International Journal of Assessment Tools in Education
2019, Vol. 6, No. 2, 218–234
https://dx.doi.org/10.21449/ijate.564824
Published at http://www.ijate.net http://dergipark.org.tr/ijate *Research Article*

Investigating a new method for standardising essay marking using levels-based mark schemes

Jackie Greatorex [1*], Tom Sutch [1], Magda Werno [1], Jess Bowyer [2], Karen Dunn [3]

[1] Cambridge Assessment, Triangle Building, Shaftesbury Road, Cambridge, UK, CB2 8EA
[2] University of Exeter, St Luke's Campus, Heavitree Road, Exeter, UK, EX1 2LU
[3] British Council, 10 Spring Gardens, London SW1A 2BN, UK

ARTICLE HISTORY

Received: 18 Janauary 2019
Revised: 11 April 2019
Accepted: 23 April 2019

KEYWORDS

Comparative judgement,
Marking,
Standardisation,
Reliability,
Essay

Abstract: Standardisation is a procedure used by Awarding Organisations to maximise marking reliability, by teaching examiners to consistently judge scripts using a mark scheme. However, research shows that people are better at comparing two objects than judging each object individually. Consequently, Oxford, Cambridge and RSA (OCR, a UK awarding organisation) proposed investigating a new procedure, involving ranking essays, where essay quality is judged in comparison to other essays. This study investigated the marking reliability yielded by traditional standardisation and ranking standardisation. The study entailed a marking experiment followed by examiners completing a questionnaire. In the control condition live procedures were emulated as authentically as possible within the confines of a study. The experimental condition involved ranking the quality of essays from the best to the worst and then assigning marks. After each standardisation procedure the examiners marked 50 essays from an AS History unit. All participants experienced both procedures, and marking reliability was measured. Additionally, the participants' questionnaire responses were analysed to gain an insight into examiners' experience. It is concluded that the Ranking Procedure is unsuitable for use in public examinations in its current form. The Traditional Procedure produced statistically significantly more reliable marking, whilst the Ranking Procedure involved a complex decision-making process. However, the Ranking Procedure produced slightly more reliable marking at the extremities of the mark range, where previous research has shown that marking tends to be less reliable.

1. INTRODUCTION

General Certificate of Secondary Education (GCSE), Advanced Subsidiary (AS) and Advanced Level (A Level) are school examinations taken in the UK. Given the high-stakes nature of these examinations, it is essential that marking reliability is high and that standardisation (examiner training to accomplish uniform use of the mark scheme) is effectual, so that marks and grades are dependable. Generally, marking reliability is greater for short

CONTACT: Jackie Greatorex ✉ greatorex.j@cambridgeassessment.org.uk Research Division, Cambridge Assessment, Triangle Building, Shaftesbury Road, Cambridge, UK, CB2 8EA

ISSN-e: 2148-7456 /

answer questions than for questions requiring a long response which are marked with levels-based mark schemes[†]. Consequently, effective procedures for standardising examiners' essay marking are crucial.

It may be feasible to improve the current approach to standardisation and maximise essay marking reliability. The purpose of standardisation is to ensure that all examiners apply the mark scheme fairly and consistently, and procedures vary between Awarding Organisations, subjects and units. Traditionally, standardisation consists of practice marking, a meeting, where examiners are trained to apply the mark scheme, typically by marking a number of scripts as a group. Examiners then individually mark a sample of scripts at home, which are checked by their Team Leader, or the Principal Examiner (PE: the lead marker in charge of the examination or qualification) in smaller subjects. The Team Leader or PE may require further samples of marking to be checked.

In addition to standardisation there are several procedures for maximising marking quality including scaling (correcting consistently lenient or severe marking), and marker monitoring (observing marking post standardisation). The focus of the research is standardisation and it is beyond the scope of the research to account for these additional procedures.

Laming (2004) concludes from extensive research that people are better at comparing two objects than making absolute judgements about an object. Subsequently, a new approach to standardisation, forthwith called the Ranking Procedure, was proposed. This procedure focuses on comparing essays with one another and ranking them from the best to the worst.

The present research had two aims:

- to investigate whether the Ranking Procedure and Traditional procedure resulted in equivalent levels of marking reliability or whether the reliability from one was demonstrably better;
- to evaluate whether examiners considered the Ranking Procedure to be useful, how they conducted their marking, and whether the procedure was efficient.

1.1. Literature Review

Extended response questions are widely regarded as the most difficult questions to mark and are associated with the lowest levels of marking reliability (Black, Suto, & Bramley, 2011; Suto, Nádas, & Bell, 2011b). Consequently, there has been much research investigating why extended response questions have lower reliability, and how this can be improved.

One suggestion is that marking extended response questions entails a high cognitive load for the examiner, which could in turn lead to reduced marking reliability. Suto and Greatorex (2008) found that more complex cognitive strategies, such as evaluating and scrutinising, were used significantly more in the marking of GCSE Business Studies, which uses a levels-based mark scheme, than in GCSE Mathematics, which uses a more objective points-based mark scheme. Senior examiners in the same study suggested that it might be useful to train examiners in the use of these cognitive strategies. This may particularly benefit new examiners, as research indicates that examiners with lower subject expertise, marking and teaching experience mark extended response questions less accurately than others (Suto, Nádas, & Bell, 2011a).

[†] Levels-based mark schemes (levels-of-response mark schemes) are generally used for marking extended written responses. Such mark schemes often divide the available marks into smaller mark bands, each mark band is associated with a level and a description of the type of answer that will obtain a mark from within a given mark band. The examiner classifies a candidate's response into a level and then decides which mark from the associated mark band is most appropriate. For more detailed descriptions see Pinot de Moira (2013) and Greatorex and Bell (2008a).

Attempts to increase marking reliability in extended response questions have tended to focus on two key aspects of the marking process: standardisation (or examiner training), and mark schemes. Both are particularly pertinent to the current study.

