Using Adaptive Comparative Judgement to Obtain a Highly Reliable Rank Order in Summative Assessment

Claire Whitehouse and Alastair Pollitt

Abstract

Adaptive Comparative Judgment (ACJ) offers an alternative to marking, especially for performance assessments for which achievement can be difficult to describe in mark schemes. The ACJ system uses a web-browser to deliver pairs of candidates' work to judges. The judges apply their subject expertise to select which of the two sets of work is better in terms of meeting appropriate educational objectives. A quality parameter is estimated for each piece of candidates' work using the Rasch logistic model to analyse the judges' decisions. Placing the pieces of work on a measurement scale gives them a rank order. The adaptive nature of ACJ lies in the pairing of essays for each judge. The selection of essays is based on a maximum distance between the quality parameter estimates, allowing useful information to be gained from each paired comparison.

This paper reports a study in which ACJ was used to rank order a random sample of 564 essays on a topic in physical geography based on the judgements of a group of 23 teachers and examiners. The reliability of the rank order was 0.97. Evidence is presented for the need for judges to be teaching at the same qualification level as they are judging at. There is a discussion of the factors that need to be addressed before implementing ACJ in summative assessment.

Keywords: adaptive comparative judgement, Rasch model, summative assessment

Introduction

Comparative judgement

Adaptive Comparative Judgement (ACJ) offers an alternative to marking. It originates in the method of comparative judgement (CJ) that was first proposed by Thurstone (1927) as a description of the processes of human judgement that are not visible to the observer. In stating the model of comparative judgement, Thurstone suggested that it could be used to quantify the perceived quality of objects, including handwriting specimens and children's drawings.

Thurstone's model has five versions. The simplest and most tractable is based on three assumptions. The first assumption is psychological in nature: that the judgements being made are holistic and above all comparative. As Laming (2011) states "*There is no absolute judgement. All judgements are comparisons of one thing with another.*" When a judge is asked to make an absolute judgment about the perceived quality of an object, previous experience, level of knowledge, self-efficacy and the opinions of others all influence that judgement. In summative assessment, examiners are (and should be) greatly influenced by the mark scheme to an extent that overcomes bias as far as possible. So the absolute judgement of what mark to award is relative to the mark scheme plus any error and bias in its interpretation.

A series of absolute judgements results in a number of measurements on the scale of perceived quality, and when they are made by a group of judges the leniency or severity with which each

judge makes their judgements will be visible. This bias is one of the threats to the validity and reliability of marking. In comparing two objects and deciding which one is the 'better' of the two, however, it does not matter where a judge would locate the quality of the individual objects on a scale of quality. What matters is the difference in quality between the two objects. So, in placing students' work on a measurement scale the bias of individual judges is unimportant so long as they are able to perceive differences in quality based on a set of shared criteria. The cancelling out of the bias of individual judges is the second, statistical, assumption behind CJ. In current examining the use of mark schemes, the training of examiners, and the monitoring and adjusting of their marks are all activities aimed at achieving the same cancellation of the biases of individual examiners (Meadows & Billington, 2005).

The third assumption is that judgements are independent of each other. That is, that the outcome from one judgement will not influence the outcome of other judgements.

Thurstone's original CJ model took a normal-distribution form. Andrich (1978) recast the model in the Rasch logistic form which can be expressed in terms of log odds.

$$log\ odds\ (A\ beats\ B|v_a - v_b) = v_a - v_b \qquad (1)$$

In words, two objects, A and B have perceived quality v_a and v_b, respectively. The difference between the perceived quality of A and the perceived quality of B is equal to the log of the odds that object A will be judged to be better than object B. If different judges compare the two objects on many occasions, the data from these comparative judgements can be used to estimate the difference in perceived quality between objects A and B. When there are many objects to be compared, there will be many equations, one for each paired comparison.

Rasch (1960) showed that the maximum-likelihood procedure could be used to optimise the estimates of the quality parameters for a set of objects. An expected score is calculated for each object, using the current estimated quality parameters to predict the outcome of every comparison it was involved in, and summing these. This expected score is compared to the observed score (number of 'wins') and an updated estimate calculated. As Rasch explained, this procedure implied that the number of times an object wins a comparison is all that is needed to estimate quality parameters.

The combination of comparative judgement and the Rasch model confers a number of advantages when quantifying the perceived qualities of objects. First, the quality parameters are placed on an equal interval scale that is easily visualised and understood. Second, as the Rasch model is a strong model, it will fail if judges are in disagreement about what constitutes 'good' in a particular domain. Failure means that a measurement scale is not established. Third, the model is robust enough to cope with non-random missing data. In comparative judgement terms this means that each judge does not have to compare every object with every other object. A full design, in which every judge made every comparison possible, would result in *N(N-1)/2* comparisons, ie the number of comparisons is proportional to N^2. Applying the Rasch model means that the number of comparisons needed is proportional to *N* (Pollitt, 2012).

A final advantage lies in the utility of differences between the expected and observed outcomes of each comparative judgement. Deviations from the expectation provide statistical quality control over the judging process. In Rasch terminology this is called misfit. Problematic, or misfitting, scripts and judges can be identified whilst the judging is in progress and appropriate action taken.

Comparative judgement and assessment in education

Pollitt (2004) was the first to propose the comparative judgement (CJ) method as an alternative to marking. He argued that, in some cases, holistic judgements of performance are better than the summation of a number of "micro-judgements" at question level which tends to favour

reliability at the expense of validity. An holistic approach to assessment has the potential to re-focus assessment on the demonstration of skills. This is particularly so where the judgement of a performance or product has a large element of subjectivity. Further, mark schemes can require complex thought processes, which may create a barrier to the exercising of legitimate subjectivity by examiners. As Grant (2008) states about the use of mark schemes: "*This is difficult and requires considerable depth of knowledge and experience and can still result in different assessors judging the same work differently because they have different standards in mind.*"

On the other hand, one of holistic judgement's flaws is its lack of visible judgement criteria (Meadows & Billington, 2005). Mark schemes provide information about what is deemed to be mark worthy in an assessment to teachers and students, as well as markers. To some extent they make visible what is obscured by the marking process.

The CJ method has been used in the cross-moderation strand of comparability studies organised by the exam boards since 1996 (D'Arcy, 1997). It is now the favoured method for cross-moderation in the regulator's reviews of standards over time (QCA, 2006a). As such it has already gained credibility with a major stakeholder.

A comprehensive review of the comparability studies that have used CJ (or Thurstone pairs) is contained in Bramley (2007), and the papers of Jones (1997) and Fowles (2000) further illuminate the methodological issues that arose during the studies. Some of these issues are applicable to the use of ACJ as an alternative to marking and will be addressed later in this report.

