

An experimental investigation of the effects of mark scheme features on marking reliability

Simon Child, Jess Munro and Tom Benton

This document contains colour figures

February 2015

ARD Research Division
Cambridge Assessment
1 Hills Road, Cambridge, CB1 2EU

Contents

List of Tables and Figures	3
Acknowledgements	4
Executive Summary	5
1. Introduction	8
1.1. The role of mark schemes.....	8
1.2. The reliability of levels-based mark schemes.....	9
1.3. Optimising the features of levels-based mark schemes	10
1.4. Aims of the present study.....	11
2. Methods	12
2.1. Participants and recruitment	12
2.1.1. Target unit	12
2.1.2. Principal Examiner	12
2.1.3. Assistant Examiners	13
2.2. Design	13
2.3. Materials	15
2.3.1. Mark schemes	15
2.3.1.1. The 'original' mark scheme	15
2.3.1.2. The 'experimental' mark scheme	15
2.3.2. Scripts.....	18
2.3.3. Examiner questionnaire	18
2.4. Procedure	18
2.4.1. Meeting with PE.....	20
2.4.2. AE recruitment and mark scheme development.....	20
2.4.3. Standardisation meetings	20
2.4.4. Marking of the scripts.....	22
3. Results	23
3.1. Descriptive results.....	23
3.2. Marker consistency.....	27
3.3. Consistency between individual markers and the Principal Examiner	31
3.4. Questionnaire analysis.....	34
3.5. Results summary	36
4. Discussion	37
4.1. Limitations	38
4.2. Implications.....	39

4.3. Conclusions	40
5. References	41
Appendix A: Examination paper for unit A680/02 (June 2014).....	44
Appendix B: The 'original' mark scheme.....	47
Appendix C: The 'experimental' mark scheme.....	64
Appendix D: Post-marking questionnaire.....	74
Appendix E: The distributions of median scores under each mark scheme	79
Appendix F: Methodology used to compare intra-class correlations between mark schemes	81
Appendix G: The mark-remark reliability of the Principal Examiner.....	84

List of Tables and Figures

Table 2.1: Participant details and marking history.....	14
Table 2.2: Mark scheme manipulations	16
Figure 2.1: Summary of the experimental procedure.....	19
Table 2.3: Agenda of the standardisation meetings	22
Table 3.1: Descriptive Statistics.....	24
Figure 3.1: Score distributions across examiners and scripts for each mark scheme	26
Figure 3.2: Cumulative score distribution for total score across Q2 and Q4 for each mark scheme	27
Table 3.2: Marking reliabilities (intra-class/intra-candidate correlations) under each mark scheme	28
Table 3.3: Mean average absolute deviations under each mark scheme.....	29
Figure 3.3: Average absolute deviations <i>from the median</i> for each candidate under each mark scheme	30
Table 3.4: Mean average absolute deviations from the Principal Examiner (PE) mark under each mark scheme.....	31
Table 3.5: Correlation of candidate ranking between Principal Examiner and other individual markers.....	32
Figure 3.4: Average absolute deviations <i>from the Principal Examiner's</i> mark for each candidate under each mark scheme	33
Table 3.5: AE ratings of the two mark schemes.....	34

Acknowledgements

Several colleagues from Cambridge Assessment's Research Division provided conceptual and technical support during various stages of the project. In particular, we would like to thank Tom Bramley, Sylvia Green and Jackie Greatorex for their helpful advice on the study design, and Tom Sutch for his expertise in collating and preparing the research materials. We would also like to thank colleagues from OCR for their guidance, especially Beth Black (now at Ofqual), Brian Wilding, Sophie Maloney, Keeley Nolan, and Michelle North. Finally, we are grateful for the time and efforts of the research participants, in particular the Principal Examiner, who played a pivotal role in the research.

Executive Summary

Introduction

The quality of marking for any one item of an examination paper is influenced by several interacting factors including the question design, the marking task, and the examiner. There have been recent attempts to better understand the marking process and what influences its effectiveness, including research on standardisation, marker feedback, the Enquiries About Results (EARs) procedure, and mark schemes.

A central point of reference for examiners at each stage of the marking process is the mark scheme. Well-designed mark schemes (in addition to examiner training) should enable examiners to make an accurate assessment of candidates' responses to an item, while simultaneously not be too cognitively demanding.

The difficulties in establishing high levels of reliability using levels-based mark schemes have led to a body research investigating how mark scheme features relate to reliability and, more broadly, to overall quality of marking. From previous research, it was possible to make several preliminary recommendations concerning which features of levels-based mark schemes might be manipulated to improve marker reliability. The present study aimed to analyse experimentally how these features influenced reliability, and quality of marking more generally; measured in terms of marking distribution, the degree of agreement between examiners and a Principal Examiner, and examiners' perceptions of mark scheme usability.

Methods

The Principal Examiner (PE) for the target unit (English Language GCSE unit A680/02 – Information and Ideas) was recruited to assist in the development of two mark schemes and to offer training to the Assistant Examiners. The two mark schemes were developed after discussions between the research team, the PE and OCR colleagues. The 'original' mark scheme contained the same content and features as the mark scheme used in the live session in June 2014. The 'experimental' mark scheme contained the same content as the 'original' mark scheme, but had a number of its features manipulated. This included changes to the positioning of guidance related to specific questions, the salience of key terms, and the page formatting. The changes introduced in the 'experimental' mark scheme were informed by previous research.

Twenty Assistant Examiners (AEs) were recruited to mark 150 scripts comprising two questions from the target unit examination paper (Questions 2 and 4) from June 2014. Half of the AEs were trained to use the 'original' mark scheme, and half were trained to use the 'experimental' mark scheme. The AEs each attended a standardisation meeting, where the PE provided information on how to apply the mark scheme correctly using a sample of 10 exemplar scripts.

Results

An analysis of reliability (intra-class correlations) revealed that for Question 2 marking reliability was not significantly improved by using the 'experimental' mark scheme. However, for Question 4 (particularly Q4_AO3iii) results indicated that the 'experimental' mark scheme

may, to an extent, have improved reliability in terms of the degree of agreement between markers, and between the AEs and the PE.

This observed improvement in reliability for Question 4 appears to be explained in part by changes in the distribution of scores for the 'experimental' mark scheme compared to the 'original' mark scheme. The results indicated that the 'experimental' mark scheme seemed to encourage AEs to use a greater range of marks, and that this increase resulted in a greater proportion of variance attributed to the true score rather than to error.

Furthermore, the change in the distribution of marks observed in the 'experimental' mark scheme implied that inconsistency in marking may have a smaller effect on grade outcomes when using this mark scheme. If a wider range of marks are being used under the 'experimental' mark scheme, the gaps between grade boundaries are likely to be wider. Consequently, a difference of a given size between markers would be less likely to affect a pupil's grade in the 'experimental' mark scheme.

The questionnaire results highlighted particular mark scheme manipulations that were perceived by examiners to focus their marking decisions, including the bolding of key terms, the proximity of level descriptors to each other, and the positioning of guidance that assisted level decision-making.

Discussion and implications

The introduction of several new features in the 'experimental' mark scheme improved reliability (as measured by ICCs) for one of the target questions, but not the other. It is therefore uncertain as to whether the 'experimental' mark scheme conclusively improved marking reliability.

However, the results suggest that for practical purposes, quality of marking may have improved under the 'experimental' mark scheme. Examiners that were using the 'experimental' mark scheme used a wider range of the marks, compared to the examiners using the 'original' mark scheme. Although it is not a straightforward task to know for certain how a marking distribution should look (Pinot de Moira, 2013), a fair assumption is that no marks in the mark scheme should be underutilised (Pinot de Moira, 2011).

This increased discrimination between scripts may have the advantage of increasing the distance between the grade boundaries in terms of marks. This is advantageous for levels-based mark schemes as it means that differences between examiners are less likely to influence the final grade for a candidate for a paper.

Finally, the examiners in the present study related specific features of the 'experimental' mark scheme to their perceptions of usability and cognitive processing, including the bolding of key terms, the close proximity of level descriptors to guidance, and the one page formatting of the 'experimental' mark scheme. In this context, these results indicated that simple changes to mark schemes can potentially improve their usability without detriment to their reliability. AEs articulated that mark scheme features present in the 'experimental' mark scheme facilitated their focus on the salient elements of the mark scheme, and in some cases reduced cognitive load.

In conclusion, the present study provides evidence to suggest that changes to some mark scheme features are worthy of future consideration, with respect to increasing mark scheme usability and consequently the overall quality of marking.

1. Introduction

The quality of marking for any one item of an examination paper is influenced by several interacting factors including the question design, the marking task, and the examiner. Broadly defined, the term 'quality of marking' refers to aspects related to marking reliability, accuracy or agreement, in addition to factors that influence the experience of examiners within the examination process (e.g. the usability of mark schemes). These influences can be broadly categorised into within-*task* and within-*marker* factors. Within-task factors that influence the quality of marking are determined early on in the development of an item and its accompanying mark scheme. Examination items are typically written in tandem with the development of its mark scheme, and proceed through a series of reviews and revisions (Cambridge Assessment, 2013). The features of the item influence the candidate's response (Massey & Raikes, 2006; Raikes & Massey, 2007) and subsequently the mark scheme, which is finalised only after students have sat the examination (Cambridge Assessment, 2013). Within-*marker* factors are those that contribute to marker expertise (Bramley, 2008), and include examiner marking experience (Brooks, 2012, Suto & Greatorex, 2008), teaching experience (Suto & Nadas, 2008) and knowledge of the examination subject (Suto, Nadas & Bell, 2011).

With the various influences on quality of marking outlined above, there have been recent attempts to better understand the marking process and what influences its effectiveness (Ofqual, 2014a,b), which has led to a number of literature reviews (e.g. AlphaPlus, 2014; Tisi, Whitehouse, Maughan & Burdett, 2013). There has also been research on specific elements of the marking process including standardisation (Chamberlain & Taylor, 2011) and marker feedback (Johnson, 2014a,b).

1.1. The role of mark schemes

A central point of reference for examiners at each stage of the marking process is the mark scheme (Ofqual, 2014b). Chairs of Examiners and Principal Examiners involved in the creation of examination papers face the important challenge of developing mark schemes that facilitate reliable and valid assessment (Meadows & Billington, 2005). Whilst mark schemes (in addition to examiner training) should enable examiners to make an accurate assessment of candidates' responses to an item, they should simultaneously not be too cognitively demanding (Black, Suto & Bramley, 2011).

There are three broad categories of mark scheme: objective (or constrained); points-based; and levels-based (Massey & Raikes, 2006; Tisi et al, 2013). Objective mark schemes are used when there is a short response (typically one or two words) where there is an unambiguously correct answer. Points-based mark schemes list "objectively identifiable words, statements or ideas" (Tisi et al., 2013, p.94). When using this type of mark scheme, the marker has to read the entire response to determine which sections are related to the mark scheme. A mark is awarded for each point candidates produce that matches a creditworthy response in the mark scheme. Finally, levels-based mark schemes are typically used when the question requires an extended written response. *Holistic* levels-based mark schemes require the marker to make an overall judgement of performance. Each level of performance may have several elements contained within the description, but the markers attach their own weighting to each feature (Nadas & Suto, 2011). *Analytic* levels-based marks schemes comprise descriptions for each of the aspects of interest at different levels.

In other words, this type of mark scheme weights different features of the response (Hamp-Lyons, 1991; Tisi et al., 2013).

As mentioned above, the development of mark schemes is intertwined with the development of the question item it supports, and is influenced by the anticipated (and actual) responses from candidates. Indeed, this approach is encouraged by the examinations regulator Ofqual (2011). Ahmed and Pollitt (2011) suggested that mark schemes should be developed with a full understanding of the potential range of responses candidates may produce. By the end of its development, the mark scheme will refer to the responses that were expected when the question was written *and* some responses observed in the early stages of marking. For example, items that require a one word response use objective mark schemes that may comprise several appropriate responses. In contrast, extended response items typically utilise levels-based mark schemes, which comprise descriptions of the standard of response required for entry to each level.

1.2. The reliability of levels-based mark schemes

The reliability of a mark scheme is closely related to its form. Bramley (2007) defined *reliability* as “the ratio of true-score variance to observed score variance” (p.26) and suggested it is best used when referring to a set of scores. He outlined several ways that reliability could be measured, including intra-class correlations which calculate the proportion of the variance in scores that are attributable to candidates as opposed to markers. *Marker agreement* is conceptually different (Bramley, 2007), and is based on the comparison between a mark provided by an examiner for an item, and a mark provided by a more senior examiner for the same item (assumed to be the ‘true’ mark).

Black et al. (2011) suggested that one of the primary influences on the cognitive demand of marking, and thus its reliability, was the type of mark scheme adopted (objective, points-based or levels-based). For example, Massey and Raikes (2006) analysed marking data for five examinations, and found that across all subjects one mark items had consistently higher levels of marker agreement compared to items that had multiple marks attributed to them. Further to this finding, Black (2010, cited in Tisi et al., 2013) separated *question* constraint (e.g. objective, short-answer, extended response) and *mark scheme* constraint (objective, points-based and levels-based). It was found that objective mark schemes had the highest level of marker agreement, with levels-based mark schemes having the lowest.

There are several reasons for why levels-based mark schemes are in general less reliable than objective or points-based mark schemes. First, the structure of the examination item and the mark scheme influence the cognitive processes underlying examiners’ marking behaviours (Suto & Greatorex, 2008; Suto & Nadas, 2008). For example, Suto and Nadas (2008) found that examination items that required examiners to scrutinise unexpected responses from candidates, or required examiners to evaluate a response using knowledge from several sources were more likely to result in lower reliability, compared to simpler items. They concluded that certain types of marking task are fundamentally more challenging for the marker in terms of cognitive load compared to others, perhaps due to the emphasis on interpretation within certain scripts (see also Newton, 1996). However, whilst increasing the amount of detail that is included in the mark schemes would potentially limit the number of responses markers encounter where they have to make an unguided

judgement, it could also increase demands on working memory and thus reduce reliability (Black et al, 2011).

Secondly, the necessity to use terms to distinguish between each level in levels-based mark schemes may be problematic, as they are open to interpretation from individual markers (Harsch & Martin, 2013; Johnson, 2008). Johnson (2008) found that when assessing portfolios, markers had different interpretations of what constituted a 'detailed' or 'basic' description and what was deemed to be 'good'. This difficulty is compounded by the requirement when using levels-based mark schemes to make a holistic judgement. Laming (2004) suggested that all human judgements are relative and "scarcely better than ordinal" (p.8). In the context of marking, this implies that markers would find it challenging to mark a response without guidance as to what, for example, a 'good' response looks like (allowing them to make a relative judgement on the response they are marking). Crisp (2010) suggested that markers are likely to compare the response they are marking to their own personal representation of a prototypical response. Even with sufficient contextual information, markers would also find it difficult to determine *to what degree* the response they are marking is better or worse (e.g. how many marks to add or remove relative to a 'good' response).

Thirdly, Fowles (2009) suggested that the number of marks within each level influenced the reliability of levels-based mark schemes. In her study, she found that markers differed in their inclination to use extreme marks, and that this effect was exaggerated the more marks there were within levels. In other words, the mark scheme adopted acts to change markers' reasoning about the quality of a response. Indeed, Pinot de Moira (2011a,b) suggested that *most* papers have examples of 'underutilised' marks within the mark scheme that could potentially influence discrimination between candidates.

1.3. Optimising the features of levels-based mark schemes

The research outlined above suggests there are significant challenges in establishing high levels of reliability using levels-based mark schemes. These issues, and a perceived lack of confidence in the examinations system more broadly (as evidenced by a recent large increase in EARs requests, Ofqual, 2014c), have led to several research studies that have attempted to establish how the reliability (and more broadly the quality of marking) when using levels-based mark schemes can be improved, without detriment to item validity. From these studies, it was possible to make several preliminary recommendations concerning which features of levels-based mark schemes may contribute to improved marker reliability. Analyses of data from live marking sessions have only found weak relations between mark scheme features and marker reliability (AlphaPlus, 2014; Pinot de Moira, 2013). Pinot de Moira (2013) examined item level data from 133 units (a total of more than 27,000 items) in an attempt to model which of nine mark scheme features predicted either marker agreement or absolute marker differences. It was found that while some features, such as the number of marks within a band and the level of detail in a band descriptor specific to the item, were related to improved reliability, these effects were not statistically significant. However, more than 85 per cent of the items included within this study were from just seven subjects (French, Philosophy, Spanish, Geography, Media Studies, General Studies and German) with more than a third of the items from modern languages. Subjects that have high rates of entry and use levels-based mark schemes, such as English and English Literature, were not explored within this research, and History only provided a tiny percentage (3.5 per cent) of

the items studied. Furthermore it is difficult for observational studies such as these to fully control for the effects of all of the factors that may influence marking reliability.

1.4. Aims of the present study

The research outlined above suggests a number of ways in which a levels-based mark scheme could be manipulated to be 'optimised' in terms of quality of marking, as defined by marking reliability, marker agreement or examiners' perceptions of mark scheme usability. However, so far there has not been any research that has analysed experimentally how these features influence quality of marking according to these measures.

The present study aimed to fill this gap in the literature. It compared the reliability of marking using two mark schemes that were matched in terms of content within the levels descriptors and the number of marks within each mark band¹ but differed in their presentation. Theoretically, experimentally controlling the target items, the examiner profiles, the scripts marked, and the training the examiners receive will isolate the effect that the mark scheme has on quality of marking. If two mark schemes with the same *content* (with different mark scheme features) score differently in terms of reliability or on other outcomes such as mark distribution, it would suggest that particular features (or a combination of features) contribute to overall quality of marking.

¹ See Pinot de Moira (2011b) for a review concerning how mark scheme band width affects marker bias.

2. Methods

In this section, we outline several related procedural stages including participant recruitment, the study design, the development of the materials, and the data collection phase. We first describe the recruitment of participants, after the initial selection of the target unit for the study. This is followed by an outline of the study design, which guided our approach to the recruitment of the Principal Examiner and Assistant Examiners. The materials subsection provides a description of the two mark schemes used in the study. Finally, the procedure for the study is outlined, including details of the standardisation meetings that were run to provide the Assistant Examiners with guidance about how to interpret the mark schemes.

