Bramley's initial challenge, linked to the problem observed by Digital Assess (the developers) that led to the launch of a thorough investigation of the reliability problem. A number of causes were considered for this result, including looking within the algorithm itself and particularly within that part of the algorithm that dealt with the distribution and matching of pairs.

A solution is proposed

The investigation was presented in a report "Addressing the issue of bias in the measurement of reliability in the method of Adaptive Comparative Judgment" (Rangel-Smith and Lynch 2018), and the challenge for Rangel-Smith and Lynch was two-fold. First, to find the source and extent of any bias arising from the general ACJ method and then to modify the 'pairs engine' algorithm so as to avoid the problem. Their report was launched at the PATT 36 International Conference: Ireland, and identifies the critical interplay of three factors.

(i) *The Standard Deviation (SD) of the items (portfolios)*: This is a measure of discriminability; the range of quality represented in the items. For low SD the quality range is small (say between − 2 and + 2) so discriminating judgements are difficult to make. But for bigger SD the quality range can be more like − 10 to + 10 and discrimination is much easier. Discriminability interacts with the expertise of judges as inexperienced judges will be less able to distinguish the quality of objects, whereas expert judges can achieve higher consensus and discrimination power.

(ii) *Level of adaptivity*: This concerns making use of the parameter value of a portfolio (estimated after the previous round) to choose the portfolio that it will next be

judged against. How far apart should they be on the putative scale of quality? The "gap" between portfolios becomes critical; a big gap (eg. comparing a very good piece with a very poor piece) produces easy judgements but less information for the system, whereas a small gap produces difficult judgements but more information for the system.

(iii) *Scale Separation Reliability (SSR)*: The resulting reliability statistic generated within the algorithm. It was this figure that Bramley challenged, claiming a 'bias' (reliability inflation) generated by the adaptivity.

In order to test the effect of adaptivity, Rangel-Smith and Lynch simulated the judging process many times with different starting hypotheses about the SD of the items and the level of adaptivity (the gap). In all the simulations, the "true quality" parameters of 100 objects were generated and were judged approximately the same number of times, and the session was simulated to last 40 rounds of judgments. Each hypothesis was simulated independently 40 times to reduce the effects of any statistical fluctuations in the result. Four central findings emerged.

First, for all hypothesized variables, the system shows bias in the reliability values in early rounds (less than 10), where there is not enough data. This arises because there is too much uncertainty for the information function to work effectively. This bias reduces as the data expands through later rounds. Second, an adaptive algorithm maximizes the performance of the system where there is higher discrimination (expert judges). On the other hand, the adaptivity process brings a bias in the measurement of the reliability in cases where the consistency in the judges is poor (inexperienced judges). Third, with non-adaptive (random) allocations, the bias of the "SSR" metric is smaller than in highly adaptive systems. Fourth however, the reliability performance of non-adaptive (random) selection is significantly poorer than in an adaptive system, where (with high discrimination) the system can reach a value of "True Reliability" 10 rounds earlier than the random allocation system. (See also Crompvoets et al. 2020; Bramley and Vitello 2019).

So adaptivity brings advantages and disadvantages. It produces a good SSR result quicker (fewer rounds of judging) especially when using expert judges. But in early judging rounds there will be SSR inflation, and when using inexpert judges the system will need significantly more rounds to generate a reliable result.

The Rangel-Smith/Lynch paper therefore recommended an approach that combines a 'controlled' level of adaptivity (1.5–2.5 logits) rather than the 0.67 of the original e-scape 'pairs engine' algorithm. This increased 'gap' between the selected portfolios makes it easier for judges to distinguish the winner. The second element of the solution involved the Standard Deviation (SD) of the portfolios. An SD of 0.0 logits is no discrimination (like tossing a coin); whereas 1.9–2.5 logits is medium discrimination (non-experienced judges), and 6.6 logits is high discrimination (expert judges). So the Scale Separation Reliability generated by the algorithm is based on three factors; the 'gap' between selected portfolios; the SD of the items (expertise of the judges), and the number of rounds of judging.

The key chart that illustrates this in the Rangel-Smith/Lynch paper is shown here. It shows the SSR value, as a function of the number of rounds in a judging session. Assuming an SD of 1.9 and a gap of 1.5 logits, then the red dots show the SSR in relation to the number of rounds. At 12 rounds the value is 0.87, at 15 rounds it is 0.89, and at 20 rounds it is > 0.9 (Fig. 2).

