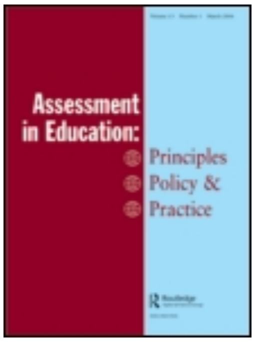


Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

The method of Adaptive Comparative Judgement

Alastair Pollitt

To cite this article: Alastair Pollitt (2012) The method of Adaptive Comparative Judgement, Assessment in Education: Principles, Policy & Practice, 19:3, 281-300, DOI: 10.1080/0969594X.2012.665354

To link to this article: https://doi.org/10.1080/0969594X.2012.665354

Published online: 15 Mar 2012.

Submit your article to this journal

Article views: 1607

Citing articles: 39 View citing articles

Assessment in Education: Principles, Policy & Practice
Vol. 19. No. 3, August 2012, 281–300

The method of Adaptive Comparative Judgement

Alastair Pollitt*

Cambridge Exam Research, Cambridge, UK

(*Received 17 July 2011; final version received 7 February 2012*)

Adaptive Comparative Judgement (ACJ) is a modification of Thurstone's method of comparative judgement that exploits the power of adaptivity, but in scoring rather than testing. Professional judgement by teachers replaces the marking of tests; a judge is asked to compare the work of two students and simply to decide which of them is the better. From many such comparisons a measurement scale is created showing the relative quality of students' work; this can then be referenced in familiar ways to generate test results. The judges are asked only to make a *valid* decision about quality, yet ACJ achieves extremely high levels of *reliability*, often considerably higher than practicable operational marking can achieve. It therefore offers a radical alternative to the pursuit of reliability through detailed marking schemes. ACJ is clearly appropriate for performances like writing or art, and for complex portfolios or reports, but may be useful in other contexts too. ACJ offers a new way to involve all teachers in summative as well as formative assessment. The model provides strong statistical control to ensure quality assessment for individual students. This paper describes the theoretical basis of ACJ, and illustrates it with outcomes from some of our trials.

Keywords: judgement; marking; reliability; assessment methods

Introduction

Adaptive Comparative Judgement (ACJ) uses professional judgement by teachers to replace the marking of tests. A judge is asked to compare the work of two students and simply to decide which of them is the better. From many such comparisons a measurement scale is created showing the relative quality of each student's work; the scale can then be referenced in familiar ways to generate test results.

ACJ is a scoring method applicable in many educational contexts including those where inter-marker reliability is often a problem. The judges are asked only to make a *valid* decision about relative quality, yet ACJ achieves extremely high levels of *reliability*, often considerably higher than any practicable operational marking process has ever achieved.

ACJ does not require any change from the point of view of students. Once the test is over, their work is digitised if necessary, and uploaded for presentation online to judges. The procedure requires little training, and has proved very popular with assessors and teachers in several subjects, and in several countries.

This paper describes the theoretical basis for the method, and illustrates the outcomes from some of these explorations. The question is no longer whether the method can work, but how widely could it, or should it, be used.

*Email: alastair@camexam.co.uk

ISSN 0969-594X print/ISSN 1465-329X online

http://dx.doi.org/10.1080/0969594X.2012.665354
http://www.tandfonline.com

History

Thurstone and CJ

ACJ is derived from the well known 'method of comparative judgement' (CJ) originally proposed by Louis L. Thurstone for work in psychophysics (Thurstone 1927a). Although he noted that the method might also be used for quantifying purely subjective properties such as the perceived quality of things 'such as handwriting specimens, English compositions, sewing samples, Oriental rugs' (Thurstone 1927b, 376), he does not seem to have pursued these educational ideas very far, preferring to explore instead the assessment of beliefs and attitudes (e.g. Thurstone 1931).

Andrich (1978) showed that Thurstone's model could be restated as a Rasch logistic model with essentially the same results. This is also known as the Bradley-Terry-Luce model (Bradley and Terry 1952; Luce 1959), and it is in this form, rather than Thurstone's original normal-distribution form, that CJ is generally used today.

This CJ model can be written:

$$prob(\text{A beats B} \mid v_a, v_b) = \frac{exp(v_a - v_b)}{1 + exp(v_a - v_b)} \tag{1a}$$

Two 'objects' A and B, which may be (amongst other possibilities) portfolios, exam scripts, or recordings of performances, are compared according to some criteria for 'goodness'. If the true quality of each is given by their parameters v_a and v_b, then Equation 1 calculates the probability that object A will be chosen as the better of the two. The meaning may be clearer if this is changed from probability to odds, and converted to a logarithm:

$$log\,odds(A\,beats\,B|v_a, v_b) = v_a - v_b \tag{1b}$$

The difference between the two quality parameters is equal to the log of the odds that A will be judged better than B. If these two objects are compared several times, by different judges, the resulting data allow us to estimate the gap between the two parameters. In practice several, or many, objects will be compared, giving many equations like Eq 1, and many estimates of the parameter differences. Rasch (1960) showed that a straightforward maximum-likelihood (ML) procedure could optimise a set of parameters as follows. For a particular object, A, and a set of data:

The score for that script, S_a, is given by:

$$S_a = \sum_{i \neq a}^{n} X_{a,i} \qquad [\text{X} = 1\,if\,A\,wins,\ 0\,if\,B\,wins] \tag{2}$$

The expected score is the sum of the probabilities for all comparisons with A:

$$E(S_a) = \sum_{i \neq a}^{n} p_{a,i} = \sum_{i \neq a}^{n} \left[\frac{(exp(v_a - v_i)}{1 + exp(v_a - v_i)} \right] \tag{3}$$

If preliminary estimates are made for all the parameters v_i, then the predictions $E(S_a)$ can be compared to the observed scores, and Newton's method used to update the parameter estimates:

$$v'_a = v_a + \frac{S_a - \sum_{i \neq a}^{n} p_{a,i}}{\sum_{i \neq a}^{n} (p_{a,i}(1 - p_{a,i}))} \tag{4}$$