Comparative Judgement as an alternative to marking

In Australia a number of studies have applied CJ to the teacher assessment of students' work in a variety of domains that involved open-ended tasks and performances (Heldsinger & Humphry, 2010; Newhouse, 2011). Heldsinger and Humphry (2010) focused on a study involving twenty staff (mainly teachers) from one school in Western Australia who judged thirty narrative texts from students aged six to twelve years old. Apart from reporting the high reliability of the rank order (0.982), the authors also noted strong concurrent validity. This was based on a high correlation (r = 0.921) between the quality parameters estimated from the paired comparisons and independent estimates obtained from an experienced examiner using a well-established mark scheme. The staff who carried out the judging commented that the process "*force[d] consideration of the qualitative characteristics that distinguish one performance from another*".

Recent reports of the piloting of ACJ in the assessment of candidates' work also show high reliabilities for the rank order (Kimbell *et al.*, 2009; Pollitt, 2012; Pollitt & Poce, 2011). For example the E-scape project, which was motivated by a desire to bring innovation to the assessment of performance in GCSEs, in particular for design and technology, reported reliabilities of 0.95 with ACJ "*work[ing] very effectively*" (Kimbell *et al.*, 2009, pg. 67). In their study, a total of 28 teachers from design and technology, science and geography made paired comparison judgements in the online environment. As well as high reliabilities, nearly all the teachers who took part in the judging thought ACJ could be of benefit to assessment of their subjects: "*With this, good Design and Technology can be rewarded as a result of (and not in spite of) the examinations*" (Kimbell *et al.*, 2009, pg. 154).

The E-scape project drew two further conclusions. First, that the one day training was sufficient to enable judges to produce a highly reliable rank order. The training comprised a half-day spent matching criteria to portfolio evidence and a half-day learning to use the ACJ system and making paired comparisons. Second, that the paired comparisons method was able to spread out the quality parameter estimates on a measurement scale such that the standard unit of the

scale was almost four and a half times larger than the uncertainty of the measurement. This represents a high level of measurement accuracy, especially for a GCSE assessment.

Information Technology enables Adaptive Comparative Judgement

At the start of the judging process the ACJ system has no information about the quality of the scripts. To gather information, in the first round of judgement the system selects pairs of scripts randomly and presents them to the judges. Due to the random nature of the selection the scripts in a comparison could be some distance apart or very close together in terms of quality of performance. In subsequent rounds the system selects scripts that have won the same number of judgements. These are 'Swiss' rounds of comparison, a label borrowed from tournament chess. Each script takes part in at least one comparison in a round.

Information, the proportion of comparisons that a script has won, is collected over a few Swiss rounds. When sufficient information is available the ACJ system starts to estimate the quality parameters for the Rasch model. An algorithm then uses these estimates to select pairs of scripts that will provide the most information for increasing the reliability of the rank order. This is the adaptive nature of ACJ. The benefit of adaptivity is that comparisons between a very good script and a very poor script in which the decision is obvious (and so no further information is added to the mix) are avoided. By stating a maximum distance between the quality parameters of a pair of scripts the ordinal nature of human judgements is used to best effect (Laming, 2011, pg. 67). Improvements in the selection algorithm have reduced the number of Swiss rounds needed from six to four (Pollitt, 2012) and in the current study to three.

ACJ exploits greater computing power and faster connection speeds to calculate quality parameters as comparative judgement is in progress. Networking allows judges to participate by remote working, which may increase the potential pool of judges. More judges make more judgements and estimates of quality parameters improve with increasing numbers of judgements.

Using ACJ is analogous to applying a computer adaptive test (CAT). In a CAT a candidate is presented with items that are selected based on the calculation of his or her ability. The test then adapts to the calculated ability of the candidate, presenting increasingly more difficult items to candidates who demonstrate high ability and items of lower difficulty to candidates who demonstrate lesser ability. However, ACJ is not subject to the disadvantages associated with CATs. CATs need large item banks and pre-testing to calibrate the items in the bank; both of these have cost implications and are avoided with ACJ. Nevertheless, in ACJ there is likely to be a cost associated with exemplification of judgment criteria and the establishment and maintenance of a professional community that is willing and able to take on the role of judge.

The current study

The aim of this study was to evaluate the ACJ method as a feasible alternative to marking using candidates' responses from a summative assessment in a low risk environment. The key questions were: (1) could judges with different backgrounds make consistent judgements; (2) what is the optimum volume of candidates' work to judge using ACJ; and, (3) what is the nature of the task judges are asked to carry out? These questions are similar to those asked in the methodological reviews of the inter-awarding body comparative studies. A fourth question, about how the results of the rank ordering should be reported, is not addressed in the current study. Additionally a self-administered online questionnaire was used to explore the judges' experiences of the ACJ method; the results of this questionnaire will be reported separately.

Responses to an essay question in a GCE Advanced Subsidiary (AS) geography question paper sat in the summer of 2011 were chosen for this study because they were already in electronic form and easy to 'clean', ie to remove marks, examiners' comments and any features

that could identify candidates. Also, the essay questions in this paper were known to have variable inter-rater reliabilities (Whitehouse, 2010a).

Entire scripts were not used as this particular question paper was printed in a question-and-answer booklet that was 48 pages long. The large number of pages was due to the need to print every optional question; the rubric asked candidates to respond to four questions out of eight. Scrolling on screen through two or three pages of a response is one of the main sources of complaint from examiners who mark on-screen (Whitehouse, 2010b). The essays chosen for use in the present study did not exceed two pages.

Method

Essays

Responses to the compulsory essay question on physical geography in one of the AS units of GCE Geography were used for this study. The paper comprises two sections, one on physical geography and the other on human geography; the maximum mark for each section is 60. Candidates respond to one compulsory question and one optional question (from three) from each section. The maximum raw mark for the question paper is 120; the maximum raw mark for an essay is fifteen. A random sample of 564 essays was selected. The essays were responses to the following question:

> *'Soft engineering is a better river flood management strategy than hard engineering.'*
>
> *Discuss this view.*

The essays in the sample (mean mark = 8.3, sd = 3.0) and the scripts (mean total mark = 59.4, sd = 14.9) from which they were extracted were representative of the essays (mean mark = 8.4, sd = 3.0) and scripts (mean mark = 59.5, sd = 15.5) of the entire cohort. The candidates whose scripts formed the sample were also representative of the entire cohort in terms of the type of school or college they attended, age, sex and grade distribution.

Participants

Teachers and examiners were recruited by an email invitation. The examiners had marked geography scripts at GCE AS level for AQA and were graded A, B or C[1] in the summer 2011 series. The teachers of geography were a more geographically convenient sample being based in the London and south east area. Payment of a fee at the standard rate was offered as an incentive to participate in the study.

In total 23 participants were recruited to the role of judge. Sixteen of the judges had recent examining experience with AQA, and five of these had marked responses to the question used in this study. The other seven participants had no relevant examining experience. Only two participants had no current or recent teaching experience at GCE level; they were teaching at GCSE level however.

The judges were reasonably representative of the schools and colleges that form the entry for GCE Geography, although there was an under-representation of the independent schools sector. Just over two-thirds (16) of the participants were female. This proportion is higher than

[1] At the end of an exam series all expert examiners are graded according to their performance on two criteria: marking and administration. Examiners graded A, B or C may be employed again; those graded D would be employed again after additional training; grade E examiners would not be employed again.

that likely to be found in the general population of teachers of geography in secondary schools, i.e. 49% according to data for Scotland presented by Riddell *et al.* (2005, pg. 26). The literature does not appear to contain an equivalent figure for England.