1.2. Improving Standardisation

The greatest recent change to standardisation procedures (also called examiner or rater training in the literature) has been the transition from face-to-face to online standardisation. Some research indicates that online standardisation may be very slightly more effective in increasing marking accuracy, although in the context of English as a Second Language (ESL) assessment (Knoch, Read, & von Randow, 2007; Wolfe, Matthews, & Vickers, 2010; Wolfe & McVay, 2010). Research into the use of online standardisation in UK high-stakes examinations indicates that online standardisation is equally as effective as face-to-face standardisation (Billington & Davenport, 2011; Chamberlain & Taylor, 2010).

There is evidence to suggest that face-to-face or online meetings are not particularly effective for increasing marking accuracy on their own (Greatorex & Bell, 2008b; Raikes, Fidler, & Gill, 2009), and should be combined with additional feedback (Greatorex & Bell, 2008b; Johnson & Black, 2012). The type of feedback received does not affect marking reliability, whether it is iterative or immediate, personalised or prewritten, or targeted at improving accuracy or internal consistency (Greatorex & Bell, 2008b; Sykes et al., 2009). However, Johnson and Black (2012) found that examiners find feedback most helpful when it is immediate, refers to the mark scheme and focuses on specific problems with scripts.

Standardisation is of greatest benefit for new or less experienced examiners, whilst having little or no effect on the marking accuracy of experienced examiners (Meadows & Billington, 2007; Meadows & Billington, 2010; Raikes et al., 2009; Suto, Greatorex, & Nádas, 2009). Additional background effects such as subject knowledge and expertise also affect marking reliability (Suto & Nádas, 2008). Despite this, with adequate training, some examiners with no or little teaching and marking experience can become as reliable as the most experienced examiners, although these individuals may be difficult to identify before the standardisation process begins (Meadows & Billington, 2010; Suto & Nádas, 2008). It is noteworthy that research by Meadows and Billington (2010) and Suto and Nádas (2008) related to questions requiring short or medium length responses rather than essays, and therefore the findings may not generalise to examinations marked with levels-based mark schemes.

1.3. Mark Schemes

Alternative studies have investigated whether changes to levels-based mark schemes could improve marking reliability. A key factor is that levels-based mark schemes are often lengthy and contain a lot of information. Whilst more constrained mark schemes are associated with higher levels of reliability (Bramley, 2009; Pinot de Moira, 2013; Suto & Nádas, 2009), they do so by restricting the number of creditable responses and thus would compromise the validity of extended-response assessment (Ahmed & Pollitt, 2011; O'Donovan, 2005; Pinot de Moira, 2011a; Pinot de Moira, 2011b).

An alternative is to use a holistic, rather than analytic, mark scheme. Holistic mark schemes are where one overall mark is given to a response. The mark scheme may specify different elements of performance, but the examiner attaches their own weighting to each feature. In analytic levels-based mark schemes, the examiner awards separate marks for individual elements of a response. There is no clear consensus as to which is more valid and reliable.

The evidence suggests that inter-rater reliability is higher in holistic scoring (Çetin, 2011; Harsch & Martin, 2013; Lai, Wolfe, & Vickers, 2012). However, analytic scoring is particularly helpful in diagnostic English as a Second Language (ESL) assessment as features

of a student's writing can be assessed individually and that information then be fed back to the candidate to guide future learning (Barkaoui, 2011; Knoch, 2007; Lai et al., 2012; Michieka, 2010). Holistic mark schemes, on the other hand, can obscure differences in individual traits of a student's response, as well as how examiners weigh and apply different assessment criteria (Harsch & Martin, 2013).

A style of holistic marking that has particular relevance for the current study is Comparative Judgement (CJ). This method entails deciding which is the better of two scripts, thus making holistic but also *relative* judgements about script quality. Examiners make a series of these judgements, until each script has been judged a number of times. A rank order of all scripts is then statistically compiled, usually by fitting a Bradley-Terry model to the paired comparison data.

If the pairs are presented online and the data can be analysed in 'real time' it is possible to make the presentation of pairs 'adaptive'. This means that as more judgements are made, examiners are given scripts that appear to be closer together in quality, in order to make more nuanced distinctions between scripts and reduce the overall number of comparisons that need to be made. This process is known as 'adaptive comparative judgement' (ACJ), Pollitt (2012a) and Pollitt (2012b).

Whilst CJ is most often used for comparability studies, it is argued that it is more valid and reliable than traditional marking, as examiners are simply making overall judgements about script quality (Kimbell, 2007; Kimbell, Wheeler, Miller, & Pollitt, 2007; Pollitt, 2009, 2012a, 2012b; Pollitt, Elliott, & Ahmed, 2004). A project, which used ACJ to assess Design and Technology portfolios, found a reliability coefficient of 0.93 (Kimbell, 2007), whilst Whitehouse and Pollitt (2012) found a reliability coefficient of 0.97 when using ACJ in AS level Geography papers. However, adaptivity can inflate the reliability coefficient, so the high reliability found in these studies is disputed (Bramley, 2015; Bramley & Vitello, 2018). Moreover, there is empirical evidence that the strength of CJ lies in multiple judgements and a strong statistical model, rather than comparing one script directly with another (Benton & Gallagher, 2018). Also the process is very time-consuming (Whitehouse & Pollitt, 2012). Consequently, there are serious doubts as to whether it is practicable in large-scale, high-stakes assessment.

1.4. Research Questions

The experiment tested the following hypotheses:

H0: The Traditional and Ranking Standardisation result in equivalent levels of marking reliability.

H1: The Traditional or Ranking Standardisation result in more reliable marking.

Examiners' perspectives were collected to answer the following questions:

- How did the examiners undertake the Ranking Procedure?
- Was the Ranking Procedure (in)efficient and (un)suitable for upsacling or digitising?

2. METHOD

The project utilised candidates' scripts from an OCR AS level History examination. This examination was chosen for several reasons: firstly, the entry was large enough to select a wide range of scripts. Secondly, the questions required essay responses and were marked using a levels-based mark scheme.

The Principal Examiner from live examining was used as the Principal Examiner for this study. Ten Assistant Examiners (examiners) participated in the study. They had not marked this paper

in live examining, but they had either been eligible to mark it or had marked a similar examination (e.g. another A level History paper).

