Report

for Ofqual

A review of literature on marking reliability research

Jo Tisi
Gillian Whitehouse
Sarah Maughan
Newman Burdett

Ofqual/13/5285

June 2013

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

Acknowledgements

The report writing team at NFER are very grateful to all those who have contributed to this study:

- Ben Styles, Alistair Pollitt and Dougal Hutchison for their expert input and advice on the content of the paper;
- Frances Brill for commenting on the readability of the report; and
- Pauline Benefield, Emily Houghton and Alison Jones for their support in conducting the searches and sourcing the documents.

Published in January 2013
by the National Foundation for Educational Research,
The Mere, Upton Park, Slough, Berkshire SL1 2DQ
www.nfer.ac.uk

How to cite this publication:
Tisi J., Whitehouse, G., Maughan S. and Burdett, N. (2013). *A Review of Literature on Marking Reliability Research (Report for Ofqual).* Slough: NFER.

Contents

1 Executive summary

1.1 Context

The purpose of any assessment is to measure the level of some defined trait, quality or ability within the subject being tested. This may be a body of knowledge, competence in a skill, or estimation of future potential. In all cases, it is important to know how accurately, and with what repeatability, the test measures the feature of interest.

A candidate's actual score on any particular occasion is made up of their 'true' score plus a certain amount of measurement error. On a given day, a candidate might score higher or lower than their true score, depending on how they are feeling, what questions they are asked or who marked the paper.

Reliability refers to the consistency of a measure. A test is considered reliable if we would get the same result if the test or examination were taken on hypothetical, multiple occasions. Many people accept the fact that a test result could have been different if the candidate had taken the exam on a different day or if different questions had come up. This uncertainty in the system appears to be accepted as "the luck of the draw". However, when it comes to human error in the process of assessment, including marking variability, the general public are, understandably, much less tolerant. If the factors that affect marking reliability can be better understood, this information could potentially be used to improve marking and/or to set realistic expectations for levels of marking reliability.

1.2 Background and focus

Meadows and Billington (2005) produced a comprehensive review of the literature on measurement error that stems from variability in marking. So as not to repeat their work, this paper considers the literature that has been published since Meadows and Billington's review, with particular attention to advancements that have been made in the fields of quantification and improvement of marking reliability.

The focus of this review is marking reliability, which is affected by factors relating to the marking process, such as mark scheme design, individual marker behaviours and different marking processes. We review evidence of reliability at both item and whole paper level, as well as reports describing new methods for measuring reliability and new ways of improving the reliability of test and examination results. To this end the report aims to identify the main advances that have been made in improving and quantifying marking reliability. As such, this review of the literature forms part of the Ofqual Quality of Marking Project.

1.3 Methodology

The NFER carried out a robust and systematic review of the best available, relevant literature which has been published during the period 2004 to 2012.

This involved systematic searching and a consistent best evidence approach to the selection of literature, including reviewing selected literature against the frequency

with which it addresses the defined research questions. We focused on empirical and practice-based evidence and factual documentation, such as reports and key policy papers[1].

The search focused on documents:

- Published since 2004;
- Published internationally;
- Available in English;
- Covering national curriculum tests, general qualifications including GCSE and A level, and summative teacher assessment.

The systematic searching identified 240 sources for the review. Screening was carried out against the research questions, which identified 28 key items for inclusion. Ofqual were able to recommend a further six documents which addressed the research questions and aligned with their reliability programme to further strengthen the evidence base.

It should be noted that this literature review has reviewed several other pieces of literature and has followed standard good practice when referring to the work of one author as quoted in that of another, by citing both works in the reference list (or bibliography entry). This does not imply that NFER have read or appraised the texts that have been quoted in the work of other authors.

1.4 Advances in quantifying marking reliability

For some examination questions, the correct answer(s) can be unambiguously defined and each candidate response can be assigned a definitive mark. For these types of questions, it should theoretically be possible for markers to assign marks with zero variation. Mistakes may still occur, but investigations of marking this type of question have shown that high levels of marker accuracy can be achieved. As questions, and student responses, become more complex it is harder to determine exactly how 'good' a response is. Detailed mark schemes can go a long way to clarifying what makes a credit-worthy response, and so are important in improving marking accuracy. However, lower levels of marker agreement on essay questions may be a result of legitimate differences in opinion between markers. There is a large body of literature that researches this area and argues that the use of questions with longer responses is an important part of the assessment process and so an educational system may choose to accept the lower levels of reliability.

To measure marking accuracy a suitable data set is needed. The marks of individual examiners must be compared with some estimation of the 'correct' mark for each response, or the 'true' score for each candidate. The correct mark is sometimes defined as the mark awarded by a senior examiner, or the average mark awarded by

[1] Full details of the search strategy and of the search terms are included in Appendix 2.

multiple markers. Therefore, studies of marking reliability generally require that the candidate responses are marked at least twice. This is difficult and expensive to organise, particularly during 'real time' examination marking. In addition, marking reliability should always be calculated using data from independent marking events. If the second (or subsequent) marker can see the marks and comments of the first marker there is evidence that the annotations will bias their marking, resulting in an overestimation of the level of marker agreement.

Advancements in technology have resulted in many units/components being marked on screen. The marker monitoring processes employed in on-screen marking gather multiple marking data as a matter of course. These data can be used to estimate marking reliability and, with some adjustment, could be a very valuable source of information about marker accuracy. This is an important advancement in the quantification of marking reliability.

Although numerous studies of marking reliability have been conducted, our review highlighted a lack of consensus regarding terminology and statistical techniques. This can be a problem because it is difficult to compare reliability statistics that have been generated through different methods of data collection and analysis. In addition, studies of marker agreement do not always use the same definition of 'true' score, which also makes comparison difficult.

Statistical techniques, such as Generalizability theory and the Multi-Facets Partial Credit Rasch Model, can be used to analyse multiple sources of measurement error and estimate the relative effect of each one. In other words, researchers can determine how far unreliability in marking affects test results in comparison with errors stemming from questions, candidates, occasion etc. Generalizability theory can also be used to model the effects of changing parameters of the test, for example increasing the number of markers. The output of the model can be used to improve the test design and, subsequently, the test's reliability.

As there is no current agreement about the best means of quantifying unreliability, the assessment community are not any closer to achieving a key recommendation from the Meadows and Billington review, which was to publish reliability information alongside the results of examinations. As there is no clear answer, it may be useful for an organisation, such as Ofqual, to take the lead and state which method could be used and which statistics should be provided.

1.5 Advances in improving marking reliability

The second major theme of this review is the advances made since 2005 in improving marking reliability. The literature reviewed does demonstrate that the assessment community's understanding of the factors affecting reliability has broadened, leading to a number of approaches that can be used to improve the reliability of results. Evidence is included that demonstrates that marking accuracy is affected by various features of the individual test questions and the mark schemes, characteristics of the markers, and the processes that are used for marking. There appears to be some consensus that reliability can be improved by adapting the mark schemes, by for example making them more constrained, and by selecting markers

with appropriate characteristics, for example using more experienced markers to mark more complex questions.