Training

Sixteen of the judges participated in a training day. This provided context for the use of ACJ in the scoring of responses; a hands-on session with the software in which all judges made a number of judgements; and, a feedback session on the results of their judgements. This is in contrast to standardisation meetings in which examiners are trained in how to mark each question according to a rubric. The training session used 60 essays covering the range of responses; none of these essays was used in the main study.

Due to unforeseen circumstances a number of judges were unable to attend a training day. Consequently seven judges took part in one-to-one telephone briefing sessions that focused on how to use the software and offered general advice on how to make comparative judgements. The telephone briefings lasted approximately 40 minutes. This provided an opportunity to consider whether the type of training affected the quality of judgements.

Procedure for the judging session

The judging session lasted fifteen days. Participants were asked to complete a minimum of 150 judgements of pairs of essays during this time. Other than this they had complete freedom as to how they scheduled their workloads.

TAG Development provided the ACJ system which presented each judge with pairs of cleaned essays. After reading through both essays, the judge decided which essay demonstrated the stronger performance and recorded his or her decision within the ACJ system. He or she could type in a comment (a word, phrase or single sentence) about what was in their minds at the point at which they made their decision before moving on to the next pair of essays. Care was taken to emphasise that judges should not offer considered comments as this would detract from the holistic nature of the judgement and extend the process unnecessarily. Time taken to type in any comment was not included in the time taken to make a judgement.

Guidance on making judgements was offered in the form of two importance statements (see Appendix A) and a question. The first importance statement comprised the aims of an AS and A Level specification in geography as shown in the regulator's subject criteria (QCA, 2006b). Selecting this as a guidance statement established a link between the rigour of the GCE specification and making holistic judgements without reference to a mark scheme. The second importance statement was from the programme of study at key stage 3 for geography in the National Curriculum (QCA, 2007); unfortunately there is no equivalent statement for key stage 4, which would be more appropriate for GCE. The question, positioned after the importance statements, was: *Based on these statements, which of the essays shows more evidence of a higher level of development of what is deemed important in Geography?*

The first three comparison rounds were Swiss rounds. The analyses at the end of these rounds sorted the essays into an increasing number of groups based on the number of times they had 'won' a judgement. Adaptivity started from round 4, as did 'chaining'. A judge made a decision about the first pair of essays, A and B, presented at the start of round 4. With the decision recorded, the judge moved on to the second pair of essays in which script B from the previous pair was now essay A and a new essay replaced script B from the previous pair. Each judge's chain would start anew with the start of a new round of judgement. The process went on until each judge had completed 150 judgements which was after 12½ rounds.

Results

The outcomes of this ACJ trial are arranged in the order of activities undertaken in a marking series: the evolution of the rank order; quality control of judging; and the resulting quality parameter estimates of the essays and their validity.

The judges' decisions from the paired comparisons were analysed using bespoke code from Alastair Pollitt. This software sets up 'dummy' essays: one essay that had won all the comparisons it took part in and the other that had lost all the comparisons it took part in. The dummy essays did not appear in the results. Using dummy essays reduces the standard error associated with any actual essays that are at the extreme ends of the measurement scale.

The alternative is to use an adjustment, as FACETS software does, whenever a script has won or lost all its comparisons. Values for the essay parameter estimates, standard errors and the misfit statistics are higher using the adjustment method. By the end of the judging session, though, the correlation between estimates for the quality parameter produced by the dummy method and the adjustment method was positive and strong (r = 0.99, n = 564, p < 0.01, one-tailed).

The evolving reliability of the rank order of essays

It took the 23 judges 12½ rounds to make 3,519 comparative judgements between pairs of essays (i.e. on average 153 judgements per judge). This is 2.2% of the total of 158,766 possible judgements needed for a complete design. The evolution of the rank order and how the standard errors decreased with each successive round are shown in the plots of the quality parameters against rank order at the end of rounds 3, 6, 9 and 12 (Figure 1). Round 3 was the first round from which there were sufficient judgements to calculate parameters using the Rasch model. The unit of measurement for the quality parameters is logits. The higher the positive value of the quality parameter, the better the essay was judged to be. The best essay had a quality parameter of 10.5 logits and the worst essay had a quality parameter of -12.7 logits.

Reliability statistics

At the end of each round of judgement a set of reliability statistics was calculated (Equations 2 and 3). The first statistic, the separation coefficient (*G)*, is the ratio of the average spread of the quality parameters of the essays in logits (sd_v) to the average amount of uncertainty in the position in the rank (as given by the root mean square of the errors in the parameter estimates). Therefore a larger separation coefficient indicates that measurement error is a small proportion of the quality parameter. Ideally this statistic should work in tandem with the reliability (α) so that when the separation coefficient is larger than 2 (or 3) the reliability is larger than 0.8 (or 0.9).

Separation coefficient: $$G = \frac{sd_v}{rmse} \quad (2)$$

The second reliability statistic, the reliability coefficient (Equation 3), is analogous to Cronbach's alpha (Andrich, 1988). It is an index of the proportion of variance between the essay quality parameter estimates that is due to measurement error (Bramley, 2007) and shows how consistent the judges are with each other in terms of standard and rank order (Pollitt, 2012). This is because the average error variance of the scripts is used rather than the error variance of an average script (Wright & Stone, 1999). In practice the minimum value of reliability is 0 and the maximum value is 1. Higher values indicate greater consistency of judgement and lower measurement error.

Reliability: $$\alpha = \frac{(G^2 - 1)}{G^2} \quad (3)$$

where *G* is as given in Equation 2.

The round by round values for the separation coefficient and the reliability coefficient are shown in Figure 2. By the end of round 7 the reliability coefficient exceeded 0.9 and the separation coefficient was 3.6. By the end of the judging session (12½ rounds) the scale properties were:

Standard deviation of the quality parameters (sd_v) = 4.96

Root mean square estimation error (*rmse*) = 0.81

Separation coefficient (*G*) = 6.10

Reliability coefficient (α) = 0.97

The separation and reliability coefficients had high values meaning that tests of fit and conventional tests of significance could be interpreted with some confidence (Bramley, 2007).