The estimates normally stabilise after about three iterations. Eq 3 is critical since, as Rasch (1960) explained, it implies that the number of wins an object makes in a data set is a *sufficient* statistic for estimating its parameter. The consequence is that deviations from these expectations give us very powerful statistical quality control over the assessment procedure. This point will be developed later. Standard errors for these parameters are also calculated, in the standard ML way. The *information* in each decision is calculated, and summed over all decisions involving each script:

$$\text{Information on script A:} \quad I_a = \sum_{i \neq a}^{n} (p_{a,i}(1 - p_{a,i})) \tag{5}$$

from which the standard error for the estimate of v_a is given by:

$$\text{Standard error for script A:} \quad se_a = \frac{1}{\sqrt{I_{a.}}} \tag{6}$$

CJ in education

To my knowledge, the first modern application of CJ in education was an investigation of a foreign language speaking assessment (Pollitt and Murray 1993). This was original, too, in using Kelly's Personal Construct Theory (Kelly 1955) to explore what aspects of the performance judges were focusing on as they made their decisions; this combination of methods has been used in several comparability studies since then (e.g. Fearnley 2000; Gray 2000; discussed in Pollitt, Ahmed, and Crisp 2007).

Thurstone's method was introduced to the British examination boards in 1995, as a tool for exploring the comparability of standards between different boards, and it was first reported in D'Arcy (1997). For a complete review of these inter-board studies, see Bramley (2007). One significant conclusion from this work is that, in some contexts at least, judges are able to make consistent comparisons between performances on *different* tasks or in *different* examinations, meaning that CJ may be able to provide a test-equating function at the same time as a scoring one. Whether judges can do this in any particular context would, of course, need to be examined before it could be relied on to maintain a standard.

Heldsinger and Humphrys (2010) showed that CJ could be used to obtain reliable internal assessments from teachers, a domain in which reliability tends to be low and costs rather high.

ACJ

Pollitt (2004) was the first to outline how CJ could be used as an alternative to marking in a wide range of examinations and tests, suggesting the possibility of running it adaptively. The first opportunity to try this came with an innovative assessment in Design & Technology, in which pupils used hand-held electronic devices to collect evidence of their design process as they developed a prototype product. With help from an examination board the team had developed a mark scheme for the resulting e-portfolios, but were not satisfied that such an analytic approach could properly measure the evidence of what was most important in D&T education. They wanted a scoring method that used, 'the holistic approach to assessment that we have always advocated for design & technology' (Kimbell et al. 2007, 91). Their project was based on a concern for authenticity in assessment, and validity which: 'for the purposes of our work may be summarised as the extent to which the activity, as we have designed it, represents "good" design & technology' (28). This was an ideal test-bed for ACJ.

We developed a web-based system in which pairs of portfolios are presented on-screen to a judge who studies both and chooses the 'better' one; these judgements are accumulated into a data file, and a rudimentary adaptive system was created for choosing the best pairs to present. The improved versions of these are described in Kimbell et al. (2009). With an assessment procedure designed wholly with validity in mind, and with no explicit marking scheme, the concern was that the results might not be sufficiently reliable to be used for certification. Kimbell et al. (2009) report the outcome when 28 judges made 3067 judgements to assess 352 e-portfolios: the overall score reliability was 0.96.

What this means will be discussed later, but this level of reliability is almost certainly higher than any operational marking system could have achieved. It was clear that ACJ *can* deliver highly reliable assessment of the relative quality of students' work in some contexts, and was worth exploring further.

Estimation in ACJ

The real power of CJ is only realised when it is made adaptive. In the next sections I will discuss the principles of Adaptive Comparative Judgement (ACJ), and illustrate with evidence from one large study, with occasional reference to others.

In accordance with the British convention, I will generally use the term 'script', but the reader should remember that the objects being compared are digital files of any kind – word-processed documents, scans, images, audio or video recordings, or any combination of these. The comparability studies mentioned earlier were designed as experiments, with no more than 30 scripts in each, and it would not have been practicable to extend this to assess hundreds or thousands of students.

ACJ is an online procedure, for reasons that will become clear. It operates through a website, and the system used in these studies was developed by TAG Learning in collaboration with the author. TAG Learning is a division of Sherston Software Ltd.

The 'primary writing' pilot

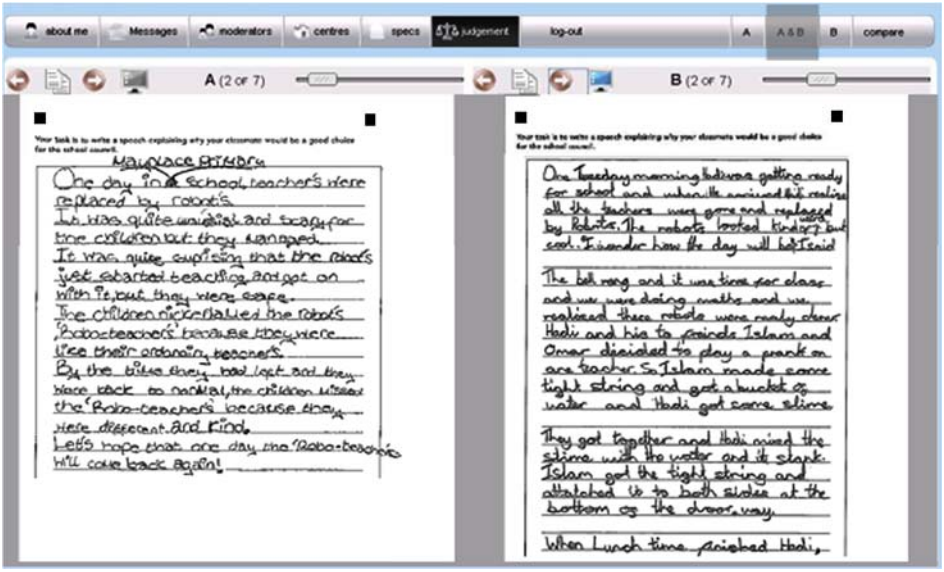
This study involved 1000 writing samples from children aged 9–11 years, selected as being approximately at Level 4 for English in England's national curriculum system – rather more homogeneous than a full one-year cohort. They each wrote two pieces, one persuasive and one creative narrative, and the scripts were scanned and uploaded to a website operated by TAG Learning. Before the upload, code numbers were given to the scripts to preserve anonymity. The aim was to test the feasibility of ACJ as an operational scoring system for large-scale writing assessment.