The Rangel-Smith/Lynch study therefore recommends as follows:

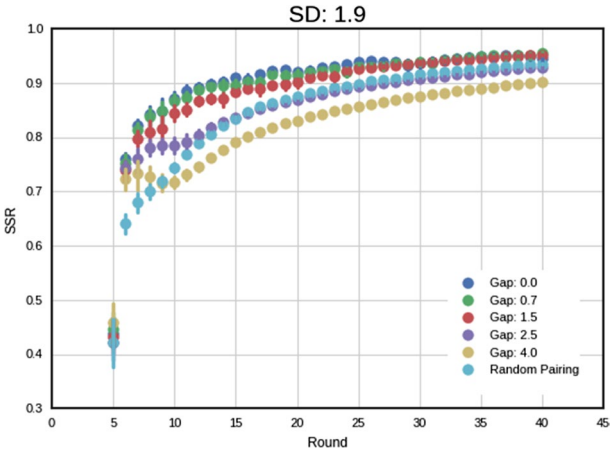

Fig. 2 Scale Separation Reliability (SSR) improves through 'rounds' of judging

> This study advises against using the highest level of adaptivity, where the pair of objects allocated are the closest in its parameter values…. It is recommended to run an ACJ session with a "controlled" level of adaptivity, translated by using a minimum "Gap" size to separate the allocated objects (1.5 to 2.5 logits)…. Depending on the chosen "Gap" value used in the session, there is a minimum number of rounds that have to occur before trusting the "SSR" metric as a reliability measurement. For a "Gap" vale of 1.5 logits there should be a minimum of 15 rounds, while for a separation of 2.5 logits it can be 12 rounds. (Rangel-Smith and Lynch 2018, p. 386).

In March 2020 and then again in Oct 2020 I had a discussions with Tom Bramley in which I asked him whether he thought the solution presented in the 2018 paper would deal with the concerns that he had raised in his 2015 paper. In extended and frank discussions he made three points in relation to the proposed solution. First, he thought that the SD values for discriminability should be set at what the paper describes as 'medium' discrimination. He believes a value of 1.9 is realistic. Second, he thinks the 'gap' is a sensible approach since (as the paper argues) it reduces the extent of the adaptivity and of any reliability inflation. He thought the gap value of 1.5 logits was sensible. Third, he agrees with Rangel-Smith/Lynch that this would then require 15 rounds of judging to remove the bias (reliability inflation) and 20 rounds to generate an SSR value > 0.9.

So the two-part question that launched this paper has its answers. Bramley and Rangel-Smith/Lynch agree that (i) the original 'pairs engine' algorithm did cause SSR inflation in particular conditions; with low SD, with inexpert judges and particularly in early rounds of judging. However they also agree that (ii) the new algorithm, in the conditions discussed here, will eliminate bias (reliability inflation) and will produce a secure SSR value of > 0.9.

Conclusions

Since the Rangel-Smith/Lynch paper was written, the company RM Education (a leading supplier of learning and assessment resources to the education sector) has acquired the original 'pairs engine' algorithm from Digital Assess and has already implemented the recommendations of the Rangel-Smith/Lynch paper. The RM Education product 'RM

Compare' now optimises the algorithm to offer the advantages of an adaptive approach to comparative judgement, whilst minimizing any reliability inflation.

Beyond the three specific points that Bramley made about the Rangel-Smith/Lynch proposal, he added an important, and more generic question. All parties agree that adaptivity works both ways … it enables a more efficient selection of pairs but it runs the risk of reliability inflation. (Crompvoets et al. [2020]). So Bramley's question is ".. is adaptivity worth it?". It's a balance … a trade-off. Why not just use random selection, which will give a secure result—but a bit slower (more rounds of judging)?

One of the features emerging from the research into the method of ACJ is that it illustrates the mistake of thinking in terms of *using* or *not-using* adaptivity. Such binary (on/off) thinking is less helpful than thinking about *degrees of adaptivity* (turning it up/down). All tools have advantages and disadvantages in the jobs that they do—but we do not say that therefore we won't use any tools. Rather we seek to use tools for what they are good at whilst taking steps to avoid their disadvantages. In the case of ACJ, an effective algorithm should seek to take advantage of the efficiency benefits of adaptivity (a proven useful tool in computer-based assessments) whilst controlling the tendency towards bias in early rounds.

In the years since the e-scape project at Goldsmiths first launched the ACJ 'pairs engine' into the arena of educational assessment, many more comparative judgement tools have emerged. 'No-More-Marking' have developed a comparative judgement tool particularly to help teachers and learners with writing tasks; Microsoft research have developed 'TrueSkill' using ranking in the context of doctors' judgments of videos of patients; in Belgium, Digital Platform for Assessing Competencies (D-PAC) has a digital tool to help in the assessment of video and image; and Bramley and his colleagues at Cambridge Assessment have developed 'Cambridge CJ scaling'. Even the UK Government—in the form of Ofqual—is 'running pilot studies involving comparative judgement methods for capturing expert judgement for the purpose of standard maintaining' (Ofqual 2019).

The method of Adaptive Comparative Judgement has existed for only a very short time. It is about 15 years old and has emerged into fields of educational scholarship (assessment and pedagogy) that have histories spanning centuries. In that 15 years it has started to open many new doors at the interface of assessment with learning.

References

Angoff, W. H. (1974). Criterion-referencing, norm-referencing and the SAT. *College Board Review, 92*(Summer), 2–5, 21.