2.1. Participants and recruitment

2.1.1. Target unit

Before proceeding with the recruitment of participants, it was first necessary to decide the target unit from which to develop the mark schemes. The target unit for the study had to meet four main criteria. First, the unit had to have one examination paper that included at least two extended response items that used levels-based mark schemes. Secondly, the unit had to have a Principal Examiner (PE) who was willing to be a ‘confederate’ in the study, and available to assist the research team in the development of the mark schemes. Thirdly, the unit had to have a large enough pool of Assistant Examiners (AEs) from a different (but related) unit for recruitment purposes. Fourthly, the unit had to have sufficient candidate entries so that enough scripts were available for marking. Following these criteria, one unit from OCR’s GCSE in English Language (A680/02 – Information and Ideas) was selected, and confirmed after discussion with colleagues in OCR. The higher tier examination was selected from this unit because OCR colleagues determined this unit to meet the above criteria more closely than the foundation tier.

The examination paper unit A680/02 (June, 2014) comprised five questions across two sections (A and B). Section A focused on reading skills, whilst Section B assessed writing skills. For section A, candidates were required to answer three compulsory questions (numbered 1 to 3) based on source materials provided in the exam. The range of marks for each of these questions was between 12 and 14 marks. For section B, candidates were required to select one question from a choice of two (numbered 4 and 5). Each question from section B was worth 40 marks in total (see Appendix A for the examination paper).

For the present study, Question 2 from section A (worth 14 marks) and Question 4 from Section B were selected. These questions were selected for the following reasons. First, the questions varied in terms of the number of marks available. Secondly, the PE reported that these questions encouraged a range of different candidate responses, thus representing a sufficient challenge for examiners marking the scripts. Thirdly, it was determined that the mark schemes for these questions could be similarly manipulated in several ways in the development of an ‘experimental’ mark scheme. Fourthly, the questions related to different skills and assessment objectives that were assessed in the examination paper.

2.1.2. Principal Examiner

The Principal Examiner was the current PE for unit A680/02, and had over 20 years of marking experience. After discussions with OCR colleagues about the target unit (see target

unit section above), the research team approached the PE outlining the aims of the study, their role in the design of the ‘experimental’ mark scheme and the training meetings, an approximation of the time commitment required, and details of payment. Once participation in the study had been agreed, the PE was sent a contract and consent form to sign.

2.1.3. Assistant Examiners

The Assistant Examiners (AEs) were recruited from a cohort of examiners who had marked A664/02 – Literary Heritage and Poetry (a GCSE English literature unit) in 2013. This unit was selected because it had a relatively high number of examiners, was the same qualification level as the target unit (GCSE higher tier) and was in the same subject area. Examiners who had experience of marking the target unit (either foundation or higher tier) were removed from consideration for participation in the study. This was to make sure that the recruited examiners had comparable levels of experience with the target unit in the study.

A subset² of the examiners eligible to participate in the study was sent an introductory email which outlined the broad aims of the research, their potential role in the study, and information regarding payment. The examiners that expressed an interest in the research were then sent a follow-up email with a contract and consent form attached. They were also asked for information related to the administration of the study, and about their marking history.

In total, 20 examiners were recruited to participate in the study. All had at least two years’ experience marking for OCR, had achieved a Grade 3 or above³ when marking unit A664/02, and were Assistant Examiners for this unit.

2.2. Design

The study utilised a between-participants (matched pairs) design. Each examiner was assigned to one of two conditions (the ‘original’ mark scheme condition or the ‘experimental’ mark scheme condition). The condition that each AE was assigned to determined which mark scheme they were given to use in their subsequent training and marking. Each AE was paired with a second examiner, who had the same marking grade for unit A664/02. The marking grade was deemed an appropriate measure to estimate markers’ recent success in marking according to the guidelines set by the PE and the mark scheme. Where possible the AEs were also matched in terms of their overall marking experience (in years), and their marking history with CIE (all participant pairings were matched on at least one of these additional criteria). The details of each participant pairing are provided in Table 2.1.

² In the first instance, 20 examiners were contacted. Once this set of examiners had responded, a smaller group of examiners were contacted to supplement the initially recruited examiners. This method was adopted to prevent oversubscription for the study and to control participant pairings (see design section).

³ The grade for any one examiner is determined by the examiner’s Team Leader, in discussion with the Principal Examiner. The grade is an overall judgement on the examiner’s marking accuracy and their efficiency with regards to the administrative process.

Table 2.1: Participant details and marking history

Participant pairing	Original mark scheme					Experimental mark scheme				
	Participant ID	Current marker role	Recent marking grade	Marking experience	Previously marked for CIE	Participant ID	Current marker role	Recent marking grade	Marking experience	Previously marked for CIE
1	1A	AE	1	7 years	N	1B	AE	1	31 years	N
2	2A	AE	1	2 years	Y	2B	AE	1	2 years	Y
3	3A	AE	2	3 years	N	3B	AE	2	2 years	N
4	4A	AE	2	11 years	N	4B	AE	2	8 years	N
5	5A	AE	3	5 years	N	5B	AE	3	2 years	N
6	6A	AE	2	3 years	N	6B	AE	2	2 years	N
7	7A	AE	2	3 years	N	7B	AE	2	4 years	N
8	8A	AE	3	9 years	N	8B	AE	3	2 years	N
9	9A	AE	1	3 years	Y	9B	AE	1	3 years	N
10	10A	AE	2	3 years	Y	10B	AE	2	21 years	Y

Highlighted – did not complete marking

2.3. Materials

2.3.1. Mark schemes

There were two mark schemes developed in the present study: the mark scheme used in the live session in June 2014 for the two target questions (the 'original' mark scheme); and a mark scheme which contained the same content as the 'original' mark scheme but had some of its features manipulated (the 'experimental' mark scheme).

2.3.1.1. *The 'original' mark scheme*

This was identical to the mark scheme used in the live session for the target unit. To prevent examiner confusion, the sections related to the questions not used in the study were removed. Additionally, as this was a paper-based study, instructions related to on-screen marking were removed. Overall, the 'original' mark scheme document comprised 16 pages and contained the following information (see Appendix B for full version of the mark scheme):

- Front cover
- Marking instructions - for marking on-screen and paper based marking. This included general information on how to determine a mark within a band, appropriate marking annotations, and how to record marks.
- Mark scheme specific to Question 2 (notes on task, band descriptors and annotation guidance)
- Mark scheme specific to Question 4 (assessment objectives, notes on task, and band descriptors)

2.3.1.2. *The 'experimental' mark scheme*

This mark scheme was developed as part of a collaboration between the research team, and the PE. The research team first developed a list of potential mark scheme manipulations that was guided by previous research and the existing features of the mark scheme document used in the live session. The research team then discussed the potential manipulations with the PE, and each manipulation was either confirmed or removed from the list. This discussion with the PE was essential because the PE had to use the both forms of the mark scheme during the study, and thus the research team needed to be certain that the PE was comfortable with the changes that had been made. Indeed, the PE provided several justifications for why some features of the original mark scheme were included, and some supplementary suggestions for changes that they felt would improve the mark scheme.

Changes to the *content* of the mark scheme (e.g. mark scheme bandwidth or level descriptor information) was discounted at an early stage, as the primary focus of the study was on how mark scheme *features* impacted marker reliability. Furthermore, it was determined that changes in mark scheme word content would make it impossible for the standardisation phase of the study to be sufficiently similar across the experimental conditions.

Once the potential manipulations were agreed with the PE, the researchers developed a draft version of the 'experimental' mark scheme, which they then sent to the PE for comments. The final list of manipulations and the justifications for their introduction to the 'experimental' mark scheme is reported in Table 2.2.

Table 2.2: Mark scheme manipulations

Possible mark scheme manipulation	Manipulation type	Change agreed with PE
Insertion of full question in MS instructions	Question specificity	Yes
Insertion of AO criteria for questions 2 and 4 more explicit	AO referencing	Yes
Bracketing of Assessment Objectives within level descriptors	AO referencing	No
Use of term 'guidance' formalised across mark schemes	Guidance	Yes
Guidance related to annotation of scripts provided on first page only	Guidance	Yes
One-page formatting of level descriptors	Formatting	Yes
<i>Notes on task</i> moved onto same table as level descriptors for Question 2	Guidance/formatting	Yes

Possible mark scheme manipulation	Manipulation type	Change agreed with PE
Bolding of key terms in each level descriptor	Formatting	Yes
Reverse ordering of bands	Formatting	No
Shading of lower bands	Level descriptors/question specificity	Yes
Introduction of a 0 mark criteria	Level descriptors	No
Horizontal presentation of criteria	Formatting	No

Overall, the document comprised 10 pages and contained the following information (see Appendix C for the full version of the 'experimental' mark scheme):

- Front cover
- Marking instructions - for marking on-screen and paper based marking. This included general information on how to determine a mark within a band, appropriate marking annotations, and how to record marks.
- Mark scheme specific to Question 2 (question, assessment objective, guidance and band descriptors)
- Mark scheme specific to Question 4 (question, assessment objectives, guidance, and band descriptors)

2.3.2. Scripts

The scripts were selected from the entire batch of entries for unit A680/02 in June 2014. As one of the target questions for the study was an optional question (Question 4) candidates who had selected the alternative option (Question 5) were removed from possible selection. From this subset of candidates' scripts, 160 scripts were selected at random. These scripts were then downloaded and checked to confirm that each candidate had answered each of the two target questions.

Once the 160 scripts had been confirmed, attempts at Questions 1 and 3 were removed for each script, and each script was anonymised.

2.3.3. Examiner questionnaire

The AEs were also asked to complete a brief questionnaire that aimed to gather their views on the ease of use, clarity, and layout of the mark scheme for both Questions 2 and 4. The questionnaire comprised three sections, and included Likert scale items, multiple choice questions and questions that required written responses (see Appendix D for the full version of the questionnaire). The first section asked the examiners for their views on the mark scheme for Question 2, while the second section asked examiners about Question 4. The final section asked examiners for their views on the entire mark scheme document, and offered examiners the opportunity to write down any further comments they had. The questionnaire was identical for each of the two mark scheme conditions.

2.4. Procedure

The procedure for the recruitment, mark scheme development, standardisation meetings, and data collection is summarised in Figure 1.

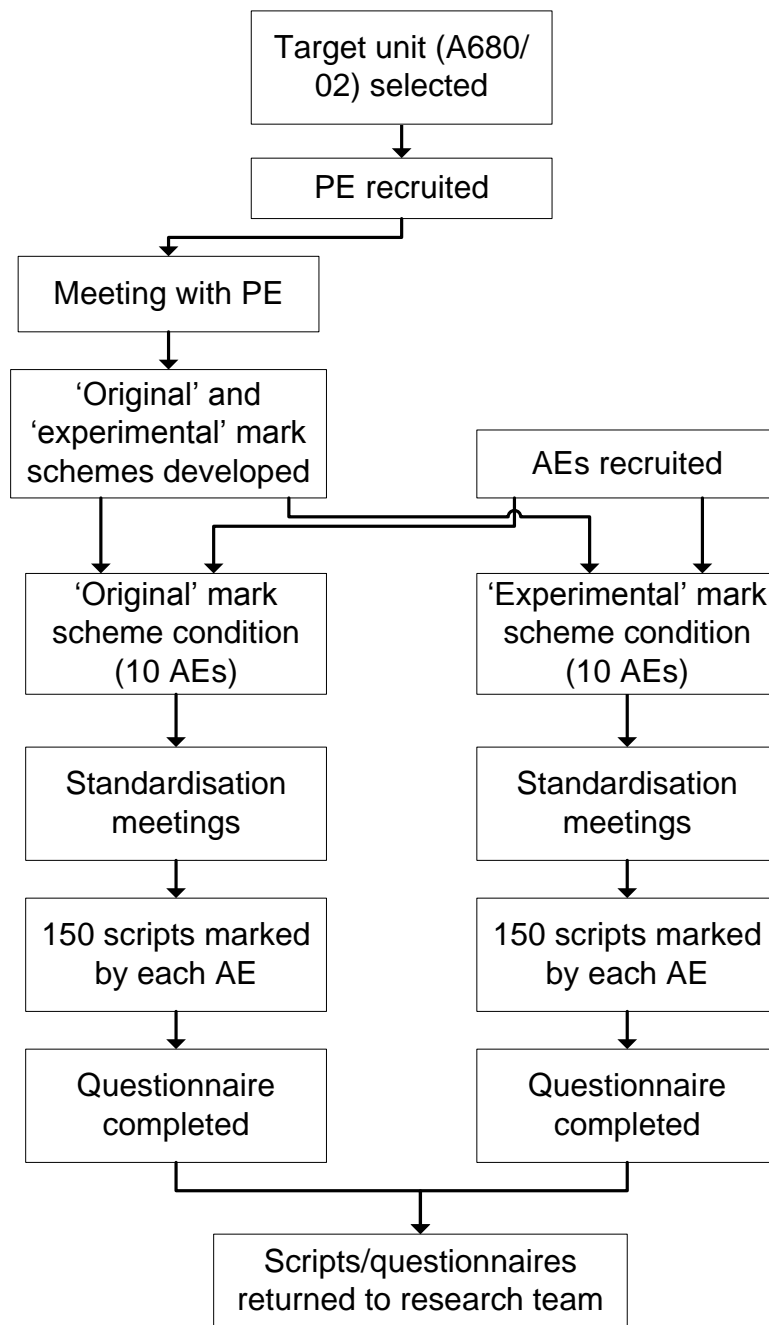


Figure 2.1: Summary of the experimental procedure

2.4.1. Meeting with PE

Once the PE for the target unit had been recruited, an introductory meeting was arranged between the PE and the research team. The meeting confirmed the project plan and timeframe to the PE, allowed the PE to be active in the process of determining the manipulations to be introduced in the 'experimental' mark scheme, and confirmed potential dates for the standardisation meetings to be held.

2.4.2. AE recruitment and mark scheme development

Once the meeting with the PE was completed, the research team then began the process of recruiting the AEs for the study, and assigning them to either the 'original' mark scheme condition, or the 'experimental' mark scheme condition. The researchers also developed the two mark schemes that were to be used in the two experimental conditions, and confirmed these changes with both the PE and OCR colleagues.

2.4.3. Standardisation meetings

To improve the ecological validity of the study, it was necessary to replicate the processes that examiners undergo when learning how to use a mark scheme. As the AEs recruited for the study had no previous experience of marking the target unit, it was appropriate for them to receive training from the PE (in the form of a standardisation meeting) on how to interpret the mark scheme. After consultation with the PE about the timing and size of the meetings, the researchers organised four standardisation meetings that took place over two weeks, which were led by the PE. Two of the meetings used the 'original' mark scheme, and two used the 'experimental' mark scheme. Participants were registered to attend one of the meetings that matched the mark scheme condition they were assigned to, with five participants in each meeting.

Each standardisation meeting was exactly four hours in length, and had the same overall structure. This was to ensure that the four standardisation meetings were as similar as possible to each other, to prevent differences in marking being attributed to differences in marker training across meetings or experimental conditions. Due to unforeseen circumstances on the day of the standardisation meetings, three Assistant Examiners withdrew from the study at this point, leaving a total sample of 17 AEs. To achieve a standard format for each meeting, the research team and PE agreed a formal structure that was to be followed for each meeting (see Table 2.3). The PE was also instructed by the research team to present the practice scripts in the same order for all four of the meetings, to take care to deliver the same commentary on each script, and to spend approximately the same length of time discussing each script. Before the meetings, the PE was asked to select 10 scripts from the available 160 for use in the standardisation meetings. The remaining 150 scripts were used as the set of scripts for the AEs to mark post-standardisation.

The research team attended each meeting to ensure that the administrative elements of the meeting were accounted for, to offer support to the PE when required, and to check that the meeting structure was adhered to.

Each AE was sent the relevant mark scheme, the question paper, the source booklet and administrative details one week before their meeting was due to take place. They were instructed to familiarise themselves with these materials before their meeting.

All four of the meetings were held in the same meeting room. The room was set up in a 'boardroom' format, with the PE at the head of the table and the AEs distributed around the remaining three sides. The research team were positioned at one end of the room so that they could observe the session without intruding on the activities of the meeting. When the AEs sat down at the table, they were given copies of the appropriate mark scheme, the question paper, the source booklet and the 10 standardisation scripts.

Following the standardisation phase of the meeting, the AEs were given a materials pack which contained the following items: 150 scripts for the AEs to mark using the mark scheme provided; a marking sheet for the AEs to write the marks given for each question; a copy of the questionnaire for AEs to complete after they had finished marking the 150 scripts; expenses forms and return envelopes; and an instruction sheet detailing how to return the materials back to the research team. It was at this point that the AEs were thanked for their attendance and the meeting was concluded.

The AEs were not given any further instruction from the PE (which might often happen during a live examination procedure). This was to enable the desired amount of experimental control over the interactions between the AEs and the PE.

Table 2.3: Agenda of the standardisation meetings

Time spent	Meeting activity
15 minutes	Welcome and introductions (research team)
5 minutes	Administrative briefing by research team – collection of contracts/consent forms
3 hours 30 minutes	Standardisation begins (PE): <ul style="list-style-type: none"> • Initial reading and comment on mark schemes by PE. • Read through of three example scripts, with comment by PE. • Short break (10 minutes) • Read through and discussion of four further scripts as a group, with guidance as required. • Personal read through and marking of three scripts, then group discussion on marks given. • Final questions
10 minutes	End of standardisation. Examiners given marking packs (research team)
	Meeting ends

2.4.4. Marking of the scripts

The AEs were each given two months following their respective meetings to mark the 150 scripts and return the materials back to the research team. The AEs were given the same batch of 150 scripts to mark, no matter which experimental condition they were assigned to. This was to ensure maximum comparability between the two mark schemes. They were instructed not to contact the PE or each other during this time.

The PE was required to mark the 150 scripts twice; once using the ‘original’ mark scheme and once using the ‘experimental’ mark scheme. This was to cater for the theoretical possibility that differences in the mark schemes may result in a difference of ‘true’ mark. The PE was asked to mark the 160 scripts using the ‘original’ mark scheme before the standardisation meetings (as part of the process of selecting standardisation scripts for use in the meetings), and the remaining 150 scripts using the ‘experimental’ mark scheme after the meetings had concluded. To reduce the likelihood that the PE would remember scripts from the first set of marking, they were asked to leave a gap of one month between marking using the ‘original’ mark scheme and the ‘experimental’ mark scheme. All of the marking sheets and scripts were also collected once the PE had completed marking using the ‘original’ mark scheme, so that they could not be referred to during their marking of the second allocation of scripts. As the PE also participated in a live session for the unit in between marking the two batches it was expected that they would be unable to remember the scripts or the marks given.