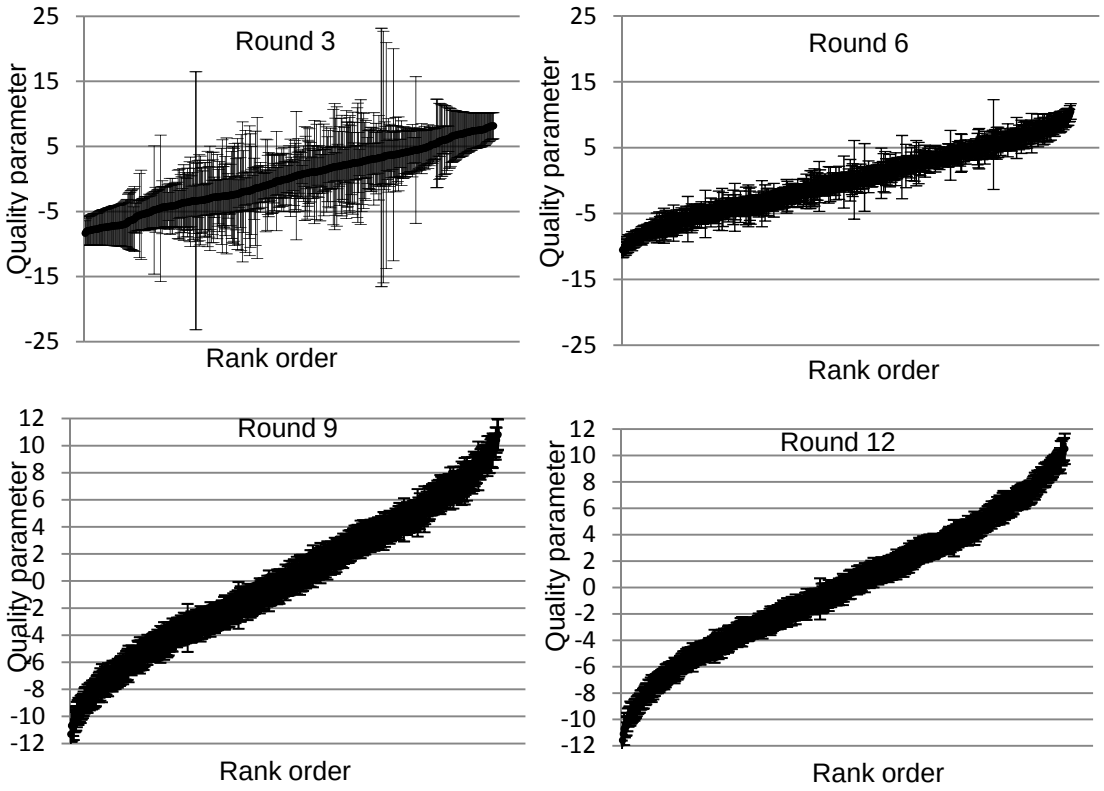

Figure 1: Evolution of the rank order of 564 geography essays with decreasing standard error over 12 rounds of judgement. Note that the scales on the y-axis for rounds 9 and 12 go from -12 to +12 so that essays with the highest standard errors are visible.

The methodology of ACJ removes differences in severity between judges by cancelling out their internal standard. It also removes differences between judges in their ability to differentiate between essays of similar quality. This is in contrast to the traditional inter-rater reliability used in marking reliability studies, which is based on the correlation between a marker who is taken to embody the marking standard and a second marker. Inter-rater reliability correlations do not include bias, or ability to differentiate at the small scale, as sources of measurement error. Therefore, they tend to over estimate marking reliability.

With this in mind, a reliability of 0.97 is very high when compared with an inter-rater reliability of 0.59 (n = 173, $p < 0.001$) found for a similar question on physical geography from the winter 2010 series during a different study (Whitehouse, 2010a).

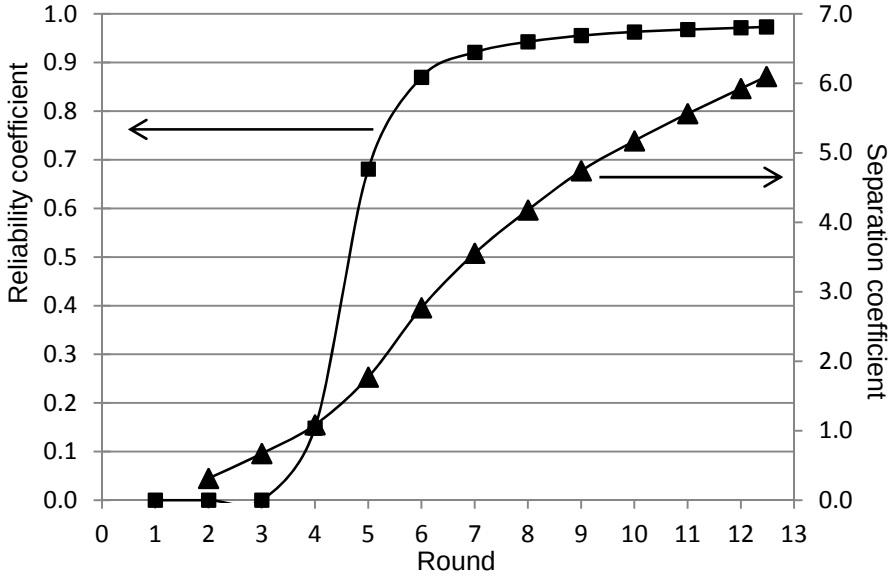

Figure 2: Plot showing how the separation coefficient and reliability of the rank order increase round by round

Quality control of the judging process by using misfit statistics

Misfit statistics

As mentioned earlier, one of the advantages of using the Rasch model is that it allows misfit statistics to be calculated for both judges and scripts. The basis of the misfit statistics are residuals or the extent to which an actual score deviates from that expected by the model. Every judgement has a residual; therefore every judgement can be evaluated for its consistency with all the other judgements. The criterion value for a judge to be considered misfitting is conventionally taken to be 2.

Squared, standardised residuals (SSRs) can be averaged across either judges or scripts. In either case they are summarised into an information-weighted mean square (WMS), where 'information' refers to the statistical variance in each comparison and is largest for comparisons between scripts of similar quality. The WMS is the infit mean square in a Rasch analysis; its theoretical mean is equal to 1 and the possible range of values is from zero to infinity (Bond & Fox, 2007). A WMS value greater than 1 indicates that the judge (or script) is behaving inconsistently with other judges (or scripts) and is deviating from the model. The concept of overfitting the model is introduced when the WMS is lower than 1 and may indicate some redundancy. Here the judge (or script) is showing too little variation and following the model too closely. A reasonable range of values for WMS in testing contexts is 0.4 – 1.2 (Wright & Linacre, 1994), but what is reasonable for ACJ is not yet clear. Certainly, above 1.2 or 1.3 a judge or a script should be suspected of distorting the rank order.

There are other misfit statistics produced by the Rasch analysis. Experience has shown that these can produce erratic values due to the adaptive nature of ACJ (Pollitt, 2012). Therefore, they are not discussed here.

Misfitting judges

In this study 61 judgements had standardised residuals with values higher than 2. This was 1.7% of the total number of comparisons and was in line with the proportions of misfitting judgments, usually less than 5%, recorded for inter-board comparability studies (Bramley, 2007).

On average each judge took part in 153 judgements and viewed 186 unique essays. By the end of 12½ rounds of comparative judgements the mean of the WMS for all judges was 0.78 and had a standard deviation of 0.18. A mean WMS indicating overfit is not unsurprising as agreement on the quality of the essays was being sought through the judging process. Using the definition of mean plus two times the standard deviation as the criterion for misfit gives a value of 1.14(8). As Figure 3 shows, only one judge (number 82) exceeded this criterion. Applying a more stringent criterion of mean plus one standard deviation identified another three judges (numbers 77, 83 and 91) as misfitting. These four judges, and particularly judge number 82, were making judgements that were inconsistent with the decisions of their fellow judges. An alternative way of expressing this is to say that these misfitting judges were basing their judgements on criteria that were different to those that the other judges were using.

