Research and Analysis

Marking reliability studies 2017

Rank ordering versus marking – which is more reliable?

ofqual

Authorship

This report was written by Stephen Holmes, Beth Black and Caroline Morin from Ofqual's Strategy, Risk and Research Directorate

Contents

Executive summary

The English examination system values the use of extended response questions as valid ways of assessing important higher level skills, but these responses are harder to mark than shorter, or more constrained question types. This can impact upon the validity of the rank order of candidate work. It is therefore worth considering constructing rank orders of candidates by means other than marking, and testing them to determine whether they may be more reliable, thus supporting a valid candidate rank order.

This study looked at 2 different methods for placing AS history extended responses in a rank order and comparing these methods to 'traditional' marking using a mark scheme. Teams of 12 to 15 examiners per question generated rank orders for 60 responses for 3 questions taken from 3 papers using (i) paired comparative judgement and (ii) a comparative rank ordering approach. In addition, traditional marking had been carried out in a parallel exercise. The rank orders, internal reliability, predictive value of the rank orders and time efficiency of the different approaches were then compared.

One innovation in this study was the construct maps used by the expert judges to evaluate the quality of the responses in the 2 comparative approaches. The senior examiners for the papers from which these questions came devised the construct maps, which captured the elements of the response a candidate would need to demonstrate to give a good response to the question.

To further aid examiners in the rank ordering exercise, anchor responses were also used. These were responses that were clear exemplars of different levels of quality and were used in a training meeting, and the ranking task, to define the scale of quality. This approach was designed to make the task of creating a rank order of 60 responses more manageable and consistent across examiners than it might otherwise be.

The 3 methods produced rank orders that were highly consistent with each other, even though marking used the mark scheme to define response quality and the other 2 approaches used the construct maps.

When comparing the reliability of the rank ordering and marking approaches by looking at the similarity of individual examiner's rank orders to that of the principal examiner's rank order, there was no difference between the 2 methods. For rank ordering, there was a suggestion that more experienced examiners were more consistent with the principal examiner's rank, although this was not statistically significant.

Because paired comparative judgement combines judgements from all the examiners into one statistical model fit giving a single rank order it is not possible to make a direct comparison of reliability to the rank ordering and marking data. However, the statistical model fitted the data well, with high internal consistency, suggesting the examiner judgements were generally consistent with each other.

In terms of time efficiency, rank ordering took a similar length of time to carry out to marking. Although the full paired comparative judgement study took much longer to complete, there is evidence that the number of judgements collected could be reduced to take a similar time to complete as marking or ranking, whilst retaining the same predictive value as these other methods.

Although this study does not definitively identify one method as more effective than the others in producing a reliable rank order, it does suggest that there is scope to further investigate alternatives to marking without risk of creating invalid rank orders.

Alternative approaches such as these would need careful thinking about how they would work in practice. For example, consideration would be required as to how to scale up to a full entry size, and of how such alternative methods could have necessary features of a high stakes assessment system, such as transparency in how the rank order placing was determined for any individual candidate.

1. Introduction

Extended response questions are valued in the English qualifications system because they are seen as the most valid way to assess important skills such as synthesising material, constructing an argument or conducting an in-depth analysis of given information. They are thus likely to remain a feature of assessments in England. However, extended written responses to tasks and questions, whether in coursework, controlled assessment or examinations, are less reliably marked than responses to short answer questions, with objective questions (such as multiple choice) the most reliably marked. For open response questions, there is generally a relationship between the maximum mark (or tariff) of a question and the extent of disagreement between markers, as shown in the companion reports on marking reliability and marking metrics (Ofqual, 2018a,b; but see also Meadows and Billington, 2005 for a review). Although there may be changes to the marking process which reduce the absolute magnitude of disagreement between markers, this relationship is likely to remain. One of the fundamental aims of assessing candidates' performance is to produce a final rank order of their work that fairly reflects the quality of the performances demonstrated. Extended response questions capture highly valid evidence of the required skills, but lower marking reliability can undermine the validity of the rank order. Although this trade-off is well known and accepted – Gove (2013) acknowledged it in setting out part of the reform agenda – consideration of alternative ways to construct rank orders that may be more reliable (and thus more valid) are of interest to Ofqual.

Awarding marks by using a mark scheme may also sometimes produce unwanted effects. Sometimes over or underused marks are apparent in mark distributions, producing 'cliffs' in these distributions, often where there are implicit hurdles in the descriptors for the levels of response defined in the mark scheme. It may be that many candidates are missing certain features of the response vital to access a level and so are grouped at the top of the level below, or conversely, when many responses contain a feature which lifts them into a level, but they lack all the other characteristics of that level and so are clustered at the lowest mark of that level. There may also be difficulties in balancing the different aspects of a response where, for example, the mark scheme is not clear enough on how to deal with inconsistent responses. Finally, with the pressure on results through accountability, the presence of mark schemes provides detailed guidance to teachers on what will be rewarded in examinations, and this may wash back into teaching, creating an overly narrow focus in lessons on what is needed to get marks, rather than focussing only on learning and understanding. Approaches to creating a rank order that do not require mark schemes may therefore offer some advantages, although there will always be a need to define in some way what will be rewarded.

This study compared 3 different methods of creating a rank order of candidate work on extended response questions, in this case AS history questions taken from 3 different papers. For each question, the set of responses was ordered in 3 ways:

- using the marks given under traditional marking against a mark scheme (reported in a companion report on marking consistency, Ofqual 2018a)
- rank ordering using holistic judgement of responses and comparing them to 'anchor' responses and to one another to define the response quality scale
- paired comparisons in online comparative judgment

This report details the 2 comparative approaches used and the training and holistic construct map of response quality that we devised in order to help the participants make their judgements using these approaches, and compares the reliability of the ranking in the three separate approaches.

1.1 Using comparative judgement to construct a rank order

Comparative judgement is a technique that involves making relative judgements between items against a chosen criteria, rather than making an absolute judgement of the item in isolation. Laming (2004) stated that there are not really any truly absolute judgements, as even these can be viewed as relative comparisons between a concrete item and an abstract internal mental standard. In the case of awarding marks with a mark scheme, the mark scheme represents the internal standard against which work is being judged. However, the principle still applies that humans are better at making comparisons between two concrete items than they are at comparing a concrete item to an abstract standard, for which Laming (ibid) provides extensive evidence.

The foundations for the comparative approach lie with Thurstone (1927), whose law of paired comparisons first suggested that people may be more consistent when making relative judgements than when making absolute judgements. This was supported by Thurstone's investigations into ranking various sensory and psychological stimuli. Thurstone also provided the original mathematical formulation for how to place items along a scale (based on a normal distribution) constructed from a set of these paired comparisons.

Later mathematical formulations used logistic scales along which items are placed (Bradley & Terry, 1952; Rasch, 1960) but they are still based upon the collection of many individual paired comparisons between items, where judges are asked to decide in each case which of the 2 items presented better meets the chosen criteria (see Bramley, 2007, for a description of the use of paired comparison methods). The construction of a scale by fitting the mathematical model to the paired comparisons not only produces the rank order, but it allows properties of the model such as the consistency of judgements to be evaluated. Difficult to rate items (ones that have been viewed differently by different judges) can be identified and perhaps undergo additional scrutiny, and judges whose ratings diverge from the overall consensus, or are not consistent within their own judgements, can be picked out, and if necessary, excluded.

As well as this richness of data, comparative judgement offers several potential (or actual) advantages over absolute judgements, such as marking. As stated above, this is a natural psychological task that humans are good at, and this may promote greater consistency between judges. Any severity or leniency effects for particular markers are automatically eliminated due to the relative nature of the comparisons. Likewise, any reluctance on the part of examiners to use the extremes of the mark range is avoided. Most comparative judgement studies use a number of participants to make the judgements, and with modern online systems a large number of judges can make their judgements in a distributed fashion, which has the advantage of

capturing a group consensus and thus minimising the influence of idiosyncratic individuals.

In assessment, comparative judgement has a history going back to the 1990s and 2000s when it was used as a method for various comparability studies (eg D'Arcy, 1997; Fearnley, 2000). Pollitt and colleagues then began to promote the method from the early 2000s as an alternative to traditional marking (Pollitt, 2004, 2012). In more recent years there has been a growth of interest in this approach. Comparative judgement has been used in experimental studies to produce highly reliable rankings for a wide variety of item types such as design and technology portfolios (Kimbell, 2012), early years creative writing (Heldsinger and Humphry, 2010) and several types of mathematical work (eg Bisson et al, 2016; Jones & Alcock, 2014).

Several studies have looked at using comparative judgement to produce rank orders of extended responses or whole exam scripts. Raikes, Scorey & Shiell (2008) found correlations of 0.91 to 0.95 between marks and Rasch model parameters for whole AS biology scripts. These high correlations may have been related to the nature of the papers, which are typically much more reliably marked than those containing predominantly extended constructed responses. Whitehouse and Pollitt (2012) investigated ranking of an essay question on an AS geography paper and found very high reliability in the model fit[1], but only a moderate correlation of 0.63 between the model essay quality estimates and the marks awarded in traditional marking. Similarly, Furlong and Glanville (2015) carried out a study which rank ordered the quality of complete International Baccalaureate Organisation Diploma Programme English Literature scripts. They also obtained a highly reliable model fit, but the correlation of this rank order to that obtained with traditional marking was only 0.54. Although these rather moderate correlations could indicate different judgement criteria across the two methods, the correlations are not much different from typical inter-marker reliability measures for traditional marking of these kinds of extended responses (eg Meadows and Billington, 2005).