54 judges were enlisted, of whom 31 were professionally trained test/examination markers; the others were teachers with no specific marker training. Their preparation consisted of two hours or so of familiarisation with the website and judgement interface, and about two hours of practice and discussion in making comparative judgements. They were then issued with IDs and passwords, and left to carry out their work on home computers at any time during the following 10 days.

Collecting data

The judge's task was intrinsically simple: just to consider two scripts and decide which one was 'better'. When they logged in, they were presented with two scripts in a panel like that shown in Figure 1. It includes various tools to facilitate making judgements by reading one or both scripts at a time, zooming in and out and, perhaps, making notes or annotations. They then chose a 'winner', and were given a new pair to judge – until all 1000 scripts had been scored satisfactorily.

The data file consisted of a list of the comparisons made and the judgements. For example,'*3 231 452 231*' meant that *Judge 3* compared *Script 231* and *Script 452*, and judged that *Script 231* was the 'winner'. In the first round scripts were matched randomly, and 500 matches were needed so that every one of the 1000 scripts was judged once – after this every script had a 'score' of 1 (for a 'win') or 0 (for a 'loss'). In the second round, scripts were matched by score, so that winners were compared to winners and losers to losers – this is more efficient than just random matching. Matching by score was repeated for the third and fourth rounds, but from then on there was sufficient data to apply Thurstone's model exactly, and to estimate more accurate scores, now called *quality parameters*, by iteratively re-solving Equation 4 for each script, using all the data collected up to that point. These parameters were re-estimated whenever another round of 500 comparisons had been



Your task is to write a speech explaining why your classmate would be a good choice for the school council.

Mayplace Primary
One day in a school, teacher's were
replaced by robot's.
It was quite unusual and scary for
the children but they managed.
It was quite suprising that the robots
just started teaching and got on
with it, but they were safe.
The children nicknamed the robot's
'Robo-teacher's' because they were
like their ordinary teachers.
By the time they had left and they
were back to normal, the children missed
the 'Robo-teacher's' because they
were different and kind.
Let's hope that one day the 'Robo-teacher's
will come back again!

Your task is to write a speech explaining why your classmate would be a good choice for the school council.

One Tuesday morning Hadi was getting ready
for school and when he arrived he realised
all the teachers were gone and replaced
by Robots. The robots looked kind of weird but
cool. "I wonder how the day will be?" I said

The bell rang and it was time for class
and we were doing maths and we
realised these robots were really clever.
Hadi and his two friends Islam and
Omar decided to play a prank on
our teacher. So Islam made some
tight string and got a bucket of
water and Hadi got some slime.

They got together and Hadi mixed the
slime with the water and it stank.
Islam got the tight string and
attached it to both sides at the
bottom of the door way.

When Lunch time finished Hadi,

Figure 1. The judging interface.

added to the data file. The new, improved, quality parameters were then used to set up the next round of comparisons, with each script matched with another of reasonably similar quality: these pairings were immediately sent out for judgement. Round by round, the estimates became more accurate as they were based on more and more data.

This matching is the ‘adaptive’ aspect of ACJ, and it is analogous to computer-adaptive testing (CAT: e.g. Straetmans and Eggen 1998). It is well known that CAT is considerably more efficient than traditional testing (e.g. Wainer 2000), but there are several significant problems with it (e.g. Reckase 2011; van der Linden and Glas 2000), such as transferring to on-screen testing, the need for a very large bank of items, for pre-calibration, for optimal selection algorithms, and for strategies to avoid over-exposure of some items: above all, different students take different tests. None of these problems apply to ACJ because it only changes the *assessors’* experience; the *students’* experience is not affected in any way by the switch from marking to judging. It is therefore much simpler to introduce ACJ than CAT.

Reliability

As judgements come in, a chief assessor or manager is able to monitor progress, and the system generates periodic reports. One key feature is the calculation of reliability statistics at the end of each round. These are calculated as follows. First, the standard deviation of the current parameter values is calculated, and the root mean square of the estimation errors. The *separation coefficient* is defined as the ratio of these:

$$\text{SC:} \qquad \mathrm{G} = \frac{sd_v}{rmse} \tag{7}$$

This is converted into a form analogous to Cronbach’s alpha (Wright and Masters 1982):

$$\text{Reliability:} \qquad \alpha = \frac{G^2}{(1+G)^2} \tag{8}$$

While alpha is the familiar reliability statistic, and is reported in the TAG system, SC is more intuitively meaningful, being the ratio of the spread of the parameters to the average uncertainty in their position. Figure 2 shows how the SC grew after each round in the primary writing study.