2.1. Design

The experiment had two conditions.

- The control condition was a simulation of a traditional standardisation process. This is used within some current awarding organisations. Each examiner marked a Provisional Sample. They attended a standardisation meeting where their marking was standardised using the Traditional Mark Scheme. After the meeting, the examiners marked the Standardisation Essays and received feedback on their marking from the PE. The PE decided whether each examiner could proceed to the next stage of the experiment, re-mark the Standardisation Sample or mark a further Standardisation Sample. Finally, the examiners marked the Allocation.
- The experimental condition, called Ranking Standardisation, broadly followed the same process as the control condition. Each examiner ranked a Provisional Sample. They attended a standardisation meeting at which they learnt how to rank responses (with no marks) and then learnt how to mark the ranked essays. After the meeting each examiner rank ordered the Standardisation Sample from the best to the worst response. They received feedback on their ranking from the PE. The PE decided whether each examiner could proceed to the next stage of the experiment or re-rank the Standardisation Sample. Subsequently, each examiner marked the Standardisation Sample. The PE decided whether each examiner could proceed to the next stage of the experiment, re-mark the Standardisation Sample or rank and mark a further Standardisation Sample. Finally, the examiners marked the Allocation.

The design was within subjects. Marking reliability was measured at the end of the experiment. Counterbalancing was achieved by:

- Allocating examiners to groups based on which date they were available to attend a meeting[‡]
- Conducting conditions in the order determined by a 2x2 Latin Square:
  - group 1 control condition then experimental condition
  - group 2 experimental condition then control condition.

This was to guard against the order of conditions affecting the results.

2.2. Materials

2.2.1. *Scripts*

All the essays involved were responses to the two most popular questions for that examination paper (referred to here as questions A and B). Each had a maximum of 50 marks available. Scripts were anonymised for use by examiners.

The experiment involved four samples of essays: The Provisional, Meeting and Standardisation samples, as well as the examiners' final marking allocation. This was intended to mirror the real standardisation process. Participants marked the Provisional Sample before the standardisation meeting; the Meeting Sample was used in the standardisation meeting to teach participants about applying the mark scheme correctly; and the Standardisation Sample was marked after the meeting to ensure participants were applying the mark scheme correctly and to gain a measure of inter-rater reliability. After the standardisation process was completed, participants were asked to mark an allocation of 50 essays. The essays covered a range of quality of performance.

[‡] It was assumed that availability would be as random as any other way of allocating examiners to groups.

2.2.2. *Mark Schemes*

Traditional Mark Scheme

The Traditional Mark Scheme was the live mark scheme for question A or B. The live mark scheme included level descriptors and a description of content for each question.

Ranking Mark Scheme

In the Ranking Mark Scheme the level descriptors from the live mark scheme were replaced with a brief description of the characteristics of quality of performance. The Ranking Mark Scheme was written by the PE and reviewed by OCR. After the standardisation meeting, an indication of the marks given to each Meeting Essay was added. For an example of the Ranking Mark Scheme see Figure 1.

2.2.3. *Examiner Questionnaire*

A questionnaire was developed that included open and closed questions. The questionnaire focused on the usefulness of aspects of standardisation, how the Ranking Procedure might be upscaled or conducted on-screen, and related merits and limitations.

2.3. Controls

Control mechanisms were in place. First, examiners were allocated to groups based on availability. Secondly, group 1 completed each stage of the control condition using the Traditional Mark Scheme on question A before completing the parallel stage of the experimental condition using the Ranking Mark Scheme for question B. Group 2 completed each stage of the experimental condition with the Ranking Mark Scheme on question A before completing the parallel stage of the control condition with the Traditional Mark Scheme for question B. Thirdly, all examiners marked the same essays at each stage of the experiment. Fourthly, none of the examiners marked the question paper in live marking, as such participants would have violated the crossover design by experiencing the control condition before the experimental condition.

These controls and the within subjects design enabled a direct comparison between the reliability of marking generated by the two experimental conditions.