1.2 Rank ordering packs

Most of the more recent comparative judgement studies have used an online system to collect paired comparisons of items. An alternative to this is to rank order 'packs' of scripts/responses and convert these ranks to a set of paired comparisons by assuming that higher ranked scripts always 'win' the comparison against lower ranked scripts (eg Bramley, 2005; Black and Bramley, 2008). Although these inferred paired comparisons are not independent, Bramley (ibid) suggests that this makes little difference to the data apart from inflating the reliability measures. The Rasch model can then be fitted to these inferred paired judgements as per paired comparison judgement data.

This work focussed on maintaining standards across years, rating the quality of whole scripts across exam series. It allowed comparison of the standard across years and offered a way to equate assessments and set grade boundaries. As well as achieving high model reliability for whole A level psychology scripts, Black and Bramley (2008) obtained correlations of 0.81 to 0.88 between model parameters and

[1] It is worth noting that this study used an adaptive process to select the pairs of essays in the paired comparisons, and that recent work (Bramley and Vitello, 2018) has suggested that the internal reliability statistics are inflated using this adaptive method.

original marks awarded across several different conditions. The task of comparing the quality of scripts is similar to that in the present study. Using a 'pack' approach, with 10 scripts ranked within each pack, lessened the possibility of judges having to read through the same script more than once, as may occur in random paired comparisons. As well as potentially leading to boredom, this also has efficiency considerations in terms of time taken to construct the ranking. However, in order to tie all of the various packs together onto one large scale, scripts were mixed up between packs and participants, with each pack containing linking scripts.

1.3 Anchored rank order

In the current study we used a different approach to rank ordering. Rather than work in packs and use linking items to generate a larger rank order, we chose to rank the full set of responses in one large set. The larger the set of items to rank, the less manageable it will become, as participants have a lack of reference points to help place new responses in the growing set of ranked ones, and excess cognitive load or confusion can arise. To help overcome this problem, we used anchor responses to define the quality scale more clearly (see Figure 1).

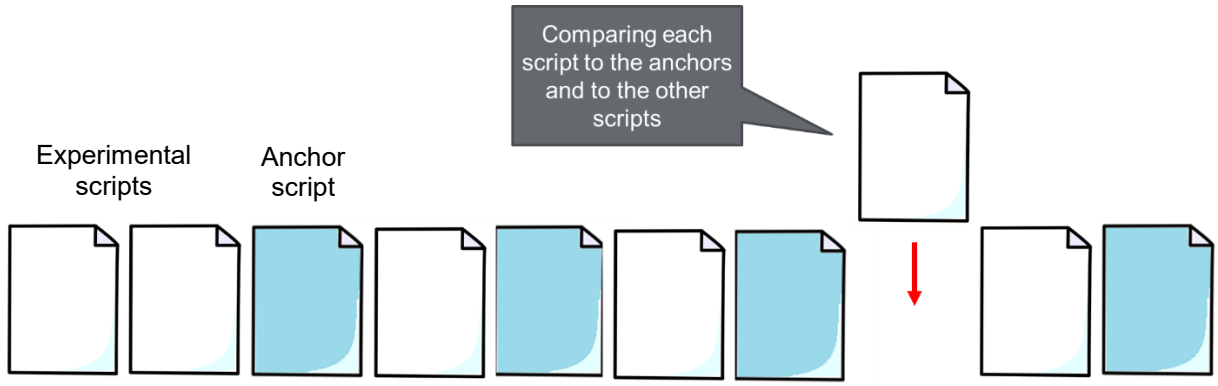

Figure 1: *The anchored rank order approach.*

Anchor responses should have been chosen as clear exemplars of the typical kind of response that might be expected at various points along the scale of quality. Ideally, they would be approximately equally spaced along this scale, and they should be straightforward responses to evaluate, with a consistent quality across different aspects of an answer. Responses that are hard to evaluate will define the scale poorly. A set of anchors for which all participants have a clear understanding of how their quality varies along the scale would provide a constant starting point for the decision of where to place each new response. The initial decision should be 'where does this response sit in the scale defined by the anchors?' In other words, which anchors lie either side of it. This decision then locates the new response into a narrow region of the rank order, and may therefore require fewer comparisons than placing a response into an unanchored set, where it may be hard to know where in the rank to start. Following the placing of a response into the anchor order, comparisons can then be made against the adjacent (usually non-anchor) responses to refine the final position of a response in the rank order.

As well as aiding the initial localisation of responses in the rank order, the anchors should help create a common scale across participants. If they are thoroughly discussed within the team, who all understand the rationale for the anchor rank order, there will be an increased probability of consistent placement of new responses, compared to each individual trying to place them in a relatively undefined scale.

1.4 Holistic rating of responses

In addition to trialling different methods of ranking responses, we also wanted to change the way in which the construct was defined with which the responses would be judged, away from mark schemes and towards a more holistic view of the response. We wanted our participants to consider the qualities of analysis and evaluation which make a good historian. To create this new construct the principal examiners (PEs) responsible for the selected questions worked with us to devise a construct map (eg Wilson, 2005) representing the elements required to produce a good quality answer, and selected anchors which clearly illustrated this construct. This development process is described in Section 1.2.3.

Having devised this construct map, a training meeting took place where the PE trained the examiners on how to evaluate the relative quality of responses by applying the construct map. This focussed on the anchor responses selected by the PE, as described in Section 1.2.4.

1.5 What gives comparative judgement such high reliability?

Before moving on to the current study, it is worth reflecting on a recent study which considered what it is about comparative judgement, particularly in the context of online paired comparisons, which leads to the high reliability measures reported. Model fits in comparative judgement studies typically have reliabilities over 0.8, and often above 0.9, indicating very high internal consistency, much higher than traditional marking in most subjects (although based on a different measure of reliability). Benton and Gallacher (2018) considered which aspects of comparative judgement give rise to these high reliability measures. They identified 3 possible factors that distinguish online paired comparisons from traditional marking:

- the use of multiple views of the essays by multiple markers
- the simple holistic mark scheme and relative judgements required
- the statistical model used to place items on a common scale.

Several aspects of this study are directly relevant to the current work, and so we will consider it in some detail. To try to disentangle the three listed factors, they analysed paired comparison data from an online comparative judgement study and marking data from 17 markers collected on a common set of GCSE English essays. Through comparison with external benchmarks of concurrent performance on related tests, they found that the paired comparison model gave better predictive power than single marking, similar predictive power to double marking, but was less predictive than the full 17-marker multiple marking.

They then used the multiple marking data to model a pseudo comparative judgement study. Paired judgements for the same pairs of responses that occurred in the comparative judgement study were decided by the scores assigned to the responses in the marking. Each judge's paired comparisons were matched to a different marker's scores, so that this was not simply a reflection of the rank order of one set of scores, but captured the multiple judges required of comparative judgement. This pseudo comparative judgement study had predictive accuracy at least as good as the true comparative judgement study. This suggested that there may be nothing special about making relative comparisons using a simple mark scheme in a comparative judgement study, but that it is the accumulation of evidence from a large number of individuals that is key.

In addition they found only a moderate effect of the use of a powerful statistical model. By applying similar statistical models to the marking data (ie Rasch or individual marker scaling) they obtained moderate increases in predictive accuracy. This was mainly related to the removal of individual marker severity/leniency, and probably also to the spread of their marks (their willingness to use the extremes of the mark range), which scaling or the Rasch model equalises.

When comparing comparative judgement to the mean mark of all the markers, there was a reasonably strong correlation between the two rank orders of 0.80, but given that the model fit in the comparative judgement study was quite weak (a scale separation reliability of 0.72), this is likely to have attenuated the correlation. A disattenuated value of 0.85 was suggested, which is much higher than the correlations for essay responses detailed in Section 1.1.

Finally, the PE mark was found to be a better predictor than individual markers, or even double marking, but had less predictive accuracy than full (17-marker) multiple marking, showing that the consensus mark was the best predictor of concurrent performance of all.

An analysis of time efficiency was carried out. Benton and Gallacher's estimate was that each paired comparison took around half the time of traditional marking (3.5 minutes vs 7 minutes), and so was gathering evidence 4 times faster (each judgement being on 2 items). However, the full comparative judgement study involved just under 14 views of each item, so 23.5 minutes per item, while double marking of equal predictive power took 14 minutes.

They therefore concluded that comparative judgement is "just a form of multiple marking combined with a simple mark scheme and fancy statistics" (Benton and Gallacher, 2018, p.22). Overall, paired comparisons gave similar predictive power as double marking but took around the same amount of time as triple marking, at least for the work assessed in these studies.

2. Methods

2.1 Design

Three questions from the AS history summer 2016 assessments were used, one question from each of 3 exam boards. We obtained a sample of scripts from the

exam board and selected the highest-tariff question for which we had responses for every candidate. These were all source based questions, one targeted AO3 and the other two targeted AO2 (see Table 1 and Figure 2). All 3 questions were on papers that had been marked by a team of 6 to 8 examiners in a marking reliability study (Ofqual 2018a) giving reliability measures for traditional marking. Some of these examiners also took part in the current study.

Table 1: *Questions included in the study*

Question	Nature of sources	Tariff	Assessment Objectives
AQA	two contrasting sources	25	All marks AO3
OCR	three sources	20	All marks AO2
Pearson	one source	12	All marks AO2

AO2: Analyse and evaluate appropriate source material, primary and/or contemporary to the period, within its historical context.