Technological advances in on-screen marking have enabled a number of the advances in improving reliability, such as by allowing question papers to be split up into individual items/questions. The items that are harder to mark can then be sent to the examiners with the most experience (who, in general, will mark complex items more reliably than less experienced examiners); while those that are easier to mark can be sent to less experienced markers.

Theoretically, item-level marking, enabled by onscreen marking, should also improve reliability because when different examiners mark each item on an examination script any marking errors that exist will be not be related and are likely to cancel one another out. Whereas, when the same examiner marks the whole paper, it is likely that any errors will be related (e.g. the examiner is consistently slightly too harsh) and they will compound rather than cancel out. In addition, item-level marking reduces the effects of biases caused by the context in which an item is marked. For example, when one examiner marks all the questions on a script, the mark that they allocate to one item may be affected by the student's responses to other, unrelated, questions. This is known as the halo effect and is eliminated in item-level marking. In addition, it has been shown that an examiner's mark for a given response can be affected by the quality of the immediately preceding responses, such that a lower mark may be given if the preceding responses are of particularly high quality and vice versa. In item-level marking, the responses are presented in a random order, and so any bias stemming from comparisons with preceding items will not be systematic. That is, the bias affecting one item from a script is likely to be different to that affecting another item on the same script.

On-screen marking further improves marking reliability by allowing regular marker monitoring. This means that inconsistent or inaccurate marking can be detected early and that either the marker can be retrained or their marking can be reallocated. The continuous monitoring can ,furthermore, detect markers whose accuracy changes over time. On-screen marking also eliminates errors resulting from incorrect addition or transcription, prevents items being left unmarked, and removes the economic and logistic burdens of transporting paper scripts to and from examiners.

A number of the papers included in this review considered the benefits of double or multiple marking. Multiple marking has the potential to improve marking reliability for some question types, in particular those that require some level of subjectivity in their marking. However, there are a number of logistic and financial obstacles to introducing multiple marking and it is unclear whether the benefits will outweigh the problems. There is also a theoretical consideration – the combination of double/multiple marks to produce a final score is an acknowledgement that legitimate differences in opinion can exist between examiners. This is fundamentally different from the current system, in which the marks of the most senior examiner are considered to be the most 'true'.

There have been some advancements in the field of computer-based marking. However, with the exception of some objectively marked item types, much work is

still needed before computer marking becomes a viable alternative to human marking.In summary, much work has been done since Meadows and Billington published their review in 2005. Useful descriptions and comparisons of the methods for quantifying and presenting information about marking reliability now exist. In addition, technological advances have facilitated the measurement of marking reliability and led to an increase in our understanding of the influencing factors. This, in turn, has produced tangible methods for improving marking reliability that can be implemented in high stakes examinations in England.

2 Introduction

2.1 Background

This report presents the findings of a literature review into marking reliability commissioned by Ofqual from the National Foundation for Educational Research (NFER) in October 2012. The purpose of the review is to update the findings about reliability methodologies and measures since the Meadows and Billington (2005) report. Meadows and Billington provided a thorough history of, and overview of the findings from, research into marking reliability for the then National Assessment Agency. However, much progress has been made in this area since 2005, not least in the Ofqual Reliability Programme (Opposs and He, 2011).

The Ofqual Reliability Programme (Opposs and He, 2011) was launched in 2008. The programme aimed to gather evidence to develop regulatory policy on reliability with a view to improving the assessment system in England further. The programme included research conducted by teams and individuals from both the UK and abroad focussing on generating evidence of the reliability of results, interpreting and communicating reliability evidence, and researching the public's perceptions of reliability. Many of the published reports are included in the documents cited in this review.

This review has been written for a broad range of audiences, including awarding organisations, school staff, parents and students, as well as the general public. Many of the concepts related to reliability are technical, and we have, wherever possible, tried to explain the technical aspects in the simplest way.

What is meant by reliability?

A useful definition of reliability is provided by Ofqual as part of the Reliability Programme:

> *'"Reliability" in the technical context means how consistent the results of qualifications and assessments would be if the assessment procedure was replicated – in other words, satisfying questions such as whether a student would have received the same result if he or she happened to take a different version of the exam, took the test on a different day, or if a different examiner had marked the paper.*[2]

Meadows and Billington (2005) and Baird *et al*. (2012) both present detailed descriptions of the different levels at which reliability can be measured and the statistical methods that can be used, and discuss their pros and cons in an operational setting.

In his discussion of National Curriculum testing, Newton (2009) highlights the difference between the 'reliability' and 'accuracy' of an assessment as a whole. He argues that "reliability coefficients do not estimate the accuracy of assessment results, per se, because they fail to model the impact of systematic error" (p. 184).

[2] http://www2.ofqual.gov.uk/standards/reliability/

That is, they quantify the effect of random error on the measurement but give no information as to how close the measurement is to what we were hoping to measure. Baird *et al.* (2011) also raise this point, noting that "we are not very explicit about what our tests are trying to measure and this causes problems for clarity in quantification of whether our measures are consistent" (p. 19).

Most of the published literature on reliability has investigated the issue at the level of whole question papers. More recently, data allowing the measurement of reliability at item level has become more readily available. Evidence about reliability at these levels is helpful for improving assessment design and monitoring marking; however, reliability statistics at the level of a whole qualification are likely to be of most interest to users because they provide information on grading reliability (Baird *et al*., 2011). Qualification level reliability will usually be higher than the reliability levels of the component parts because measurement error is random and aggregation of the results from multiple components can act to cancel out errors. However, qualification level reliability can be difficult to calculate especially in modular examinations and those with many options or shared units (Bramley and Dhawan, 2010; He, 2009).

To interpret the results of reliability studies we need to consider how reliable we can reasonably expect examination results to be. Baird *et al*. (2011) comment that:

> *Few guidelines are available for interpreting the value of reliability evidence. The obvious reason for this is that the reliability indices will largely depend on the population for which the test is used, and the conditions for the administration (e.g. how much time is available, what is the motivation of the respondents etc). For example, it is hard, if not impossible, to design a writing assignment that has a valid marking scheme and for which at the same time the reliability index is relatively high.* (pp. 15 -16)

The *Dutch Association of Psychologists* (Evers *et al*., 2009) give some general rules. For reliability of high stakes tests they suggest that a reliability index above 0.9 is good, between 0.8 and 0.9 is sufficient, and below 0.8 the reliability is considered insufficient. For a sub-test (e.g. a single component or unit) these figures can be 0.1 lower than for the total test. Baird *et al.* (2011) go on to suggest that:

> *Although it is not currently possible to specify acceptable values for reliability for assessments regulated by Ofqual, it would be possible with more information. Standards for reliability of particular qualifications should be empirically grounded in data on reliability for assessments of different formats.* (p.16)

The focus of this review is marking reliability, and so it is concerned with the factors relating to the marking process that affect the consistency of the results, such as mark scheme design, differences in individual marker behaviours, and different marking processes. We review evidence of reliability at both item and whole paper level, as well as reports describing new methods for measuring reliability and new ways of improving the reliability of test and examination results.

Summary of Meadows and Billington (2005) [3]

As stated above this review builds on the work produced in 2005 by Meadows and Billington. They produced an extensive review of the literature on marking reliability spanning nearly 100 years. They covered "the levels of marking reliability achieved in different forms of assessment and research into methods of improving marking reliability" with a focus on "the marking of externally assessed examination scripts, rather than on the assessment of coursework, performance or of competence..." (p. 4).