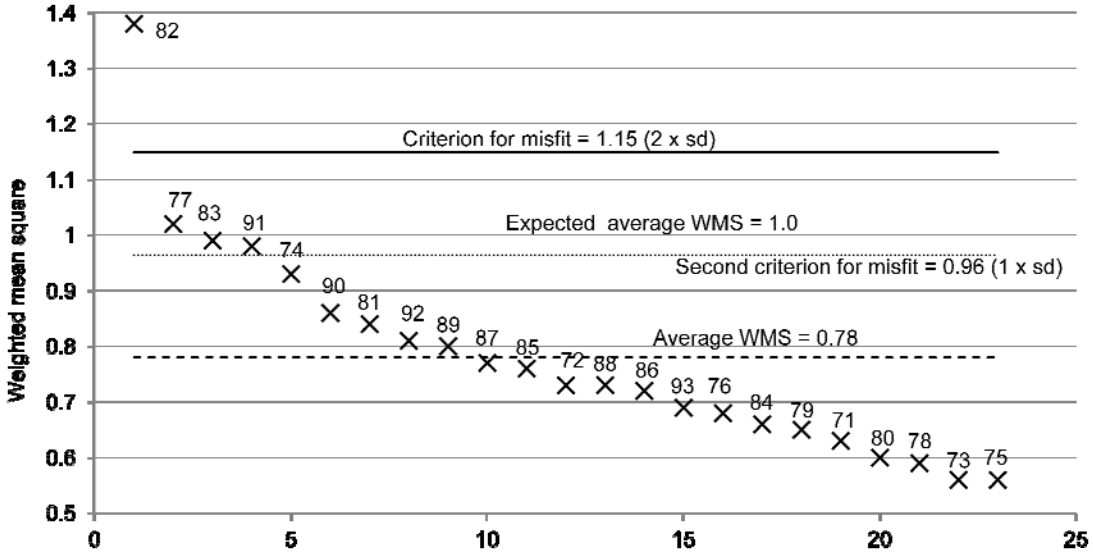

Figure 3: Showing the misfitting judges based on weighted mean square (WMS) values and two criteria for misfit

Three of the four judges who displayed inconsistency had no previous relevant examining experience and two of these taught only at GCSE level. Analyses of variance (ANOVAs) can indicate which factors may affect the quality of the decisions made by the judges. The effects of current teaching level, previous relevant examining experience and type of training on the weighted mean square values were investigated. Current teaching level did exert a statistically significant effect on the difference between WMS values ($F_{(3,11)} = 15.35$, $p < 0.005$, $\text{eta}^2 = 0.719$, power = 0.994). Judges who were teaching at GCSE level only (mean = 1.19, sd = 0.28) demonstrated greater inconsistency in their decisions than did teachers who were teaching at GCE level only (mean = 0.67, sd = 0.10) and teachers who were teaching at both levels (mean = 0.82, sd = 0.12). It is unlikely that these differences arose by chance.

There were no statistically significant differences between the mean WMSs for different groups of judges based on either previous relevant examining experience ($F_{(2,11)} = 0.17$, $p < 0.848$, $\text{eta}^2 = 0.027$, power = 0.070) or type of training ($F_{(1,11)} = 0.008$, $p < 0.931$, $\text{eta}^2 = 0.001$, power = 0.051).

Another potential indicator of the quality of judgement is the amount of time spent on judging a pair of essays. Holistic judgements should be speedier than those in which a mark scheme intervenes between judge and assessment. Intuitively, though, stakeholders are suspicious of quick judgements, especially when they are high stakes judgements.

The median duration of judgements was 133 seconds (inter-quartile range = 132 seconds). The median is used here because judges left some judgements open for lengthy periods. The ACJ software requires the judge to close a judgement with a decision before issuing a timestamp to mark the end of a judgement. It is not possible to know what proportion of judgement time was spent judging and what was spent off-task.

Median judgement durations ranged from 53 seconds over 150 judgements to 492 seconds over 155 judgements. The plot of WMS versus median judgement duration in Figure 4 shows the four misfitting judges with the marker ×. The most misfitting judges tended to take a long time to make their judgements.

In this study there was a moderate positive and statistically significant, relationship between median duration of judgement and the WMS values (r = 0.60, n = 23, p < 0.01, two-tailed). The longer the judgement took to make the more likely it was that the judge's interpretation of quality was different from that of the other judges; see Figure 4. The outlier in Figure 4, a judge with a median judgement time of almost 500 seconds, was not misfitting. This judge was an experienced examiner who does not mark physical geography, but who used participation in the study to critique the assessment objectives. The additional activity increased the judgement time. Removing this outlier increased the strength of the correlation between time and misfit (r = 0.79, n = 22, p < 0.01, two-tailed). A possible conclusion is that a consistently long judgement time may be indicative of a judge whose criteria for judgment are not aligned with the features that they are seeing in the essays.

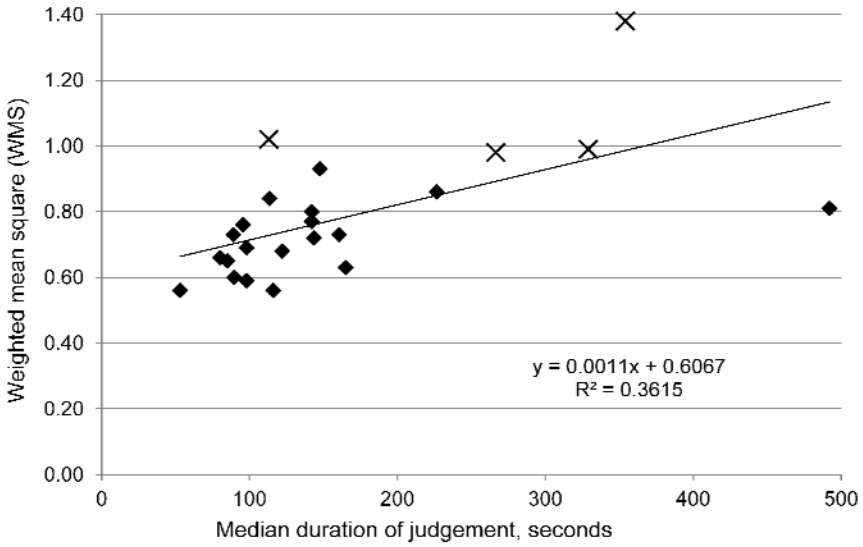

Figure 4: Showing the relationship between misfit for judges and the duration of judgements

In an operational environment the WMS values would identify judges whose interpretation of 'good' was misaligned. If the operation was high stakes summative assessment there could be three strategies to handle this situation: provide timely feedback and allow the judge to continue; provide feedback and remove the judge from the session; and, remove the judge's decisions from the data file and have other judges repeat the judgements.