AO3: Analyse and evaluate, in relation to the historical context, different ways in which aspects of the past have been interpreted.

Figure 2: *Assessment objectives targeted by the questions in the study*

The data from the companion marking reliability study suggests that the source-based questions are amongst the least reliably marked on each paper, offering the most challenge to reliable application of the mark scheme. This is likely to be because of the high tariff and possibly because there is scope for the candidates to give atypical responses to, and interpretations of, the sources, making marking more subjective.

2.2 Examiners

Each set of responses was evaluated by 12 to 15 examiners including the PE (see Table 2). Most of the examiners had not taken part in live marking of the summer paper, but had either marked a different paper, or were from the 'eligible-to-mark' reserve list held by the exam board – they had been cleared as qualified to mark but had not been employed previously. Although familiar with mark schemes, they would not have been trained to apply the summer 2016 mark scheme for this paper, and so would not have internalised that mark scheme to any extent. The other examiners had taken part in live marking in summer 2016, and also our marking reliability study. They would have been familiar with the mark scheme and may potentially have found it harder to avoid using this knowledge in their judgements.

Table 2: *Participants in the study*

Board	Total no. of examiners	Also took part in study 1 (inc PE)	'Eligible to mark' examiners

AQA	15	4	11
OCR	15	5	10
Pearson	12	2	10

2.3 Construct map development meeting

The 3 PEs attended a joint meeting with the Ofqual researchers to devise a construct map illustrating a set of related qualities, concepts and skills that good answers are likely to exhibit in answering the questions. The intention was for the responses in this study to be compared using this construct map to decide on their relative qualities.

The meeting began with a discussion of the qualities of 'what makes a good historian', before each question was discussed in turn with an analysis of 'what makes a good response to the question'. Through group discussion, with the respective PE taking the lead, a representation of the processes involved in constructing an answer was devised. For each question, a slightly different construct map was created. The final construct maps are shown in Figure 3 to Figure 5. The 2 AO2 construct maps (Figure 4 and Figure 5) are identical except for 1 word change. For each construct map there is a general progression from beginning the analysis in the lower left and coming to final conclusions in the upper right, and the way the different elements relate to one another are indicated by the overlap of the elements. The size of the bubbles for each element were intended to be broadly representative of the importance of each aspect in the holistic evaluation of the response.

At all times the PEs were encouraged to think beyond the mark scheme, and to consider any aspects of answering the question which the mark scheme might not be able to fully capture. All 3 PEs embraced this, recognising that although the mark scheme captured most aspects well, the need to encourage highly reliable marking could sometimes make capturing certain qualities more difficult. Generally, the construct maps included many aspects explicit in the mark schemes, but also contained some aspects not explicitly stated.

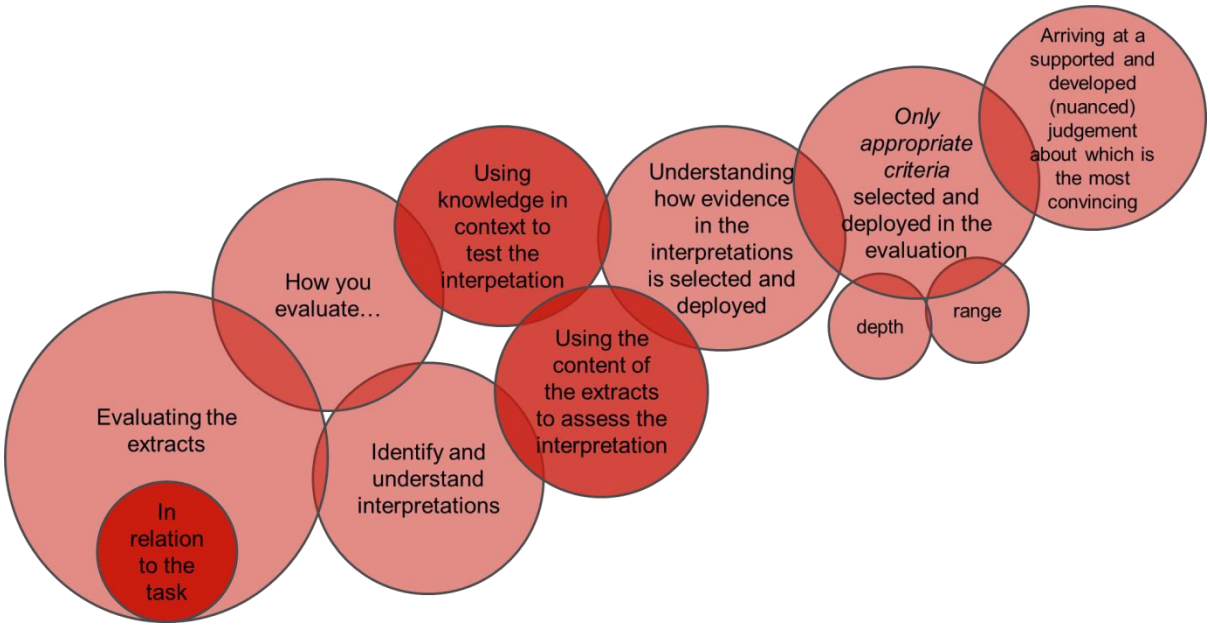

Figure 3: *AQA AO3 construct map*

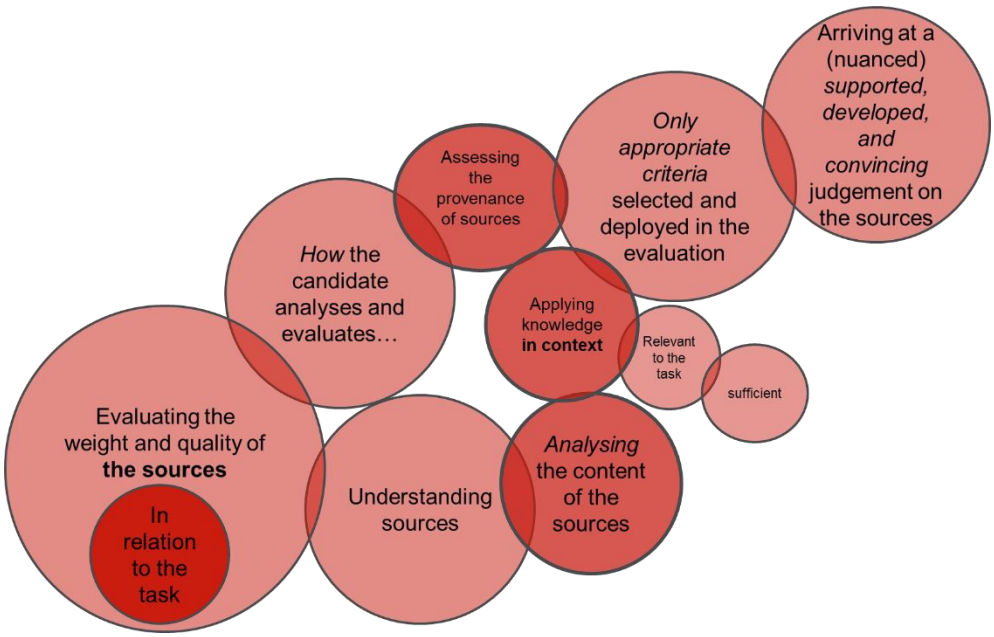

Figure 4: *OCR AO2 construct map*

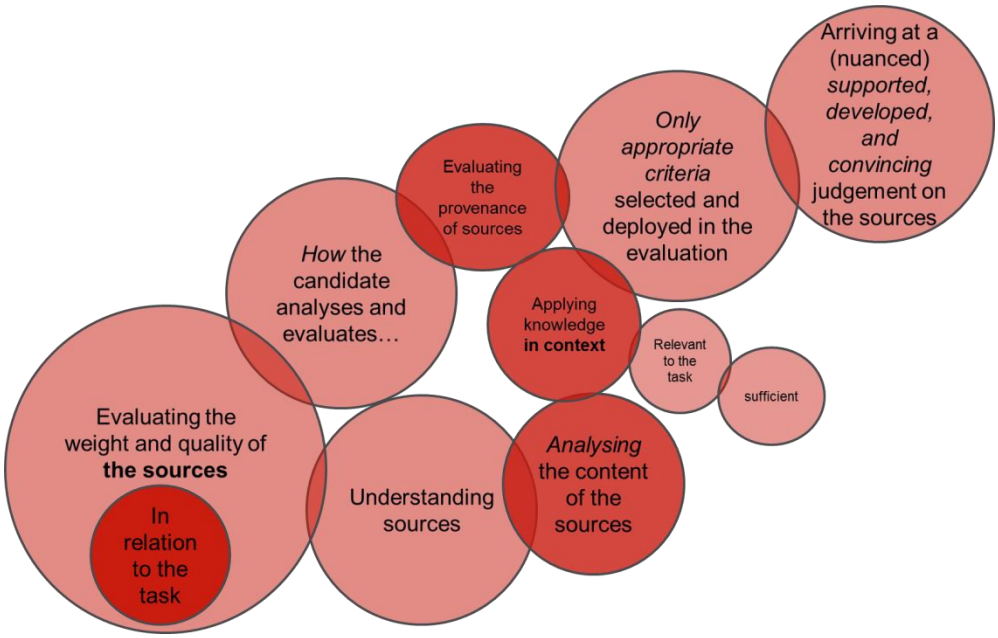

Figure 5: *Pearson AO2 construct map*

The training meeting for examiners was also planned, including how the PEs should select the anchor responses, discussion of approximately how many they might select, and some planning of the actual training the examining teams would receive at the meeting.