3. Results

The analysis of the data is presented below in four subsections. The first subsection presents the descriptive statistics for both the 'original' and 'experimental' mark schemes, and information related to the score distributions for Questions 2 and 4. For the second subsection, marker reliability for both the mark schemes is analysed using intra-class correlations, and an analysis of the average absolute deviation of examiners from the median scores for each question. For completeness, the third subsection presents analyses related to the comparison between the marks given by the AEs and the definitive marks provided by the PE for each candidate. It should be noted that for this subsection the analysis is no longer strictly measuring reliability. That is, we are no longer looking at how much the marks pupils are awarded would change if the process of marking were simply repeated. Instead, the discrepancies between individual markers and the PE will be affected by both the extent of agreement amongst the AEs, and the extent to which the average scores awarded across AEs agree with the PE's marks. Finally there is a brief analysis of the questionnaire data.

3.1. Descriptive results

For ease of analysis it was necessary that there was a complete set of marks for each script for all markers. This was achieved within the data collection with the exception of one mark for one script that was awarded a mark of 15 for Q4_AO3iii despite only 14 marks being available. This one unusual mark was replaced with 14 (the maximum for the scale). This was plausible as the marker had also awarded very high marks (25 out of 26) for this script for the first part of the same question (Q4_AO3i/ii)⁴.

Having made this slight amendment to the data we first examined the differences in the marks awarded under the two different mark schemes. To begin with the distributions of marks awarded to each question under each mark scheme were examined (see Table 3.1). Descriptive statistics are shown for all three of the elements that were marked as part of the study (Q2, Q4_AO3i/ii and Q4_AO3iii) as well as for the total scores awarded to Question 4 and for the total score awarded across both of Questions 2 and 4 combined.

There was little overall change in the mean score awarded to candidates under each mark scheme with the overall mean changing from 33.2 to 33.6. None of the changes in the mean scores awarded are statistically significant⁵. However, more striking are the differences in the standard deviations across different mark schemes⁶. For both Q2 and Q4_AO3i/ii the standard deviation increased by more than a tenth between the two experimental conditions. Similarly, for the overall total across both questions the standard deviation of scores increased by more than half a mark from 7.9 to 8.5 marks. This indicates that, under the 'experimental' mark scheme, markers were more likely to award a wider range of marks and more likely to award marks at the extremes of the available range. In other words, it appears that the 'experimental' mark scheme encouraged markers to be less conservative in their

⁴ Analysis was also run completely excluding *all* marks from the offending candidate and marker (meaning a total of 166 sets of marks were ignored). This made no difference to results overall but seemed a fairly heavy-handed approach. For this reason, the approach of imputing a sensible value for the one missing mark is preferred.

⁵ Verified using t-tests from cross-classified multilevel modelling in R.

⁶ Furthermore, some simple multilevel modelling based upon the mean scores awarded to each pupil indicated that the changes in the standard deviations of the scores were statistically significant.

marking and to reward the best answers with higher marks whilst awarding marks to lower quality answers more stringently. Table 3.1 also displays the mean of the standard deviations of marks awarded by each individual marker to each of the questions – i.e. the average amount of variation in scores by individual markers under each mark scheme. These show similar changes to the overall standard deviations, confirming that under the ‘experimental’ mark scheme markers tend to use a broader range of marks. Bearing this change in the score distributions between the mark schemes in mind will be crucial to the interpretation of the results presented later in this section.

Table 3.1: Descriptive Statistics

Statistic	‘Original’ mark scheme	‘Experimental’ mark scheme
Number of markers	8	9
Number of candidates marked	150	150
Number of set of marks awarded	1200	1350
Q2		
Min	0	0
Max	14	14
Mean	6.90	6.74
Median	7	7
Standard Deviation	3.02	3.54
Mean SD for individual markers	2.94	3.31
Q4_AO3i/ii		
Min	6	3
Max	26	26
Mean	17.26	17.44
Median	17	18
Standard Deviation	3.57	3.90
Mean SD for individual markers	3.35	3.69
Q4_AO3i/ii		
Min	2	1
Max	14	14
Mean	9.05	9.39
Median	9	9
Standard Deviation	2.34	2.26
Mean SD for individual markers	2.18	2.20
Q4 (Total)		
Min	8	5
Max	40	40
Mean	26.31	26.83
Median	26	27
Standard Deviation	5.78	5.93
Mean SD for individual markers	5.41	5.67
Total (across Q2 and Q4)		
Min	9	7
Max	53	54
Mean	33.21	33.56
Median	33	34
Standard Deviation	7.88	8.45
Mean SD for individual markers	7.54	7.97

A visual presentation of the differences in scores distributions between the two mark schemes is shown in Figure 3.1. The difference in the distribution of marks awarded to Q2 is

particularly striking although the increased spread of scores is also visible for Q4_AO3i/ii, Q4 (Total) and for the overall total score distribution.

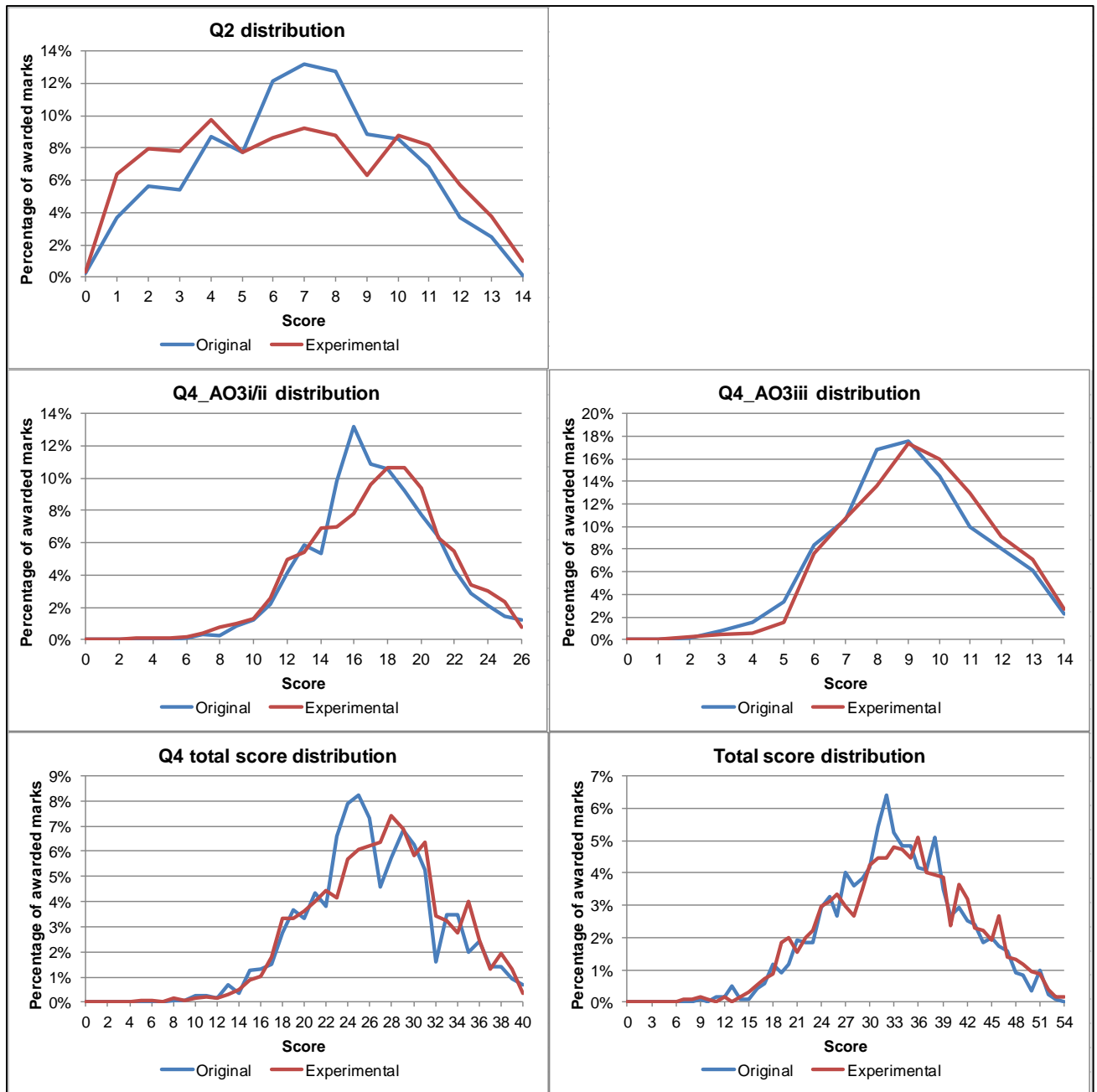


Figure 3.1: Score distributions across examiners and scripts for each mark scheme

The change in the distribution of total scores is illustrated further, via the cumulative distributions of scores (Figure 3.2). This chart reveals the differences in the score distributions more clearly (see also Appendix E for these distributions by question). Assuming that the pass rates at each grade in the sample of candidates were the same as for the population of candidates taking A680/02 overall, then the C grade boundary would be positioned at 28 marks under the 'original' mark scheme and 1 mark lower (at 27 marks) under the 'experimental' mark scheme. In contrast, the A grade boundary would be placed at 41 marks under the 'original' mark scheme and 1 mark higher (42 marks) under the 'experimental' mark scheme. In other words, it is likely that the grade boundaries would be further apart under the 'experimental' mark scheme than under the 'original' mark scheme.

This in itself can be helpful in terms of improving the reliability of the grades awarded to candidates.

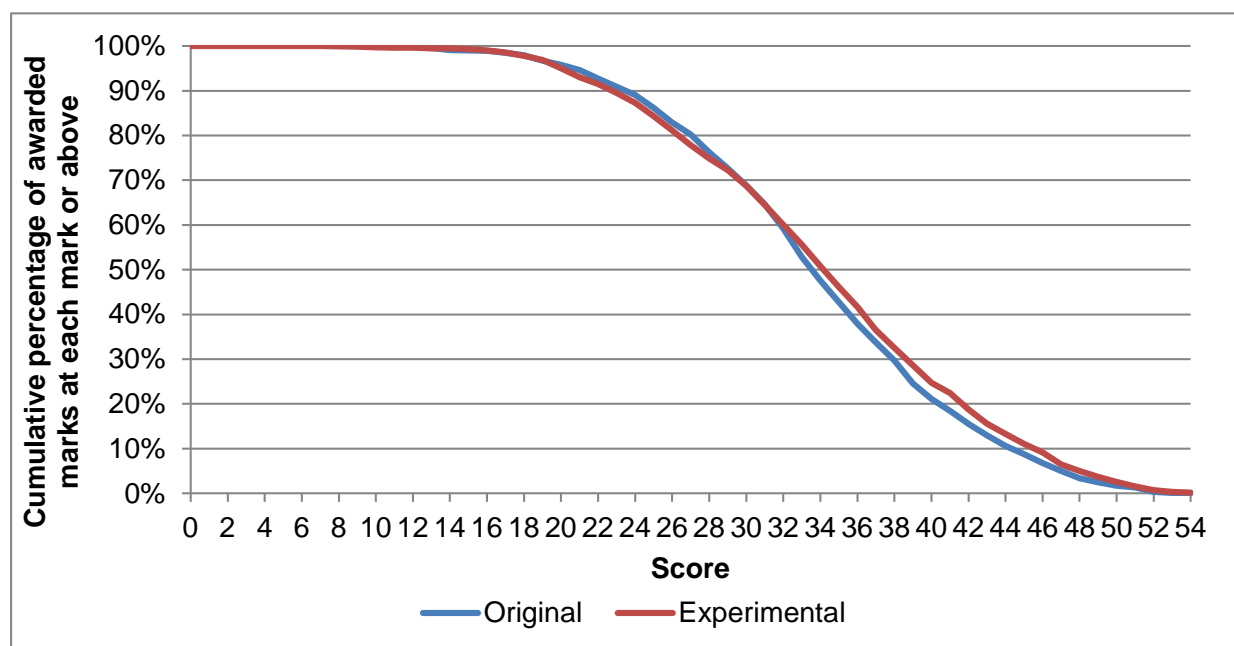


Figure 3.2: Cumulative score distribution for total score across Q2 and Q4 for each mark scheme

3.2. Marker consistency

With the above results in mind we then examined the empirical reliability of marking under each mark scheme. To begin with, marking reliability is quantified in terms of intra-class correlations; that is, the percentage of the variance in scores attributable to candidates as opposed to markers⁷. Ideally all of the variance in scores would be associated with the answer that has been provided and none of it would be associated with who has marked the paper. However, for some components marking reliability may be lower than this as can be assessed using intra-class correlations.

The main advantage of assessing marking reliability via intra-class correlations is that it automatically accounts for the overall variation in scores. That is, if the marks awarded to candidates are more spread out then differences of a given amount between markers become less important as they are less likely to affect the rank order of candidates. Similarly, as the gaps between grade boundaries become wider then differences of a given amount between markers are less likely to have an impact on the grades awarded to candidates. As outlined above, because the ‘experimental’ mark scheme appears to have affected the score distribution, such considerations are likely to be important for our analysis.

The estimated marking reliability of each mark scheme for each element of the assessment is shown in Table 3.2. As can be seen, the marking reliability under the ‘original’ mark scheme is extremely low. In particular for all elements of Q4 the marking reliability is below

⁷ See Appendix F for details on the method used to compare intra-class correlations.

0.5 indicating that who marks a given answer is a greater influence on the mark that is awarded than what has been written⁸. The reliability of Q2 appears a little higher under the 'original' mark scheme, with 60 per cent of the variance in scores attributable to candidates. Marking reliability under the 'experimental' mark scheme appears somewhat better. Specifically, although there is little change for Q2, substantially improved marking reliability can be seen for both elements of Q4 as well as for the total score awarded across both items. The difference is particularly striking for Q4 where an additional 10 per cent of the variance in scores is attributable to candidates rather than markers when the 'experimental' mark scheme is used. The statistical significance of differences was calculated using Fisher's z-test using the formulae detailed by Donner and Zou (2002). As can be seen, the improvement in marking reliability is statistically significant for all of the elements considered where an increase in reliability was found.

Table 3.2: Marking reliabilities (intra-class/intra-candidate correlations) under each mark scheme

Score	'Original' mark scheme marking reliability	'Experimental' mark scheme marking reliability	Significance of difference (p-value)
Q2	0.601	0.583	0.396
Q4_AO3i/ii	0.481	0.535	0.032
Q4_AO3iii	0.452	0.599	0.000 ⁹
Q4 (Total)	0.482	0.585	0.000 ¹⁰
Total (Q2+Q4)	0.586	0.627	0.049

To provide some context to the positive findings above, we also examined the average absolute deviation (AAD) of marks from the median mark awarded to each candidate. In order to do this we first calculated the median mark awarded to each candidate under each mark scheme (across all markers). Then the absolute difference between the marks awarded by each individual marker and these median marks was calculated. The mean of these differences is the average absolute deviation. This provides a fairly simple measure of the level of inconsistency between markers.

The mean of these average absolute deviations is shown in Table 3.3. This shows, for example, that on average, under the 'original' mark scheme, the total mark awarded by any individual marker was 3.7 marks away from the median mark awarded to that candidate. As can be seen, this value is essentially unchanged when using the 'experimental' mark scheme (3.8 marks). Indeed, across all of the different questions only Q4_A03iii shows a substantial reduction in the extent of differences between markers (in terms of marks) the 'experimental' mark scheme was used. Furthermore, for Q2 the mean average absolute deviation is actually somewhat greater under the 'experimental' mark scheme than under the 'original' mark scheme.

⁸ Note that analysis of seeds scripts marked during the live session does not reveal quite such a disappointing story with item level marking reliabilities of roughly 0.7 estimated for each item. However, only 9 seed scripts were available in this analysis compared to the 150 used for analysis in our research here.

⁹ Exact p-value was 1.04×10^{-9} .

¹⁰ Exact p-value was 1.65×10^{-5} .

Table 3.3: Mean average absolute deviations under each mark scheme

Score	Mean average absolute deviation (across candidates)		Percentage of candidates where average absolute deviation is lower for experimental mark scheme
	'Original' Mark Scheme	'Experimental' Mark Scheme	
Q2	1.37	1.66	31%
Q4_AO3i/ii	1.84	1.89	44%
Q4_AO3iii	1.25	1.01	69%
Q4 (Total)	3.00	2.72	58%
Total (Q2+Q4)	3.67	3.76	45%

Further details on these results are given in Figure 3.3. This shows the average absolute deviation (from the median) for each candidate under each mark scheme. As can be seen (and as is also detailed in the final column of Table 3.3), the average absolute deviation only has a tendency to be lower under the 'experimental' mark scheme for Q4_AO3iii; in this case the average absolute deviation is lower under the 'experimental' mark scheme for 69 per cent of candidates. In contrast, the average absolute deviation tends to be *larger* under the 'experimental' mark scheme for Q2 (for 69 per cent of candidates¹¹). These results indicate that, except for Q4_AO3iii, the 'experimental' mark scheme does not improve reliability in terms of the agreement of the AEs with each other. It is only once we take into account the changes in the distribution of scores that the 'experimental' mark scheme can be viewed as improving marking reliability.

¹¹ That is, 100% minus the 31% displayed in Table 3.3.