The review discusses different definitions of reliability, and marking reliability in particular, in some detail and the arguments for and against the different ways in which reliability can be measured. It also describes the different sources of unreliability, including context effects, text effects (such as handwriting), the candidate and the examiner, as well as the design of the question paper. The review considers a significant number of articles and concludes that "it is often difficult to draw conclusions about the factors that influence reliability. … because the studies often vary in so many important respects (the training of the markers, the type of assessment, the mark scheme, the subject assessed and so on)" (p. 20). In part, this conclusion informed the development of the Ofqual Reliability Programme which aimed to collect evidence about reliability in a more systematic way.

Meadows and Billington (2005) concluded that a measure of the reliability of a test should be published alongside the results in order for the results to be fully understood. They also found, as might be expected, that reliability is strongly associated with the type of question being used. Tightly defined questions with definite answers can be marked much more reliably than, for example, essay questions. They did note, however, that we may choose to accept the lower levels of reliability associated with certain question types, where we believe the question type to add value over more tightly constrained questions. However, for questions that do traditionally have lower levels of reliability, it may be possible to make improvements by, for example, refining the mark scheme or by improving marker training.

Why this new study is needed

Reliability of the results from tests and examinations is an increasingly important issue, with results being used for a variety of purposes with very high stakes for students, teachers, schools and government. Questions about the extent to which the results can be trusted are raised much more frequently as the accountability pressures are increased.

Newton (2009) stated that we"need more openness and transparency about the uncertainty that is an inevitable feature of national curriculum testing" (p. 208). He went on to note that this is not just about making the information publicly available, but he highlighted the need "to identify the best mechanisms for communicating uncertainty, the most appropriate ways to present reliability evidence" (p. 208). He

[3] A detailed summary of the Meadows and Billington review (2005) is included as Appendix 1.

also called for ”more public debate on how much error constitutes too much error for the various uses of results” (p. 208). It is clear that Newton’s views apply equally to many other forms of tests and examinations.

In this context, it is important that assessment developers are able to provide evidence of the level of reliability that users of the results can expect from their tests and examinations. Ofqual (2012a) published its Corporate Plan for 2012-2015 in May 2012. The plan notes that “over 15 million scripts are marked for each summer examination series alone, and few people know how marking works, and what is achieved. Confidence is understandably undermined by the small proportion of cases where the student or school believes an injustice has been done” (p. 15).” Ofqual therefore committed to undertake a programme of work looking at quality of marking in GCSE and A-level examinations. The work will set out publicly how marking happens, whether the current system is good enough and what improvements could be made. In addition to this literature review, Ofqual will be gathering detailed evidence from exam boards and examiners, and using external research to understand the perceptions and expectations of teachers, parents, students and the public.

Initial studies of public perceptions of reliability (Burslem, 2011) show that, in general, understanding of reliability issues is low. In addition, and perhaps of more relevance here, members of the public have different degrees of tolerance for different sources of unreliability. For example, unreliability caused by learners having an ‘off-day’ and performing less well than expected was seen as acceptable, whereas unreliability caused by variability in the marking was seen as much less acceptable.

While it is probably impossible to remove all unreliability caused by marking from assessment results, it is likely that reliability levels can be improved at least in some forms of assessment. However, in order for this to be possible the assessment community must first:

- develop and agree robust means of measuring reliability in a range of different contexts, in a way that can be compared across different assessments;
- measure the current level of reliability in different subject areas, different qualifications, and in different forms of assessment;
- conduct research into the different ways that reliability can be improved, such as through mark scheme refinements;
- use the lessons learned to improve the reliability of existing and new assessments.

Changes since 2005

This review is published at a time of considerable change in the test and examinations system in England. The Conservative/ Liberal Democrat coalition government, which came into power in 2010, is introducing reform to national curriculum tests, to GCSEs and to A levels, building on a period of comparable change under the previous Labour government. The current changes in the test and examinations system are, in part, influenced by the debates about the reliability of the

results produced. The changes include a reduction in the elements of the assessments which are believed to be less reliable, such as the assessment of writing at key stage 2 about which the Bew Review (Bew, 2011) stated that ”perhaps the most significant point is the frequently-made criticism over the inconsistency and subjectivity of the external marking” (p. 60). Similarly, the coursework components of GCSE have been replaced by Controlled Assessments (Johnson, 2011). More recently, proposed changes to qualifications at 16 include the option of a 100 per cent externally marked qualification (DfE, 2012).

In addition to the changes to the education system at the qualifications level, improvements to the test and examinations systems made possible by new technologies have driven a large number of changes since the Meadows and Billington (2005) report. The largest change has been in the introduction of on-screen marking, in which the student work is scanned and presented to the examiners online. On-screen marking has enabled changes firstly to the processes, so that monitoring of marking can take place at more points throughout the marking period; secondly to the staff involved, so that different groups of items can be marked by different types of markers and, finally, to the outputs, so that a much larger number of item level marks are available. A number of the papers that are reviewed as part of this study have been produced in this context of e-marking.

2.2 Aims

The overall aim of this research is to review literature on marking reliability, with a particular focus on **advances in quantifying and improving marking reliability** since the 2005 review conducted by Meadows and Billington (Meadows and Billington, 2005).

The report aims to address the following research questions:

1. What advances have been made in **quantifying marking reliability**? What does the evidence say about:
    - terminology for marking reliability
    - methods for quantifying reliability.
2. What advances have been made in **improving marking reliability**? What does the evidence say about:
    - factors affecting marking accuracy
    - cognitive processes used in marking
    - technological advances in marking processes
    - multiple marking
    - teacher assessment.
3. What does the evidence say about detecting and correcting unreliable marking when looking at:
    - externally marked components
    - internally assessed components

- detection of errors after marking is complete
- tolerance levels
- methods for adjusting scores.

These research questions have been used to focus the selection of texts and to guide the structure of the report.

2.3 Methodology

The NFER carried out a robust and systematic review of the best available, relevant literature which has been published during the period 2004 to 2012.

This involved systematic searching and a consistent best evidence approach to the selection of literature, including reviewing selected literature against the frequency of addressing the defined research questions. We focused on empirical and practice-based evidence and factual documentation, such as reports and key policy papers[4].

The search focused on documents:

- Published since 2004;
- Published internationally;
- Available in English;
- Covering national curriculum tests, general qualifications including GCSE and A level, and summative teacher assessment.

The systematic searching identified 240 sources for the review. An initial screening was carried out on the documents to exclude items that did not meet our review parameters. A team of internal and external experts triple-screened all the sources and triangulated the results, producing a hierarchical list of documents. This list was then mapped against the research questions to identify frequency of coverage. Low frequency resulted in additional documents being identified from the original sources, which filled the gaps and improved coverage.

This resulted in a list of 33 documents which were then sourced and read in full. Based on this review, 28 key documents were selected for inclusion. Each selected document was rated for quality and 18 were classified as high, which is defined as large scale quantitative studies; or in-depth case studies that cover a range of institutions and a wide range of stakeholders, where views are triangulated; or a meta-analysis or systematic review. The remainder were classified as either medium or modest quality.