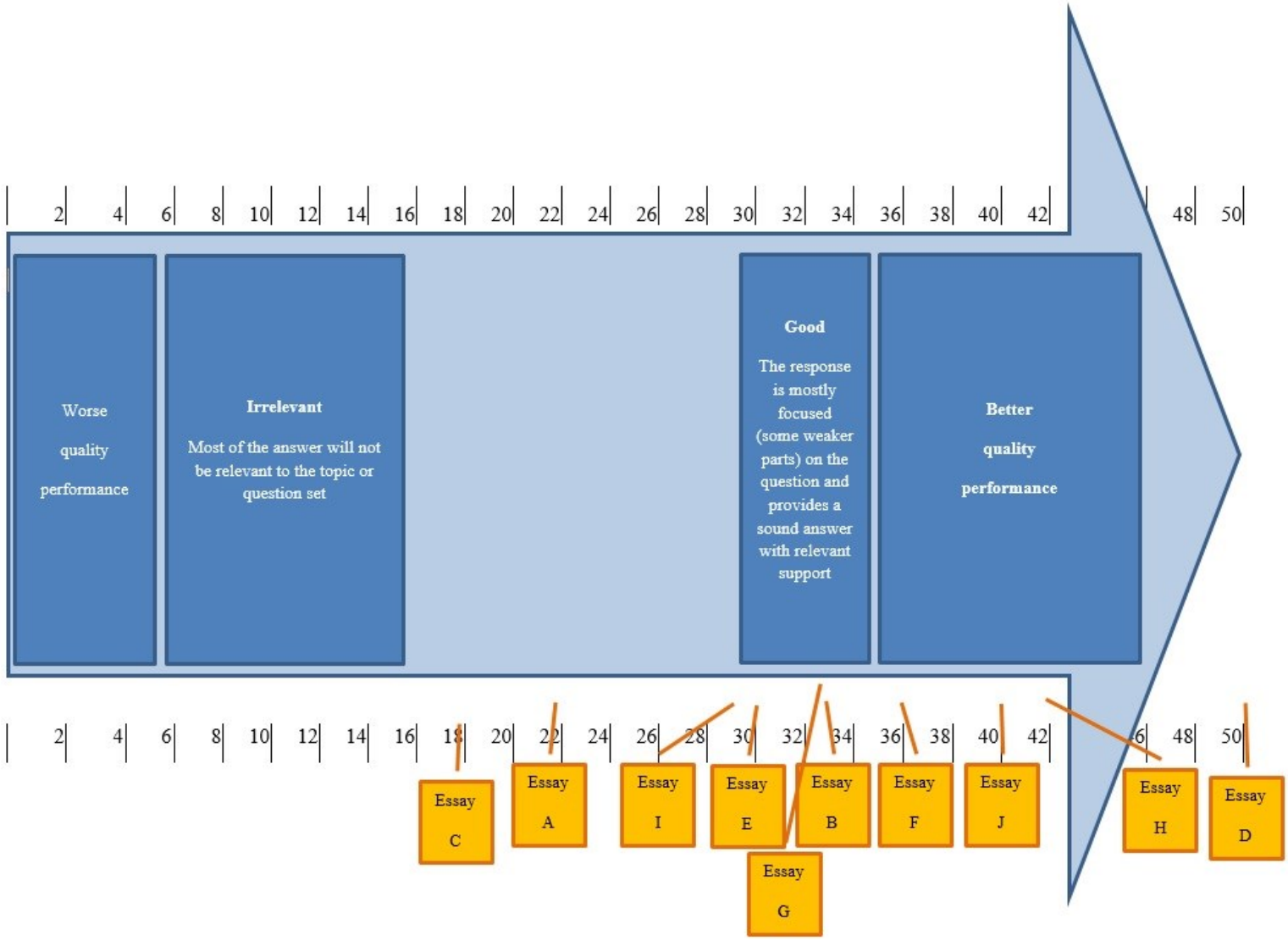

Figure 1. Ranking Mark Scheme Question A

2.5. Analysis

2.5.1. *Quantitative*

The within-subjects aspect of the design was statistically efficient because each examiner served as their own control, so examiner-level variation was isolated.

The experiment was more complex than a standard crossover design, because the measurement taken for each examiner for each condition also had a structure; an average of 50 essays. This was exploited in the analysis by modelling for candidate level effects.

The difference from the definitive mark§ for each marked essay was computed. The appropriate definitive mark was used depending on whether the essay was marked with the Traditional or Ranking Mark Scheme.

Different mark schemes were used for the two procedures, therefore the definitive marks for a given response could vary depending on procedure. In turn, this could mean that the effective mark range may vary and thus differences between examiners could be affected. To take an extreme example, if a mark scheme strongly fostered marking towards the centre of the mark range, the examiners would comply and the differences between examiners would be small, giving a false impression of high reliability. In order to address this, standardised differences were computed, by dividing the difference between the examiner and PE marks (definitive marks) by the standard deviation of the definitive mark for the given question under the appropriate procedure.

Analyses of variance were calculated using a variety of dependent variables. A standard analysis for a 2×2 crossover design, as described by Senn (2002) for example, would be possible at the examiner level (using the mean difference under each procedure), but this would not exploit the multiple measurements for each examiner and the variation at a candidate and response level. As a result, a more complex model was applied:

$$Y_{ijk} = \mu + m_{(i)k} + q_j + \tau_{d[i,j]} + c_l + cr_{jl} + e_{ijkl}$$

where the terms are as follows, notation based on Jones and Kenward (1989):

Y_{ijk} : Random variable representing marking difference (with observed values y_{ijk}) – either actual difference, or absolute difference as appropriate

μ : General mean

$m_{(i)k}$: The effect of examiner k in group i

q_j : The effect of question (and also period) j

$\tau_{d[i,j]}$: The direct effect of the treatment (procedure) used in period j for group i

c_l : The effect of candidate l

cr_{jl} : The effect of response by candidate l to question j

e_{ijkl} : A random error for candidate l, examiner k, period/ question j and group i, assumed to be independently and identically normally distributed with mean 0 and variance σ^2.

Note that no carry-over effect (denoted by Jones and Kenward (1989) as λ, and capturing any effect of the method used for the first period on the results in the second period) was included in the model. We followed the advice of Senn (2002) in not testing for this and carrying out a two-stage analysis, as was once common**.

§ There are several legitimate ways to calculate the definitive mark. For the purposes of this study, the Principal Examiner's marks were used as the definitive marks.

** A two-stage analysis first tests for the presence of a carry-over effect, then if such an effect is found, only data from the first period are used to test for the treatment (in our case, procedure) effect. As Senn (2002) explains, this

In our experiment it was not possible to separate any effect of period (that is, whether examiners' reliability changed between the first and second sets of responses marked) from the question marked (any effect on reliability due to the essay question, or the History topic) because all examiners in both groups were standardised on and marked question A first, followed by question B.

2.5.2 *Qualitative*

A thematic analysis of the qualitative responses to the questionnaire was guided by advice from Braun and Clarke (2006). There were four themes in the data, however, our focus is:

- Decision Process for the Ranking Procedure (including an Initial Sorting stage)
- Ranking is Time-consuming

Regarding the Decision Process a diagram was drawn to represent the data. A second researcher checked the diagram against their reading of the data.

3. FINDINGS

3.1. Quantitative

In the six analysis of variance models most effects were strongly statistically significant, reflecting the size of the sample. For the purposes of brevity these figures are not included.

The focus of the research was the procedure effect, which was significant at the 5% level for all analyses except standardised absolute difference from the average examiner mark (Table 1). The unstandardised differences were more easily interpreted as they are articulated in terms of raw marks, while the standardised differences are the number of standard deviations (of the definitive mark distribution). Using the measure of actual difference, the examiners were more severe (by 0.9 of a mark), and further away from the definitive mark, under the Ranking Procedure. When absolute difference from the definitive mark is considered, there was greater marking error (by 0.5 of a mark on average) using the Ranking Procedure than the Traditional Procedure; the difference was somewhat smaller (0.3 marks) when the average examiner mark was used as the comparator.