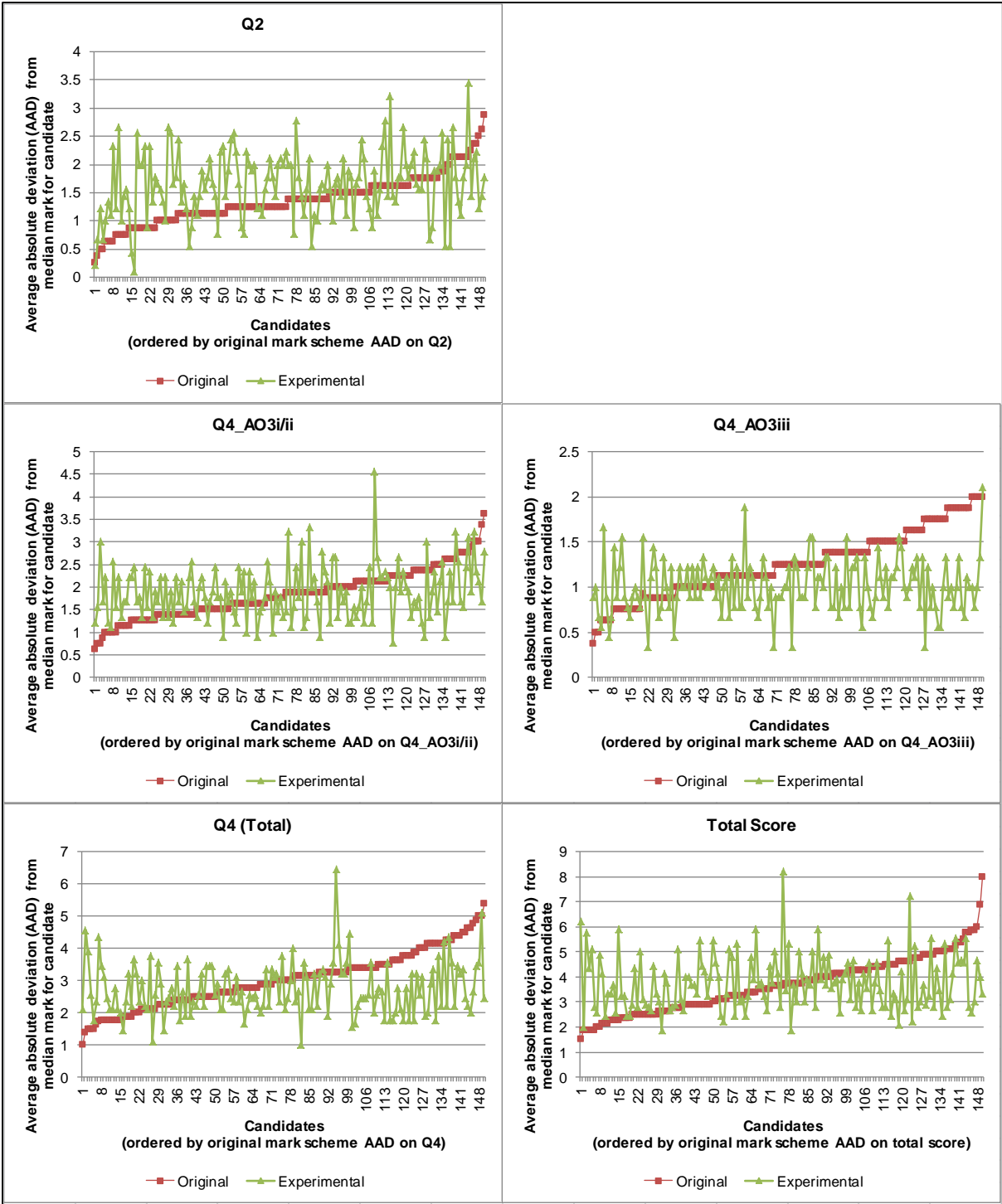


Figure 3.3: Average absolute deviations from the median for each candidate under each mark scheme

3.3. Consistency between individual markers and the Principal Examiner

For the second stage of the analysis the marks awarded by individual markers under each mark scheme to the marks awarded by the Principal Examiner (PE) were compared. This, in Bramley's (2007) sense, is an analysis of the degree of *agreement* between the AEs and the PE under each mark scheme. As part of the method the PE marked each script twice - once for each mark scheme. This meant that the marks awarded by individual examiners could be compared to the PE marks that were awarded under the same mark scheme. However, as explored further in Appendix G, the marks awarded by the PE were extremely similar regardless of which mark scheme was used.

Table 3.4 and Figure 3.4 explore the average absolute difference (AAD) between the marks awarded by individual markers and the PE. The results in Table 3.4 are marginally more positive than those examining pure reliability in terms of marks (Table 3.3). Specifically, the size of differences between the PE and individual markers is lower under the 'experimental' mark scheme for all of Q4_AO3i/ii, Q4_AO3iii, the total score for Q4 and the total score across both questions. Paired t-tests on these differences showed that the reduction in the size of differences is statistically significant for Q4_AO3iii and for the total score across Q4.

On the other hand, Table 3.4 also shows that for Q2 the discrepancy between individual markers and the PE is significantly higher for the 'experimental' mark scheme. However, this result should be set in the context of the way in which the distribution of marks awarded has widened under the 'experimental' mark scheme. That is, we know that a wider range of marks are being used under the 'experimental' mark scheme and that, as a result, the gaps between grade boundaries are likely to be wider (see Figure 3.1). This implies that a difference of a given size between markers is less likely to affect a pupil's grade in the 'experimental' mark scheme. In this context, the increase in the size of differences between markers and the PE for Q2 may make little practical difference.

Table 3.4: Mean average absolute deviations from the Principal Examiner (PE) mark under each mark scheme

Score	Mean average absolute deviation from PE mark (across candidates)		P-value of difference (paired t-test)	Percentage of candidates where average absolute deviation from PE mark is lower for experimental mark scheme
	'Original' Mark Scheme	'Experimental' Mark Scheme		
Q2	1.72	1.99	0.000	34%
Q4_AO3i/ii	2.97	2.91	0.404	47%
Q4_AO3iii	1.86	1.52	0.000	70%
Q4 (Total)	4.72	4.27	0.000	63%
Total (Q2+Q4)	5.33	5.19	0.288	53%

To further understand the reason for the slight improvement in agreement between individual markers and the PE we also examined the extent of agreement in the ranking of candidates. In order to undertake this analysis each marker's set of marks (including for the PE) were replaced with their ranking of each candidate (between 1 and 150). Where more than one

candidate was awarded the same mark by an individual marker (so that their ranking would be the same) the mean ranking across all such candidates was used. The correlation between the rankings awarded by individual markers and the rankings awarded by the PE was then calculated.

The mean of the correlations between the PE and the AEs for each item under each mark scheme are shown in Table 3.5. As can be seen the correlation between the rankings was greater under the 'experimental' mark scheme in every case other than Q4_AO3i/ii. This implies that the ranking of candidates agreed more closely with that of the PE under the 'experimental' mark scheme. However, the sizes of these improvements are quite small (at most 0.05) and some exploratory further analysis¹² showed that none of these improvements was statistically significant. For this reason, it does not appear that the marginally positive results in Table 3.4 can be entirely explained by an improvement in the consistency of the ranking of candidates.

Table 3.5: Correlation of candidate ranking between Principal Examiner and other individual markers

Score	Correlation of individual markers rankings and PE ranking	
	'Original' Mark Scheme	'Experimental' Mark Scheme
Q2	0.74	0.78
Q4_AO3i/ii	0.65	0.65
Q4_AO3iii	0.65	0.70
Q4 (Total)	0.67	0.69
Total (Q2+Q4)	0.74	0.76

In fact, at least part of the explanation for the improvement is again found in the change to the score distributions associated with the 'experimental' mark scheme. Specifically, the distribution of the median marks awarded to candidates becomes closer to the distribution of definitive marks under the 'experimental' mark scheme. For example, looking at the total score awarded across Q2 and Q4, under either mark scheme, the distribution of PE marks had a mean of roughly 36.7 with a standard deviation of 7.4. Under the 'original' mark scheme the median marks for individual candidates across markers had a somewhat lower mean (33.3)¹³ and a substantially lower standard deviation (6.5). In contrast, for the 'experimental' mark scheme, the mean of the median scores is slightly closer to that of the PE's marks (33.7) and the standard deviation is much closer at 7.0. In other words, the changes to the score distribution bring the scale of the marks more into line with those awarded by the PE.

In assimilating all of the results provided in this subsection perhaps the key piece of information, that we should not lose sight of, is from the final row of Table 3.4. This shows no

¹² Using multilevel modelling and looking at the change between mark scheme in the regression coefficient (rather than the correlation) between the PE ranking and individual markers' rankings.

¹³ The substantially lower mean could potentially be explained by the fact that a full, formal standardisation process (of the type that would take place for live marking) had not occurred. This is another reason to prefer focussing on pure marker reliability (see the previous section) over comparisons of each marker to the PE.

significant change in the consistency of individual markers with the PE. In other words the findings are essentially unchanged from the earlier analysis (section 3.2). However, the results from this analysis are entirely consistent with the direct assessment of marking reliability presented earlier.

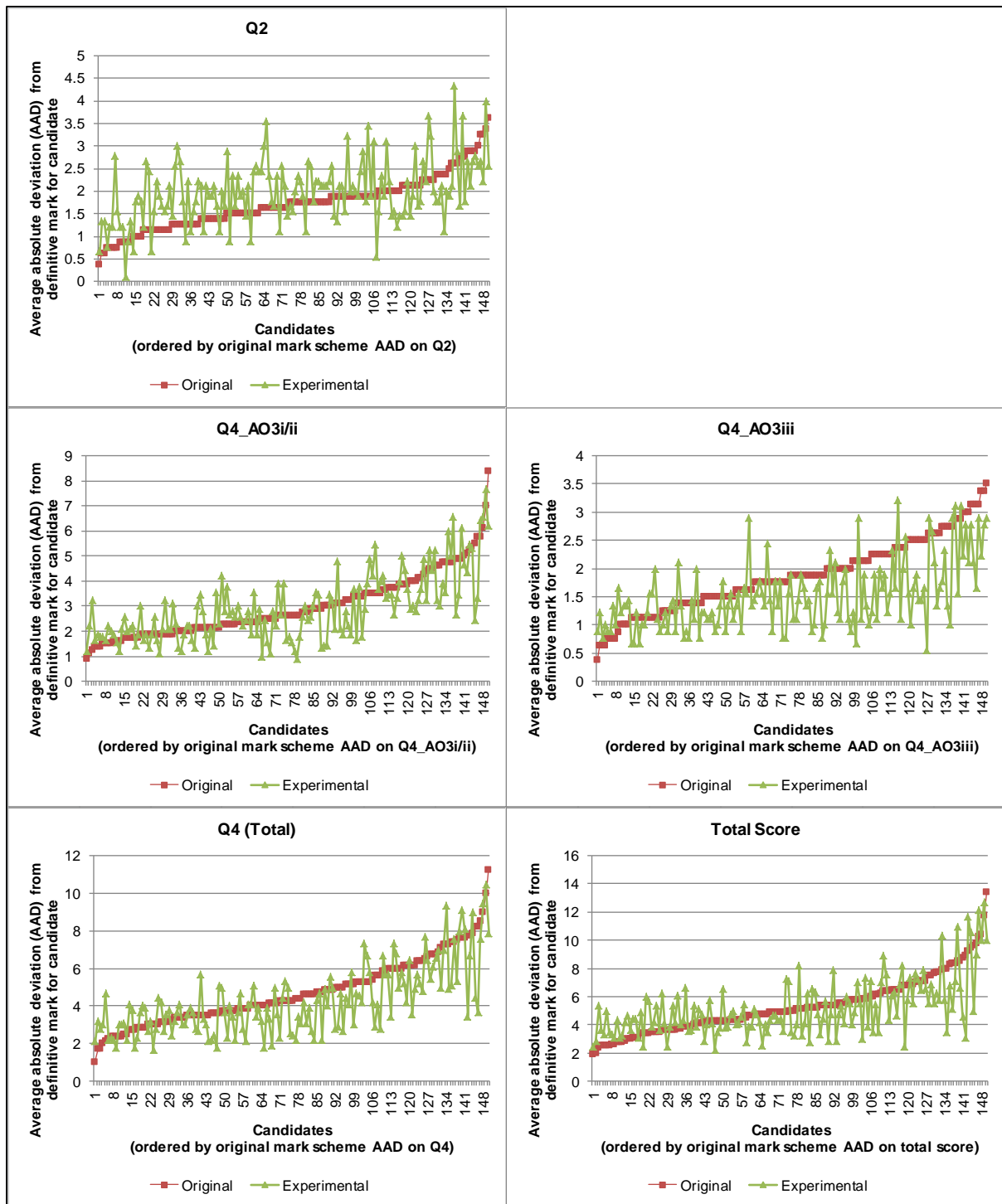


Figure 3.4: Average absolute deviations from the Principal Examiner’s mark for each candidate under each mark scheme

3.4. Questionnaire analysis

Once the AEs had completed their allocation of scripts, they were asked to complete a short questionnaire, which attempted to gather their views on the mark scheme they had used. Due to the small sample size for a questionnaire, it was not expected that statistically significant differences in ratings between the mark schemes would be found, although some trends were possible to observe.

The AEs were first asked to rate the mark scheme they used for its ease of use, clarity, and layout on a five point Likert scale (point five being the top rating). Although none of the differences were statistically significant (see Table 3.5) a reasonably large difference was found in the rating of the layout of Question 4, with the 'experimental' mark scheme scoring 0.5 more compared to the 'original' mark scheme. This appears to suggest that the layout of the 'experimental' mark scheme is a particular strength, particularly for the question that contains the most text within the mark scheme.

Table 3.5: AE ratings of the two mark schemes

Question	Mark scheme	N	Mean rating	Likert frequencies				
				1	2	3	4	5
Q2 - Please rate the mark scheme in terms of its ease of use	Original	8	3.75	0	1	1	5	1
	Experimental	9	3.78	0	0	2	7	0
Q2 - Please rate the mark scheme in terms of its clarity	Original	8	3.75	0	0	4	2	2
	Experimental	9	3.89	0	0	2	6	1
Q2 - Please rate the mark scheme in terms of its layout	Original	8	4.38	0	0	0	5	3
	Experimental	9	4.11	0	0	2	4	3
Q4 - Please rate the mark scheme in terms of its ease of use	Original	8	3.38	1	0	4	1	2
	Experimental	9	3.00	0	3	3	3	0
Q4 - Please rate the mark scheme in terms of its clarity	Original	8	3.38	0	1	5	0	2
	Experimental	9	3.44	0	2	2	4	1
Q4 - Please rate the mark scheme in terms of its layout	Original	8	3.50	0	1	3	3	1
	Experimental	9	4.00	0	0	4	1	4

At various points in the questionnaire, the AEs were given the opportunity to articulate their perceptions of using the mark schemes. They were asked for their views on the content of the mark scheme, mark scheme features that they thought helped or hindered their marking, and for any changes they would make to the mark schemes they used.

There were some trends observed in the AE comments related to specific mark scheme features. For Question 2, over five out of the nine AEs who used the 'experimental' mark scheme thought that the 'notes on task' section of the mark scheme was useful in making marking decisions. Some comments are provided below:

'I found the additional notes on task useful in sharpening my focus.' (Participant 5B)

'Notes on [the] right side helpful when undecided for checking.' (Participant 2B)

In contrast, the 'notes on task' section of the mark scheme was not mentioned at all by the AEs using the 'original' mark scheme. Although appropriate caution should be maintained here, the unprompted positive views about the 'notes on task' for Question 2 by users of the 'experimental' mark scheme is likely due to its close proximity to the levels descriptors (i.e. on the same page) compared to the 'original' mark scheme.

There were also multiple positive references to the bolding of key terms and phrases in the 'experimental' mark scheme. The AEs who used the 'experimental' mark scheme felt that the bold words for each level improved their ability to focus on the task:

'The bold text helped me to focus on what I felt I should look for in each script.'
(Participant 3B)

Question: *Were there any features of the mark scheme that you felt helped you mark more effectively?*

'Words in bold. Especially the analytical comments as they were a key factor in choosing bands.' (Participant 2B)

The main issue across both mark schemes was the level of detail, particularly for Question 4. Seven of the 17 AEs reported to there being too much detail in the mark scheme for Question 4, compared to two out of 17 for Question 2. One of the AEs using the 'original' mark scheme commented that a reduction of content to reflect key points would have been helpful:

'Too many filler words. Short bullet points with key words or phrases would have meant more bands on a page and a mark scheme physically easier to deal with.'
(Participant 4A).

The physical proximity of the level descriptors appeared to be linked to AEs views of the mark scheme. In the examples below, AEs who used the 'original' mark scheme suggested changing the presentation of the mark scheme so that the most relevant information for each question was accessible on one page:

Question: *Would you make any changes to the layout of the mark scheme?*

'Note form. Perhaps a full descriptive + detailed section and a one page (A4 for Q2 and A3 for Q4) chart for reference. A flow chart could be interesting.' (Participant 4A)

'Place the most relevant band descriptors for reading/writing closer together in the package, so that one can quickly flip between the two, would be better.' (Participant 5A)

'Any mark scheme for English will necessarily be lengthy and detailed; therefore, possibly a fold out version would be more user friendly. The double sided booklet format means I spend ages photocopying and making my own.' (Participant 9A)

This is interesting given that the AEs who used the 'original' mark scheme were unaware of the change in formatting in the 'experimental' mark scheme. The AEs noted that while the use of bolding helped them quickly identify differences between level descriptions, the use of multiple bullet points made it challenging to gauge which points were most important in selecting an appropriate level.

3.5. Results summary

The consistency of marking, as measured by intra-class correlations, was not clearly improved by using the 'experimental' mark scheme. While for Question 4 (particularly Q4_AO3iii) results indicate that the 'experimental' mark scheme may, to an extent, improve reliability in terms of the degree of agreement between the AEs and the PE, there were no clear differences between mark scheme types for Question 2.

It is only once we take into account the changes in the distribution of scores that the 'experimental' mark scheme can be viewed as improving marking reliability. The results indicate that the 'experimental' mark scheme seemed to encourage AEs to use a greater range of the mark scheme and that this is achieved without any loss of consistency between examiners. This change in the distribution of marks observed in the 'experimental' mark scheme implies that inconsistency in marking may have a smaller effect on grade outcomes. If a wider range of marks are being used under the 'experimental' mark scheme, the gaps between grade boundaries are likely to be wider. Subsequently, this implies that a difference of a given size between markers is less likely to affect a pupil's grade in the 'experimental' mark scheme. Thus, although there is no clear overall improvement in reliability in terms of numbers of marks, this does imply that for practical purposes the quality of marking was improved when using the 'experimental' mark scheme.

However, it should be noted here that even under the 'experimental' mark scheme it was found that approximately 40 per cent of the variance in scores was associated with which marker has marked a script rather than the answer that had been provided – a far from ideal situation. Nonetheless, these results indicate that simple changes to mark schemes can indeed help improve their usability without detriment to the quality of marking. The questionnaire results highlighted potential mark scheme manipulations that were perceived by examiners to focus their marking decisions, including the bolding of key terms, the proximity of level descriptors to each other, and the positioning of guidance that assisted level decisions.

4. Discussion

The present study makes an important contribution to the expanding body of research related to the development of mark schemes. This study was an attempt to experimentally test the role that mark scheme features have on the statistical reliability of a matched set of examiners using levels-based mark schemes. Previous research has only investigated examiners' *perceptions* of mark scheme features, or had used statistical models that could not tease apart aspects related to the item type from those related to mark scheme features. The present study controlled the items (and scripts) marked, the mark scheme content, and the training provided to the AEs in each mark scheme condition. This made it possible to attribute differences in reliability, marker agreement, and views on the usability of the mark schemes between examiners, to the differences in the two mark schemes in the study. In this section, we first elucidate key findings from the statistical comparison of the two mark schemes, and offer some potential explanations for these findings. Potential limitations of the research are also described and discussed, followed by a section that outlines possible implications of the current study.