As this is an Ofqual commissioned report, Ofqual were able to recommend a further six documents which addressed the research questions and aligned with their reliability programme to further strengthen the evidence base.

[4] Full detail of the search strategy and of the search terms are included in Appendix 2.

2.4 Report structure

Sections 3, 4 and 5 of this report address the different research questions outlined above. Section 6 draws conclusions from the reviewed literature in answer to the research questions posed. Section 7 and appendices 4 and 5 provide a glossary and explanation of commonly used terms and techniques.

3 Advances in quantifying marking reliability

3.1 What does this section cover?

Our review found various papers relating to the quantification of marking reliability. Some researchers focused on terminology and methodology of quantification, while others estimated marking reliability in operational settings. Technological advances in recent years have led to a proliferation in on-screen marking, which produces a convenient source of multiple-marking data and data at the item level that can be used for reliability estimates. In this section we discuss some of the statistical techniques that have been used in recent studies of marking reliability, including Generalizability Theory (G-theory), Multi-Facets Partial Credit Rasch Models (MFRM), and Intra-class Correlations (ICCs). We also review the use of data produced during on-screen marking for measuring reliability.

3.2 Key Findings

Our key findings in this area are:

- If meaningful comparisons are to be made between the results of marking reliability studies, a consensus must be found on the terminology and statistical techniques used for measuring marking reliability;
- Data from routine monitoring of on-screen marking has the potential to be used for regular measurement of marking reliability;
- Generalizability theory can be used to separate sources of error in test results and to model the effects on reliability of changing parameters, such as the number of markers.

3.3 What is the evidence base?

Evidence for this section came from a variety of the papers that we reviewed. Bramley (2007), Newton (2009) and Baird *et al.* (2011) all comment on terminology. Baird *et al*. (2011; 2012), Baker *et al.* (2008), Billington (2012), Bramley (2007), Bramley and Dhawan (2010), Brooks (2004), Curcin (2010), Dhawan and Bramley (2012), Fearnley (2005), Kim and Wilson (2009), Massey and Raikes (2006), Meadows and Billington (2007), Newton (2009) and Vidal Rodeiro (2007) all contribute to recent discussions about the methods of measuring marking reliability.

3.4 Standard terminology for marking reliability

Bramley (2007) comments that the lack of standard terminology and statistical indicators of marking reliability can sometimes make it difficult to compare results from different studies. He argues for the importance of distinguishing reliability of marking as a whole from measures of agreement between a marker and some form

of 'correct' mark. In other words, researchers need to be clear about whether they are comparing sets of observed scores, or observed scores with true scores (or as close to true scores as we can get operationally) (Newton, 2009; Baird *et al.*, 2011).

Bramley suggests that for objective marking, or cases where there is an unambiguously correct answer, the agreement between a marker's mark and the correct mark should be described as '*accuracy'*. For questions requiring longer responses or essays, where an examiner's interpretations of the mark scheme may legitimately differ from another, Bramley (2007) suggests that the term '*agreement'* should be used.

Bramley (2007) suggests that the term '*reliability*' be reserved for the relationship between 'true' score variation and observed (true+error[5]) score variation. We have followed Bramley's (2007) conventions in this report.

3.5 Methods for quantifying reliability

3.5.1 Data collection

Bramley and Dhawan (2010) point out that the way in which marker agreement is defined has implications for the choice of indicator used to quantify it. "From a measurement perspective, the ideal scenario for quantifying marker agreement is when two or more markers mark the same piece of work without knowledge of what marks the other markers have given to it (referred to as 'blind' double or multiple marking). The marks can then be treated as independent in the statistical sense which is usually an assumption of most common methods of analysing agreement" (p. 50).

The main obstacle to quantification of marking reliability is that it is often difficult to get the necessary data. Double or multiple-marking is difficult and expensive to include in live examination situations. Therefore, many of the estimates of marking reliability have been conducted as research exercises. However, the routine procedures for monitoring marking in live examinations can potentially provide a source of double-marking data. The monitoring of paper-based marking requires examiners to send samples of marked work to a more senior examiner for re-marking. The data from this exercise are not routinely collected for use in estimating marking reliability but some studies have used these data to calculate levels of marker agreement (e.g. Bramley and Dhawan, 2010). The problem with using this data to investigate marker agreement is that the senior examiner can see all the annotations of the first examiner. Thus the two marks are not independent, which has been shown to lead to an overestimation of marker agreement (e.g. Murphy, 1979; Newton, 1996; Vidal Rodeiro, 2007; Billington, 2012).

The proliferation of on-screen marking has produced a new and potentially rich source of double-marking data. On-going marker monitoring is conducted using 'seed' items/scripts. These seed items are pre-selected and marked by the senior examiner (or a senior examiner panel) and introduced at regular intervals into the

[5] '*Error'* is a technical term, which here refers to variation from sources that we would like to exclude.

marking. The marker is not aware that a particular item is a seed item and has no knowledge of the 'definitive' mark that has been awarded. The same seed items are marked by many markers and the data are automatically stored. Thus, the data set comprises multiple independent marks for each seed item.

Bramley and Dhawan (2010) investigated the mean and standard deviation of the distribution of differences between the marks awarded by examiners and the definitive mark assigned to seed scripts in on-screen marking of 276 GCSE and A-Level units/components. They found that: test-related unreliability (i.e. variability in outcomes arising from the particular questions or tasks in the assessment) was generally higher than marker-related unreliability; on average, markers tended to be neither severe nor lenient compared to the definitive mark; and systematic differences in severity among markers made the smallest contribution to score variability – less than systematic differences among seed scripts and much less than random (non-systematic) error. Similar levels of on-screen marking consistency were found by Dhawan and Bramley (2012) in a later study of GCSE and A-Levels.

These findings come with the caveat that they are based on the kind of component that was marked on screen at the time of the research, which tended not to include extended responses or essays. In addition, the power of the data set is compromised by the fact that there are few seed items and that they have been selected for the purpose of marker monitoring rather than reliability estimation and so they may not be fully representative of the whole population (Bramley and Dhawan, 2010; Baird *et al.,* 2012). For example, very low scoring scripts with many blanks would not be included. In addition, the senior examiner may purposefully include scripts with 'problematic' responses that are difficult to mark, to check that markers have fully understood how the mark scheme should be applied. If there are a large proportion of these 'problematic' responses the data would probably underestimate the levels of marker agreement that could be attained with non-seed scripts. This means that generalisation is limited in value. In addition, the seed items are often from different candidates and do not comprise the complete work of an individual, thus they have limited use in estimating reliability at component level. It should, however, be relatively easy to expand the seed item selection strategy so that the items are representative of the cohort as a whole and include the entire work of individual candidates. This would allow the seeds to be used in their current monitoring context and also provide more accurate component-level reliability statistics (Baird *et al.,* 2012).

A further issue is the extent to which the definitive mark assigned to the seed item is indeed the correct mark. In the majority of cases it almost certainly is, but in cases where the most frequent mark awarded by the markers differs from the definitive mark there is a possibility that the definitive mark is wrong (Bramley and Dhawan, 2010). Despite these limitations, Bramley and Dhawan (2010) suggest that "[because the statistics from this data] can be calculated across the full range of examinations marked on screen, [they give] a useful snapshot of the whole system" (p. 71).