Bartholomew, S. R., Strimel, G. J., Garcia Bravo, E., Zhang, L., & Yoshikawa, E. (2018). Formative Feedback for Improved Student Performance through Adaptive Comparative Judgment. In *Paper presented at the Paper presented at the 125th ASEE conference*, Salt Lake City, Utah.

Bartholomew, S. R., Strimel, G. J., & 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.
Black, P. (1993). The shifting scenery. In P. O'Hear & J. White (Eds.), *Assessing the National Curriculum*. London: Paul Chapman.
Bramley, T. (2015). *Investigating the reliability of Adaptive Comparative Judgment*. Cambridge, UK: Cambridge Assessment Research Report.
Bramley, T., & Oates, T. (2010). *Rank ordering and paired comparisons—the way Cambridge Assessment is using them in operational and experimental work*. Cambridge, UK: Cambridge Assessment.
Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(1), 43–58.
Crompvoets, E. A., Béguin, A. A., & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics, 45*(3), 316–338.
Department of Education and Science (DES) (1988). *Task Group on Assessment and Testing – A Report*. London, Department of Education & Science
Dyson, J. (2001). *Against the odds: An autobiography*. Knutsford Cheshire: Texere Publishing.
Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713–734.
Eisner, E. W. (1981). *The methodology of qualitative evaluation. The case of educational connoiseurship and educational criticism*. Stanford: Stanford University.
Finn, M. (Ed.). (2015). *The Gove Legacy: Education in Britain after the Coalition*. Basingstoke UK: Palgrave Macmillan.
Gipps, C. (1994). *Beyond Testing: Towards a theory of educational assessment*. London & New York: Routledge Falmer.
Golley, J. (2009). *Jet: Frank Whittle and the Invention of the Jet Engine*. Liphook Hampshire: Datum 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.
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 mathematical problem solving using comparative judgment. *International Journal of Science and Mathematics Education, 13*(1), 151–177.
Kimbell, R. (1997). *Assessing Technology: International Trends in curriculum and assessment*. Buckingham UK: Open University Press.
Kimbell, R., Wheeler, T., Miller, S., & Pollitt, A. (2007). *E-scape portfolio assessment: Phase 2 Report (p57/58)*. Technology Education Research Unit, Goldsmiths University of London.
Kimbell, R., & Stables, K. (2007). *Researching Design Learning. Issues and findings from two decades of research and development*. (See esp Ch 5), Dordrecht NL: Springer
Kimbell, R., Wheeler, T., Stables, K., Sheppard, T., Martin, F., Davies, D., et al. (2009). *E-scape portfolio assessment: Phase 3 report*. Technology Education Research Unit: Goldsmiths University of London.
Marshall, B. (2017). The politics of testing. *English in Education, 51*(1), 27.
Office of Qualifications and Examination Regulation (Ofqual) (2019) *Improving awarding: 2018/2019 pilots* (Curcin, M., Howard, E., Sully, K., and Black, B.) Coventry https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/851778/Improving_awarding_-_FINAL196575.pdf.
Polanyi, M. (1958). *Personal Knowledge*. London, UK: Routledge & Kegan Paul.
Polanyi, M. (1966). *The tacit dimension*. London, UK: Routledge & Kegan Paul.
Pollitt, A. (2004) Let's stop marking exams. In *Paper presented at the IAEA Conference*, Philadelphia, June 2004.
Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281.
Pollitt, A., Derrick, K., & Lynch, D. (2009). *Single level tests of KS2 Writing: The method of paired Comparative Judgement*. Sherston, Malmesbury, Wiltshire: BLI Education.
Rangel Smith, C., & Lynch, D. (2018). Addressing the issue of bias in the measurement of reliability in the method of Adaptive Comparative Judgment. In *PATT36 international conference. Research & Practice in technology education: Perspectives on Human Capacity and Development* (pp. 378–388) Athlone Ireland, 18–21st June 2018.
Schon, D. (1983). *The reflective practitioner: How professionals think in action*. New York, NY: Basic.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286.

Thurstone, L. L. (1959). *The measurement of values*. Chicago, Illinois: University of Chicago Press.
Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.
Wiliam, D. (1998a). Enculturating learners into communities of practice: Raising achievement through classroom assessment. In: *Paper presented at the European conference for educational research*, University of Ljubljana, Slovenia September 17th to 20th 1998.
Wiliam, D. (1998b). Construct-referenced assessment of authentic tasks: Alternatives to norms and criteria. In: *Paper presented at the 24th annual conference of the international association for educational assessment—testing and evaluation: confronting the challenges of rapid social change*, Barbados, May 1998.
Williams, P. J., & Kimbell, R. (eds) (2012) Special Issue on 'e-scape'. *International Journal of Technology and Design Education, 22*(2)
Williams, P. J., & Newhouse, C. P. (Eds.). (2013). *Digital representations of student performance for assessment* (pp. 169–195). Rotterdam NL: Sense Publishers.