The analysis of reliability (intra-class correlations) revealed that for Question 2 marking reliability was not significantly improved by using the 'experimental' mark scheme. However, for Question 4 (particularly Q4_AO3iii) results indicated that the 'experimental' mark scheme may, to an extent, have improved reliability in terms of the degree of agreement between markers, and between the AEs and the PE. It is therefore uncertain as to whether the 'experimental' mark scheme conclusively improved marking reliability. This observed improvement in reliability for Question 4 appears to be explained in part by changes in the distribution of scores for the 'experimental' mark scheme compared to the 'original' mark scheme. The results indicated that the 'experimental' mark scheme seemed to encourage AEs to use a greater range of marks, and that this increase resulted in a greater proportion of variance attributed to the true score rather than to error.

This observed difference between the mark schemes suggests some improvement in overall quality of marking under the 'experimental' mark scheme. Examiners that were using the 'experimental' mark scheme used a wider range of the mark scheme, compared to the examiners using the 'original' mark scheme. Although it is not a straightforward task to know for certain how a marking distribution should look (Pinot de Moira, 2013), a fair assumption is that no marks in the mark scheme should be underutilised (Pinot de Moira, 2011). Therefore, it appears that the 'experimental' mark scheme improved examiners' quality of marking in this respect across all the questions analysed. This increased discrimination between scripts may have the advantage of increasing the distance between the grade boundaries in terms of marks. This is advantageous for levels-based mark schemes as it means that differences between examiners are less likely to influence the final grade for a candidate for a paper.

It is not certain as to which changes introduced in the 'experimental' mark scheme affected this observed improvement in the mark distribution, although the comments from the examiners in the present study may offer some enlightenment. The examiners that used the 'experimental' mark scheme thought that the bolding of key terms and phrases was useful in helping them focus on key script features. This was deemed particularly useful for Question 4, where each level descriptor contained multiple bullet points for each assessment objective. Taken together, this suggests that examiners may benefit from a reduction in the overall content of mark schemes; containing more concise level descriptors focused on key

phrases. Achieving the optimum balance between detail in the level descriptors and supplementary oral instruction (e.g. from standardisation meetings) is a potential area for future investigation (see Black et al, 2011).

It may also be the case that presenting the mark scheme in a one page format meant that less information related to the extreme mark bands had to be held in examiners' working memory. Indeed, some examiners who used the 'original' mark scheme commented that they would change its layout so that the level descriptors for each individual question were presented on one page each (in line with the 'experimental' mark scheme). Levels-based mark schemes require examiners to adopt cognitively demanding marking strategies (Suto, Crisp & Greatorex, 2008; Suto & Greatorex, 2008) which may impinge on examiners' abilities to internalise the mark scheme effectively. This is potentially most problematic in the early stages of marking, when examiners are low on expertise and thus make slow, reflective and effortful judgements on scripts (known as *System 2* judgements, Kahneman & Frederick, 2002; Suto & Greatorex, 2008). As examiners gain experience with a particular mark scheme, their judgements become more intuitive (known as *System 1* judgements). The presentation of mark schemes (e.g. the proximity of level descriptors) may influence the 'migration' process between *System 2* and *System 1* judgement processes, for example by encouraging examiners to internalise a wider range of the mark scheme.

Crucially, the improvement in mark distributions observed under the 'experimental' mark scheme was not detrimental to the overall reliability of the mark scheme for each question. Relative to the 'original' mark scheme, the 'experimental' mark scheme was just as reliable, both in terms of marker consistency and in terms of examiners' deviation from the mark given by the PE. Indeed, there was some limited evidence to suggest that for one of the questions (Question 4), there were improvements in reliability for the examiners using the 'experimental' mark scheme compared to those using the 'original' mark scheme. As Pinot de Moira (2013) and others suggest, it is likely that marker reliability is influenced by an interaction between the item type and the mark scheme. It remains a possibility that the mark scheme for Question 2 was easier to internalise given its length relative to Question 4, and so the mark scheme manipulations had less of an effect on examiners' processing of this question. However, the increased mark distribution for Question 2 under the 'experimental' mark scheme suggests that the changes did contribute to the overall quality of marking.

4.1. Limitations

Before discussing the implications of this research, it is appropriate to address some potential limitations of the study. It is important to note that the overall reliability of the examiners across both mark schemes was lower than what has been observed in live marking for the target unit. The lower than hoped for reliability overall means that we cannot be certain that positive effects of the experimental mark scheme would be retained with a more experienced group of markers. The lower reliability is likely due to the recruitment of examiners with experience in marking English Literature as opposed to English Language, and differences between the marking procedure in the present study and the typical procedure in live marking. In the present study there was no practice stage (which typically comprises feedback from a more senior examiner) and no use of seeding scripts to monitor marking quality during the live marking period. However, the removal of these stages was necessary in the present study, as differences in feedback for the AEs could have had an

influence on marker reliability measures (see Johnson, 2014b). In this sense, the design of the study attempted to ensure that instruction given to the AEs was standardised across conditions. Similarly, while steps were taken to standardise the structure, content and delivery of the standardisation meetings, their face-to-face nature meant that there were occasional small differences between the meetings.

Another concern is that, based upon the intra-class correlations, only one question (Question 4) out of the two studied showed any statistically significant improvement from the application of the 'experimental' mark scheme. This leaves us with some doubt as to the proportion of items more generally where the suggested changes to the mark scheme would be beneficial, and the likely extent of any benefit. Replication of the findings in the present study would be beneficial in establishing the role of specific mark scheme features on the marking of items across a wider range of mark scheme types and subject disciplines.

4.2. Implications

The main implication of the present research is that the perceptions of examiners in previous research converted into some statistically recognisable differences in measures of reliability in the expected direction, in addition to notable improvements in marker discrimination between scripts and their perceptions of mark scheme usability. In other words, the results suggest that for practical purposes overall quality of marking may have improved under the 'experimental' mark scheme. The examiners in the present study related specific features of the 'experimental' mark scheme to their perceptions of usability and cognitive processing. This suggests that careful consideration should be given to understanding how levels-based mark schemes are presented, as this appears to have some influence on quality of marking. When using levels-based mark schemes, examiners have the cognitively challenging task of synthesising the mark scheme and the feedback provided by senior examiners, and applying this complex information to make an appropriate best-fit judgement. We suggest here that it is potentially beneficial for mark schemes to follow particular 'principles' of design that may facilitate the marking process. These principles may include introducing features that may reduce examiners' cognitive load, improve levels-based mark scheme usability, increase the salience of key phrases related to each level, or increase the salience of item-specific information (e.g. assessment objectives).

To give an illustrative example, one feature that was identified as beneficial to quality of marking in previous research was the highlighting of key terms in bold. This may act to reduce examiners' cognitive load associated with their initial evaluation of a response before a level is determined (Crisp, 2010), as they are encouraged to interpret item responses based on smaller chunks of text. However, this approach would require PEs and Team Leaders to sculpt their training and feedback to examiners appropriately by focusing their guidance on what these terms mean and how they are represented in candidate responses (Harsch & Martin, 2013; Johnson, 2008; Pollitt & Ahmed, 2008). In the present study, the Principal Examiner articulated several idiosyncratic aspects of the target unit that influenced what was possible to change in the mark scheme. It is important that this kind of expertise and judgement is maintained in the development of mark schemes. It may be the case that some of the feature changes introduced in the 'experimental' mark scheme are not appropriate for other units (or already present). However, the principles briefly outlined above may offer some guidance as to how mark scheme usability can be improved within current frameworks.

4.3. Conclusions

The mark scheme represents the crucial point of reference for examiners when considering a candidate's response to an item. Examiners who are tasked with marking extended response items also have to learn to use mark schemes that comprise multiple levels, containing information related to numerous assessment objectives. Given this challenge for examiners, it was appropriate to consider how aspects of mark scheme design could contribute to improving examiners' quality of marking. The present study provides some evidence to suggest that changing mark scheme features is worthy of future consideration with respect to increasing mark scheme usability. It is important to note, however, that no mark scheme is an island; their design is intimately related to the target item, the expected responses from candidates, and the guidance provided by examiners involved in training and standardisation. Future research could consider how senior examiners refer to and use mark schemes in standardisation meetings, and whether different approaches relate to different marking outcomes.

5. References

- Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy and Practice*, 18(3), 259-278.
- AlphaPlus. (2014). *Standardisation methods, mark schemes, and their impact on marking reliability*.
- Black, B., Suto, 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 and Practice*, 18(3), 295-318.
- Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement*. Paper presented at the annual conference of the British Educational Research Association, Edinburgh, Scotland.
- Brooks, V. (2012). Marking as judgment. *Research Papers in Education*, 27(1), 63-80.
- Cambridge Assessment. (2013). *A simple guide: Assuring OCR's marking accuracy*. Available at: <http://www.cambridgeassessment.org.uk/Images/143058-a-simple-guide-to-marking-and-grading.pdf>. (accessed 19th September 2013).
- Chamberlain, S. & Taylor, R. (2011). Online or face-to-face? An experimental study of examiner training. *British Journal of Educational Technology*, 42(4), 665-675.
- Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education*, 36(1), 1-21.
- Donner, A. & Zou, G. (2002). Testing the Equality of Dependent Intra-class Correlation Coefficients. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 51(3), 367-379.
- Fowles, D. (2009). How reliable is marking in GCSE English? *English in Education*, 43(1), 49-67.
- Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts*, 241-76. Norwood, NJ: Ablex.
- Harsch, C. & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy and Practice*, 20(3), 281-307.
- Johnson, M. (2008). Assessing at the borderline: Using a mixed-method approach to find out what assessors attend to when they holistically judge a vocationally-related portfolio. *Issues in Educational Research*, 18(1), 26-43.

- Johnson, M. (2014a). Insights into Contextualised Learning: How do professional examiners construct shared understanding through feedback? *E-learning and Digital Media*, 11(4), 363-378.
- Johnson, M. (2014b). A case study of inter-examiner feedback from a UK context: Mixing research methods to gain insights into situated learning interactions. *Formation et pratiques d'enseignement en questions*, 17, 67-88.
- Kahneman, D. & Frederick, S. (2002) Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin & D. Kahneman (Eds). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.
- Laming, D. (2004). *Human Judgement: The Eye of the Beholder*. London: Thomson.
- Massey, A.J., & Raikes, N. (2006). Item-level examiner agreement. Paper presented at the annual conference of the British Educational Research Association, Warwick, UK.
- Meadows, M. & Billington, L. (2005). *A Review of the literature on marking reliability: Report to NAA*. Available at: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf (accessed 6th February 2015).
- Nadas, R., & Suto, I. (2011). *Assessing the Extended Project Qualification: A model of early challenges for marking accuracy*. A Cambridge Assessment Report.
- Newton, P. (1996). The reliability of marking General Certificate of Education scripts: Mathematics and English. *British Educational Research Journal*, 21, 404-420.
- Ofqual. (2011). *GCSE, GCE, Principal learning and project code of practice*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/371268/2011-05-27-code-of-practice.pdf (accessed 18th December 2014).
- Ofqual. (2014a). *Corporate plan 2014-2017*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/379021/2014-08-19-corporate-plan-2014.pdf (accessed 18th December 2014).
- Ofqual. (2014b). *Review of quality of marking in exams in A levels, GCSEs and other academic qualifications*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/393832/2014-02-14-review-of-quality-of-marking-in-exams-in-a-levels-gcses-and-other-academic-qualifications-final-report.pdf (accessed 6th February 2015)
- Ofqual. (2014c). *Enquiries about results for GCSE and A level: Summer 2014 examination series*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/386109/enquiries-about-results-for-gcse-and-a-level-summer-2014-exam-series.pdf (accessed 19th December 2014).

- Pinot de Moira, A. (2011). *Effective discrimination in mark schemes*. Available at: https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP-RP-APM-05042011.pdf (accessed 19th December 2014).
- Pinot de Moira, A. (2011). *Levels-based mark schemes and marking bias*. Manchester: AQA Centre for Education Research and Practice.
- Pinot de Moira, A. (2013). *Features of a levels-based mark scheme and their effect on marking reliability*. Available at: https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_TR_APM_03042013.pdf (accessed 19th December 2014)
- Pollitt, A., & Ahmed, A. (2008). Outcome Space Control and Assessment. Presented at the 9th Annual Conference of the Association for Educational Assessment, Hissar, Bulgaria.
- Raikes, N. & Massey, A. (2007). Item-level examiner agreement. *Research Matters: A Cambridge Assessment Publication*, 4, 34-38.
- Revelle, W. (2014). Psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 1.4.8. Available at: <http://CRAN.R-project.org/package=psych>. (accessed 9th February 2015).
- Suto, I., Crisp, V. & Greatorex, J. (2008). Investigating the judgemental marking process: An overview of our recent research. *Research Matters: A Cambridge Assessment Publication*, 5, 6-8.
- Suto, I., & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, 34(2), 213-233.
- Suto, I., & Nadas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education*, 23(4), 477-497.
- Suto, I., Nadas, R., & Bell, J. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, 26(1), 21-51.
- Tisi, J., Whitehouse, G., Maughan, S. & Burdett, N. (2013). *A Review of literature on marking reliability research* (report for Ofqual). Slough, NFER.

Appendix A: Examination paper for unit A680/02 (June 2014)

OCR 

H

Tuesday 3 June 2014 – Morning

GCSE ENGLISH/ENGLISH LANGUAGE

A680/02/QPI Information and Ideas (Higher Tier)

QUESTION PAPER INSERT

Duration: 2 hours



INSTRUCTIONS TO CANDIDATES

- This Insert is for your reference only.
- Answer **all** the questions in Section **A** and **one** question in Section **B**.
- Read each question carefully. Make sure you know what you have to do before starting your answer.

INFORMATION FOR CANDIDATES

- The number of marks is given in brackets [] at the end of each question or part question.
- The total number of marks for this paper is **80**.
- This document consists of **4** pages. Any blank pages are indicated.

INSTRUCTION TO EXAMS OFFICER/INVIGILATOR

- Do not send this Question Paper Insert for marking; it should be retained in the centre or recycled. Please contact OCR Copyright should you wish to re-use this document.

Answer **all** the questions in Section **A** and **one** in Section **B**.

SECTION A – Non-Fiction and Media

You are advised to spend about one hour on Section A.

Read carefully the two passages, *GREED IN THE GLOBAL WHALING INDUSTRY (Greenpeace)* and *ON WHALING (Stephen Fry in America)* and then answer questions 1, 2 and 3. These questions will be marked for reading.

1 GREED IN THE GLOBAL WHALING INDUSTRY (Greenpeace)

Outline **concisely** the **key points** Greenpeace makes against commercial whaling.

Use **your own words** as far as possible.

[12]

2 GREED IN THE GLOBAL WHALING INDUSTRY (Greenpeace)

How does Greenpeace try to make the case against commercial whaling convincing?

In your answer you should comment on the effectiveness of the **presentation** and the use of **information** and **language** in the text.

(Presentation may include reference to headings and pictures and the way the article is structured.)

[14]

3 ON WHALING (Stephen Fry in America)

How does Stephen Fry present his thoughts about whale hunting and traditional whale hunting communities in these extracts?

In your answer you should refer to the **language** used and the **tone** created.

[14]

SECTION B – Writing

You are advised to spend about 50 minutes on Section B.

Answer EITHER question 4 OR question 5.

This answer will be marked for writing. Plan your answer and write it carefully. Leave enough time to check through what you have written.

Either

4 ‘And it made me change my mind...’

Write an entry for either a personal diary or a blog giving an account of an experience which made you change the way you thought about something – perhaps a visit to a place or a meeting with a person. **[40]**

Or

5 ‘We must learn from the past as we move towards the future.’

Write your views on this statement. **[40]**

Appendix B: The 'original' mark scheme



H

Tuesday 3 June 2014 - Morning

GCSE ENGLISH/ENGLISH LANGUAGE

A680/02 Information and Ideas (Higher Tier)

MARK SCHEME

Duration: 2 hours

MAXIMUM MARK 80

Post-Standardisation Version

(FOR OFFICE USE ONLY)

This document consists of 16 pages

MARKING INSTRUCTIONS – FOR MARKING ON-SCREEN AND FOR PAPER BASED MARKING

1. Mark strictly to the mark scheme.
2. Marks awarded must relate directly to the marking criteria.
3. Crossed Out Responses and Rubric Error (Incorrect texts)

Crossed Out Responses

Where a candidate has crossed out a response and provided a clear alternative then the crossed out response is not marked. Where no alternative response has been provided, examiners may give candidates the benefit of the doubt and mark the crossed out response where legible.

Rubric Error Responses – Incorrect texts

Candidates are expected to answer text based questions on the text specified in the question. Use of another text cannot be given credit and will score 0.

4. Always check the additional pages (and additional objects if present) at the end of the response in case any answers have been continued there. If the candidate has continued an answer there then add a tick to confirm that the work has been seen.
5. There is a NR (No Response) option. Award NR (No Response)
 - a. if there is nothing written at all in the answer space
 - b. OR if there is a comment which does not in any way relate to the question (e.g. 'can't do', 'don't know')
 - c. OR if there is a mark (e.g. a dash, a question mark) which isn't an attempt at the question













Note: Award 0 marks - for an attempt that earns no credit (including copying out the question)

6. For answers marked by levels of response:

- a. **To determine the level** – start at the highest level and work down until you reach the level that matches the answer
- b. **To determine the mark within the level**, consider the following:

Descriptor	Award mark
On the borderline of this level and the one below	At bottom of level
Just enough achievement on balance for this level	Above bottom and either below middle or at middle of level (depending on number of marks available)
Meets the criteria but with some slight inconsistency	Above middle and either below top of level or at middle of level (depending on number of marks available)
Consistently meets the criteria for this level	At top of level

7. These are the annotations, (including abbreviations), including those used in scoris, which are used when marking:

Annotation	Meaning
	Blank Page – this annotation must be used on all blank pages within an answer booklet (structured or unstructured) and on each page of an additional object where there is no candidate response.
	Unclear
	Error
	Misreading
	Not Answering Question
	No Example
	Extensive Error
	Repetition
	Tick
	Strong point/Apt Ref
	Blurred point
	Omission

8. Subject-specific Marking Instructions

Marking and Annotation of Scripts After the Standardisation Meeting

All scripts must be marked in accordance with the version of the mark scheme agreed at the Standardisation meeting.