The standardisation of the differences had a small influence on the direction of the results, but did reduce the significance of the procedure effect when considering the absolute difference. When focusing on the absolute difference with respect to the average examiner mark the difference became statistically insignificant (at the 5% level) when standardised differences were used.

In short, the results supported the hypothesis (H1: The Traditional or Ranking Standardisation result in more reliable marking) and the Traditional Standardisation procedure yielded greater marking reliability.

Figure 2 and Figure 3 show the mean error for each of the responses to question A and B (the results for each procedure originated from a different group of examiners). The x-axis shows the responses arranged by definitive mark, and the 10 Meeting Essays from the experimental condition are shown as vertical lines[††]. The three panels for each question show the same results with different y-axes:

- actual difference from definitive mark
- absolute difference from definitive mark
- absolute difference from average examiner mark[‡‡].

approach is flawed because the test based on the first period only is highly correlated with the pre-test for carry-over, and is thus heavily biased.

[††] Note that these vertical lines are not necessarily the definitive marks for the control condition, but they are retained to enable comparison between the two halves of each graph.

[‡‡] Average actual difference from average examiner mark is not shown, as it is zero for each response.

Table 1. Effect sizes for procedure, and estimates of mean

Response	Estimates of means under each procedure		Effect of procedure			
	Traditional	Ranking	Estimate	Standard Error	t value	Pr > |t|
Actual difference	-1.628	-2.518	-0.890	0.247	-3.60	0.0003
Absolute difference	3.780	4.326	0.546	0.178	3.07	0.0022
Actual difference (standardised)	-0.2449	-0.3622	-0.1173	0.0363	-3.23	0.0013
Absolute difference (standardised)	0.5688	0.6237	0.0550	0.0261	2.11	0.0353
Absolute difference (average examiner mark)	2.68	3.01	0.332	0.135	2.47	0.0138
Absolute difference (average examiner mark) (standardised)	0.4084	0.4179	0.00949	0.01949	0.49	0.6266

There were few discernable trends. The proximity between the marks for Meeting Essays and the definitive mark of the target essay had no clear effect on the marking reliability of question A or question B. For actual difference in question A, the negative gradient suggests a slight tendency for examiners to be harsher for higher marks (that is, they mark closer to the middle of the mark scale than the PE) in both conditions.

For the bottom panel (absolute difference from average examiner mark) there were a few indications that the Ranking Procedure yielded greater marking reliability at the extremes of the mark range. For question A, there was more consensus among examiners using the Ranking than the Traditional Procedure at the upper end of the mark range. For question B a similar effect was observed at the lower end of the mark range, although, the responses in the allocation had definitive marks lower than the lowest Meeting Essay, G.

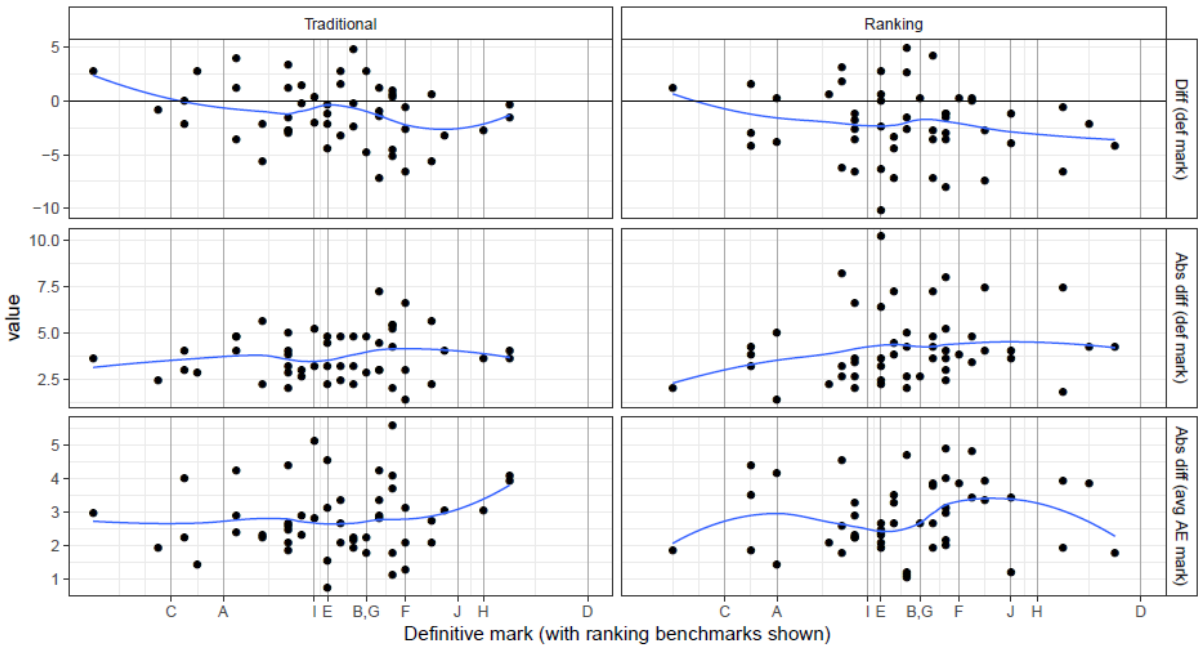

Figure 2. Mean error for each response along with definitive mark and Meeting Essays: question A