Any estimation procedure needs some data to get started. In this study, the first six rounds used a simplified process known as ‘Swiss tournament’, a procedure used in chess competitions. The first round consisted of random pairs, but in rounds 2 to 6 all pairings involved two scripts with the same number of wins, providing very rough sorting by quality. After round 6 the method in Eq 1 was used, and the reliability rose sharply, reaching an alpha of 0.96 after 16 rounds:

Summary of the scale properties:

Standard deviation of the Object parameters	=	3.85
Root mean square estimation error	=	0.74
Separation coefficient	=	5.20
Reliability coefficient	=	0.96

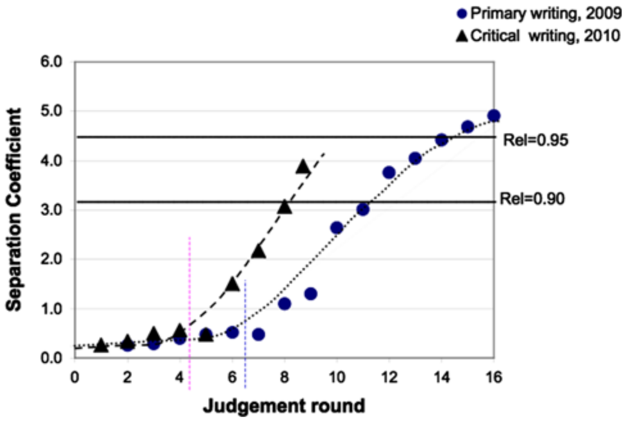


Figure 2. How reliability increases round by round.

This is an extraordinarily high value for the reliability of a writing test; for comparison Royal-Dawson reported values ranging from 0.75 to 0.80 in a study of marking reliability in English national curriculum writing at age 14 (Royal-Dawson 2005; Royal-Dawson and Baird 2009).

It is worth considering in more detail. We were dealing with 11-year-olds rather than Royal-Dawson's 14-year-olds, but otherwise the tests were similar. Rather than comparing marks given by about 16 markers, the ACJ figure is based on comparisons of each script to about 16 others, made by about 12 judges. The judges simply decided that one script was 'better' or 'poorer' than another. The notion of 'better' was formally defined as 'showing more evidence of what it means to be good at writing', with reference only to the 160-word-long official statement of '*The Importance of English*' (QCDA 1999–2011; notice this refers to *English*, not specifically to *writing*). Most judges, when asked, said they rarely or never referred to it, relying instead on their professional skill and experience. Yet they were extremely consistent, as the reliability shows (there is a more detailed consideration of individual consistency in the next section). Not only did they rank-order the scripts consistently, but the final measurement scale represented the average construct of all the judges.

Five important threats to the reliability of writing assessment can be identified, and the ACJ coefficient deals with them as follows. Differences in the *severity* of judges are removed methodologically since the comparative process cancels out each judge's standard; the same is true for differences in judges' ability to *spread out* scripts of similar quality. Traditional marker reliability studies that are based on correlations between markers ignore these two sources of error, and therefore report coefficients that are spuriously high compared to ACJ reliability. Of the three traditional sources of marker error in writing tests, only differences in *rank order* by judges remain as a source of error. Fourth, *internal inconsistency* (candidate by component interaction) also contributes to the ACJ estimation error since a judge's evaluation of each script must deal with any variability across component traits of writing skill and across the two tasks. Finally, ACJ takes no account of *equivalent forms* unreliability, but this, regrettably, is ignored in almost all reliability research.

The most comparable picture of reliability in the marking of writing comes from Baker et al. (2008). While our primary writing study was a pilot of an operational system, theirs was an experimental investigation of reliability in the tests for 14-year-olds, in which three groups of five markers each marked common batches of 40 scripts. The report gives an average *relative* coefficient of 0.93, which was considerably higher than those found by Royal-Dawson, as reported above. Baker et al. (2008) also reported *absolute* reliability measures that averaged 0.85.

The relative/absolute reliability distinction comes from Generalisability Theory (Cronbach et al. 1972; Brennan 2001): relative reliability refers to the consistency of rank ordering amongst raters, but does not consider any absolute differences in severity between raters as part of its error term, so allowing each judge to rate on a scale reflecting their own expectations of standards. Absolute reliability estimates how well each judge conforms to the average of all the judges in standard as well as in rank order. In Baker et al. (2008) the absolute reliability estimates of 0.85 include three extra error terms, arising from differing marker standards, internal inconsistency, and the interaction of these two. Since the estimation errors in ACJ take all three of these into account as well as rank order differences, the ACJ coefficient is more like an absolute than a relative coefficient. Our value of 0.96 would have been exceptional were it just a relative coefficient but in comparison to their absolute coefficient it is extremely high.

The second plot on Figure 2 shows results from a different study, carried out a year later, of English critical writing. In this test, students had a wide choice of which questions to answer, and also which book, poem or play to use in their critical response. It had proved extremely difficult to mark reliably. Using ACJ, four judges with experience in marking judged 110 scripts and reached a reliability of 0.93 after less than 9 complete rounds. As the figure shows, we had reduced the ‘Swiss’ component from six to four rounds, and had improved other features of the pairing algorithm.

The ACJ process can be monitored as it happens, and terminated when the reliability is considered high enough, or when some pre-determined number of judgements has been made, or more elaborate stopping criteria can be used. This flexibility highlights an important principle: every judgement is independent, and it is always possible to add more data without compromising the integrity of the estimation procedure. More data can be collected specifically to increase the information about particular scripts, for example, as will be discussed in the next sections.

Robustness

It is worth noting that, in Thurstone’s derivation of the CJ method, several or many judgements were made of every possible pairing. For assessment this would make the amount of work needed proportional to n^2, where n is the number of scripts. Adaptivity makes the work proportional to n, by discarding all the possible pairings where the information $I_{j,a,b}$ would be very small. In the primary writing study, there were 499,500 possible comparisons, but only 8161 decisions were made, or 0.016 per possible pairing. To create a consistent measurement scale with so little data shows the remarkable robustness of ACJ.

Quality control

The strength, or simplicity, of the model in Equation 1 enables powerful statistical quality control. The basis for this is that every judgement can be evaluated for consistency. The *residual* is the difference between the observed result and the probability predicted, by Equation 1a, from the final parameter estimates:

$$\text{Residual}_{\text{j,a,b}}: \qquad X_{j,a,b} - p_{a,b}|X = 1, 0 \tag{9}$$

The residuals can provide information about the consistency of each of the scripts involved, A or B – or about the judge J. First it is standardised and squared:

$$\text{SqStdRes}: \qquad z^2_{j,a,b} = \frac{\text{Residual}^2_{j,a,b}}{p_{a,b}(1 - p_{a,b})} \tag{10}$$

These SSR values can be averaged across any focus of interest.