Recording of marks

- Show evidence that you have seen the work on every page of a script on which the candidate has made a response
- Cross through every blank page to show that it has been seen
- Follow the current guidance on crossed-out work.

Handling of unexpected answers

The Standardisation meeting will include discussion of marking issues, including:








- consideration of the mark scheme to reach a decision about the range of acceptable responses and the marks appropriate to them
- the handling of unexpected, yet acceptable, answers.

MARK SCHEME: SECTION A READING

Question 2 Greed in the Global Whaling Industry (Greenpeace)

INSTRUCTIONS TO EXAMINERS – 2

1. We are not marking writing in Section A unless the expression is so bad that it impeded communication.
- 2 Use the Band descriptors in conjunction with the standardisation scripts to arrive at your mark.
- 3 Indicate the band and mark with a brief comment, taken from the band descriptors, if appropriate.

Notes on the Task		Marks	Guidance
2	<p>General: Candidates may comment on the ways in which the article delivers factual information while seeking to convey a sense of threat to a ‘beautiful and intelligent’ mammal from human greed and dishonesty. They may offer comment on the image – the visual and emotional impact of a free swimming live whale as the background to the Greenpeace mission statement. They may explore the structure, which moves from ‘catastrophe’ to ‘potential hope’.</p> <p>Candidates may explore the effect of both stating and correcting the ‘myths’ and possibly consider the emotional impact around the attack on ‘half-truths and outright lies’. They may comment on the addition of personal testimony from a man closely involved in whaling who later became a Greenpeace supporter and explore what the article gains from this.</p> <p>Candidates may consider how the selection of facts and statistics demonstrate both the scale of the threat to the whale and, in the closing section, offers reinforcement to the idea that it makes economic sense to preserve the species. Information about species extinction may be seen as emotive, as might mention of toxic blubber, brutal slaughter and ‘Blood money’.</p> <p>Some more detailed comment on specific language devices may be offered, most obviously in the headings and subheadings but also throughout the text, as in direct address to the audience (‘you might want to think again’), in the use of inverted commas around ‘protection’ and ‘tradition’ and the use of alliteration and triplet (‘consumption, contamination and catastrophe’). Although ‘tone’ is not specifically requested here, candidates may comment on the rather assertive tone conveyed via short, ‘punchy’ paragraphs and clipped, very direct sentences.</p>	14	<p> in the body of the answer indicates clear relevant points.</p> <p> in the margin indicates use of supporting reference.</p> <p> indicates points not fully expressed.</p> <p> indicates strong thoughtful comment</p> <p> may be used to indicate lack of supporting examples</p> <p> indicates misunderstanding.</p> <p> indicates irrelevant comment.</p>

Notes on the Task		Marks	Guidance
	<p>Higher Band (1+2) Responses may offer insightful comment on the manipulation of the reader's perceptions. They will make consistently analytical and more developed comments on the language used, supported by fully appropriate references. Comments about presentation will show good overview and understanding of the way the article is structured and how the images reinforce the text. Candidates may also refer to the way the writer's opinion is implied through the use of direct quotations.</p> <p>Middle Band (3+4) responses are likely to show some appreciation of the ways in which the passage informs and persuades. There may be some consideration of how the article seeks to engage the reader's emotions. There is likely to be some comment on how the information is presented and some comment on how the visual images contribute to this. There may be some attempt to explain language effects, but it is unlikely to be sustained, and not always firmly linked to the writer's purpose.</p> <p>Lower Band (5+6) responses are likely to show only a rudimentary understanding of the task and will make general, mainly unsupported comments about the writer's use of language, possibly achieving little more than the naming of a device. There may be some misunderstanding of the text and responses at this level will probably consist mainly of paraphrase/summary of the content and description of the images.</p>		

Question 2**GENERIC band descriptors******Be prepared to use the FULL range*****The band descriptors which are shaded reward performance below that expected on this paper.*

BAND	MARKS	DESCRIPTOR
1	14 13	<ul style="list-style-type: none"> • Excellent range of points showing perceptive appreciation of the ways in which information, language and structure convey the text's purpose • Very effective use of apposite supporting references in a full, relevant and consistently analytical response • Complete understanding of text and task
2	12 11	<ul style="list-style-type: none"> • Wide range of points showing clear and thoughtful appreciation of the ways in which information, language and structure convey the text's purpose • Judgments are supported convincingly by appropriate textual references • Clear understanding of text and task
3	10 9 8	<ul style="list-style-type: none"> • A good range of points showing a secure understanding of the ways in which information, language and structure contribute to the text's purpose • Careful supporting references and some analytical comment • Sound awareness of text and task
4	7 6 5	<ul style="list-style-type: none"> • A range of points showing a sound understanding of the ways in which information, language and structure contribute to the text's purpose • Appropriate supporting references and an attempt at an analytical approach • Task has been addressed for the main part
5	4 3 2	<ul style="list-style-type: none"> • Easier information points show some understanding of the text's purpose • Comments tend to be descriptive rather than analytical, and references may be inert • Some focus on the task
Below 5	1 0	<ul style="list-style-type: none"> • Points likely to concentrate on simpler information and basic language features • Assertions predominate, with minimal or no textual evidence in support • A little evidence that the task has been understood






SECTION B: WRITING

CRITERIA

Candidates should demonstrate that they can:

- Write to communicate clearly, effectively and imaginatively, using and adapting forms and selecting vocabulary appropriate to task and purpose in ways that engage the reader (AO3i)
- Organise information and ideas into structured and sequenced sentences, paragraphs and whole texts, using a variety of linguistic and structural features to support cohesion and over coherence (AO3 ii)
- Use a range of sentence structures for clarity, purpose and effect, with accurate punctuation and spelling (AO3 iii).

INSTRUCTIONS TO EXAMINERS – 4

1. Use  for good ideas and  for merits of expression to show how you have formed your judgement. Use a wavy line  underneath the candidate's writing, or in the margin  to show awkward or incorrect syntax/unclear expression. Circle any errors of spelling or punctuation. Use a caret  to show omission.
2. You should write a brief summative comment drawn from the wording of the descriptors to show how you have arrived at your final marks.
3. For writing tasks, LENGTH is not in itself a criterion.

Short answers (50-100 words) may well be self-penalising in terms of the marking criteria (eg control and development of ideas; structure; maintaining the reader's interest), but may still demonstrate significant qualities. Very short answers (fewer than 50 words) should not normally be marked higher than Band 7.
4. Award TWO separate marks, one for AOs 3(i) + (ii), one for AO3 (iii), using the appropriate instructions and Band Descriptors. Be prepared to use the full range of marks in each sub-set.

5. Use the standardisation scripts as guides to your assessment.
6. The generic marking criteria for Writing appear after the Notes on the Task.

Question		Notes on the Task	Marks	Guidance
4		<p>Candidates have been asked to write with form and tone suitable for a diary or blog. Expect a wide range of responses and various interpretations of the diary/blog format. Both are forms of informal, personal writing but with diary having connotations of more private writing, where the intended audience is the writer alone and blog implying a wider, online audience. The response should show some awareness of audience, appropriate to the candidate's chosen form.</p> <p>Look to reward those responses that are well crafted and clearly focused on the task. There should be some intention to use language to create effects.</p> <p>Please note that there is a free choice of topic. There is no expectation that candidates selecting this question will continue the themes of the texts.</p>	40	

Generic Marking Criteria for Section B: Writing

Band	Marks	Descriptors AO3i & AO3ii	Marks	Descriptors AO3iii
1	26 25 24	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • shows sophisticated control of the material and makes effective use of linguistic devices. • demonstrates a sophisticated understanding of the task, addressing it with complete relevance and adapting form and style with flair to suit audience and purpose. • uses precise vocabulary which is fully suited to the purpose of the writing, conveying subtlety of thought and shades of meaning, and where appropriate is imaginative and ambitious in scope. • uses structure to produce deliberate effects, developing the writing coherently and skilfully from a confident opening which engages the reader to a very convincing and deliberate ending. • is organised into coherent paragraphs which are clearly varied for effect and used confidently to enhance the ideas and meaning. 	14	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • uses a wide range of sentence structures to ensure clarity and to achieve specific effects relevant to the task. • uses ambitious vocabulary with very few spelling errors. • uses punctuation consciously and securely to shape meaning, with very few errors.
2	23 22 21	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • shows full control of the material and makes some effective use of linguistic devices. • demonstrates a confident understanding of the task, addressing it with consistent relevance and adapting form and style with assurance to suit audience and purpose. • uses imaginative vocabulary which is appropriate to the purpose of the writing, conveying some subtlety of thought and shades of meaning, and where appropriate may show some ambition in scope • uses structure consciously for effect, developing the writing 	13 12	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • uses a range of sentence structures to ensure clarity and to achieve specific effects relevant to the task. • uses more complex and irregular vocabulary, almost always securely spelled • uses punctuation to shape meaning, mainly securely, with errors only in more complex, irregular structures.

		<p>coherently from an opening which engages the reader to a convincing and deliberate ending.</p> <ul style="list-style-type: none"> is organised into paragraphs which have unity, are varied for effect and are used to control the content and achieve overall coherence. 		
Band	Marks	Descriptors AO3i & AO3ii	Marks	Descriptors AO3iii
3	20 19 18	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> shows generally competent control of the material but may not always convey meaning clearly when using more ambitious linguistic devices and structures. demonstrates a secure understanding of the task, addressing it in a relevant way and adapting form and style with confidence to suit audience and purpose. uses varied vocabulary to create different effects which are mainly appropriate to the purpose of the writing, conveying thought and meaning clearly. uses structure deliberately and with direction - a focused and interesting opening, events and ideas developed clearly and in some detail, an appropriate ending. uses paragraphs of varying length and structure for effect, which effectively organise and link ideas and create an overall sense of coherence. 	11 10	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> uses generally well controlled sentence structures which are varied in length and type and show evidence of being used deliberately to create specific effects appropriate to the task. shows secure spelling of complex regular words and generally secure spelling of irregular or more complex vocabulary. uses punctuation to enhance or clarify meaning - is accurate both within and between sentences, but may make some errors in complex sentence structures.
4	17 16 15	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> shows general control of the material; the response may be straightforward and controlled but linguistically unambitious or may lose some control in attempting something ambitious. demonstrates an understanding of the task, addressing it in a mainly relevant way with some evidence of adapting form and style to suit different audiences and purposes. uses some variety of vocabulary to create different effects and to suit the purpose of the writing, but which may be imprecise or fail to convey shades of meaning. uses structure with a sense of direction - a clear and focused 	9 8	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> uses sentence structures which show some variety. May tend to repeat sentence types, lose control of more ambitious structures, or make some syntactical errors. usually spells complex regular words securely; may make errors with irregular or more complex vocabulary. uses punctuation in an attempt to create some specific effects; is usually accurate for sentence separation and sometimes within sentences, but may make less secure use of

		<p>opening, straightforward development of ideas, an attempt to achieve an appropriate ending.</p> <ul style="list-style-type: none">• is organised into paragraphs which may be varied for effect and which are carefully linked together to make the sequence of events or development of ideas clear to the reader.		<p>speech marks, colons and semi colons.</p>
--	--	--	--	--

Band	Marks	Descriptors AO3i & AO3ii	Marks	Descriptors AO3iii
5	14 13 12	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • may not always show control of the material; the response may be simple and controlled but linguistically unambitious, or may attempt something ambitious but tend to lose control. • demonstrates some understanding of the task, addressing it in a sometimes relevant way and with some attempt to adapt form and style to suit audience and purpose. • uses vocabulary to create some limited effects, which may however be too simple to convey shades of meaning, not fully understood or not appropriate and may contain some idiomatic errors. • uses structure with some sense of direction - a generally clear and focussed opening, some development of ideas, a limited attempt to achieve an appropriate ending. • uses paragraphs which may occasionally be varied for effect and/or are linked together to make the sequence of events or development of ideas fairly clear to the reader. 	7 6	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • uses sentence structures which show a little variety; may tend repeat sentence types, lose control of more ambitious structures, and/or include syntactical errors. • usually spells simple regular vocabulary securely but may make errors with complex regular vocabulary. • uses punctuation which sometimes helps clarify meaning, usually accurately for sentence separation and sometimes successfully within sentences

Band	Marks	Descriptors AO3i & AO3ii	Marks	Descriptors AO3iii
6	11 10 9	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • does not always show control of the material; the response may have a level of linguistic error that distracts the reader from the merits of the content. • demonstrates a limited understanding of the task and addresses it with some relevance, making a limited attempt to adapt form and style to suit audience and purpose. • uses vocabulary which is sometimes chosen for variety and interest but likely to be limited in range, sometimes inappropriate and may contain some idiomatic errors. • structures writing with some sense of direction which may not be sustained; a fairly clear opening, some limited development of ideas, some sense of an ending. • uses paragraphs which create some sense of sequence for the events or the development of ideas but which may lack unity or have little or no evidence of links between them. 	5 4	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • uses repetitive sentence structures, which are mainly simple or compound, or lengthy with some sense of control. • usually spells simple regular vocabulary accurately but may make a number of typical errors. • sometimes uses punctuation accurately for sentence separation but has limited success with attempts to use it within sentences to clarify meaning.

Band	Marks	Descriptors AO3i & AO3ii	Marks	Descriptors AO3iii
7	8 7 6	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • shows limited control of the material; the level of linguistic error may require the reader to re-read some sentences before the meaning is clear. • demonstrates a very limited understanding of the task, addressing it with occasional focus and making limited attempts to adapt form and style to suit audience and purpose. • uses vocabulary to create occasional variety and interest but which is likely to be very limited in range and often inappropriate with some idiomatic errors. • shows some signs of organisation and some sense of direction - a limited attempt to create an opening, very simple or rambling development of ideas, may come to a stop rather than achieving a deliberate ending. • uses paragraphs which may signal only obvious development of events or ideas, or which may be haphazard and lack clear links or overall unity. 	3 2	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • uses simple repetitive sentence structures with little control of more complex ones and frequent syntactical faults. • spells some simple regular vocabulary accurately but makes random errors. • uses some basic punctuation with some success between sentences but, within sentences, usually misuses or omits it.

Band	Marks	Descriptors AO3i & AO3ii	Marks	Descriptors AO3iii
8	5 4 3	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • offers occasional relevant and comprehensible content, but density of linguistic error may require the reader to re-read and re-organise the text before meaning is clear. • demonstrates a little awareness of the task, addressing it with intermittent focus; form and style may occasionally be appropriate to audience and/or purpose, but this is unlikely to be deliberate. • uses vocabulary which is very occasionally chosen for variety and/or interest but which is very limited in range and often inappropriate, with obvious idiomatic errors. • shows occasional signs of organisation and a very limited - if any - sense of direction. • uses paragraphs occasionally to signal very obvious changes in the direction of events or ideas, but which may need to be re-read or re-organised before the meaning is clear. • 	1	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • uses recognisable sentence structures, with some accuracy in the use of more simple ones. • uses erratic spelling which may be recognisable for most words but is accurate for only a limited number. • uses punctuation which is occasionally successful but is inconsistent and likely to be inaccurate.
Below band 8	2 1 0	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • is very short or communicates very little, with some sections making no sense at all; may gain some marks where there is occasional clarity. • shows almost no awareness of task, audience or purpose. • uses vocabulary which is seriously limited. • shows almost no signs of organisation or sense of direction. • uses paragraphs -if at all - in a haphazard way such that, in spite of re-reading and re-organising, very little sense emerges. 	0	<p><i>In this band a candidate's writing:</i></p> <ul style="list-style-type: none"> • uses spelling and punctuation so imprecisely that very little meaning is communicated.

Appendix C: The 'experimental' mark scheme



H

Tuesday 3 June 2014 - Morning

GCSE ENGLISH/ENGLISH LANGUAGE

A680/02 Information and Ideas (Higher Tier)

MARK SCHEME

Duration: 2 hours

MAXIMUM MARK 80

Post-Standardisation Version

(FOR OFFICE USE ONLY)

This document consists of 10 pages

MARKING INSTRUCTIONS – FOR MARKING ON-SCREEN AND FOR PAPER BASED MARKING

9. Mark strictly to the mark scheme.
10. Marks awarded must relate directly to the marking criteria.
11. Crossed Out Responses and Rubric Error (Incorrect texts)

Crossed Out Responses

Where a candidate has crossed out a response and provided a clear alternative then the crossed out response is not marked. Where no alternative response has been provided, examiners may give candidates the benefit of the doubt and mark the crossed out response where legible.

Rubric Error Responses – Incorrect texts

Candidates are expected to answer text based questions on the text specified in the question. Use of another text cannot be given credit and will score 0.

12. Always check the additional pages (and additional objects if present) at the end of the response in case any answers have been continued there. If the candidate has continued an answer there then add a tick to confirm that the work has been seen.
13. There is a NR (No Response) option. Award NR (No Response)
 - a. if there is nothing written at all in the answer space
 - b. OR if there is a comment which does not in any way relate to the question (e.g. 'can't do', 'don't know')
 - c. OR if there is a mark (e.g. a dash, a question mark) which isn't an attempt at the question













Note: Award 0 marks - for an attempt that earns no credit (including copying out the question)

14. For answers marked by levels of response:

- a. **To determine the level** – start at the highest level and work down until you reach the level that matches the answer
- b. **To determine the mark within the level**, consider the following:

Descriptor	Award mark
On the borderline of this level and the one below	At bottom of level
Just enough achievement on balance for this level	Above bottom and either below middle or at middle of level (depending on number of marks available)
Meets the criteria but with some slight inconsistency	Above middle and either below top of level or at middle of level (depending on number of marks available)
Consistently meets the criteria for this level	At top of level

15. These are the annotations, (including abbreviations), including those used in scoris, which are used when marking:

Annotation	Meaning
	Blank Page – this annotation must be used on all blank pages within an answer booklet (structured or unstructured) and on each page of an additional object where there is no candidate response.
	Unclear
	Error
	Misreading
	Not Answering Question
	No Example
	Extensive Error
	Repetition
	Tick
	Strong point/Apt Ref
	Blurred point
	Omission

16. Subject-specific Marking Instructions

Marking and Annotation of Scripts After the Standardisation Meeting

All scripts must be marked in accordance with the version of the mark scheme agreed at the Standardisation meeting.