Misfitting essays and validity of judgements

By the end of 12½ rounds the mean weighted mean square (WMS) value for the 564 essays was 0.75 and the standard deviation was 0.32. The criterion for an essay to be misfitting was an MWS value higher than 1.38. Using this value it appeared that the judges were making inconsistent decisions towards twenty essays when carrying out their paired comparisons; see Figure 5. Or, that these twenty essays contained features that caused the judges to vary in their decisions. Further qualitative work is needed to identify the features of these essays that caused difficulty for the judges.

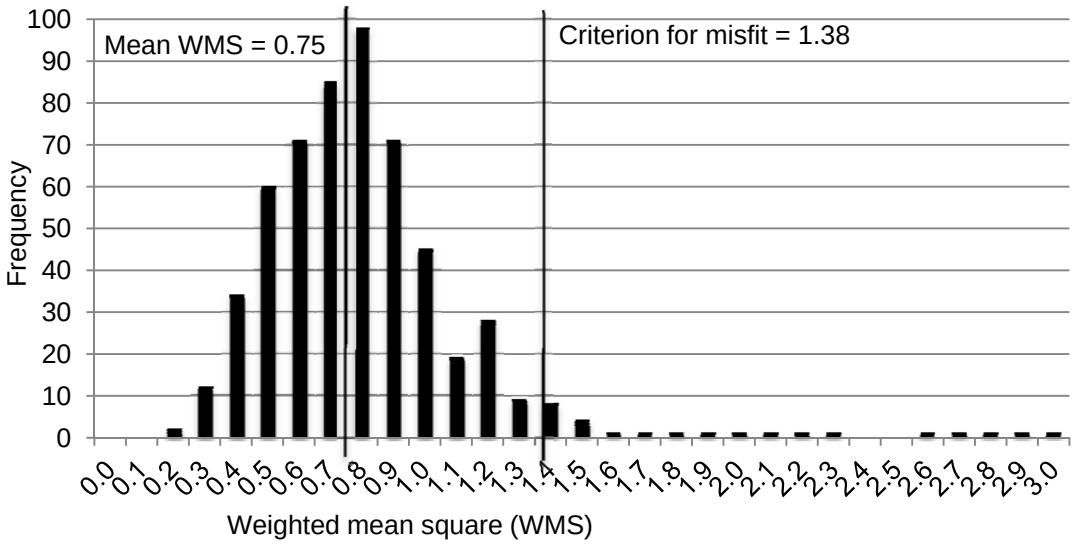

Figure 5: Showing the misfitting scripts based on weighted mean square (WMS) values and the criterion for misfit

Once a new measurement scale has been established for an existing assessment its validity needs to be checked. In the current study this is possible to do by comparing the estimates of the quality parameters with the marks awarded during live marking using Pearson's correlation coefficient. The relationship between the two variables is moderate and statistically significant (r = 0.63, n = 564, p < 0.01, one-tailed). A correlation of 0.63 between marks and quality parameters is very low in comparison to others quoted in the literature on this topic and may be peculiar to the subject of geography. For example, a correlation of 0.92 between measures of judgement and marking was found for narrative texts from students aged 6 to 12 years old (Heldsinger & Humphry, 2010). Both judgements and marking were based on guidance from the Western Australian Literacy and Numeracy Assessment.

Comparison of marks with quality parameters is, however, not possible during live marking. Before declaring a script or scripts as misfitting data from the judgements of misfitting judges could be removed and the remaining data refitted to the model. Bramley and Gill (2010) carried out this exercise whilst using the rank ordering method for standard setting for GCSE English scripts. They found that up to four judges (out of seven) could be removed before the quality parameters could no longer be estimated on a single measurement scale. This was not just a function of the number of judges, but also whether the removed judges under- or over-fit the model and how much overlap there was between the scripts rank ordered by each judge.

So it is likely that after refitting the model, the same essays will be identified as misfitting. In this circumstance such essays could be sent to one or more senior judges who are recognised as being consistent in their judgements. A sample of scripts with low errors of measurement would

be needed for these judges to compare with the problematic essays. A similar strategy could be employed for essays that arrive late and for post-results enquiries. In the latter case there is an argument that post-results enquiries are redundant if an essay has been judged many times by different judges.

In the current study the judges using ACJ may have been making judgements using different criteria to those used by markers who had applied the mark scheme to the same essays. However, given that the inter-rater reliability for a similar question from the previous study in geography was 0.59 (Whitehouse, 2010a), it may be concluded that, notwithstanding a common mark scheme, markers of geography essays do not share a set of criteria on which to base their marking decisions. The mark scheme (na, 2011) may provide clues as to why this could be. Examiners need to hold in their minds three schemes for marking an essay: a three-level generic scheme that is applied to all essays; a list of the geographical knowledge that is specific to the question that candidates may demonstrate; and, a levels mark scheme that is specific to the question. This is a large amount of information for an examiner to synthesise and apply to a candidate's response. It is repeated for the three optional essays and this is in addition to remembering the codes and sigils used for on-screen marking.

Examiners are told seven marks are available with which to reward candidates for demonstrating knowledge and understanding. Eight marks are available for rewarding the higher order skills of analysis, interpretation and evaluation. There is, however, very little guidance on how these marks are arranged in the mark scheme. For example, should knowledge and understanding be rewarded in levels one and two only; can evaluation be rewarded only in level three?

It would not be surprising if individual examiners developed their own slightly different interpretations of how the three schemes combined. Or, if they applied different weightings to the assessment objectives at different points in the range of performance offered by the candidates. Thus, a mark scheme may become a barrier to reliable marking.

Having noted the potential for the mark scheme to act as a barrier, it is also worth stating that the post-decision comments made by the judges are based to some extent on the language of the assessment objectives. "Knowledge", "understanding" and "analyse" all appear in the comments as does reference to the use of examples, case studies and geographical terminology which do appear in the question-specific mark scheme. This lends validity to the judgement process. However, a number of judges on a number of judgements made reference to the strength or depth of a candidate's argument: "*complexity of argument*", "*uses knowledge to make argument*" and "*argument holds together a bit better*". The words 'argue' and 'argument' do not appear in the mark scheme.

Discussion and conclusions

This exploratory study looked at the use of the adaptive comparative judgement method to produce a rank order for a random sample of 564 essays on a topic in physical geography. The process presented images of pairs of scripts to 23 judges through a web browser, enabling the judges to work remotely. The judges were asked to make holistic judgements about which was the better of a pair of essays rather than using analytical marking.

After 12½ rounds of judgements (or approximately 3,500 paired comparisons), the ACJ system produced a rank order with a reliability coefficient of 0.97 and a separation coefficient of 6.10. The reliability value showed that only 3% of the variance of the essay measures was due to measurement error. The high value of the separation coefficient indicated that the average uncertainty of the positions of the essays in the rank order was small in comparison with the

spread of the essay measures. The ACJ system also identified problematic judges and essays. From this perspective this trial of ACJ can be considered a success.