Judge misfit

For example, all of the SSRs from a particular judge can be averaged, which gives a measure of the quality – the consistency – of that judge's decision-making. Since Eq 6 effectively defines an element of a chi-square statistic, the average of them for each judge can be interpreted as a mean chi-square. We use an *information-weighted mean square* for this purpose – what is known as the *Infit* mean square in Rasch analysis (Wright and Masters 1982). Each SSR is weighted by the amount of information in the judgement it came from, then the sum of these is divided by the sum of the weights:

$$\text{Weighted mean square, Judge J}: \qquad wms_j = \frac{\sum_{j..} p_{a,b}(1 - p_{a,b}) * z^2_{j,a,b}}{\sum_{j..} p_{a,b}(1 - p_{a,b})} \tag{11}$$

The measures of consistency are reported in a table like Table 1 from the primary writing exercise (though for illustration here they are sorted into order of increasing *wms*).

Conventionally, any figure greater than the mean plus two standard deviations – 1.14 here – is taken to indicate significant misfit. In this case, only Judge '46 N-M 2' exceeded the criterion. To a significant degree, this judge interpreted the trait 'good writing' differently from the others. From a research point of view it might be interesting to explore why; feedback to '46 N-M 2', with perhaps a visit from a

Table 1. Judge misfit statistics.

Judge	N Judgements	wms
20 N-M 3	120	0.91
33 M 2	186	0.93
38 N-M 2	98	0.95
18 M 3	186	0.96
…	…	…
49 N-M 3	115	1.11
6 N-M 2	90	1.12
7 M 2	179	1.13
46 N-M 2	93	1.24
Mean =	151.1	1.02
SD =		0.06

local adviser, might ensure that this judge will not continue to train students according to criteria that assessors in general do not value. But from a measurement stance, reliable assessment requires judges that are consistent with each other, and at any time, misfitting judges can be stopped, and their share of decisions shared amongst the other judges. If necessary, their data can be removed from the file.

In general, misfit statistics like these should be interpreted relatively rather than absolutely; that is, the largest statistics will indicate which are the most misfitting judges, or scripts, but are not a basis for saying a particular judge or script does or does not misfit. This is more important in an adaptive context than otherwise: since the differences $v_a - v_b$ are targeted by the algorithm to be quite small, the residuals from which the statistics are calculated are not distributed as widely as they would be in traditional test analyses. Experience has shown that the weighted mean square is more robust than some other statistics in these circumstances.

Script misfit

A similar statistic is calculated for every script. A misfitting script is one that judges find difficult for some reason, one that some judges value more than others do. Again, the research potential is obvious, but to ensure fair assessment the system should send misfitting scripts out for extra judgements – perhaps, in the end, to specific 'senior' judges who carry the responsibility for difficult decisions. Of course, they do not need to know that any particular script is a 'problem' one, if they are making other, more routine judgements too.

Focusing ACJ

It is also possible to use the standard errors to decide when more information is needed on certain scripts. If there is a critical value representing pass/fail or a grade boundary then sending out scripts near it for more data will reduce their measurement error and so increase classification accuracy, again a property that ACJ shares with CAT. *Focusing* the pairing algorithm on scripts that need more information and away from those that are already well enough measured has the potential to significantly increase the efficiency of ACJ.

Bias

The misfit statistics for judges and scripts are special cases of the more general issue of *bias*, the presence of any systematic excess variation in parts of the data set. It is possible to summarise into a weighted mean square the residuals from any sub-set of the data; various hypotheses of a possible source of bias can be tested, at least approximately, by the mean square. For example, in a CJ comparability study, Pollitt and Elliott (2003) investigated the hypothesis that a judge might favour scripts from their 'home' board against scripts from other, 'away', boards. By comparing the weighted mean squares for 'home/away' pairs to those for 'away/away' pairs they concluded that only one out of the nine judges involved showed a statistically significant but substantively small amount of 'home' bias. The statistical power of this analysis of course depends on the number of judgements made in each sub-set, but the numbers of judgements collected in most ACJ studies so far are large enough to support interpretations like this.

Range of application

ACJ has also been applied to the assessment of projects in Science (Davies 2008), and fieldwork in Geography (Martin and Lambert 2008), both of which involved extended reports that matched the D&T portfolios in complexity. In a very different context, Jones, Pollitt, and Swan (in review) have shown that judges are able to use ACJ in assessing the current exams in Mathematics that are sat by most 16-year-olds in England and Wales, even though these consist mainly of 1, 2 or 3 mark questions designed for reliable marking. The aim in the next phase of that project will be to show that larger mathematics tasks can be used to assess everything that is important in learning maths with high reliability and, it is hoped, higher validity.

The non-adaptive CJ method has been used more widely, especially in the inter-board comparability studies referred to earlier. These have included English, foreign languages, history, sociology, citizenship, sciences and mathematics, business and media studies; the examinations were designed for students ranging from 14 to 18 years. In each of these, a whole examination was judged, which might mean about three components in different formats from each student; the instruction to the judges was simply to 'nominate which script they think represents the higher level of achievement' (Fearnley 2000; Greatorex, Elliott, and Bell 2002). As Bramley (2007, 247) states, CJ is now 'the preferred method in inter-board comparability studies' in England, which suggests ACJ might, in principle at least, be useful in a wide range of contexts.

Costs

The cost of running an ACJ system for certification depends largely, of course, on the time taken to make enough judgements. In the examples described, the number needed for each script was reduced from 16 to about 9, by improvements to the pairing algorithm, and several recent trials have shown that a reliability of 0.93 is usually reached in fewer than 10 rounds. For ACJ to take the same total time as marking, this means that each decision must take no more than one fifth of the time marking needs. If we include the time needed for training markers, for second marking to monitor markers' performance, and for remarking when things go wrong, this will be more like one quarter.