3.5.2 Definition of 'true' score

In many studies, marking reliability is quantified as the agreement between the awarded mark and a 'definitive' or 'true' score for the response. There are, however, various different methods for determining the 'true' score (see Dhawan and Bramley, 2012; Baird *et al*., 2011). Fearnley (2005) found significant differences between six different estimates of 'true' marks in a study of two GCSE and GCE AS level components. Similarly, Meadows and Billington (2007) found differences in the reliability estimates calculated for GCSE English using different measures of 'true' score. Therefore, the way in which the true score is determined will have implications for the levels of marking reliability that can be measured.

In classical test theory (refer to appendix 5 for explanation), the true score is defined as the mean mark awarded by a theoretical infinite pool of markers under all possible test conditions. Under item response theory (IRT) we calculate true score by summing the expected score on each item given the student's ability (underlying latent trait). True score hence corresponds to a point on the underlying latent trait.

Obviously, both these definitions are theoretical constructs and cannot be measured, so measurable approximations are needed. When considering a true score across markers in classical test theory, the average of the marks of a number of different examiners is used. There are a number of problems associated with average marks. The mean or median awarded mark may not be a whole number. The mode can be problematic because, with a small number of markers or a large total mark, there may not be a mode, or it may reflect a chance coincidence of the total marks given by a very few markers (Dhawan and Bramley, 2012). Dhawan and Bramley (2012) suggest that the 'true' mark for a whole script should be calculated as the sum of the most common marks for each item, based on the assumption that the most common mark is most likely to be the 'correct' mark. For item response theory, an ability scale is created through statistical analysis and this scale is used for estimating person ability, item difficulty and marker severity. The ability measure itself represents performance in the trait of interest and this can be converted back to a true test score if desired. Using IRT in this way takes into account errors at the level of each marker and can lead to better representations of the construct of interest (Stemler, 2004).

Another widely used method is to deem the most senior examiner's marks as true. This hierarchical approach is used in general qualifications in England, and is often used for measuring examiner agreement in reliability studies. In many cases, where an unambiguous correct answer exists, a single, experienced examiner can almost certainly assign the 'correct' mark. However, where subjectivity is required in interpreting and evaluating the candidate's response, it is debatable whether a single examiner can hold normative standards within their head (see Blood, 2011; Taylor, 1992). Pilliner (1969) argued that two or more markers will produce a 'truer' score than a single marker. But consideration must be given to the way in which the marks are combined to produce the final score. The mean mark of all the markers is usually used, but the median or modal mark may also be appropriate (Bramley and Dhawan, 2010). It should be noted that using the combined scores of two or more independent markers is an acknowledgement that legitimate differences in opinion can exist

between markers. This is a philosophically different approach from the hierarchical system that is currently operated in England (Brooks, 2004).

The consensus of a panel of examiners is sometimes used to decide the correct mark for the seed items used in on-screen marking, and in some reliability studies. Such panels may help to introduce different perspectives where there is legitimate professional disagreement, but the group dynamics within a panel can influence the final marks that are assigned (Curcin, 2010).

3.5.3 Statistical methods

There is currently little consensus in the literature as to the most useful and dependable measures of marking reliability, although it is possible that consensus will emerge over time (Baird *et al.,* 2011). In the meantime, a variety of perspectives exist.

The use of a single statistic to quantify marker agreement loses information and can make interpretation and comparison of studies difficult (Bramley, 2007). For example, the correlation between the scores of two markers independently marking the same set of scripts is often used as an estimate of reliability. However, the correlation coefficient tells nothing of where differences lie in the distribution of marks and even a high correlation can mask large systematic differences in marks (see review in Meadows and Billington, 2005). Similarly, although low values of internal consistency measures, such as Cronbach's alpha (refer to glossary), can (sometimes) be attributed to unreliable marking (Bramley and Dhawan, 2010), they can fail to indentify inter-marker differences in cases where markers are consistently lenient or severe across all items (Newton, 2009). This is because the marks awarded by a marker who is consistently lenient/severe will correlate with one another (almost) as well as if the marking had been accurate and so a measure of internal consistency would not detect the problem.

Bramley (2007) suggests that "simple statistics, based on the distribution of differences between marker and correct mark, or marker and [senior examiner], are the easiest to interpret and communicate" (p. 27). The mean and standard deviation of the differences provide the simplest way of quantifying the disagreement (Bramley and Dhawan, 2010; Bramley, 2007), and the statistic P_0 states the proportion of cases where there is complete agreement between the marker and the correct mark (Bramley, 2007). In situations when data from multiple markers is available, these statistics also have the advantage of being closely related with classical test theory.

Bramley (2007) and others (e.g. Baird *et al.*, 2011; Cronbach and Shavelson, 2004) argue that the standard error of measurement (SEM)[6] is the most easily understood

[6] The standard error of measurement (SEM) is defined as the square root of the average error variance across a group of test takers. That is, it gives a measure of the size of the error in a set of observed scores relative to the true scores. SEM can be used to provide a band of marks, around the observed score, in which the true score is likely to lie. We can be 68% confident that the true score will lie within ±1 SEM of the observed score, and 95% confident that it will lie within ±2 SEM of the observed score. The SEM is calculated as SEM = $\sigma_x\sqrt{1-r}$ where σ_x is the standard deviation of the test results and *r* is the reliability of the test.

measure of marking reliability. Dhawan and Bramley (2012) point out that SEM cannot be properly interpreted without reference to the maximum mark of the test or the number of items. They suggest that the ratio of the number of marks between grade boundaries to SEM is a useful statistic because it allows comparison in terms of grades – the higher the ratio the more repeatable the grade outcomes are likely to be.

It is also of interest to quantify the proportion of variance in marks that can be attributed to differences among the markers (Bramley and Dhawan, 2010). Massey and Raikes (2006) suggest the use of item-level intra-class correlations (ICCs) for reporting examiner agreement. In contrast to other correlation coefficients that can mask consistent severity/leniency, the ICC reflects the extent of agreement between examiners. The ICC may be interpreted as the proportion of variance in the set of candidates' marks that is due to the candidates, i.e. after examiner effects have been controlled for. So, if ICC=1 there is perfect examiner agreement and if ICC=0 there is no examiner agreement and the marks appear random. Bramley (2007) notes that different ICCs are applicable in different data collection scenarios and so expert statistical advice is essential for choosing the correct method.

Newton (2009) suggests that more use could be made of Generalizability theory. G-theory uses analysis of variance to quantify the proportion of score variability resulting from different sources. It was developed specifically as a means of investigating measurement reliability and it is a useful model because it makes very few assumptions about the data (see Baird *et al.,* 2012, for a useful description of G-theory). G-theory enables quantification of the contribution of different sources of variation to measurement error and so allows test-related and marker-related error to be studied in a single model. (Baird *et al.,* 2012; Bramley and Dhawan, 2012).