In and of themselves, these scale summary statistics are not surprising within the context of comparative judgements. Pollitt (2012) reported a reliability coefficient of 0.96 and a separation coefficient of 5.20 after sixteen rounds of judgement in a pilot of ACJ that used 1,000 writing samples in English, 54 judges and 8,161 paired comparisons (which means that, of the roughly half a million possible comparisons between two scripts, only one in every sixty was actually needed). The samples, from 9-11 year old children, were in two genres: persuasive and creative narrative. The judges' task was to decide which of two pieces of writing better fulfilled the purpose of its genre (either persuasive or creative narrative). Their task was narrower than that of the judges in the present study who had to take into account specific subject knowledge and higher order skills.

The current study demonstrated that ACJ is able to use the Rasch misfit statistics to identify problematic judges and scripts. However, further work is needed to verify this classification.

The key question is: is ACJ practicable for summative assessment? The present study is the second largest to date according to the information available in the public domain. However, it was limited to 564 essays, not scripts, and 23 judges. The AS geography unit from which the essays were taken had approximately 23,000 entries in the summer of 2011 and 60 markers. Using the median judgement duration of 133 seconds means 60 judges would spend nearly 15 hours each making judgements of pairs of essays over two to three weeks **for one essay**. This needs to be compared with the time taken to mark an essay.

In ACJ the greater the number of judgements, the higher the reliability of the rank order becomes or the faster the reliability reaches an acceptable level. This can be leveraged by increasing the number of judges, as long as these judges share a common set of criteria from which to make judgements.

There are four other questions that this trial either provided only partial answers to or did not address. These questions are similar to those asked in the reviews of the CJ methodology used in inter-board comparability studies.

Who are the judges?

This trial provided some evidence that current teaching level is important in being able to make consistent judgements. Judges teaching at the same level as the assessment were more likely to make consistent judgements. Meadows and Billington (2007) showed that PGCE students with subject knowledge and some teaching experience, were able to mark sections of a GCSE English paper almost as reliably as experienced examiners who had been teaching for at least three years. However, all of the participants in their study received a full day's training from the Principal Examiner in the marking of responses before undertaking marking. In the present study the participants received training only in the use of the ACJ system and guidance on making judgements was limited to the importance statements in Appendix A. Their judgements were based on their expertise in their subject.

ACJ is web-based and as such it may increase the number of teachers and other subject experts who are willing to assess national exams in the role of judge. All judges need to share a set of criteria on which to base their judgements. Is it sufficient to leave the creation of these shared criteria to chance? For a licensed awarding body, the answer is a resounding 'no'. But there is a dilemma: using mark schemes and training meetings similar to those for examiners to embed a standard is antithetical to holistic judgement making. At first glance ACJ appears to negate the need for training examiners at traditional standardisation meetings. Thus there appears to be a cost saving. Over time, though, the professional standards of judges may

diverge or be eroded. Thought needs to be given to how shared criteria can be exemplified and disseminated.

What should be judged?

What is the optimum size of assessment artefact that can be judged using ACJ in a high stakes qualification? The inter-board comparability studies used full sets of components and gave judges between 2½ and 5 minutes to make their judgements. These judgements, however, were significant only for the awarding bodies, not for the students, teachers and schools for whom results day had been and gone. Studies using ACJ have asked judges to consider short pieces of writing of a single genre or across many genres. Intuitively there is a maximum volume of a student's work that can be handled by judges in the current short marking period whilst also 'doing the day-job'. Establishing what this volume is would need further work, but a rough estimate can be made using the figures collected in this study. Assuming each of 23,000 candidates responds to four essays implies a total of 575,000 comparative judgements over 12½ rounds. In this study the 23 judges took a total of 194.6 hours to make 0.61% of these judgements. Therefore, the same 23 judges would need 31,800 hours to judge the essays of an entire cohort or a total of 180 judges could complete the work in 21 consecutive days working 8 hours a day. With approximately 2,000 centres making entries for AS geography, ACJ would be feasible as an alternative to marking for high stakes assessment only if a substantial proportion of centres provided judges.

The foregoing calculation assumes that each essay from each candidate must be judged and a quality parameter estimated. Without changing the assessment that the candidates face, assessors could be asked, and may prefer, to judge all four essays in one go. Or to judge the two essays on physical geography together and the two essays on human geography together. Either approach would significantly reduce the amount of time needed for judging. However, additional time would still be needed to mark the low tariff items. An alternative approach would be to modify the assessment so that it is better suited to holistic judgement. This is likely to involve fewer questions requiring longer responses.

There is also the question of which assessment artefacts are most suitable for ACJ. For example, responses that are marked in a more holistic than analytic way may benefit from ACJ, especially so where mark schemes are found to be restrictive or confusing. In contrast, ACJ would be inappropriate for assessing low tariff questions, but there may be scope for evaluating its use with *clusters* of low tariff questions.

What should judges be asked to do?

Some judges experienced difficulty without a mark scheme; others said that they did not use the importance statements, choosing instead to rely on their professional instinct. This echoes what Kimbell et al. (2009) found in the E-scape project. It is insufficient for an awarding body to offer no guidance on how to assess its high stakes exams, thus further work is required to find out what sort of guidance is most effective at the point of judgement.

The holistic comparative judgement of the essays produced a rank order with a reliability of 0.97. This is very high in comparison with the inter-rater reliability of 0.59 found for a similar essay in a previous study (Whitehouse, 2010a). Again, in the current study there is a difference in the rank orders established by comparative judgement and by marking (r = 0.63). This difference is greater than in reports of other studies that compared the two methods of producing a rank order. It appears that markers and judges carried out their role as assessors in quite different ways that may have involved different or overlapping criteria. The question is: which is the (more) valid set of criteria?

In a study that used essays of English as a Second Language (ESL) Barkaoui (2011) also found that holistic marking favoured higher levels of consistency between markers in comparison with analytic marking. However, he noted that a single holistic mark was unable to capture the uneven nature of candidates' performances, which has value when reporting results for different competencies and in providing feedback. In contrast, analytic marking favoured higher levels of internal consistency because the markers had a mark scheme to guide them.

High levels of inter-rater reliability and transparency in the awarding of marks are valued in high stakes assessment in the UK. This, however, can lead to assessments comprised of short-response questions that are divorced from the more holistic objectives of the study of a subject. Further work is needed to identify what is valued more highly in different subjects.

How should the rank order and the quality parameters be reported?

Quality parameter estimates are not immediately understandable; however, they can be converted to marks or grades as demonstrated in the E-scape project. The latter may be reasonable when ACJ is used to assess performance or portfolios of practical work, but position in rank order or percentile occupied are also alternatives.

The study reported here has a number of limitations, not least the small sample sizes of essays and judges, although both were representative of their respective populations. A key limitation is that the effects of allowing judges to manage their own workloads was not investigated. About four judges dominated the judgements made in the Swiss rounds; other judges started judging in the later chaining rounds and had the easier task. The effects of judge behaviour should be explored to assess how behaviour may affect rank order and achieving acceptable reliabilities.

As with most marking reliability studies that take place outside live marking, the motivation of the participants to make the correct judgment is likely to be lower than in live marking. However, the ecological validity of the current study is probably strong in terms of what the judges were asked to do in making comparative judgements and working remotely.