Following this meeting, the PEs worked at home selecting anchor responses from the full set of 100 responses that were marked in the reliability study. They were

instructed to choose responses that clearly represented the construct map at well-spaced intervals along the quality scale. Each PE decided how many anchors were required and all 3 PEs selected 6 or 7 anchors. The intention was for the anchors to exemplify features of the construct map and provide clear reference points along the scale to help with initial slotting of the experimental responses in the main ranking exercise. They also prepared any additional training materials for introducing the construct map, and selected additional practice responses for the examiners to use in the meeting. The PEs then carried out any preparation required for delivering the training.

2.4 Training meeting

All examiners attended a day-long training meeting for the task. After an overview of anchored rank ordering and comparative judgement, and an instruction to try to think beyond mark schemes, each marking team (as listed in
Table 2) was trained by their PE on how to evaluate the relative quality of answers using their construct map. It was entirely up to the PE how this training was delivered. It generally began with an introduction to the question and what makes a good or weak response, followed by the introduction of the construct map and a discussion about it. The anchor answers were then introduced one by one by the PE, with their characteristics (their strengths and weaknesses) analysed and discussed by the group with reference to the construct map. Where time allowed, a variable number of additional answers were analysed and slotted into positions in the rank order as defined by the anchors, and examiners were given opportunities to practise on more examples in groups. At the end of the meeting all of the materials required for the task were handed out.

2.5 Materials

Sixty responses for each question, free of any annotations or marks were randomly selected from the set of responses marked in study 1 for the ranking exercise. In addition, there were 6 anchor responses for AQA and Pearson and 7 anchor responses for OCR. Three additional responses per question were provided to be used in a pre-ranking qualification test.

For the anchored rank ordering task paper copies of all the responses (including pre-test and anchors) were used, while pdf versions were used for the online comparative judgement.

2.6 Procedure

The examiners and the PEs all took part in both rank ordering exercises, in a random counter-balanced order (see Table 3). The counter-balancing equally split experienced and eligible to mark examiners across the 2 orders. This ensured that the results for rank ordering and comparative judgement were not biased by practice effects. At the end of the training meeting each examiner was told the order they would be carrying out the tasks and were free to start their first task the next day.

Table 3: *Counter-balanced design of tasks*

Group X	Group Y
1. Rank order	1. Paired comparison
2. Paired comparison	2. Rank order
3. Survey	3. Survey

2.6.1 Anchored rank ordering task

The examiners were provided with paper copies of all 60 responses to be ranked, 3 test responses and the 6 or 7 anchor responses on coloured paper to make them stand out. Examiners also had a pack of sticky labels on which the construct map was printed. They were encouraged to annotate this label for each answer, indicating the qualities of the answer against the different elements of the construct map, as an aid-memoire when returning back to a response during the ranking process.

As a pre-test, all examiners were asked to place the 3 test responses into the rank defined by the anchor responses and indicate to their PE where they had placed them. The PE then had a discussion with them about their placing of the test responses in the rank order, particularly where they disagreed with the PE's own placing. This was a final opportunity for the PE to check that they understood the application of the construct map to the responses. Regardless of agreement with the PE at this stage, all examiners continued on to the main ranking exercise.

The main task involved placing the 60 responses into the rank order initially outlined by the anchor responses. The suggested process was to take each new experimental response in turn and slot it into the place they perceived it to fall on the quality scale defined by the construct map and the anchors, then to use previously slotted experimental responses to refine the position. The examiners used the construct map and the scale defined by the anchors, together with what they had learnt at the training meeting to make their decisions.

2.6.2 Online paired comparisons task

Pdf copies of the 60 responses plus the 6 or 7 anchor responses were uploaded to the No More Marking comparative judgement system. Judges were given detailed instructions on how to use and access the platform. Pairs of responses were presented side by side on the screen and the judges were prompted to click on a button on the screen indicating:

> 'Which is the better response to the question?'

Unlike most comparative judgement studies, which do not closely define the criteria for making the holistic judgement of quality, we clearly defined the judgement criteria in this exercise using the training and the construct map. The judges were encouraged to read the responses carefully and evaluate both against the construct before finalising their judgements. Of course, some pairs were relatively easy to distinguish, and some responses will have been seen before so may have been

partially recalled. Each judge was asked to complete 60 random paired comparisons and they were free to work to their own schedule, just with a deadline by which the judgements had to be completed. The responses were randomly distributed among judges so that all responses were seen a similar number of times. For the examiners in Group X (Table 3), a link to access the system was sent once the examiners submitted their ranking of the responses from the anchored rank ordering task.

3. Results

3.1 Traditional marking

The correlations[2] between examiners from the marking reliability study are given here for every question on the 3 papers in order to place the 3 experimental questions in context (see Table 4). Each question was marked by 6 to 8 examiners.

Two correlations are calculated for each question:

- average inter-correlation of the rank orders between all pairs of examiners. Individual correlations are calculated and then a mean obtained across the matrix
- average correlation between each examiner's rank orders (both assistant examiners and team leaders) and the PE rank order. Individual correlations are calculated and then a mean obtained

Table 4: *Traditional marking: average inter-correlation of the rank orders for all examiners, and the average correlation of examiners' ranks with the PE rank, for the experimental questions and the other questions on the paper.*

Paper/ question	Average inter-correlation	Average correlation with PE
AQA		
Experimental question	**0.52 (range 0.33 to 0.66)**	**0.53 (range 0.43 to 0.61)**
Other questions	0.71 to 0.72	0.66 to 0.75
OCR		
Experimental question	**0.54 (range 0.35 to 0.73)**	**0.56 (range 0.49 to 0.63)**
Other questions	0.54 to 0.72	0.57 to 0.75
Pearson		
Experimental question	**0.54 (range 0.33 to 0.71)**	**0.62 (range 0.39 to 0.71)**
Other questions	0.45 to 0.75	0.55 to 0.81

[2] All correlations reported are Spearman rank order correlations as the rank order is of primary interest in the study.

The reliability for the questions tested here are all in the range 0.52 to 0.62 indicating moderate agreement between markers. This is lower than the reliability for the non-source based essay questions, which are typically around 0.7. The agreement of individual markers with the PE rank is slightly higher than agreement amongst all markers, although there was quite a lot of variation in the correlations between individual markers.

3.2 Anchored rank ordering

For the anchored rank ordering task, the average rank order correlations (comparing individual examiners to all others, or to the PE) for the 3 questions are shown in Table 5.

Table 5: *Anchored rank ordering: average inter-correlation for all examiners, and the average correlation of examiners with the PE.*

Question	Average inter-correlation	Range	Average correlation with PE	Range
AQA	0.50	0.15 to 0.86	0.57	0.36 to 0.72
OCR	0.57	0.16 to 0.81	0.63	0.25 to 0.79
Pearson	0.50	0.16 to 0.68	0.58	0.31 to 0.68

In all cases the correlation with the PE's definitive rank is higher than the agreement amongst all examiners. This is a similar effect to that seen in the traditional marking data. The correlations with the PE we obtained for the AQA and OCR questions show a very small increase relative to traditional marking, while for the Pearson question it is reduced. The averaged correlation with the PE across the 3 questions is 0.59 for rank ordering and 0.57 for traditional marking, a non-significant difference ($t(55) = 0.69$, $p = 0.50$). The mean of the all marker inter-correlation is similar across traditional marking and rank ordering for all 3 questions (rank order 0.52 vs marking 0.53, n.s.).

From the range of correlations it is clear that some pairs of examiners were not consistent with each other, even with the anchors in place, with correlations down to 0.15. However, agreement with the PE was slightly better, with correlations usually above 0.3. There were also some very high correlations between examiners. It is noteworthy that these individual correlations are more widely spread than the corresponding range of correlations for marking shown in Table 4. This large range

is probably a consequence of the larger team size for rank ordering, which will naturally produce a larger range of correlations. However it is worth remembering that the examiners who took part in the rank ordering were a more varied group than the experienced examiners in the marking group, as they included both experienced and inexperienced examiners on the unit. There may also have been a contribution to this spread from the degree to which individuals were able to internalise the new construct in the short time available.

3.2.1 Effects of marker experience

Although we only had a relatively small number of experienced examiners (see Table 2), we split the examiners by whether they had previously marked the item in summer 2016 or were new, eligible-to-mark examiners. We obtain the mean correlations with the PE shown in Table 6. The number of examiners in each category is shown in brackets in the table.

Table 6: *Anchored rank ordering: average correlation of examiners with the PE for the 2 groups of experienced or new examiners for each question. The number of examiners per group is shown in brackets.*

Question	Experienced markers		New markers	
AQA	0.58	[3]	0.57	[11]
OCR	0.72	[4]	0.60	[10]
Pearson	0.65	[1]	0.57	[10]

Experienced markers had a slightly higher average correlation across the 3 questions, 0.66 as opposed to 0.58, which might suggest greater familiarity with the way the PE thinks, or greater familiarity with this material, but this was not a significant difference ($t(37) = 1.55$, $p = 0.13$). The lack of a significant effect may have been due to a lack of statistical power (we had only a few experienced markers in this study), or there may have been no underlying effect as the construct map used to assess the quality of responses was new to all markers. However, we may tentatively take the size of the correlation for the experienced examiners are more representative of the potential of this overall approach as their judgements were not compromised by any limitations in their experience.