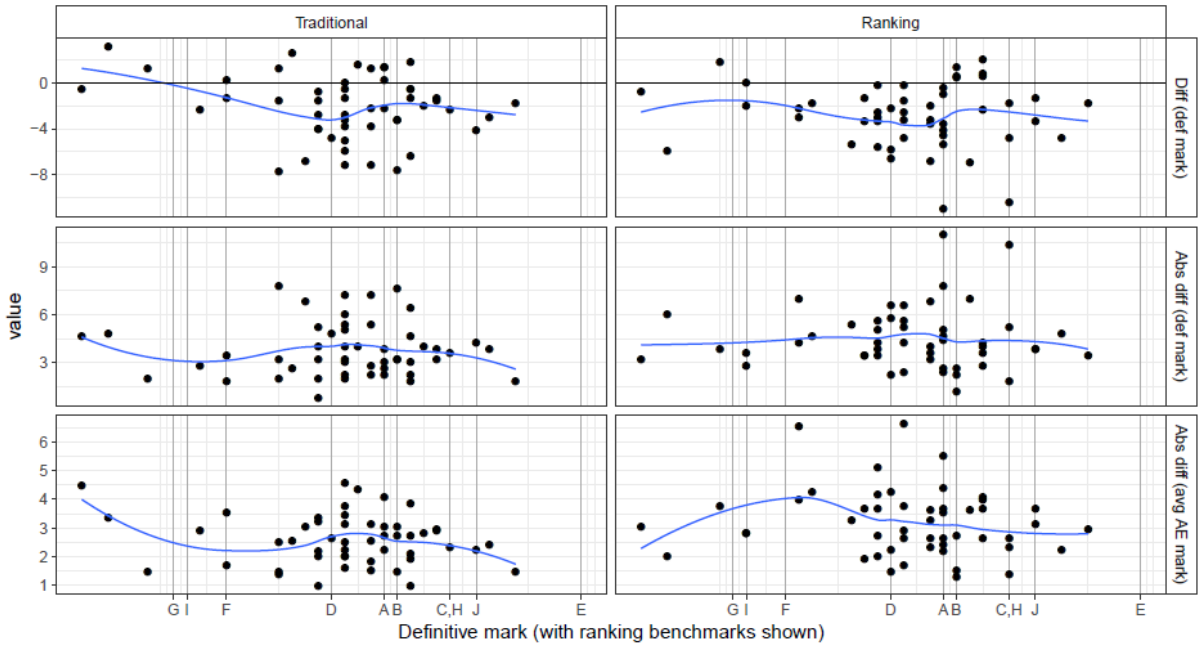

Figure 3. Mean error for each response along with definitive mark and Meeting Essays: question B

3.1. Qualitative

The Decision Process for the Ranking Procedure was a complex, multi stage process comprising several paired comparisons of essays (Figure 4). The core Decision Process was preceded in some instances by the Initial Sorting of the essays.

Nine examiners said that the Ranking Procedure was more time-consuming than the Traditional procedure. Reasons cited by examiners included:

- the difficulty of judging how essays compared to one another
- re-reading essays
- dealing with the accumulation of essays available to compare with the target essay
- lack of familiarity.

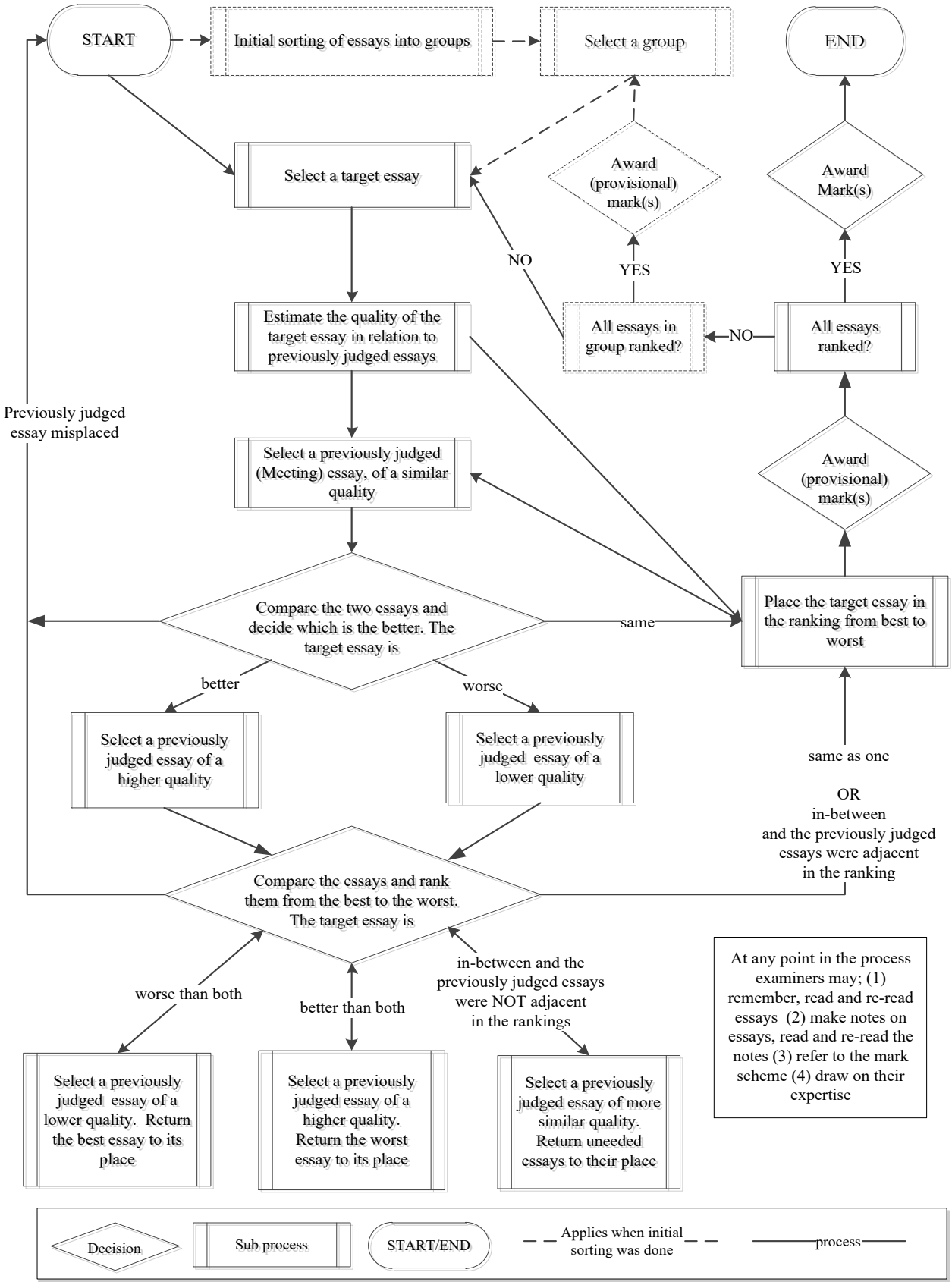

Figure 4. Decision Process for the Ranking Procedure

4. DISCUSSION

Prior studies illustrate that people's ability to compare two objects surpasses our ability to make absolute judgements about an object (Laming, 2004). Therefore, OCR proposed investigating a new procedure, the Ranking Procedure focusing on comparing the quality of essays and ranking them according to their quality before assigning marks. This research evaluated the marking reliability resulting from both Traditional and Ranking Standardisation, and examiners' experiences of the procedures. The research has limitations, which are outlined below. However, the research generated important findings.