In the primary writing study, as Figure 3 shows, judges varied greatly in speed; the fastest averaging 88 seconds for 186 judgements against the slowest's 528 seconds for 187. Yet there was no relationship between speed and consistency, suggesting that a certification system should use speed as one factor in selecting judges. (Note, the times measured are from presenting a pair to recording a decision, and so may include 'down time'.) Looking only at the faster half, the mean time was 3 minutes 15 seconds per decision, which compares well to the 15–20 minutes we were advised was normal for marking these tests. A pilot study should consider this in each new assessment context.

Advantages and potential of Adaptive Comparative Judgement

Validity

ACJ is a method for evaluating students' performances or the products of their work; it is not a method for writing test questions and mark schemes. It can be used to evaluate the evidence from many current tests without any change whatsoever to

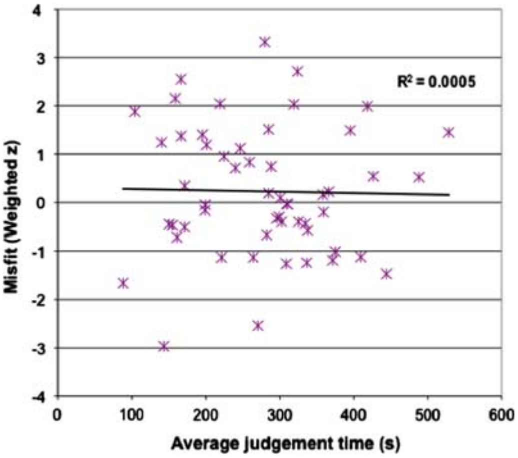


Figure 3. Judge misfit vs. judgement time.

the preparation, administering, or answering of the test. Because it relies on the judges keeping their minds constantly on the aim to maximise validity, using only an Importance Statement (or its equivalent) for guidance, it is reasonable to expect validity to be at least as high with ACJ as with any marking scheme.

By removing the need to design tasks that can be marked reliably, however, ACJ allows test designers to use tasks that may greatly improve the overall validity of their assessment process. Three simple principles can be said to control task design:

(1) the job of a test task is to elicit evidence of what is important;
(2) the job of the question is to communicate that task faithfully to the students;
(3) the job of scoring is to quantify reliably the evidence of what is important.

With ACJ instead of marking, the first of these principles need not be compromised by the third. The method enables the quantification to be extremely reliable if the evidence collected is, indeed, evidence of what is considered important in learning the subject skills and content, without restricting the task designers to what can be marked reliably. Many sources of construct-irrelevant variance are related to test-taking strategies or the need to present responses in particular forms that marking can deal with; these can be avoided when this constraint is removed. ACJ is probably more effective with – though it does not demand – holistic tasks that are as authentic as possible, tasks that aim directly to assess 'what it means to be a good scientist, writer, singer, or historian'. Test designers are therefore much freer to use whatever format they think will elicit the best evidence of the trait they are interested in measuring, keeping their focus on validity rather than reliability.

Nevertheless, if it is considered important to judge scripts analytically as well as or rather than holistically, ACJ can still be used for each component, and may still prove more reliable than marking.

Reliability

Probably the most immediate advantage of ACJ is in making reliable the assessment of skills that are currently problematic. These include various kinds of writing, from creative narrative to long essays in Politics or History where the complex mixture of criteria for content and for quality make agreement on marks difficult to reach, and a single marker – or even two – cannot be considered reliable enough for high-stakes assessment. It is even more complex to evaluate the range of kinds of evidence in project reports or portfolios, and very difficult for any mark scheme to anticipate what will appear, especially when 'creativity' is an element in the construct. For this kind of complex evaluation, no systematic procedure can be better than the human brain. But the brain has evolved to make comparisons, and is relatively poor at the kinds of numeric judgements we ask markers to make using rating scales. In an analysis of the current role of judgement in educational assessment, Laming (2004, 9) declared, 'There is no absolute judgment. All judgments are comparisons of one thing with another', and argued that examinations should therefore be limited to more objective questions.

When an assessor awards a mark to an essay using a rating scale they are, of necessity, comparing that essay to imaginary ones in their head, opening the way to the three sources of error: variations in their personal internal standard, in how finely they differentiate the qualities they are looking for, and in their individual interpretation of the rating scale. ACJ, in contrast, requires direct comparisons, relative judgements, of the kind the mind is very good at making. The first two kinds of error do not operate in comparative judgement, and the remaining one is essentially a matter of validity rather than reliability. It is therefore no surprise that ACJ achieves higher reliability than rating scales.

The principle that data are independent, and can be added whenever needed, means that there is no real limit to the level of reliability that can be reached. In CAT it is possible, in principle, to extend the length of a test indefinitely to achieve high levels of reliability, or low standard errors, but for practical reasons, or a principle of equal treatment, this is not usually acceptable. With ACJ, however, this can be achieved by adding more judgements to the same script, instead of extending the test for the student. In effect, a test agency can ask for any level of reliability they want, if they are prepared to pay for it or collect enough data by some other means.

This does highlight the fact that improving the reliability of scoring tests by shifting from marking to judging can improve only the internal consistency and inter-judge aspects of reliability, since we may end up with extremely reliable assessments only of that particular performance on those particular tasks. Arguably of more importance than these is the concept of equivalent forms reliability, or the coefficient of equivalence and stability (Baird et al. 2011, 28). But few of our educational assessments address this issue.

Appeals and similar queries

A disagreement of any kind, such as an appeal by a student or their school that they have been unfairly assessed, is similar in principle to a misfitting script. A full audit trail exists, showing which scripts the problem script has been compared to, by whom, and what the result was. This record alone, supported by the measures or grades the other scripts received, may determine the issue. If not, the script can be sent out again for extra judgements. If the appellant is seeking a higher grade, these extra comparisons can be made against scripts chosen to represent the grade boundary; if several senior judges agree that the script is not good enough, no court would uphold the appeal. Thus, dealing with appeals can be subsumed into the

Table 2. Partitioning the weighted mean squares.