3.2.2 Order effects

A check was made that there were no order effects occurring where agreement with the PE was higher for those carrying out the rank ordering exercise after completing the online comparative judgement. These correlations between the examiners and the PE are shown in Table 7. The number of examiners in each category is shown in brackets in the table.

Table 7: *Anchored rank ordering: average correlation of examiners with the PE for the 2 task order groups of examiners. The number of examiners per group is shown in brackets.*

Question	Rank order first		Rank order second	
AQA	0.56	[7]	0.59	[7]
OCR	0.61	[7]	0.66	[7]
Pearson	0.58	[6]	0.57	[5]

Although the group carrying out the rank order second had larger correlations (average across the 3 questions 0.61 vs 0.58) this was not a significant difference ($t(37) = 0.65$, $p = 0.52$). Again the statistical power may be too low to detect small differences, so it is unclear whether some participants may have benefitted from greater familiarity with the candidate responses or applying the construct map. The size of the correlations for the rank order second group may be more representative of what is possible with this rank order approach.

3.3 Online paired comparisons

All judgements with a judging time of less than 10 seconds were removed on the basis that this was either an error or represented unreasonably rapid judging of quality. Eleven judgements were each removed from the AQA and OCR studies and 12 from the Pearson study.

The R package sirt was used to estimate quality parameters for each response under the Bradley-Terry (1952) model. R code was also used to estimate item and judge infit, and scale-separation reliability (SSR). Table 8 shows the model statistics after the data cleaning.

Table 8: *Summary statistics for the 3 paired comparison model fits*

Question	Number of judgements	Judgements per item	SSR
AQA	892	27.0	0.88
OCR	892	26.6	0.87
Pearson	705	22.7	0.88

Reliability is quantified in paired comparison studies by the SSR statistic which is derived in same way as the person separation reliability index in Rasch analyses, and is analogous to Cronbach's Alpha. It is interpreted as the ratio of true variance to observed variance in the estimated scale values. Reliabilities of almost 0.9 indicate that a robust statistical scale of response quality was obtained. The model fit was consistently good across the 3 questions.

Following the fitting of the Bradley-Terry model, judge infits were checked. Infit is the information-weighted mean square of the residuals for the set of judgements for each judge. Typically, any judges with infits more than 2 standard deviations above the mean infit is considered to have judged inconsistently or to have a different internal standard to the other judges.

For the OCR study, one judge had an infit 2.53 z-scores from the mean. Removal of this judge improved the model fit to give an SSR of 0.88, so this second model fit was used for the analysis in the following section.

For the Pearson study, one judge also had an infit 2.53 z-scores from the mean. Removal of this judge reduced the model fit to give an SSR of 0.87, so this judge was not removed and the initial model fit was used for the analysis in the following section.

The anchors were presented in this study as part of the larger set of responses, and were not fixed in their rank positions in any way. We found that in all 3 final rank orders, the anchors came out in the same order that the PEs had initially placed them. This indicates that the original choice of anchors by the PE was sensible, and that the rank order in the paired comparison study was consistent with that from the rank ordering study, at least to the extent of the same ordering of the anchors.

3.3.1 Judgement time

After removal of sub-10 second trials, and the one misfitting OCR judge, across all 3 studies the median judging time was 230 seconds. Median judging times for all the judges and for the 2 different order groups within each study were then calculated. Table 9 shows that although there are differences between the groups within a question, overall there is no consistent trend of those carrying out the paired comparison task second going much more quickly due to familiarity with the responses.

Table 9: *Median judgement time (in seconds) per trial for all judges, and split by task order. The number of judges in the 2 order groups is shown in brackets.*

Question	All	Paired comparison first		Paired comparison second	
AQA	223	247	[8]	192	[7]
OCR	273	271	[8]	278	[6]
Pearson	180	165	[6]	227	[6]

It is possible that with different instructions and criteria against which to judge quality the judging time could have been reduced. Most paired comparison studies leave the judging criteria loose and allow judges to decide how much information they need before coming to a decision. However, we wanted our judges to carefully read the responses and evaluate them against the construct, and to avoid making snap judgements using incomplete information or superficial features.

3.4 Relationship between all three rank order approaches

The mean mark across all markers from the marking reliability study was converted into a rank order of responses. This marking rank order, the mean rank order from

anchored rank ordering and the model output rank order from the paired comparisons were then correlated and the coefficients obtained are shown in Table 10.

Table 10: *Correlations between the 3 different rank orders generated from the mean marks for traditional marking, the mean rank for anchored rank ordering and pooled paired comparisons, for each question.*

Correlation between approaches	AQA	OCR	Pearson
Anchored rank order vs paired comparison rank order	0.89	0.94	0.88
Anchored rank order vs marking rank order	0.86	0.92	0.91
Paired comparison rank order vs marking rank order	0.84	0.86	0.84

Across the 3 questions, the mean correlation between paired comparison and anchored rank ordering was 0.90, and that between paired comparison and marking was 0.85. The mean correlation between the anchored rank ordering and marking was 0.90. In all cases the paired comparison rank order correlates more highly with the anchored rank order than the marking rank order. For all but the Pearson question, the anchored rank order correlates more strongly with the paired comparison rank order than the marking rank order.

These findings suggest that all 3 methods are broadly consistent with each other, although there is some suggestion that the paired comparison rank order may be capturing a slightly different quality to that of the traditional marking. This is not a large effect, since the correlations are always well above 0.8.

Correlations based on the rank orders of individual examiners in the marking and rank ordering task would probably be lower than these averaged rank orders, given the variability between individuals seen in sections 3.1 and 3.2. However, the purpose of this analysis was to determine if there were gross differences in the underlying rank orders each method generated.

3.4.1 Relationship of the rank orders to a concurrent benchmark

Although all 3 methods produce similar rank orders, as evidenced by the high correlations in Table 10, there may be some differences in how well they can predict a related outcome, the rank order of the marks on the second component of the AS history qualification (this may be viewed as a measure of concurrent validity, although we will use the term predictive value). As an additional analysis we compared the effect of varying the number of examiners in the marking and rank ordering exercises, or the number of judgements in the paired comparison study, on the correlation.

Table 11: *Correlations between the 3 different rank orders generated from the mean marks for traditional marking, the mean rank for anchored rank ordering and pooled paired comparisons, compared to the rank order of the mark on the second component of the qualification.*

Correlation between rank order approaches and second component mark	AQA	OCR	Pearson
Marking rank order	0.61	0.45	0.62
Anchored rank order	0.58	0.51	0.65
Paired comparison rank order	0.53	0.50	0.66

Table 11 shows that for all 3 questions, the 3 different rank order methods produced similar correlations to the rank order of candidates' mark on the second component of the qualification, with no one method producing a more accurate prediction. The OCR question on average showed a less close relationship between rank orders than the other 2 questions, perhaps due to the nature of the split in content across the components. Averaged across the 3 questions, all the rank order methods showed a correlation of 0.56-0.58.Given the higher correlations across the methods we found for the Pearson question, we concentrate on this question when considering the way in which the correlation with the mark on the other component changes with different numbers of examiners, or different numbers of comparisons in the paired comparison study. The patterns observed below were very similar for the other 2 questions.

For marking and anchored rank ordering we calculated the mean rank order for each size of examiner team by randomly drawing the required number from the full group of examiners and taking averages of the mark or rank for each response and repeating this process 100 times. The mean rank-order correlation across the 100

repetitions was then calculated. For the paired comparison task we modelled 100 random selections of trials at each point, and averaged the correlation between the model fits to these trials and the overall qualification mark.