The experiment was designed to simulate standardisation processes. It was beyond the scope of the research to incorporate the many checks and balances that are used to achieve reliable marking in addition to standardisation, for instance scaling. Therefore, the marks and statistics from the experiment were not directly comparable to marking reliability in live marking. However, the experimental data were suitable for testing the hypotheses.

The examiners were likely to be more familiar with the Traditional Procedure than the Ranking procedure, which may result in the former yielding more reliable marking. Arguably, examiners new to both procedures should have been recruited to ensure the experiment was a fair evaluation. However, if an Awarding Organisation were to switch from the Traditional Procedure to another procedure then many examiners would, in the short term, be more familiar with the Traditional Procedure. Consequently, the experiment is an authentic comparison of reliability delivered from the Traditional Procedure and a potential new procedure.

There was insufficient time between operational activities for examiners to complete one condition after another. Therefore, a departure from a crossover design was invoked. Group 1 undertook each stage of the experiment using the Traditional Mark Scheme on question A before completing the parallel stage of the experiment using the Ranking Mark Scheme for question B. Group 2 undertook each stage of the experiment using the Ranking Mark Scheme on question A before completing the parallel stage of the experiment using the Traditional Mark Scheme for question B. Each examiner marked the same essays as the other examiners at each stage of the experiment. The interweaving of stages may have had a confounding effect on each condition. Ideally there would be a 'wash-out' period between the two conditions, to allow any effects from the first half of the experiment to dissipate before commencing the second. The lack of a wash out period was a major practical constraint to the design. It was hoped that the effect of the conditions would be large enough to overpower any confounding variables, particularly as the interweaving of stages was common to the groups (with the two procedures reversed).

Several findings emphasised the limitations of the Ranking Procedure. First, the Traditional Procedure produced greater marking reliability than the Ranking Procedure. The Traditional Standardisation resulted in smaller mean differences for all measures. Second, the procedure effect was statistically significant at the 5% level for all measures of reliability, with the exception of standardised absolute difference from the average examiner mark. This concurred with previous research. When alternatives to Traditional Standardisation were investigated they did not consistently lead to better marking reliability than Traditional Standardisation (Greatorex & Bell, 2008b). Additionally, the mark scheme and feedback to examiners improve marking reliability, but examplar scripts do not improve marking reliability in terms of absolute difference between the PE's and examiners' marking (Baird, Greatorex, & Bell, 2004). Finally, the high reliability of using paired comparisons in marking (ACJ) has been disputed (Bramley, 2015; Bramley & Vitello, 2018). Based on the reliability measures Ranking Standardisation was not as effective as Traditional Standardisation.

There were additional limitations of the Ranking Procedure. First, it was time-consuming, due to the need to re-read essays. Both using ranking of more than two objects and involving

adaptivity have the potential to reduce the time taken. Whitehouse and Pollitt (2012) maintained that ACJ with pairs was not viable as a form of summative assessment for large scale public examinations in England in its current form as it was too time consuming. This study suggested that the Ranking procedure is also likely to be too time-consuming for these purposes. Secondly, for examiners the Decision Process for the Ranking Procedure is complex, indeed more complex than the decision process in CJ with pairs. Together the thematic analysis and the literature suggested that in its current form the Ranking Procedure was too complex and too time-consuming to be used for summative assessments for large scale public examinations in England.

However, the Ranking Procedure had merits which suggest it is worth further consideration. The Ranking Procedure overcame one drawback of levels-based mark schemes: that marking can be less reliable at the extremities of the mark range. Ideally marking is reliable throughout the mark range. However, prior research showed that marking reliability was better towards the centre of the mark range and not as good towards the top and bottom of the mark scale (Pinot de Moira, 2013) and a remedy is sought. Our qualitative data included comments from examiners that the Ranking Procedure gave greater discrimination and a lower prospect of inaccurate marking. The qualitative findings aligned with the quantitative evidence that at the extremities of the mark scale the Ranking Procedure performed better than the Traditional Procedure. In question A, there was higher reliability among examiners using the Ranking Procedure than the Traditional Procedure at the upper end of the mark range, and in question B a similar effect was noted at the lower end of the mark range. There is no clear cause for this result. It is possible that the holistic marking of the Ranking Procedure outperformed the analytical marking of the Traditional procedure in supporting reliability at the extremes of the mark range. Further consideration may be given to how (features of) the Ranking Procedure can be applied to mark essays at the extremes of the mark range.

5. CONCLUSION

This study investigated the merits and limitations of the Ranking Procedure compared with the Traditional Procedure. The limitations of the Ranking Procedure were that it yielded less reliable marking than the Traditional Procedure, was time consuming and entailed an inefficient Decision Process. However, the Ranking Procedure had several merits. Examiners noted that the Ranking Procedure gave greater discrimination and a lower prospect of inaccurate marking. The quantitative results indicate that the Ranking Procedure produced reliable marking at the extremities of the mark range, whereas traditional levels-based mark schemes tend to generate more reliable marking in the middle of the mark range and less reliable marking at the extremities. The findings suggest that the Ranking Procedure is unsuitable for implementation in public examinations in its current form. However, it may be advantageous to explore techniques for refining the Ranking Procedure so that its merits may be realised, for example in awarding or research studies.