Average weighted mean square						N					
Total 0.913						8161					
	N-N 0.911	O-O 0.912	N-O 0.954				7828	4	329		
				N>O 0.943	N<O 0.966					158	171

normal procedure or, at worst, into a standard follow-up that is dealing with all disagreements, and the outcome will not be biased by the judge knowing which script is of concern, or why.

Test equating

It was mentioned earlier that some studies have given evidence that it may be possible to build test equating into the normal ACJ procedure. In the primary writing study we explored this directly. Twenty scripts from an old test were included in the set of 1000 scripts to explore passing the standard from the old test to the new one. If judges found it more difficult to be consistent in comparing performances from different tests then this should be visible in the residuals. Table 2 shows that the Overall weighted mean square for all comparisons can be partitioned into sub-sets of the data.

In theory the average mean square should always average 1.00, but in this case the overall average was 0.91. Amongst the ‘mixed’ comparisons this rose to 0.95, showing that there was some extra disturbance when judges were asked to compare scripts from different tests; the increase was small, however, and the value remained below 1.00. There was a curious difference amongst these mixed comparisons: the judges seemed slightly more inconsistent when they decided the ‘New’ script lost the comparison than when they decided it won. We cannot explain this effect, but it does show that the general effect of mixing scripts from the two tests did not cause a serious amount of inconsistency in the decisions. We concluded that, in this case, it was safe to assume the transferred standard was equivalent to the old one.

It may be possible to transfer standards like this in any ACJ application, but it would always be necessary to check, using this kind of ‘bias’ analysis to see if the judges differ significantly in how they adjust for their perceptions of how difficult each test was. Of course, no bias does not guarantee the equating is accurate, because it is possible that all the judges are equally misled in their adjustment for difficulty; the ‘Good and Cresswell effect’ is a well-known example of this problem (Good and Cresswell 1988). However unlikely this is, it will always be wise to use other methods as well in equating. But this method is simple, and may be good enough.

Judge feedback

Following the primary writing study an evaluation form was circulated and was returned by 28 of the 54 judges, including 14 trained markers. One question asked, ‘If you could choose between judging and marking, which would you prefer?’ One judge omitted this question; the others chose:

CJ	19
CJ with ‘marker training’	4
CJ but with some reservations	2
Both should be used	2
Marking	0

To understand why they were unanimously in favour of ACJ, we analysed their responses to ‘What do you consider to be the advantages of the paired comparisons method over traditional marking?’ Eighty-two comments were identified and classified into three main themes.

The largest group (42 distinct comments) described comparative judgement as more *professional* in concentrating on *holistic* judgement of quality, rather than on a fragmented view of writing. Examples included:

> uses years/decades of experience rather than minutes of brain washing training

> making a general judgement as to the level of a piece of work is what most teachers do anyway before they go through the criteria to prove what they think

> by comparing pairs, one could more easily see how well a pupil was able to express his/her thoughts and also how creatively

> releases you from having to make continuous reference to an increasingly complex mark scheme.

A second group (13 comments) described the method as *fairer* for students, or less prone to bias, while another six described this as an advantage for judges too. For example:

> each piece was judged several times and therefore eliminated problems caused by human error and opinion

> I was aware that several people would be involved in the judgement of 1 piece of work and that seemed much more likely that a fair decision would be arrived at.

The third main group (16 comments) referred to speed as an advantage. Some referred only to the ease of the decision itself, and others to the absence of tedious activities associated with marking:

> in many cases, a decision can be reached by the time (and often well before) the scripts have been read

> being able to make immediate responses – without the need for red pens etc was also a refreshing change to pouring (sic) over scripts for hours on end

There were five other comments, referring to enjoying judging, and to possible positive wash-back such as:

> it excites me to think we could actually teach children the overall value of texts rather than subject them to judged deconstruction of a text.

When asked about disadvantages, five judges could not think of any, and five others mentioned non-relevant e-assessment difficulties like the legibility of scans. The commonest real concern (seven comments) was related to subjectivity or uncertainty, such as:

> not clear what exactly we were looking for to award, aside from what was the better piece

> the major problem is that with little or no guidance the decision making can become completely subjective

> where would non specialists start without a mark scheme?

In four comments, judges reported some decisions had been very difficult to make. Finally, five comments suggested that some judges might get away with 'unscrupulous' work, or some pupils not be punished for poor spelling and grammar, or that there would be no feedback to teachers.

In general, these disadvantages can be easily addressed through training (which was deliberately minimised for this experiment) or through statistical monitoring. Since Kimbell et al. (2009, 70) reported that their D&T judges were 'unequivocal' about the ease and fairness of ACJ, it is clear at least that many teachers will support the method as an alternative to marking in some contexts.

Big bang or Steady state?

So far, all of our trials have simulated traditional marking exercises, in that all of the scripts have been available from the start for judging. There is an alternative model that might be attractive in some contexts, such as certification of competence in, for example, speaking or writing in a foreign language. The system could begin with a few hundred scripts (or recordings), and evaluate them. More scripts could be added as they become available, and be added to the collection. In a steady state model, scripts would 'drop out' when they are well enough evaluated and be replaced by new ones. Because the system is web-based, judges anywhere in the world could participate whenever they are available. If necessary, some standard borderline scripts could be used repeatedly when the 'live' pool falls too low.

Who can judge?

Ideally, all of the teachers in a state could be invited to be judges as part of their continuing professional development. The powerful statistical control means that any teacher who differed significantly from the consensus would be identified: two actions could follow. First, their judgements could be removed, and replaced by others, in order to maintain the reliability and validity of the overall procedure. Second, feedback to these teachers could be used to improve the quality of their judging – and hence of their teaching.