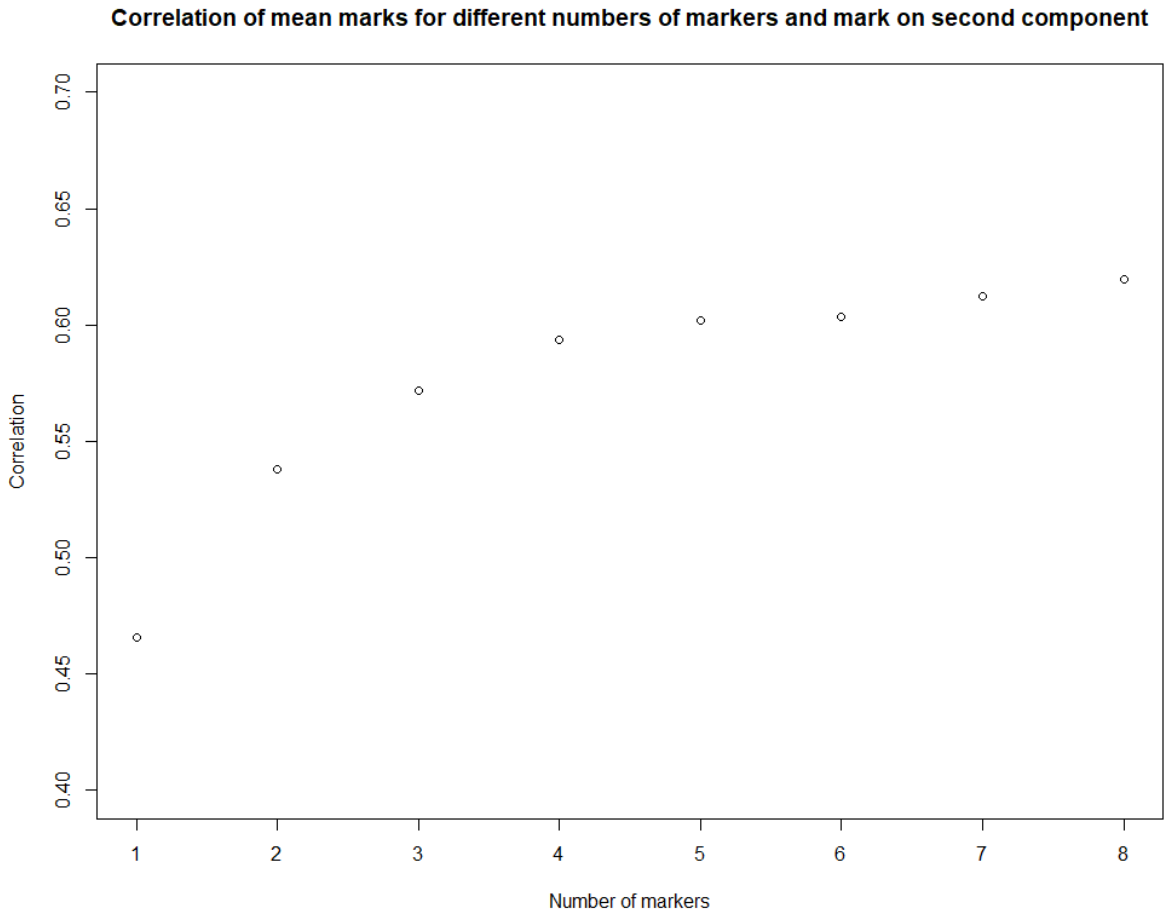

Figure 6: *Change in the correlation between the rank order derived from the mean mark and the mark on the second component of the qualification for different numbers of examiners for the Pearson question. Each data point is based on 100 repetitions with different random examiner selections each time.*

A single examiner's marking rank order correlates with the rank order of the second component in the qualification around 0.47 (see Figure 6). Double marking increases this to 0.54 and triple marking to 0.57, with a gradual rise as more examiners are averaged to the maximum correlation for all examiners of 0.62. Moving from single to double marking represents the largest jump in predictive value.

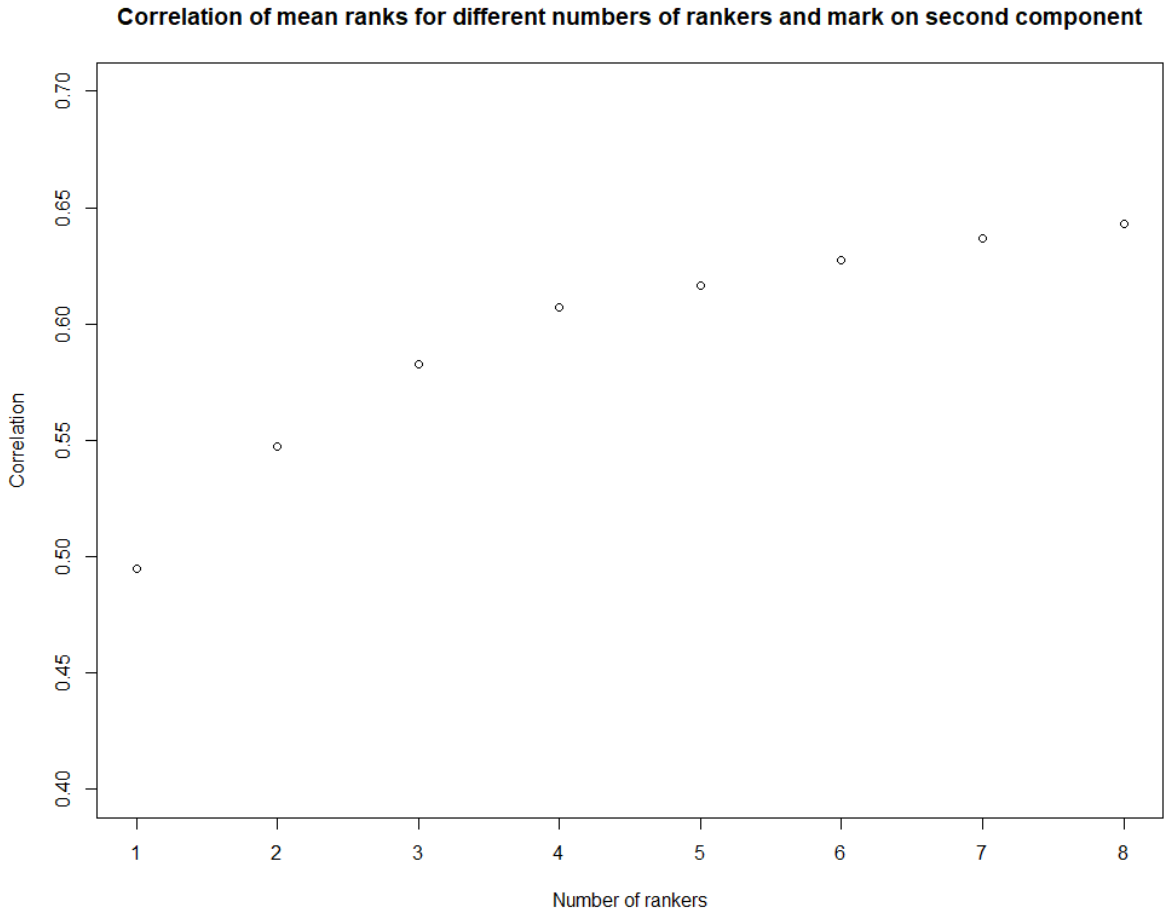

Figure 7: *Change in the correlation between the anchored rank order and the mark on the second component of the qualification for different numbers of examiners for the Pearson question. Each data point is based on 100 repetitions of different random examiner selections each time. This plot has been limited to 8 examiners for comparison to the previous figure.*

A single examiner's rank order correlates with the rank order of the second component in the qualification around 0.49 (see Figure 7). The mean rank of 2 examiners increases this to 0.55 and triple ranking to 0.58, with a gradual increase as more examiners are averaged up to the maximum correlation for all examiners of 0.65. Moving from one examiner to the average of 2 examiners represents the largest jump in predictive accuracy. This pattern is almost identical to that seen in the marking data in Figure 6.

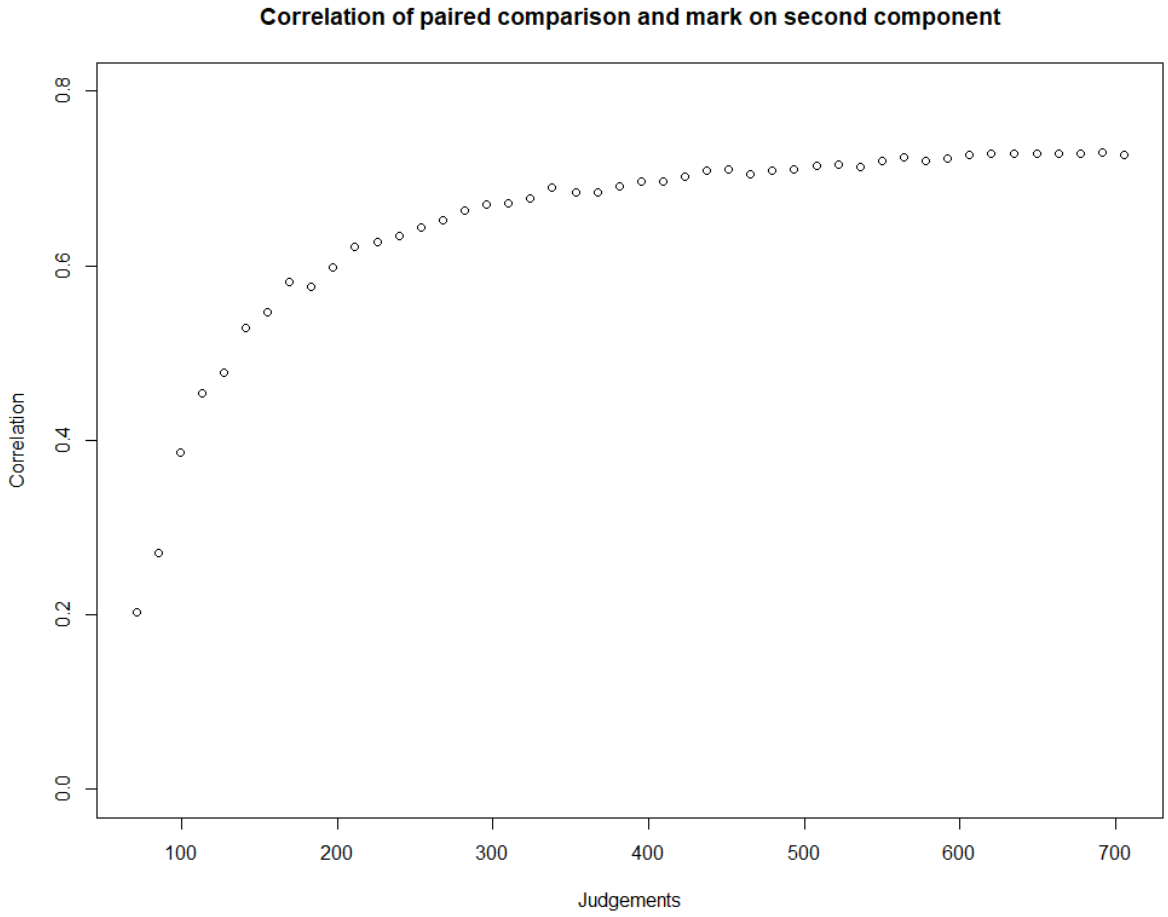

Figure 8: *Change in the correlation between the paired comparison rank order and the mark on the second component of the qualification for different numbers of judgements for the Pearson question. Each data point is based on 100 different random trial selections.*