ORCID

Jackie Greatorex https://orcid.org/0000-0002-2303-0638
Tom Sutch https://orcid.org/0000-0001-8157-277X
Karen Dunn https://orcid.org/0000-0002-7499-9895

Conflicts of Interest

The authors declared no conflict of interest.

Acknowledgements

We would like to thank colleagues in OCR and the Research Division at Cambridge Assessment for their help and advice with the study, particularly Beth Black who had the original ideas for

the Ranking Procedure. We would also like to thank the examiners who engaged with the research, especially the Principal Examiner who played a pivotal role in the project.

6. REFERENCES

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259-278. doi: http://dx.doi.org/10.1080/0969594X.2010.546775

Baird, J.-A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*(3), 331-348.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293.

Benton, T., & Gallagher, T. (2018). Is comparative judgement just a quick form of multiple marking. *Research Matters: A Cambridge Assessment Publication* (26), 22-28.

Billington, L., & Davenport, C. (2011). On line standardisation trial, Winter 2008: Evaluation of examiner performance and examiner satisfaction. Manchester: AQA Centre for Education Research Policy.

Black, B., Suto, W. M. I., & Bramley, T. (2011). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement. *Assessment in Education: Principles, Policy & Practice, 18*(3), 295-318.

Bramley, T. (2009). Mark scheme features associated with different levels of marker agreement. *Research Matters: A Cambridge Assessment Publication* (8), 16-23.

Bramley, T. (2015). Investigating the reliability of Adaptive Comparative Judgment *Cambridge Assessment Research Report*. Cambridge, UK: Cambridge Assessment.

Bramley, T., & Vitello, S. (2018). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 1-16. doi: 10.1080/0969594X.2017.1418734

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77-101.

Çetin, Y. (2011). Reliability of raters for writing assessment: analytic - holistic, analytic-analytic, holistic-holistic. *Mustafa Kemal University Journal of Social Sciences Institute, 8*(16), 471-486.

Chamberlain, S., & Taylor, R. (2010). Online or face to face? An experimental study of examiner training. *British Journal of Educational Technology, 42*(4), 665-675.

Greatorex, J., & Bell, J. F. (2008a). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education, 23*(3), 333-355.

Greatorex, J., & Bell, J. F. (2008b). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education, 23*(3), 333-355.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281-307.

Johnson, M., & Black, B. (2012). Feedback as scaffolding: senior examiner monitoring processes and their effects on examiner marking. *Research in Post-Compulsory Education, 17*(4), 391-407.

Jones, B., & Kenward, M. G. (1989). *Design and Analysis of Cross-Over Trials*. London: Chapman and Hall.

Kimbell, R. (2007). e-assessment in project e-scape. *Design and Technology Education: an International Journal, 12*(2), 66-76.

Kimbell, R., Wheeler, T., Miller, S., & Pollitt, A. (2007). E-scape portfolio assessment. Phase 2 report. London: Department for Education and Skills.

Knoch, U. (2007). ' Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12*(2), 108-128. doi: 10.1016/j.asw.2007.07.002

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26-43. doi: 10.1016/j.asw.2007.04.001

Lai, E. R., Wolfe, E. W., & Vickers, D. H. (2012). Halo Effects and Analytic Scoring: A Summary of Two Empirical Studies *Research Report*. New York: Pearson Research and Innovation Network.

Laming, D. (2004). *Human judgment: the eye of the beholder*. Hong Kong: Thomson Learning.

Meadows, M., & Billington, L. (2007). *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection*. Manchester: National Assessment Agency.

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Manchester: Centre for Education Research and Policy.

Michieka, M. (2010). Holistic or Analytic Scoring? Issues in Grading ESL Writing. *TNTESOL Journal*.

O'Donovan, N. (2005). There are no wrong answers: an investigation into the assessment of candidates' responses to essay-based examinations. *Oxford Review of Education, 31*, 395-422.

Pinot de Moira, A. (2011a). Effective discrimination in mark schemes. Manchester: AQA.

Pinot de Moira, A. (2011b). Levels-based mark schemes and marking bias. Manchester: AQA.

Pinot de Moira, A. (2013). Features of a levels-based mark scheme and their effect on marking reliability. Manchester: AQA.

Pollitt, A. (2009). *Abolishing marksism and rescuing validity*. Paper presented at the International Association for Educational Assessment, Brisbane, Australia. http://www.iaea.info/documents/paper_4d527d4e.pdf

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157-170.

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

Pollitt, A., Elliott, G., & Ahmed, A. (2004). *Let's stop marking exams*. Paper presented at the International Association for Educational Assessment, Philadelphia, USA.

Raikes, N., Fidler, J., & Gill, T. (2009). *Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology* Paper presented at the British Educational Research Association, University of Manchester, UK.

Senn, S. (2002). *Cross-Over Trials in Clinical Research*. Chichester: Wiley.

Suto, I., Nádas, R., & Bell, J. (2011a). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education, 26*(1), 21-51.

Suto, W. M. I., & Greatorex, J. (2008). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice, 15*(1), 73-89.

Suto, W. M. I., Greatorex, J., & Nádas, R. (2009). Thinking about making the right mark: Using cognitive strategy research to explore examiner training. *Research Matters: A Cambridge Assessment Publication*(8), 23-32.

Suto, W. M. I., & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education, 23*(4), 477-497. doi: 10.1080/02671520701755499

Suto, W. M. I., & Nádas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education, 24*(3), 335-377. doi: http://dx.doi.org/10.1080/02671520801945925

Suto, W. M. I., Nádas, R., & Bell, J. (2011b). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education, 26*(1), 21-51.

Sykes, E., Novakovic, N., Greatorex, J., Bell, J., Nádas, R., & Gill, T. (2009). How effective is fast and automated feedback to examiners in tackling the size of marking errors? *Research Matters: A Cambridge Assessment Publication* (8), 8-15.

Whitehouse, C., & 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.

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment, 10*(1). http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1601/1457

Wolfe, E. W., & McVay, A. (2010). *Rater effects as a function of rater training context*. New York: Pearson Research and Innovation Network.