All judges make some decisions that are contestable. In our training sessions, a group of judges are asked to choose the better of two scripts that are quite similar in quality; the resulting discussion – argument – raises most of the issues of qualities and how to aggregate them for a global decision. If this were repeated more generally, the result would be a re-professionalisation of assessment, not through the emergence of 'professional assessors', but through making assessment an integral part of every teacher's professional activity. This conforms to the concept of 'constructive alignment' proposed by Biggs and Tang (2007) as a powerful strategy for improving the quality of learning by ensuring that teaching, learning and assessment address the same learning outcomes.

This sharing of professional judgement and values can start small. There is interest in applying ACJ across school consortia in several countries, in order to keep standards and the moderation of internal assessment under the control of teachers, while spreading standards and expectations beyond the walls of a single school. A recent government review of national curriculum assessment in England has advocated 'cluster moderation', a form of local collaboration on standards (Bew 2011, 15, 67); ACJ would make this very easy and routine, and would also allow any group of schools to collaborate rather than just those who are geographically close.

There is no need to limit judgement to teachers. Any interested party could, in principle, be invited to make judgements – to try the system and so gain a better understanding of its strengths. Parents, politicians, and journalists all could be encouraged to develop an understanding of the complexity and the professionalism of comparative judgement in this way; it is more natural than trying to master a mark scheme. Of course, the judgements they make might not be included as real assessment data – if they turn out to be 'misfits'.

Conclusion

Adaptive Comparative Judgement is a method for scoring students' work in which judges are asked only to consider *validity* while making their decisions; nevertheless, the result is extremely high *reliability*. The principle is Judgement, exploiting the professionalism of teachers rather than the tangential skills of marking. The method is Comparison, the natural human way to make complex judgements. The power of the method comes from its Adaptiveness, maximising the information gained from the judgements to achieve high reliability efficiently.

The primary writing study showed that both trained markers and teachers with no marker training can use the method effectively to assess the quality of children's writing, and the e-scape project has demonstrated the same for the assessment of e-portfolios in Design & Technology. We do not yet know how widely it can be, or should be, applied. The advantages of ACJ are most obvious in three kinds of context: extended writing, whether in languages, politics or history, where the evidence examiners want to evaluate is unpredictable, with students themselves choosing how to tackle the task; in areas like music, dance or speaking, where the evidence is presented through a performance; and in assessing portfolios, projects or reports where the evidence is necessarily complex. In each of these, it may be better to rely on the controlled subjective judgements of professionals, making the kinds of evaluation they are used to doing while teaching, than to try to achieve reliability and validity through a marking process that allows markers little or no discretion in applying fixed rules.

In each new case it would be necessary to explore the practicability of ACJ. A system that required judges to listen to each complete musical performance 10 times might, for example, prove too expensive to use, or the intrusiveness of recording the evidence, perhaps by video, might rule it out. In addition, a new assessment system will need new practices to support it. For example, the evidence of how a single candidate's score is arrived at will be different when a list of judgements replaces a list of marks for each question, and new systems for delivering feedback to teachers and students, where that is appropriate, will be required.

Finally, it is possible that the availability of a convenient and reliable method for assessing by judgement will change examiners' perceptions of how other subjects should be assessed, as the experiments with mathematics mentioned earlier suggest. What has been shown so far is simply that the method works very well in some settings; how widely it should be used remains to be seen.

Acknowledgements

I am grateful to all of the collaborators who carried out the data collection for the various trials referred to in the paper. In particular I would like to acknowledge the contributions of Richard Kimbell and Tony Wheeler of TERU, Goldsmiths College, to the early development of the adaptive method, Donald Laming for suggesting the Swiss system to get us started, and Karim Derrick and Declan Lynch at TAG Learning for turning ACJ from a mere concept into a practical assessment system.

Notes on contributor

Alastair Pollitt is Director of CamExam, and an affiliated lecturer in the Department of Theoretical and Applied Linguistics in the University of Cambridge, and was formerly Director of Research and Evaluation in the University of Cambridge Local Examinations Syndicate. His research interests are in models of assessment and the cognitive functioning of students and assessors during the assessment process.

References

Andrich, D. 1978. Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement* 2, no. 3: 449–60.

Baird, J., P. Black, A. Béguin, A. Pollitt, and G. Stanley. 2011. *The Reliability Programme: Final report of the Technical Advisory Group*. Coventry, UK: Ofqual. http://www.ofqual.gov.uk/research-and-statistics/research-reports/92-articles/20-reliability.

Baker, E.L., P. Ayers, H.F. O'Neill, K. Choi, W. Sawyer, R.M. Sylvester, and B. Carroll. 2008. *KS3 English test marker study in Australia. Final report to the National Assessment Agency of England*. London: QCA.

Bew, P. 2011. *Independent review of Key Stage 2 testing, assessment and accountability.* London: Department for Education. https://media.education.gov.uk/MediaFiles/C/C/0/{CC021195-3870-40B7-AC0B-66004C329F1F}Independent%20review%20of%20KS2%20testing,%20final%20report.pdf.

Biggs, J., and C. Tang. 2007. *Teaching for quality learning at university.* 3rd ed. Buckingham: SRHE and Open University Press.

Bradley, R.A., and M.E. Terry. 1952. Rank analysis of incomplete block designs, 1. The method of paired comparisons. *Biometrika* 39, nos. 3/4: 324–45.

Bramley, T. 2007. Paired comparison methods. In *Techniques for monitoring the comparability of examination standards*, ed. P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms, 246–94. London: QCA. http://www.ofqual.gov.uk/files/2007-comparability-exam-standards-i-chapter7.pdf.

Brennan, R.L. 2001. *Generalizability theory.* New York: Springer-Verlag.

Cronbach, L.J., G.C. Gleser, H. Nanda, and N. Rajaratnam. 1972. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

D'Arcy, J., ed. 1997. *Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research. Belfast: Joint Forum for the GCSE and GCE.