For the paired comparison data, we simulated the effect of running a study with fewer judgements collected by randomly selecting paired comparisons from the full 705 trials and correlating the rank order from the resulting model fit to the rank order of candidates on the second component of the qualification. The key finding from the simulations shown in Figure 8 is how far the number of judgements can be reduced until the model loses predictive power. At around 200 judgements the decline in the correlation grows steeper, however, if we take a correlation of 0.48 as indicative of what is achievable for single marking and rank ordering, this equates to around 130 judgements (or 4.2 judgements per item). Similarly, double marking and rank ordering gave a correlation with the whole qualification mark of 0.55. This equates to around 155 paired comparison trials (5.0 judgements per item). Therefore the potential to gain additional predictive accuracy grows more rapidly for paired comparisons (a 20% increase in trials required) than the doubling of time required for 2 examiners to mark or rank. We will consider time efficiency further below.

3.5 Time required to complete the task

Examiners were asked in a post-task survey how long it had taken to complete the anchored rank order task. The mean time they reported was 14.6 hours. The range was wide, extending from 4 hours to 40 hours, with a median time of 12.5 hours. It is possible that more training and guidance would have helped to increase consistency in the time each person took. A mean time of 14.6 hours equates to around 15 minutes to place each response in the rank order. This is probably only a little greater than the time taken to mark these items against a mark scheme. We did not collect this marking data.

To calculate the time taken to complete the online paired comparison task we used the individual trial time data collected automatically by the system. We again excluded trials which took less than 10 seconds, excluded the one OCR judge who had a high infit (as in the main analysis), and then combined trials from all 3 exam boards to give 2,429 trials. Twenty-four judgements which were more than 3 standard deviations from the mean judging time were excluded on the basis that these probably represented times when the system was left logged in but not being actively used (it continues to log trial time in this case).

From the remaining 2,405 trials we calculated a mean judgement time of 304 seconds, which we will round to 5 minutes for convenience. Given that each trial generates a decision for 2 responses, the time required is 2 ½ minutes per response. As a comparison to rank ordering (and, we assume, marking) time, the 4.2 trials required to gain the same predictive value as single marking equates to 10 ½ minutes, a little less than the mean time required for rank ordering, and probably close to the time required to mark one response. Therefore all 3 methods are relatively equal on time required, by these calculations.

If we move to double marking or ranking, the time required is up to 30 minutes per response. The equivalent time required to gain the same accuracy for the paired comparison method is 12 ½ minutes, a significant gain in efficiency. So although the paired comparison study we ran took around 60 hours to complete (705 trials in the Pearson study at 5 minutes each), compared to about 15 hours for the rank ordering task (and probably slightly less for the marking), if the methods are equated for the level of precision they produce, paired comparisons may be approximately equal in time required when single marking is the benchmark, or significantly quicker at a higher level of double-marked precision.

The findings here are similar, but not identical to those of Benton and Gallacher (2018). We found roughly equal time efficiency for single ranking or marking and paired comparative judgement of the same predictive power, whereas they found that comparative judgement took slightly longer than marking to arrive at similar predictive power. However, there are differences in both the materials considered and the instructions and training given to the participants in the 2 studies, and our estimate of marking time is based on the time to complete rank ordering, which could be a slight over-estimate. It is possible, too, that comparative judgement would become quicker once the task, the constructs etc became more familiar judges.

3.6 Internal reliability of the paired comparison model

In the paired comparison approach, internal reliability is measured through the scale separation reliability (SSR) which was almost 0.9 for all the questions. The marking and ranking approaches produce lower inter-rater reliabilities (around 0.6), but it is unclear how SSR compares to inter-rater correlation (eg Verhavert et al, 2018). Therefore we carried out the predictive value comparisons in section 3.4.1. It is still interesting to see how the SSR declines as fewer and fewer judgements are made for each study. This can be simulated by randomly sampling from the full set of judgements and re-fitting the model. We repeated this exercise 100 times for each number of judgements (see Figure 9). There is a steady decline in SSR as the judgement number is reduced, which becomes increasingly steep and below about 150 judgements the model fit starts to fail as indicated by the variable SSR values. The same pattern was observed across the model fits for all 3 questions.

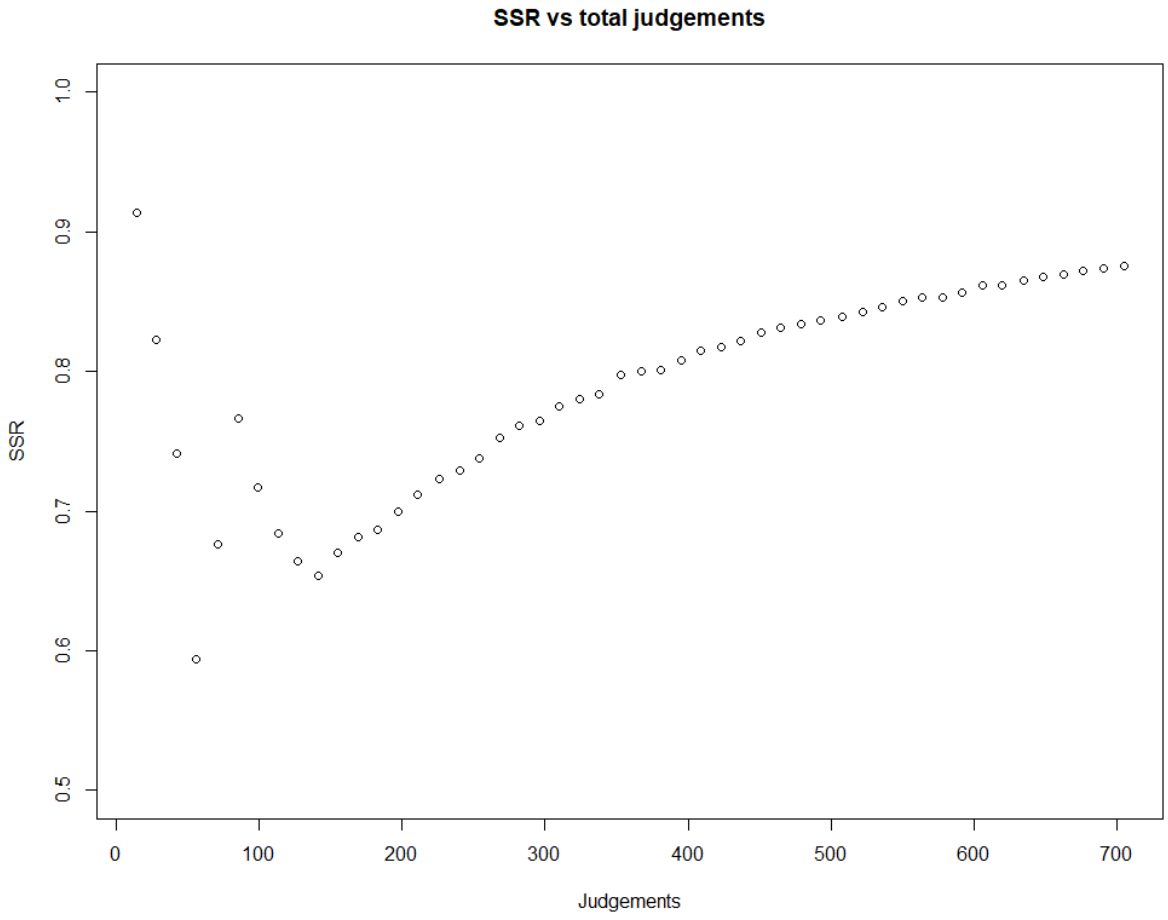

Figure 9: *Change in SSR with different numbers of judgements for the Pearson paired comparison study. Each data point is based on 100 different random trial selections.*

Compared to Figure 8, the SSR drops more steeply than the correlation to the second component on the qualification. The steadily declining SSR in Figure 9 therefore mainly indicates increasing error values for the parameter estimates for each item, and not a serious disruption of the underlying rank order. However it does show that when the model is fitted to the number of judgements that give equal predictive value to single marking (~130 judgements) it is right on the limit of fitting, and that is why the correlation to the second component reduces so dramatically below this point.

3.7 How were the relative judgements made?

One final issue is that it is not entirely clear how well our participants were able to exclusively use the construct map in making their relative judgements in both the paired comparison and rank ordering tasks, given their familiarity with mark schemes. Although our training emphasised the use of the construct map and to put the mark scheme out of mind, they may have automatically found themselves thinking about how many marks a response was worth, so the mark scheme may certainly have influenced decisions, if only unconsciously and unintentionally.

If this effect were to arise, it would probably be more evident in the experienced group of markers, given their familiarity with applying the mark scheme to these responses. Table 12 shows the correlation of the average anchored rank order for the 2 marker groups with the mean marking rank order. If the mark scheme influenced the experienced group we would expect to see a higher correlation with the marking rank order for this group.