Recording of marks

- Show evidence that you have seen the work on every page of a script on which the candidate has made a response
- Cross through every blank page to show that it has been seen.
- Follow the current guidance on crossed-out work.

Handling of unexpected answers

The Standardisation meeting will include discussion of marking issues, including:

- consideration of the mark scheme to reach a decision about the range of acceptable responses and the marks appropriate to them
- the handling of unexpected, yet acceptable, answers.

MARK SCHEME: SECTION A READING

Question 2: GREED IN THE GLOBAL WHALING INDUSTRY (Greenpeace)

- **How does Greenpeace try to make the case against commercial whaling convincing?** *In your answer you should comment on the effectiveness of the presentation and the use of information and language in the text. (Presentation may include reference to headings and pictures and the way the article is structured.) (14 marks)*

CRITERIA

Candidates should demonstrate that they can:

- Explain and evaluate how writers use presentational features to achieve effects and engage and influence the reader (AO2 iii).

INSTRUCTIONS TO EXAMINERS

1. We are not marking writing in Section A unless the expression is so bad that it impeded communication.
2. Use the Band descriptors in conjunction with the standardisation scripts to arrive at your mark.
3. Indicate the band and mark with a brief comment, taken from the band descriptors, if appropriate.

GUIDANCE

Candidates may comment on the ways in which the article delivers factual information while seeking to convey a sense of threat to a 'beautiful and intelligent' mammal from human greed and dishonesty. They may offer comment on the image – the visual and emotional impact of a free swimming live whale as the background to the Greenpeace mission statement. They may explore the structure, which moves from 'catastrophe' to 'potential hope'.

Candidates may explore the effect of both stating and correcting the 'myths' and possibly consider the emotional impact around the attack on 'half-truths and outright lies'. They may comment on the addition of personal testimony from a man closely involved in whaling who later became a Greenpeace supporter and explore what the article gains from this.

Candidates may consider how the selection of facts and statistics demonstrate both the scale of the threat to the whale and, in the closing section, offers reinforcement to the idea that it makes economic sense to preserve the species. Information about species extinction may be seen as emotive, as might mention of toxic blubber, brutal slaughter and 'Blood money'.

Some more detailed comment on specific language devices may be offered, most obviously in the headings and subheadings but also throughout the text, as in direct address to the audience ('you might want to think again'), in the use of inverted commas around 'protection' and 'tradition' and the use of alliteration and triplet ('consumption, contamination and catastrophe'). Although 'tone' is not specifically requested

here, candidates may comment on the rather assertive tone conveyed via short, 'punchy' paragraphs and clipped, very direct sentences.

Question 2 GENERIC band descriptors **Be prepared to use the FULL range**			
<i>The band descriptors which are shaded reward performance below that expected on this paper.</i>			
BAND	MARKS	DESCRIPTOR	NOTES ON TASK
1	14 13	<ul style="list-style-type: none"> Excellent range of points showing perceptive appreciation of the ways in which information, language and structure convey the text's purpose Very effective use of apposite supporting references in a full, relevant and consistently analytical response Complete understanding of text and task 	<p>Higher Band (1+2) Responses may offer insightful comment on the manipulation of the reader's perceptions. They will make consistently analytical and more developed comments on the language used, supported by fully appropriate references. Comments about presentation will show good overview and understanding of the way the article is structured and how the images reinforce the text. Candidates may also refer to the way the writer's opinion is implied through the use of direct quotations.</p>
2	12 11	<ul style="list-style-type: none"> Wide range of points showing clear and thoughtful appreciation of the ways in which information, language and structure convey the text's purpose Judgments are supported convincingly by appropriate textual references Clear understanding of text and task 	
3	10 9 8	<ul style="list-style-type: none"> A good range of points showing a secure understanding of the ways in which information, language and structure contribute to the text's purpose Careful supporting references and some analytical comment Sound awareness of text and task 	<p>Middle Band (3+4) responses are likely to show some appreciation of the ways in which the passage informs and persuades. There may be some consideration of how the article seeks to engage the reader's emotions. There is likely to be some comment on how the information is presented and some comment on how the visual images contribute to this. There may be some attempt to explain language effects, but it is unlikely to be sustained, and not always firmly linked to the writer's purpose.</p>
4	7 6 5	<ul style="list-style-type: none"> A range of points showing a sound understanding of the ways in which information, language and structure contribute to the text's purpose Appropriate supporting references and an attempt at an analytical approach Task has been addressed for the main part 	
5	4 3 2	<ul style="list-style-type: none"> Easier information points show some understanding of the text's purpose Comments tend to be descriptive rather than analytical, and references may be inert Some focus on the task 	<p>Lower Band (5+6) responses are likely to show only a rudimentary understanding of the task and will make general, mainly unsupported comments about the writer's use of language, possibly achieving little more than the naming of a device. There may be some misunderstanding of the text and responses at this level will probably consist mainly of paraphrase/summary of the content and description of the images.</p>
Below 5	1 0	<ul style="list-style-type: none"> Points likely to concentrate on simpler information and basic language features Assertions predominate, with minimal or no textual evidence in support A little evidence that the task has been understood 	

SECTION B: WRITING

Question 4:

‘And it made me change my mind...’

Write an entry for either a personal diary or a blog giving an account of an experience which made you change the way you thought about something – perhaps a visit to a place or a meeting with a person. **(40 marks)**

CRITERIA

Candidates should demonstrate that they can:

- Write to communicate clearly, effectively and imaginatively, using and adapting forms and selecting vocabulary appropriate to task and purpose in ways that engage the reader (AO3i)
- Organise information and ideas into structured and sequenced sentences, paragraphs and whole texts, using a variety of linguistic and structural features to support cohesion and overall coherence (AO3 ii)
- Use a range of sentence structures for clarity, purpose and effect, with accurate punctuation and spelling (AO3 iii).

INSTRUCTIONS TO EXAMINERS – 4

1. You should write a brief summative comment drawn from the wording of the descriptors to show how you have arrived at your final marks.
2. For writing tasks, LENGTH is not in itself a criterion.
3. Short answers (50-100 words) may well be self-penalising in terms of the marking criteria (e.g. control and development of ideas; structure; maintaining the reader’s interest), but may still demonstrate significant qualities. Very short answers (fewer than 50 words) should not normally be marked higher than Band 7.
4. Award TWO separate marks, one for AOs 3(i) + (ii), one for AO3 (iii), using the appropriate instructions and Band Descriptors. Be prepared to use the full range of marks in each sub-set.
5. Use the standardisation scripts as guides to your assessment.
6. The generic marking criteria for Writing appear after the following Guidance.

GUIDANCE

Candidates have been asked to write with form and tone suitable for a diary or blog. Expect a wide range of responses and various interpretations of the diary/blog format. Both are forms of informal, personal writing but with diary having connotations of more private writing, where the intended audience is the writer alone and blog implying a wider, online audience. The response should show some awareness of audience, appropriate to the candidate's chosen form.

Look to reward those responses that are well crafted and clearly focused on the task. There should be some intention to use language to create effects.

Please note that there is a free choice of topic. There is no expectation that candidates selecting this question will continue the themes of the texts.

Question 4 GENERIC band descriptors. **Be prepared to use the FULL range.**				
BANDS	MARKS	Descriptors AO3i & AO3ii	MARKS	Descriptor AO3iii
1	26 25 24	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> shows sophisticated control of the material and makes effective use of linguistic devices. demonstrates a sophisticated understanding of the task, addressing it with complete relevance and adapting form and style with flair to suit audience and purpose. uses precise vocabulary which is fully suited to the purpose of the writing, conveying subtlety of thought and shades of meaning, and where appropriate is imaginative and ambitious in scope. uses structure to produce deliberate effects, developing the writing coherently and skillfully from a confident opening which engages the reader to a very convincing and deliberate ending. is organised into coherent paragraphs which are clearly varied for effect and used confidently to enhance the ideas and meaning. 	14	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses a wide range of sentence structures to ensure clarity and to achieve specific effects relevant to the task. uses ambitious vocabulary with very few spelling errors. uses punctuation consciously and securely to shape meaning, with very few errors.
2	23 22 21	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> shows full control of the material and makes some effective use of linguistic devices. demonstrates a confident understanding of the task, addressing it with consistent relevance and adapting form and style with assurance to suit audience and purpose. uses imaginative vocabulary which is appropriate to the purpose of the writing, conveying some subtlety of thought and shades of meaning, and where appropriate may show some ambition in scope. uses structure consciously for effect, developing the writing coherently from an opening which engages the reader to a convincing and deliberate ending. is organised into paragraphs which have unity, are varied for effect and are used to control the content and achieve overall coherence. 	13 12	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses a range of sentence structures to ensure clarity and to achieve specific effects relevant to the task. uses more complex and irregular vocabulary, almost always securely spelled. uses punctuation to shape meaning, mainly securely, with errors only in more complex, irregular structures.
3	20 19 18	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> shows generally competent control of the material but may not always convey meaning clearly when using more ambitious linguistic devices and structures. demonstrates a secure understanding of the task, addressing it in a relevant way and adapting form and style with confidence to suit audience and purpose. uses varied vocabulary to create different effects which are mainly appropriate to the purpose of the writing, conveying thought and meaning clearly. uses structure deliberately and with direction - a focused and interesting opening, events and ideas developed clearly and in some detail, an appropriate ending. uses paragraphs of varying length and structure for effect, which effectively organise and link ideas and create an overall sense of coherence. 	11 10	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses generally well controlled sentence structures which are varied in length and type and show evidence of being used deliberately to create specific effects appropriate to the task. shows secure spelling of complex regular words and generally secure spelling of irregular or more complex vocabulary. uses punctuation to enhance or clarify meaning - is accurate both within and between sentences, but may make some errors in complex sentence structures.
4	17 16 15	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> shows general control of the material; the response may be straightforward and controlled but linguistically unambitious or may lose some control in attempting something ambitious. demonstrates an understanding of the task, addressing it in a mainly relevant way with some evidence of adapting form and style to suit different audiences and purposes. uses some variety of vocabulary to create different effects and to suit the purpose of the writing, but which may be imprecise or fail to convey shades of meaning. uses structure with a sense of direction - a clear and focused opening, straightforward development of ideas, an attempt to achieve an appropriate ending. is organised into paragraphs which may be varied for effect and which are carefully linked together to make the sequence of events or development of ideas clear to the reader. 	9 8	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses sentence structures which show some variety. May tend to repeat sentence types, lose control of more ambitious structures, or make some syntactical errors. usually spells complex regular words securely; may make errors with irregular or more complex vocabulary. uses punctuation in an attempt to create some specific effects; is usually accurate for sentence separation and sometimes within sentences, but may make less secure use of speech marks, colons and semi colons.
5	14 13 12	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> may not always show control of the material; the response may be simple and controlled but linguistically unambitious, or may attempt something ambitious but tend to lose control. demonstrates some understanding of the task, addressing it in a sometimes relevant way and with some attempt to adapt form and style to suit audience and purpose. uses vocabulary to create some limited effects, which may however be too simple to convey shades of meaning, not fully understood or not appropriate and may contain some idiomatic errors. uses structure with some sense of direction - a generally clear and focused opening, some development of ideas, a limited attempt to achieve an appropriate ending. uses paragraphs which may occasionally be varied for effect and/or are linked together to make the sequence of events or development of ideas fairly clear to the reader. 	7 6	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses sentence structures which show a little variety; may tend to repeat sentence types, lose control of more ambitious structures, and/or include syntactical errors. usually spells simple regular vocabulary securely but may make errors with complex regular vocabulary. uses punctuation which sometimes helps clarify meaning, usually accurately for sentence separation and sometimes successfully within sentences.
6	11 10 9	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> does not always show control of the material; the response may have a level of linguistic error that distracts the reader from the merits of the content. demonstrates a limited understanding of the task and addresses it with some relevance, making a limited attempt to adapt form and style to suit audience and purpose. uses vocabulary which is sometimes chosen for variety and interest but likely to be limited in range, sometimes inappropriate and may contain some idiomatic errors. is structured with some sense of direction which may not be sustained; a fairly clear opening, some limited development of ideas, some sense of an ending. uses paragraphs which create some sense of sequence for the events or the development of ideas but which may lack unity or have little or no evidence of links between them. 	5 4	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses repetitive sentence structures, which are mainly simple or compound, or lengthy with some sense of control. usually spells simple regular vocabulary accurately but may make a number of typical errors. sometimes uses punctuation accurately for sentence separation but has limited success with attempts to use it within sentences to clarify meaning.
7	8 7 6	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> shows limited control of the material; the level of linguistic error may require the reader to re-read some sentences before the meaning is clear. demonstrates a very limited understanding of the task, addressing it with occasional focus and making limited attempts to adapt form and style to suit audience and purpose. uses vocabulary to create occasional variety and interest but which is likely to be very limited in range and often inappropriate with some idiomatic errors. shows some signs of organisation and some sense of direction - a limited attempt to create an opening, very simple or rambling development of ideas, may come to a stop rather than achieving a deliberate ending. uses paragraphs which may signal only obvious development of events or ideas, or which may be haphazard and lack clear links or overall unity. 	3 2	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses simple repetitive sentence structures with little control of more complex ones and frequent syntactical faults. spells some simple regular vocabulary accurately but makes random errors. uses some basic punctuation with some success between sentences but, within sentences, usually misuses or omits it.
8	5 4 3	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> offers occasional relevant and comprehensible content, but density of linguistic error may require the reader to re-read and reorganise the text before meaning is clear. demonstrates a little awareness of the task, addressing it with intermittent focus; form and style may occasionally be appropriate to audience and/or purpose, but this is unlikely to be deliberate. uses vocabulary which is very occasionally chosen for variety and/or interest but which is very limited in range and often inappropriate, with obvious idiomatic errors. shows occasional signs of organisation and a very limited - if any - sense of direction. uses paragraphs occasionally to signal very obvious changes in the direction of events or ideas, but which may need to be re-read or re-organised before the meaning is clear. 	1	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses recognisable sentence structures, with some accuracy in the use of more simple ones. uses erratic spelling which may be recognisable for most words but is accurate for only a limited number. uses punctuation which is occasionally successful but is inconsistent and likely to be inaccurate.
Below band 8	2 1 0	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> is very short or communicates very little, with some sections making no sense at all; may gain some marks where there is occasional clarity. shows almost no awareness of task, audience or purpose. uses vocabulary which is seriously limited. shows almost no signs of organisation or sense of direction. uses paragraphs - if at all - in a haphazard way such that, in spite of re-reading and re-organising, very little sense emerges. 	0	<i>In this band a candidate's writing:</i> <ul style="list-style-type: none"> uses spelling and punctuation so imprecisely that very little meaning is communicated.

Appendix D: Post-marking questionnaire

Investigating the effects of features of mark schemes on marking reliability – questionnaire

This brief questionnaire is to gather some of your views on the mark schemes you used to mark the scripts. This will provide us with further information we can use to help improve the future design of mark schemes. The first set of questions relates to the mark scheme you used for question 2. The second question set relates to question 4, and the third set relates to the *entire* mark scheme document you were provided (including general marking and annotation instructions).

Please fill out this questionnaire once you have completed marking the allocated scripts. You can send the questionnaire back to us in the same pack as the scripts.

Your details

Name:

Current occupation:

Question set A: Mark scheme for question 2

A1. Please rate the mark scheme in terms of its *ease of use* (circle one box).

1

2

3

4

5

Very difficult
to use.

Very easy to
use

A2. Please rate the mark scheme in terms of its *clarity* (circle one box).

1

2

3

4

5

Not very
clear

Very clear

A3. Please rate the mark scheme in terms of its *layout* (circle one box).

1

2

3

4

5

Very poor
layout

Very good
layout

A4. Please rate the mark scheme in terms of the *amount of detail* (tick one response).

There was *about the right amount* of detail in the mark scheme.....

There was *too much detail* in the mark scheme.....

There was *too little detail* in the mark scheme.....

Please explain your choice in the space below:

A5. Were there any features of this mark scheme that you felt helped you mark *more effectively*?

Yes.....

No.....

If you responded 'yes' please explain your choice in the space below:

A6. Were there any features of this mark scheme that you felt made you mark *less effectively*?

Yes.....

No.....

If you responded 'yes' please explain your choice in the space below:

Question set B: Mark scheme for question 4

B1. Please rate the mark scheme in terms of its *ease of use* (circle one box).

1

2

3

4

5

Very difficult
to use.

Very easy to
use

B2. Please rate the mark scheme in terms of its *clarity* (circle one box).

1

2

3

4

5

Not very
clear

Very clear

B3. Please rate the mark scheme in terms of its *layout* (circle one box).

1

2

3

4

5

Very poor
layout

Very good
layout

B4. Please rate the mark scheme in terms of the *amount of detail* (tick one response).

There was *about the right amount* of detail in the mark scheme.....

There was *too much detail* in the mark scheme.....

There was *too little detail* in the mark scheme.....

Please explain your choice in the space below:

B5. Were there any features of this mark scheme that you felt helped you mark *more* effectively?

Yes.....

No.....

If you responded 'yes' please explain your choice in the space below:

B6. Were there any features of this mark scheme that you felt made you mark *less* effectively?

Yes.....

No.....

If you responded 'yes' please explain your choice in the space below:

Question set C: The mark scheme overall

C1. Typically, how long did it take you to mark *one script* (questions 2 and 4)? Please tick one box.

0-2 minutes.....

3-5 minutes.....

6-10 minutes.....

11-20 minutes.....

Over 20 minutes.....

C2. Typically, for how long did you mark scripts for *in one sitting*? Please tick one box.

0-15 minutes.....

16-30 minutes.....

31-45 minutes.....

46-60 minutes.....

Over 60 minutes.....

C3. Would you make any changes to the *content* of the mark scheme?

Yes.....

No.....

If you responded 'yes' please explain your choice in the space below:

C4. Would you make any changes to the *layout* of the mark scheme?

Yes.....

No.....

If you responded 'yes' please explain your choice in the space below:

C5. Please use the space below to write down any further comments you have.

End of questionnaire

Thank you for completing this questionnaire. Please enclose this document along with the marking sheet, the scripts, and other documentation.

Appendix E: The distributions of median scores under each mark scheme

Figure E1 shows the distributions of the median scores across markers awarded to each script under each mark scheme. Figure E2 shows this same information in terms of the cumulative distribution functions. These charts confirm that the ‘true’ scores for each script cover a greater range under the ‘experimental’ mark scheme than under the ‘original’ mark scheme. In particular it can be seen from Figure E2 that all of the cumulative distribution functions cross indicating that under the ‘experimental’ mark scheme there are fewer candidates achieving at least the lower marks (that is, more scripts are awarded low marks) and there are more scripts being awarded the highest marks. This reiterates the change in the score distribution between marks schemes that has been discussed at some length in the full report. Some simple descriptive statistics on the distribution of median scores under each mark scheme are given in Table E1.