Baker *et al.* (2008) conducted a large, blind, multiple-marking research study and used Generalizability theory to analyse the data. The aim of the study was to determine whether markers in countries other than England could mark the Key Stage 3 (KS3) English test consistently with markers in the national context of England. Two hundred reading scripts and two hundred writing scripts were chosen for the study. A definitive mark was assigned to each script by consensus of two senior English examiners. The scripts were independently marked approximately 10 times by English markers and four or five times by Australian markers. Thus, a very large blind multiple-marking data set was obtained.

There are five possible attainment levels for this examination. Baker *et al.* (2008) quantified overall agreement as the percentage of scripts that would have been awarded the same level using the marker's mark as using the definitive mark. Overall agreement for Australian and English markers of reading scripts was 68 per cent and 57 per cent, respectively. Agreement on the writing scripts was 41 per cent and 56 per cent respectively. These figures are similar to those found by Royal-

Dawson (2005); the range of marker agreement was 61-67 per cent on the reading test and 48-52 per cent on the writing test.

Baker *et al.* (2008) found that the patterns of Australian and English marker agreement were strikingly similar, implying that national context did not have a significant effect on marking reliability and that marking could reasonably be transferred to other national sites. However, although the markers were consistent within and across themselves, their marks did not agree well with the definitive mark, either at item or total scripts level for either reading or writing. This was true for Australian and English markers. There could be a number of explanations for this observation including a training shortfall, ambiguity in the mark scheme or error in the definitive mark (which could have been investigated by comparing the modal mark with the definitive mark – this analysis was not performed, however).

Inter-marker agreement was found to be very high for the 38 Australian markers. Baker *et al.* (2008) used Generalizability theory to determine the relative contribution of various factors to measurement error. In every Generalizability study conducted, marker variance *per se* contributed the smallest amount to measurement error – the greatest sources of measurement error were item related. A strong item by script interaction was also found, which suggests that different groups of students are performing well on different sets of items (this effect was also found by Kim and Wilson, 2009, for written compositions). This may suggest that different sets of students have learnt how to tackle different types of questions.

G-theory can also be used to model the effects of making changes to factors such as number of items, marking criteria or markers. This information can be used to improve test design and increase reliability of the test as a whole (for example, see Kim and Wilson, 2009). Thus, “G-theory is a comprehensive method for designing, assessing, and improving the dependability of measurement procedures” (p. 422).

The Many-Facets Partial Credit Rasch Model (MFRM) has also been investigated as a method for estimating marking reliability. This model can identify markers who are harsher or more lenient than others, who use the mark scheme in different ways, and who make judgments that are inconsistent with those of other markers (for a helpful summary see Kim and Wilson, 2009).

Both G-theory and MFRM can be used to analyse data with multiple sources of error, but there are differences in their approaches. G-theory analyses multiple sources of error simultaneously and compares the relative influence of each one to provide an estimate of how reliably observed scores can be generalised to make inferences about a person (the Generalizability coefficient). MFRM, however, attempts to find the simplest best-fit model to allow an unbiased person estimate. That is, the results show how each item is marked differently by each marker. Kim and Wilson (2009) recommend that G-theory should be used if the aim of the analysis is to estimate the similarity between a set of observed scores and the scores that similar groups of students might obtain under identical circumstances. MFRM is better if the aim is to estimate a measure for each student that is as free as possible from the errors that affect the raw scores.

Kim and Wilson (2009) analysed data from a writing test using G-theory and MFRM. Both methods showed a significant difference in difficulty between the two items in the test. Both analyses also found that marker severity had only a small influence on the results and that marking was generally homogeneous.

Baird *et al.* (2012) used data from Geography and Psychology AS component papers for three consecutive years to directly compare three statistical techniques for estimating marker related unreliability, namely: G-theory; MFRM; and Multilevel Modelling (MLM). Both the G-theory analyses and the multilevel modelling analyses suggested that marker contributions to item mark variance were small in comparison with those from questions and other effects. The MFRM analyses, on the other hand, revealed significant inter-rater effects and a low reliability of measurement. However, it appears that the MFRM model did not fit the data well, possibly because there were interactions between variables (identified in the G-theory analysis) that were not included in the MFRM model. These results apply only to the particular assessments investigated by Baird *et al.* and cannot be generalised; however, other authors have also found striking differences in the results obtained when using these different methods of analysis (see Baird *et al.,* 2012, for a summary).

Classification consistency statistics can also be useful for describing marking reliability, these estimate the proportion of students who might receive a different grade under a different set of circumstances (in this context, if their work were marked by a different examiner) (Newton, 2009; Baird *et al*., 2011).

3.5.4 Graphical methods

Dhawan and Bramley (2012) and Bramley (2007) suggest a number of graphical and tabular ways to present marker agreement data that allow quick and easy interpretation. These graphs/tables show at a glance how accurate a marker is, whether there is a tendency for severity/leniency, and the size and frequency of larger marker discrepancies. Bramley and Dhawan (2010) use some of these methods to present actual data from the June 2009 GCE examination session.

4 Advances in improving marking reliability

4.1 What does this section cover?

Newton (2009) notes that, "... when exploring marking consistency evidence there are all sorts of issues to investigate, from the extent of clerical errors made when transcribing marks onto mark sheets to the possibility that a single marker might mark inconsistently across the marking period" (p. 191).

Meadows and Billington (2005) comprehensively reviewed the literature on the factors that affect marking reliability. These factors can be divided into three categories: those that relate to the items and their mark schemes; those that relate to the candidates and their responses; and those that relate to the markers. Recent work in these areas has attempted to refine our understanding of which factors have the strongest effect on marking reliability. The studies have also shown that some of the factors interact to affect marking accuracy, for example examiner experience and the complexity of the marking task.

Since Meadows and Billington's review, a large amount of work has been undertaken to investigate the cognitive processes involved in marking (Suto *et al.*, 2008) and how various factors that are known to affect marking accuracy influence this process.

Technological advances have resulted in a proliferation of on-screen marking, which, in turn, has facilitated item-level marking. A number of the studies that we reviewed highlighted the potential for on-screen marking to improve marking reliability.

In this section we also discuss the literature on the benefits and constraints associated with multiple marking.

4.2 Key findings

Our key findings about ways of improving marking reliability are:

- Increased item constraint, highly specified mark schemes, lower maximum marks and questions targeted at lower grades are all associated with increased marking accuracy;
- Marker education and experience affect marking accuracy, but the relationship is not simple and depends on item type;
- Item level marking is more reliable than whole script marking because it reduces the effects of examiner biases;
- On-screen marking appears to be as reliable as paper-based marking, even for long answer and essay questions;
- On-screen marking facilitates item-level marking and all its associated benefits;
- On-screen marking allows continuous marker monitoring, which enables inaccuracy to be detected early and corrected, and it eliminates errors resulting from incorrect addition or transcription of marks and prevents items being left

unmarked;

- Multiple marking has the potential to improve marking reliability for some question types. However, there are a number of obstacles to introducing multiple marking and it is unclear whether the benefits will outweigh the problems;
- The combination of double/multiple marks to produce a final score is an acknowledgement that legitimate differences in opinion can exist between examiners and is fundamentally different from the current system, in which the marks of the most senior examiner are considered to be the most 'true'.