Table 12: *Anchored rank ordering: correlation of the mean rank order obtained from the 2 groups of experienced or new examiners and the mean marking rank order for each question. The number of examiners per group is shown in brackets.*

Question	Experienced markers		New markers	
AQA	0.82	[3]	0.84	[11]
OCR	0.91	[4]	0.90	[10]
Pearson	0.85	[1]	0.91	[10]

These correlations suggest that the experienced markers did not produce an anchored rank order more like the marking rank order than the new markers. Familiarity with the mark scheme does not appear to have unduly influenced the experienced examiners and this suggests that they were not simply rank ordering by marking, or at least no more than the new markers.

4. Discussion and conclusions

The 2 comparative rank ordering methods tested here, anchored rank ordering and paired comparative judgement, both generated rank orders consistent with those obtained from traditional marking. For anchored rank ordering the consistency of the rank orders of individual participants was at least as high as that from marking. The slightly higher agreement amongst the experienced participants, and for those

carrying out the task after more practice (the paired comparison exercise), suggests that this approach may have the potential to be more reliable than marking.

In comparing the overall mean rank orders produced by the 3 methods to the rank of the mark of each candidate on the second component of the qualification, the 3 methods appear to have roughly equal predictive value. When comparing the correlation of different numbers of markers for marking and rank ordering and different numbers of judgements for the paired comparisons, equal predictive power to one examiner marking or rank ordering can be achieved in the paired comparison approach with a number of judgements taking similar time to one person marking or rank ordering. However, paired comparisons can match the predictive accuracy of double marking or 2 examiners rank ordering with significantly less time invested. However, one caveat is that the point at which the paired comparison rank order starts to deteriorate (and the whole model fit starts to collapse) would need to be very precisely known in order to set this minimum number of judgements in advance with confidence. Given the uncertainty, there would probably be a need to include a safety net of a fair number of additional judgements to be sure the model would not fail to fit the data, thus making this approach slightly less efficient.

By basing the judgements of quality on a construct map, devised by experts in assessing historical analysis, it may be possible to capture some qualities that are hard to capture with mark schemes, which are always designed with the need to promote reliable marking. The fact that the rank orders produced here were no less reliable when the construct map was used to define quality than when a mark scheme was used suggests this more holistic alternative conception of quality can be used without leading to more disagreement between experts.

The online paired comparison approach provides a lot of data that could be used in a live situation, such as identifying responses which have a high infit error, indicating that they may be particularly difficult to classify. These responses could be passed up the hierarchy to more senior, experienced examiners, who could use a more traditional rank ordering approach to place these difficult responses in the final generated rank order. Grades could be set using something very close to the standard awarding meeting, combining statistics around proportions of candidates achieving certain grades, and expert judgement of the responses in the rank order itself.

There are of course a range of issues around the use of any form of comparative judgement as a replacement for marking, in that the outcome is less transparent, being based on a holistic judgement which is harder to audit than marking where it is easier to see how markers are awarding marks through their annotations. Examiners could be asked to write a justification of each of their decisions in a paired comparison study, but this would generate a very large set of comparative statements to collate and potentially analyse. In both online paired comparisons and anchored rank ordering an electronic implementation of the annotation of the construct map on the sticky labels we provided could be used to record how the examiner assessed the quality of a response.

However, the upside of these distributed comparative judgement approaches is that these justifications may not be required. In a live examining context, to construct a complete rank order of all candidates, anchored rank ordering would need to be based on a large ‘pack’ approach, with cross-linking of individual examiners’ allocations. So each script or response would be seen by multiple experts, much like

the distributed judgements in online paired comparisons. The rank for any script is therefore based upon a consensus of many judgements made by many judges and is highly defensible. Both approaches also offer the opportunity to identify and potentially exclude individuals whose judgements are inconsistent with the overall consensus, based on their infit error measure. All these factors mean that the existing reviews of marking processes should not be required as there is no individually-decided mark to appeal against.

In a full implementation of item comparative judgements in place of marking there are also other practical considerations around how to combine the rank orders of all the items on a paper to produce the final paper rank order. This could be a relatively simple matter of applying weights to the item ranks that correspond to the current allocation of marks to items. Combining papers into a whole assessment could be done in the same way. For anchored rank ordering or for paired comparative judgement, it may be possible to assign marks to the ordered responses by some form of expert review of a limited selection of responses, and then applying a fixed distribution around these reviewed responses. Alternatively, for paired comparative judgement the response quality parameters from the statistical model could be converted into something akin to marks, by scaling the parameters into a range, for example from 0 to 100, and then combining them. This would then retain the relative spacing between the quality of responses. However, because the item quality parameters are relative not absolute values, they are not easy to interpret, and even with scaling there would be no direct comparison of marks from year to year. So again, some of the transparency and ease of understanding of marks may be lost through ranking approaches.

Although the 3 methods generated rank orders with high correlations, we have not investigated how the rank orders differ at the level of individual responses. Given that the mark scheme and the construct map differ in what they capture and the way they weight different qualities, it might be expected that some responses would be ranked differently. A qualitative review of the responses whose ranks differ most across the methods may be required to highlight what the effect of the construct map is on the rank order.

This research was conducted on quite a specific type of question, historical source analysis. Future work could further investigate how applicable the rank order approach is to other question types across other subjects. It is certainly hard to imagine how individual low-tariff questions could easily be ranked. For these types of papers it would be more appropriate to judge whole scripts, as in Raikes, Scorey and Shiell (2008). Research would be needed to evaluate which papers are better judged or ranked as a whole, or where combining ranks for separate questions is more appropriate.

References

Benton, T. and Gallacher, T. (2018). Is comparative judgement just a quick form of multiple marking? *Research Matters: A Cambridge Assessment publication*, *26*, 22-28. http://www.cambridgeassessment.org.uk/Images/514231-research-matters-26-autumn-2018.pdf

Bisson, M.J., Gilmore, C., Inglis, M., and Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Maths Education*, *2*, 141–164. https://doi.org/10.1007/s40753-016-0024-3

Black, B. and Bramley, T. (2008) Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, *23*(3), 357-373. https://doi.org/10.1080/02671520701755440

Bradley R.A. and Terry M.E. (1952). Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons. *Biometrika*, *39*, 324–45. http://www.jstor.org/stable/2334029

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms (Eds) *Techniques for monitoring the comparability of examination standards.* London: QCA.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgement. *Journal of Applied Measurement*, *6*(2), 202-223.

Bramley, T. and Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, *online publication*, https://doi.org/10.1080/0969594X.2017.1418734

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

Fearnley, A. (2000). *A comparability study in GCSE mathematics. A study based on the summer 1998 examinations. Assessment and Qualifications Alliance (Northern Examinations and Assessment Board)*. Manchester: Joint Forum for the GCSE and GCE.

Furlong, A. and Glanville, M. (2015). *Trialling adaptive comparative judgement with long essay type responses*. Paper presented at the 16th annual conference of the Association for Educational Assessment – Europe, 5-7 November, Glasgow, UK

Gove, M. (2013). Letter to Ofqual, 7 February 2013. https://www.gov.uk/government/publications/letter-from-michael-gove-regarding-key-stage-4-reforms

Heldsinger, S.A., and Humphry, S.M. (2010). Using the method of paired comparisons to obtain reliable teacher-assessments. *The Australian Educational Researcher*, *37*(2), 1-19. https://doi.org/10.1007/BF03216919

Jones, I. and Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1774-1787. https://doi.org/10.1080/03075079.2013.821974

Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, *22*, 135-155. https://doi.org/10.1007/s10798-011-9190-4

Laming, D. (2004). *Human judgment: the eye of the beholder*. London, UK: Thomson.

Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA. https://cerp.aqa.org.uk/research-library/review-literature-marking-reliability

Ofqual (2018a). *Marking Consistency Studies: Summer 2016 and 2017 units.* (Report No. Ofqual/18/6449/3). Coventry, UK: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/759208/Marking_consistencies_-_FINAL64493.pdf

Ofqual (2018b). *Marking consistency metrics.* (Report No. Ofqual/18/6449/2). Coventry, UK: Ofqual. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/759207/Marking_consistency_metrics_-_an_update_-_FINAL64492.pdf

Pollitt, A. (2004). *Let's stop marking exams*. Paper presented at the annual conference of the International Association of Educational Assessment, June 13–18, Philadelphia, USA. http://www.cambridgeassessment.org.uk/Images/109719-let-s-stop-marking-exams.pdf

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy and Practice*, *19*(3), 281-300. https://doi.org/10.1080/0969594X.2012.665354

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Paedagogiske Institute.

Raikes, N., Scorey, S. and Shiell, H. (2008). *Grading examinations using expert judgements from a diverse pool of judges*. Paper presented at the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK. http://www.cambridgeassessment.org.uk/Images/109766-grading-examinations-using-expert-judgements-from-a-diverse-pool-of-judges.pdf

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286. http://dx.doi.org/10.1037/h0070288

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*(6), 428-445. https://doi.org/10.1177/0146621617748321

Whitehouse, C. and Pollitt, A. (2012). *Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment*. Manchester: AQA Centre for Education Research and Policy. https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_CW_20062012_2.pdf

Wilson, M. (2005) *Constructing Measures: An Item Response Modeling Approach*. Mahwah, N.J.: Lawrence Erlbaum Associates

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

January 2020 **Ofqual/18/6449/4**