The ACJ system is simple for users to access and work with; it provides robust statistics that can be operationalised for monitoring the judging as it progresses. ACJ may offer an alternative to marking in subjects where the shared criteria for assessment are limited and there is scope for subjectivity. There is a need, however, for further work to identify the assessments that are most suitable for ACJ and what is support is required in the selection of judges and the maintenance of a shared set of criteria to use when making judgements.

Claire Whitehouse & Alastair Pollitt

20 June 2012

References

Andrich, D. (1978) Relationships between Thurstone and Rasch approaches to item scaling. Applied Psychological Measurement, v2 n3 pp. 449-460.

Andrich, D. (1988) Rasch models for measurement. Beverly Hills: Sage Publications.

Barkaoui, K. (2011) Effects of marking method and rater experience on ESL essay scores and rater performance. Assessment in Education: Principles, Policy & Practice, v18 n3 pp. 279-293.

Bond, T. G., & Fox, C. M. (2007) Applying the Rasch Model. Fundamental Measurement in the Human Sciences. 2nd ed. New Jersey: Lawrence Erlbaum Associates.

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

Bramley, T., & Gill, T. (2010) Evaluating the Rank-Ordering Method for Standard Maintaining. Research Papers in Education, v25 n3 pp. 293-317.

D'Arcy, J. (1997) Comparability Studies between Modular and Non-Modular syllabuses in GCE Advanced Biology, English Literature and Mathematics in the 1996 summer examinations.: Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Fowles, D. F. (2000) A Review of the Methodologies of Recent Comparability Studies. Report on an Inter-board Staff Seminar. Manchester: Northern Examinations and Assessment Board No. RPA_00_DEF_RC_042.

Grant, L. (2008) Assessment for social justice and the potential role of new technologies. London: Futurelab.

Heldsinger, S., & Humphry, S. (2010) Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments. The Australian Educational Researcher, v37 n2 pp. 1-19.

Jones, B. E. (Ed.). (1997). A review and evaluation of the methods used in the 1996 GCSE and GCE comparability studies.: Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Kimbell, R., Wheeler, t., Stables, K., Shepard, T., Martin, F., Davies, D., et al. (2009) e-scape portfolio assessment: phase 3 report. London: Technology Education Research Unit Goldsmiths College University of London.

Laming, D. (2011) Human Judgement. The eye of the beholder. 1st ed. Andover, Hampshire: Cengage Learning EMEA.

Meadows, M., & Billington, L. (2005) A review of the literature on marking reliability. Manchester: AQA. No. RPA_05_MM_WP_05. A report produced for the National Assessment Agency

Meadows, M., & Billington, L. (2007) The Effect of Marker Background and Training on the Quality of Marking in GCSE English. Manchester: AQA No. RPA_07_MM_RP_047.

na. (2011). Geography GEOG1 (Specification 2030) Unit 1: Physical and Human Geography Post-Standardisation Mark Scheme. Retrieved 2 May 2012 from http://store.aqa.org.uk/qual/gce/pdf/AQA-GEOG1-W-MS-JUN11.PDF

Newhouse, P. (2011) Comparative Pairs Marking Supports Authentic Assessment of Practical Performance Within Constuctivist Learning Environments. In R. F. Cavanagh & R. F. Waugh (Eds.), Applications of Rasch Measurement in Learning Environments Research. (pp. 141-180). Rotterdam: Sense Publishers.

Pollitt, A. (2004). Let's stop marking exams, International Association for Educational Assessment Conference. Philadelphia PA.

Pollitt, A. (2012) The method of Adaptive Comparative Judgement. Assessment in Education: Principles, Policy & Practice, DOI:10.1080/0969594X.0962012.0665354.

Pollitt, A., & Poce, A. (2011). Adaptive Comparative Judgement and Peer Assessment., The Association for Educational Assessment - Europe 12th Annual Conference. Belfast, Northern Ireland.

QCA (2006a) QCA's review of standards. London: Qualifications and Curriculum Authority No. QCA/06/2374.

QCA. (2006b). GCE AS and A level subject criteria for geography. QCA/06/2852. London: Qualifications and Curriculum Authority.

QCA. (2007). The National Curriculum. Retrieved 13 February 2012 from http://www.education.gov.uk/schools/teachingandlearning/curriculum/secondary

Riddell, S., Tett, L., Burns, C., Ducklin, A., Ferrie, J., Stafford, A., et al. (2005, November 2005). Gender Balance of the Teaching Workforce in Publicly Funded Schools. Retrieved 3 April 2012 from http://www.scotland.gov.uk/Resource/Doc/76169/0019227.pdf

Thurstone, L. L. (1927) A law of comparative judgement. Psychological Review, v34 n4 pp. 273-286.

Whitehouse, C. (2010a) Reliability of on-screen marking of essays. Guildford: AQA No. RPA_10_CW_RP_012.

Whitehouse, C. (2010b) Marking essays on screen: a survey of examiners' experiences and opinions. Guildford: AQA.

Wright, B. D., & Linacre, J. M. (1994) Reasonable mean square fit values. Rasch Measurement Transactions, v8 n3 pp. 370.

Wright, B. D., & Stone, M. (1999) Reliability and separation. In Measurement Essentials (2nd ed.). Wilmington, Delaware: Wide Range.

Appendix A

I. Importance Statement for Geography

A student who is developing into a geographer through their course of learning is able to demonstrate that they have

- developed and can apply their understanding of geographical concepts and processes to understand and interpret our changing world
- developed their awareness of the complexity of interactions within and between societies, economies, cultures and environments at scales from local to global
- developed as global citizens who recognise the challenges of sustainability and the implications for their own and others' lives
- improved as critical and reflective learners aware of the importance of attitudes and values, including their own
- become adept in the use and application of skills and new technologies through their geographical studies both in and outside the classroom
- been and are inspired by the world around them, and gain enjoyment and satisfaction from their geographical studies and understand their relevance

Modified from Ofqual's *GCE AS and A level subject criteria for geography*, September 2006 (QCA/06/2852)

II. The importance of Geography

The study of geography stimulates an interest in and a sense of wonder about places. It helps young people make sense of a complex and dynamically changing world. It explains where places are, how places and landscapes are formed, how people and their environment interact, and how a diverse range of economies, societies and environments are interconnected. It builds on pupils' own experiences to investigate places at all scales, from the personal to the global.

Geographical enquiry encourages questioning, investigation and critical thinking about issues affecting the world and people's lives, now and in the future. Fieldwork is an essential element of this. Pupils learn to think spatially and use maps, visual images and new technologies, including geographical information systems (GIS), to obtain, present and analyse information. Geography inspires pupils to become global citizens by exploring their own place in the world, their values and their responsibilities to other people, to the environment and to the sustainability of the planet.

Geography. Programme of study for key stage 3 and attainment target. (An extract from The National Curriculum 2007) Qualifications and Curriculum Authority, 2007

Based on these statements, which of the essays shows more evidence of a higher level of development of what is deemed important in Geography?