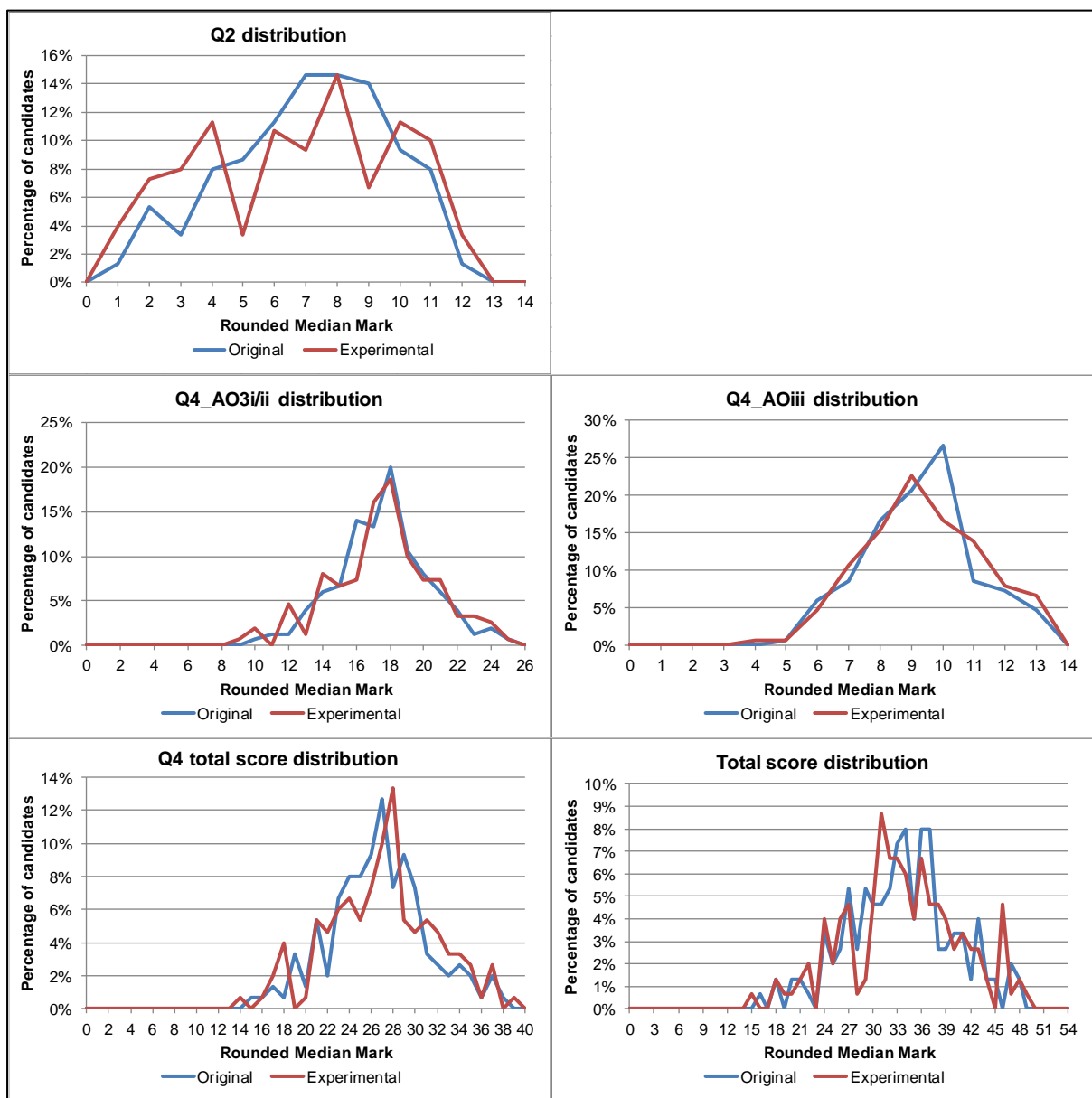


Figure E1: Distributions of median scores under each mark scheme

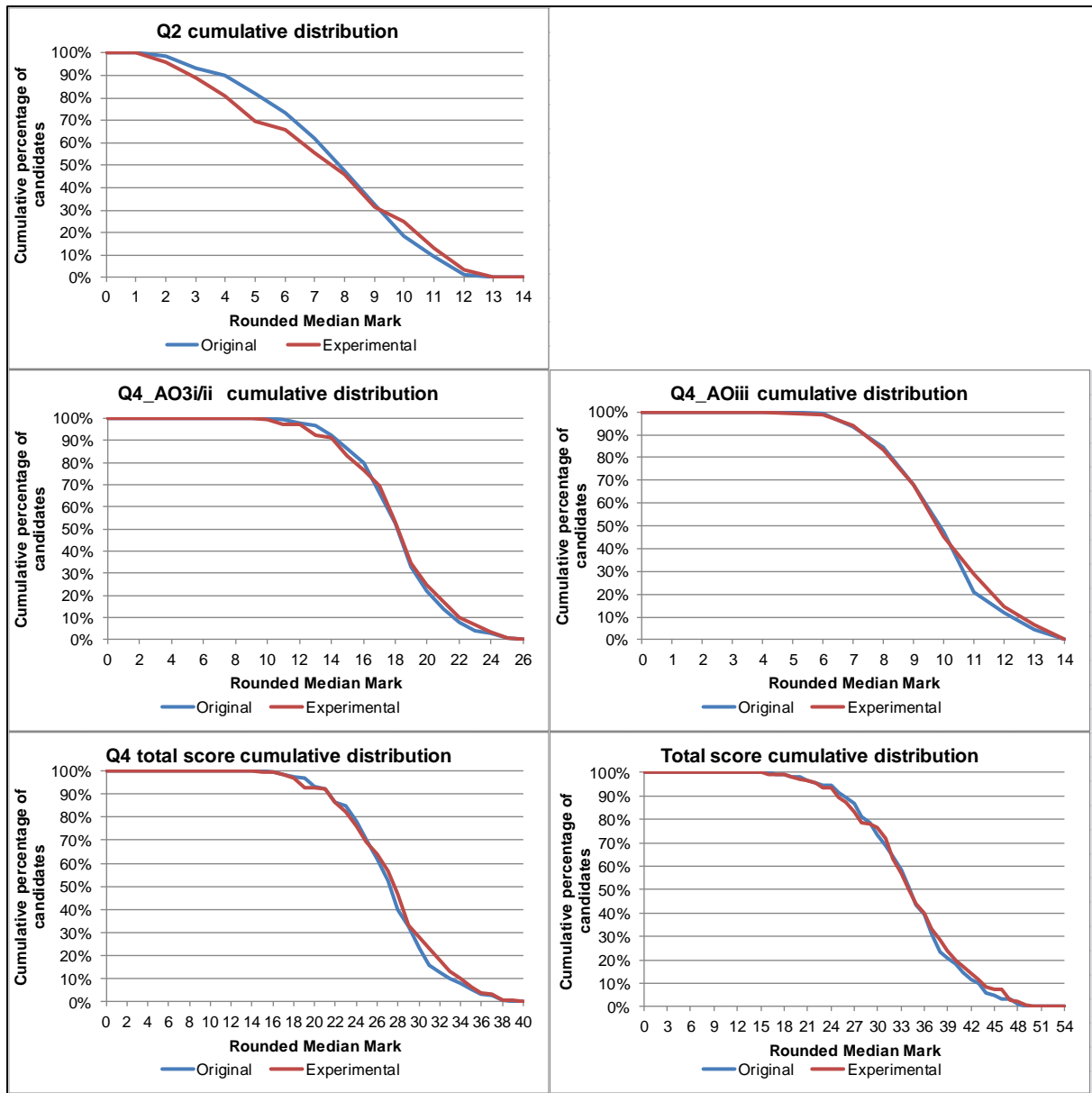


Figure E2: Cumulative distributions of median scores under each mark scheme

Table E1: Descriptive statistics on median scores across 150 scripts in each mark scheme

Question	Means		Standard Deviations	
	Original	Experimental	Original	Experimental
Q2	6.86	6.75	2.56	3.11
Q4_AOii/ii	17.33	17.58	2.78	3.12
Q4_AOiii	9.07	9.39	1.75	1.91
Q4 (Total)	26.43	26.9	4.47	4.91
Total (Q2+Q4)	33.28	33.69	6.48	7.01

Appendix F: Methodology used to compare intra-class correlations between mark schemes

The methodology used to calculate and compare intra-class correlations between mark schemes follows exactly the formulae for Fisher's Z-test described in Donner and Zou (2002). The text below largely reproduces the description within their paper although, for brevity, the justification for some of the steps has been omitted. This method assumes that:

- The scores awarded to each script follow a multivariate normal distribution, with the multivariate element of this reflecting the fact that each script is marked by a multiple number of markers.
- The correlations between the scores awarded to the same candidates by two different markers from the l th mark scheme are denoted ρ_l .
- For every marker in the l th mark scheme the expected (average) score they will award to candidates is denoted by μ_l .
- The correlation between the scores awarded to the same candidates by two markers from different mark schemes is denoted by ρ_{12} .

Using these assumptions the reliability of marking with the l th mark scheme can be estimated as an intra-class correlation using an analysis-of-variance (ANOVA) estimate via the following formula:

$$r_{lA} = \frac{MSA_l - MSW_l}{MSA_l + (k_l - 1)MSW_l}$$

Where MSA and MSW are the mean-square errors among and within subjects respectively and k_l is the number of judges (markers) using the l th mark scheme.

The above formulation means that, within the same mark scheme, the method assumes that there is no difference in the relative leniency/severity of different judges. Although this assumption is clearly violated for our data the effect of this violation on the estimates of intra-class correlations is extremely small. To verify this we compared the values of the intra-class correlation as calculated above (and, following Shrout and Fleiss (1979), labelled ICC(1,1)) to an intra-class correlation coefficient not requiring this assumption (labelled ICC(2,1) – also following the labelling of Shrout and Fleiss). Note that we know of no published method for calculating the significance of differences in ICC(2,1) in *dependent* samples¹⁴, thus leading to the necessity of relying on ICC(1,1) in the first place.

The values of ICC(1,1) and ICC(2,1) are compared for each question and for each mark scheme in Table F1. As can be seen the two intra-class correlation estimates are always extremely close (within 0.01).

¹⁴ That is, studies where all markers have marked the same scripts. Note that, the original work of Shrout and Fleiss more or less enables the calculation of the significance of differences for independent samples.

Table F1: Comparing ICC(1,1) and ICC(2,1) for each question and each mark scheme

Question	Original Mark Scheme		Experimental Mark Scheme	
	ICC(1,1)	ICC(2,1)	ICC(1,1)	ICC(2,1)
Q2	0.601	0.603	0.583	0.589
Q4_AOi/ii	0.481	0.488	0.535	0.540
Q4_AOiii	0.452	0.461	0.599	0.601
Q4 (Total)	0.482	0.490	0.585	0.588
Total (Q2+Q4)	0.586	0.590	0.627	0.631

Having established that the assumptions of the method do not have a large impact upon the estimates of reliability, we next use the method Fisher's Z-test described by Donner and Zou to calculate the statistical significance of differences in the intra-class correlations.

The first step in this process is to convert the intra-class correlations to Z values via the formula:

$$Z_l = \frac{1}{2} \ln \left(\frac{1 + (k_l - 1)r_{lA}}{1 - r_{lA}} \right)$$

Since the Z values are monotonically increasing with r_{lA} significance testing focusses on calculating the standard error of the difference in Z_l between the two marks schemes. This can then be used to calculate a t statistic for the differences in Z values and significance can be evaluated via comparison with the standard cumulative normal distribution in the usual way.

Using standard formula derived by Fisher in 1925, the variance of Z_l is then estimated via

$$V_l = \frac{k_l}{2(k_l - 1)(N - 2)}$$

Where N is the number of candidate scripts being marked by each marker.

Next we estimate ρ_{12} by making a two column matrix of each of the $N \times k_1 \times k_2$ possible pairs of scores for each script where the first column consists of scores assigned by markers in the first mark scheme and the second column consists of scores assigned by markers in the second mark scheme. ρ_{12} can then be estimated as the Pearson correlation coefficient between the scores in the two columns.

Once this has been calculated we can now estimate the covariance between Z_1 and Z_2 as

$$cov(Z_1, Z_2) = \frac{k_1 k_2 \rho_{12}^2}{2N(1 + (k_1 - 1)\rho_1)(1 + (k_2 - 1)\rho_2)}$$

If the number of judges is unequal across the different mark schemes, then in fact under the null hypothesis (that there is no difference between the intra-class correlations) we would expect a small difference in the estimated values of Z between each mark scheme. This bias correction term is given by

$$\theta = E(Z_1 - Z_2) = \frac{1}{2} \ln \left(\frac{1 + (k_1 - 1)\rho}{1 + (k_2 - 1)\rho} \right)$$

Where ρ is the estimate of the intra-class correlation based upon the full sample without distinguishing between mark schemes. Note that θ will always be equal to zero if there are the same number of markers in each mark scheme.

Finally we calculate

$$T_z = \frac{Z_1 - Z_2 - \theta}{\sqrt{V_1 + V_2 - 2Cov(Z_1, Z_2)}}$$

and compare this to the cumulative distribution function of the standard normal distribution to calculate the significance of the difference between the intra-class correlations in each mark scheme.

That the correct values for the intra-class correlations themselves had been estimated by this in-house code was verified by comparison with estimates based upon the R package *psych* by William Revelle¹⁵.

¹⁵ Note that this package does not enable the calculation of the statistical significance of differences in intra-class correlations for dependent samples – only the calculation of the intra-class correlations themselves.

Appendix G: The mark-remark reliability of the Principal Examiner

As part of the data collection the Principal Examiner (PE) marked each of the 150 scripts twice – once using the ‘original’ mark scheme and once using the ‘experimental’ mark scheme. It is worth noting that there was a considerable time gap between the two sets of marking (several weeks) so that it is unlikely that the PE would remember the marks awarded between the two occasions. This feature of the data collection provided a rare opportunity to examine the consistency of an individual PE, and thus an opportunity to examine (one aspect of) the reliability of the marks provided by PEs more generally.

Some descriptive statistics regarding the size of the differences between the marks awarded by the PE on each occasion are provided in Table E1. As can be seen, the PE was extremely consistent between occasions. Even when we consider the total score across both questions, half of the candidates received exactly the same mark on both occasions, with almost all of the others seeing a difference of just one or two marks. For only three candidates did the mark awarded change by three marks – the maximum difference seen at any point. These results are further emphasised in Figure E1 where the marks awarded on each occasion are plotted against one another showing a correlation of 0.99 between occasions.

Overall this analysis provides considerable reassurance regarding the reliability of marking by PEs. Despite the large time gap between occasions and the (largely cosmetic) changes to the mark scheme, the PE was virtually able to reproduce the original marks on a second occasion. Having said this, the marks provided by the PE are not absolutely identical between occasions. Thus we should be cautioned against treating PE marks (or possibly definitive marks) as absolutely immutable facts against which other marking can be compared.

Table G1: Descriptive statistics on the differences between the marks awarded by the PE using each mark scheme

Statistic	Question				
	Q2	Q4_AO3i/ii	Q4_AO3iii	Q4 (total)	Tot (Q4+Q2)
N	150	150	150	150	150
Minimum absolute difference	0	0	0	0	0
Maximum absolute difference	2	2	1	3	3
Median absolute difference	0	0	0	0	0.5
Mean absolute difference	0.31	0.35	0.24	0.46	0.66
Number with difference of 0 marks	106	101	114	97	75
Number with difference of 1 marks	42	45	36	38	54
Number with difference of 2 marks	2	4	0	14	18
Number with difference of 3 marks	0	0	0	1	3

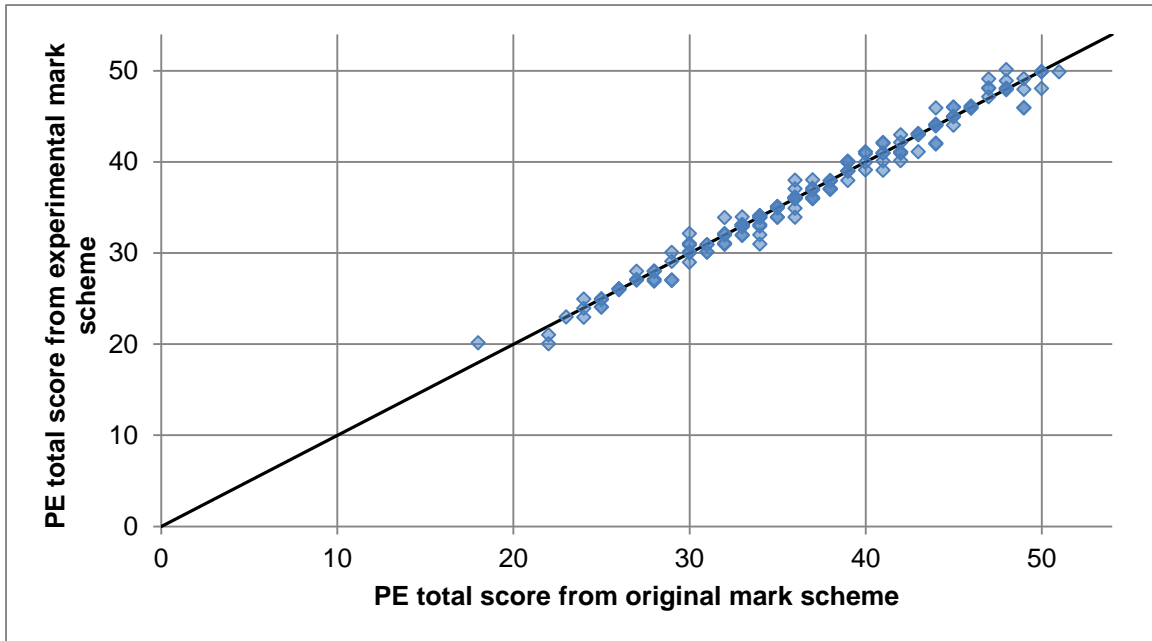


Figure G1: Scatter plot comparing marks given by the PE using the 'original' and 'experimental' mark schemes