4.3 What is the evidence base?

In addition to Meadows and Billington (2005), thirteen of the papers that we reviewed contributed to the section on the factors affecting marking accuracy: Ahmed and Pollitt (2011); Baker *et al.* (2008); Black (2010); Crisp (2010); Curcin (2010); Dhawan and Bramley (2012); Fowles (2009); Greatorex and Suto (2006); Kim and Wilson (2009); Massey and Raikes (2006); Meadows and Billington (2007); Pollitt (2012); and Suto *et al*. (2011).

Crisp (2010), Greatorex and Suto (2006) and Suto *et al.*, (2008, 2011) present a theoretical model of the cognitive processes used in marking and empirical support for aspects of the model.

Information about examiner training was included from: Baird *et al.* (2004); Meadows and Billington (2005, 2007); Raikes *et al.* (2010); and Suto *et al.* (2011).

Pinot de Moira (2011) and the review by Meadows and Billington (2005) provide the main contributions to the section on item-level marking. The discussion about on-screen marking draws from Billington (2012), Black (2010), Johnson *et al*. (2012), Myford and Wolfe (2009) and Pinot de Moira (2011). The information about automated scoring systems came from Blood (2011) and Raikes (2006). Multiple marking was discussed in a review by Brooks (2004) and we also include empirical studies by Fearnley (2005), Kim and Wilson (2009) and Vidal Rodeiro (2007). Finally, Baird *et al.* (2011) and Johnson (2008, 2011) all provide information about internal assessment.

4.4 Factors affecting marking accuracy

4.4.1 Features of items and mark schemes

Many studies have attempted to estimate marking reliability (see review by Meadows and Billington, 2005). Estimates of marker agreement from blind double-marking studies in O level, A level and GCSE range from a correlation between markers of 0.73 in English A level (Murphy, 1978) to 0.997 in GCSE Mathematics (Newton, 1996). Massey and Raikes (2006) found similar levels of marker agreement across a range of IGCSE and A level components. The Ofqual Reliability Programme commissioned a large number of empirical studies that provided information about reliability in a selection of Key Stage 2 National Curriculum tests, a range of GCE and

GCSE units, components and qualifications, and a number of vocational qualifications (Opposs and He, 2011).

In general, the highest levels of marker agreement were found in tests and examinations made up of highly structured, analytically marked questions, while the lowest levels of agreement tended to be found in examinations that placed most dependence on essay-type questions (see Meadows and Billington, 2005, for an extensive review; Baker *et al.*, 2008; Massey and Raikes, 2006; Black, 2010; Curcin, 2010; Dhawan and Bramley, 2012). Massey and Raikes (2006), however, found good levels of marker agreement for A-Level Sociology essays (mean intraclass correlation of 0.825), suggesting that it is possible to mark longer pieces of work quite reliably. The mark scheme for Sociology was much less detailed than that of Economics (for which the authors found lower mean ICCs), leading the authors to speculate that these findings support the theory that complex mark schemes challenge the examiner's working memory and inhibit the development of an accurate representation of the candidate's text.

Recent work has focused on fine-tuning the classification of items to discover more detail about the factors that affect marking reliability (e.g. Massey and Raikes, 2006; Black, 2010). Massey and Raikes (2006) were among the first to look at fine-grained features affecting marking reliability at item level. They used data from independent multiple markings of 300 scripts from each of five IGCSE and A level components. The authors found that marking accuracy decreased as the amount of time available to answer a question increased. This is probably because there is a positive relationship between the amount a candidate is expected to write and the time given to write it, and there is more scope for examiners to disagree on longer answers (Massey and Raikes, 2006).

Black (2010) used data from seed items to investigate how various features of these items influence marker agreement in five different GCSE and AS Level examinations. Overall, exact agreement ranged from 99.8 per cent (very high) to 69.2 per cent (low/moderate), and was similar to that found in other studies. Significant effects on marking reliability were found for 20 out of 21 features studied. As found in previous studies, the strongest effects were a result of item constraint and maximum mark. Marker agreement decreased as the constraint of the item decreased and the maximum mark increased (see also Meadows and Billington, 2005; Baker *et al.*, 2008; Massey and Raikes, 2006; Suto *et al.,* 2011; Dhawan and Bramley, 2012). Black (2010) separated question constraint (objective, constrained, short-answer, extended response) and mark scheme constraint (objective, points-based, levels-based). Items with objective mark schemes had highest marker agreement, followed by points-based mark schemes and then levels-based mark schemes (refer to appendix 4 for explanation of mark schemes). This finding supports the results of Massey and Raikes (2006).

Suto *et al.,* 2011, found evidence from IGCSE Biology that questions with a higher target grade, i.e. those that are more difficult for the candidates, were marked with lower accuracy.

4.4.2 Improving mark schemes

For valid assessment of a trait, the examination must elicit performance that demonstrates how good candidates are and the mark scheme must award more marks to those who are better (Ahmed and Pollitt, 2011). Ahmed and Pollitt (2011) outline a process for improving mark schemes (refer to appendix 4 for definitions of mark schemes) so that the marking is more accurate and better performances are awarded more marks. They begin by discussing the concept of '*outcome space'* which represents all the responses that candidates might produce. They argue that defining the outcome space is an important step in designing a good mark scheme. The better the overlap between the observed and expected outcome space (i.e. the more actual candidate responses, correct and incorrect, that are included in the mark scheme) the more reliable the marking will be. Black (2010) found that definition of outcome space had a strong effect on marker agreement for five components at GCSE and AS level.

Ahmed and Pollitt (2011) present a taxonomy of mark schemes, showing how they range in usefulness. Moving up the taxonomy, more help is given on how to mark borderline answers by making it clear what distinguishes better from poorer responses. The lowest level (Level 0) gives no help at all in assigning marks; it may just provide an example of a correct answer or a statement such as *'any acceptable answer*'. The next level (Level 1) describes what constitutes good performance, but gives no help with the difficult cases near the boundary. Level 2 attempts to specify the complete observed outcome space, including both good and poor responses. Level 2 may produce high marking reliability if the outcome space is well defined but only the highest level (Level 3) provides a principle for discriminating better from poorer responses.

Fowles (2009) found substantial differences in marking reliability between two different specifications for GCSE English from the same awarding body. The author suggested that differences in the nature of the mark schemes may partly explain this difference. Where there are many marks in a level, there may be less readiness to use the extreme marks, which would exaggerate differences between markers in their interpretation of the mark scheme. The author suggests that mark schemes with fewer marks in a greater number of levels may result in greater consistency.

In all of the marking we have so far described, examiners are required to assign a score based on some form of normative description of performance. Pollitt (2012) describes an alternative method whereby judges compare the work of two candidates and decide which is better. By making multiple comparisons of this sort, the candidates' work can be sorted into rank order very effectively and with high repeatability. The author argues that this is an improvement on attempting to apply an analytic marking scheme because humans are better at making comparisons than normative judgments. Crisp (2010) found evidence from Geography A-Level that suggested that, even when applying a mark scheme, examiners naturally make comparative judgments.

4.4.3 Response features

The evidence of the influence of superficial candidate response features, such as spelling and legibility, on marking reliability is mixed (Meadows and Billington, 2005). Some experimental studies involving teachers as markers have found that neater and more legible handwriting is associated with higher marks (see examples in Meadows and Billington, 2005). More recent studies of marking by experienced examiners seem to show that some of these non-relevant features have a smaller influence than previously thought (Meadows and Billington, 2005; Massey, 1983; Black, 2010; Bramley, 2009). It appears that well-defined mark schemes and good examiner training and standardisation reduce the influence of presentational style on marking reliability. However, in a large study of seed scripts from live GCSE and GCE examinations, Black (2010) found that marking accuracy was reduced by the presence of crossings-out and if part of the response was outside the designated space.

4.4.4 Characteristics of the markers

Meadows and Billington (2005) found evidence in the literature for marking biases that stem from characteristics of the marker him/herself. However, there was no consistent association between aspects of a markers' background and marking reliability. The largest body of literature related to marker experience, with a number of studies finding that inexperienced markers were more severe and/or less accurate than experienced markers. For example, Baker *et al.* (2008) found some evidence that experience of marking a state-level high-stakes examination improved the reliability with which Australian teachers marked key stage 3 English tests. However, this pattern has not been found in all studies and, where differences occurred, they could often be negated by training (Meadows and Billington, 2005).

Meadows and Billington (2007) noted that previous studies generally failed to separate the effects of markers' subject knowledge, teaching experience and marking experience on marking consistency. Thus, in their study of marking reliability in GCSE English, Meadows and Billington (2007) attempted to separate these effects. They found that both subject knowledge and some experience of teaching seemed to increase marking reliability in GCSE English. However, detailed analysis showed that the effects of marker background depended on the item that was being marked. Short answer items were marked as reliably by PGCE English students as by experienced examiners, but items that required longer answers were marked most reliably by the experienced examiners.

For GCSE mathematics and science, highest level of marker education was the best single predictor of marking accuracy, followed by teaching experience and then marking experience (Suto *et al.,* 2011; Suto and Nadas, 2008).

There have been very few studies of the relationship between personality traits and marking performance. The studies that exist tend to be small scale and use rather ambiguous personality measures, leading Meadows and Billington (2005) to conclude that these studies do not allow "sensible interpretation of the effect...on marking reliability*"* (p. 34). Meadows and Billington (2007) undertook a more rigorous investigation of the effects of personality traits and attitudes on marking accuracy in

GCSE English. They found a weak positive relationship between marking reliability and the psychometric measures of '*agreeableness*' and '*conscientiousness*'. The authors warn against using these traits to try to select more reliable markers until the findings have been replicated.

Meadows and Billington (2007) found that older participants tended to mark certain items more reliably than younger ones. The authors highlighted the need for more research to discover why this should be the case. In addition, for GCSE English, and IGCSE Biology, male participants tended to mark certain items more reliably than female participants, and vice versa (Meadows and Billington, 2007; Suto *et al.,* 2011). Again the authors emphasise that more research is required to understand these effects and warn against using age or gender to select examiners.

4.4.5 Marker training

Suto *et al.* (2011) found that whether the marker had a degree (relevant or not) made the biggest difference to marking accuracy for IGCSE Biology items and suggested that this could be used to select who should become examiners. The authors also found evidence that, with the correct training, some individuals with only GCSEs or A Levels could become accurate markers of certain items (Suto *et al.,* 2011).

There are very few empirical studies to assess which parts of training are effective and why. Of the few studies that do exist, some show a positive effect of training, while others found no lasting effects (see Meadows and Billington, 2005).

Meadows and Billington (2007), in their study of GCSE English marking, found that the effects of training differed between groups of markers with different subject expertise and teaching experience. There were also differential effects by item type. In fact, on some items the training appeared to confuse the more experienced markers. It appeared that the training had caused the markers to become more 'cautious' and less willing to use the extremes of the mark scale, thus bunching the results in the middle of the mark range. This is clearly an undesirable effect and the authors recommend that more research should be conducted to identify factors that can be used to improve training (Meadows and Billington, 2007).

Raikes *et al.* (2010) and Baird *et al.* (2004) both investigated the effectiveness of examiner standardisation, i.e. the process by which the marking criteria are explained and exemplified to the examiners before they begin marking their allocation.

Baird *et al.* (2004) found that standardisation did not improve marking accuracy for GCSE English. Three groups of 15 experienced examiners each marked the same 150 essay responses. Prior to marking, two of the groups were sent exemplar scripts, which they marked and returned to the principal examiner. The principal examiner returned the exemplar scripts to the examiners, together with the 'correct' marks and an explanation for them. One of the groups received 'prototypical' band exemplars (scripts with marks in the middle of the band) and the other group received cut score exemplars (scripts with marks at the bottom of the band: only just worthy of the mark band). The examiners in the exemplar script groups were asked to use the marked exemplar essay responses to guide their marking. Surprisingly, lack of exemplar scripts did not make marking less accurate. Baird *et al.* suggest that this may be

because the examiners who participated in the study had enough experience that they could mark an unfamiliar question paper simply using the mark scheme. However, exemplars did have some effects: cut score exemplars led to slightly generous marking, although not significantly so, while the prototypical exemplars produced severe marking. The authors suggest that this might be because examiners are accustomed to thinking about cut-score performances and so the prototypical exemplars might be interpreted as cut-score performances by the examiners in the study. As the prototypical exemplars were on higher marks than the cut-score, this would result in more severe marking.

In contrast, Raikes *et al.* (2010) found positive effects of standardisation for two GCE Psychology components. The researchers provided a test group of 24 examiners (12 new and 12 experienced) with standardisation materials comprising scripts for the examiners to mark and annotated exemplar scripts with which to compare their own marking. Half of the examiners also attended a face-to-face meeting at which the mark schemes were explained. Raikes *et al.* found that, for both new and experienced examiners marking both short-answer and structured essays, marker accuracy was improved by the standardisation process. In addition, standardisation on one set of questions produced improved marking accuracy on other, very similar questions. However, the benefit of including a face-to-face meeting was "*variable, small and questionable*" (Raikes *et al.*, 2010, p. 22).

Baird *et al*. (2004) also found that face-to-face standardisation meetings made little difference to marking accuracy. The researchers investigated the marking reliability of experienced GCSE History examiners who attended either a consensual standardisation meeting (where the examiners agreed a common interpretation of the mark scheme) or a hierarchical meeting (where the principal examiner passed his view of the mark scheme on to the examiners). The authors found no significant difference in marker agreement between examiners who had attended a meeting (of either type) and those who had not. Similarly, Greatorex and Bell (2008) found that a standardisation meeting on its own had little effect on the reliability of experienced examiners of AS Biology. However, in the study by Baird *et al.* (2004) most of the examiners who did not attend a meeting stated that they missed having it, suggesting that if examiners are deprived of the opportunity to discuss the candidates' work in relation to the mark scheme they could feel isolated (Baird *et al.*, 2004).Greatorex and Suto (2006) suggest that information about marking strategies should be made explicit in training courses for new examiners. The authors suggest that inexperienced examiners may gain insight into expert examining by listening to recordings of senior examiners 'thinking aloud' while they are marking.

Real time examination marking is monitored and examiners receive feedback during the marking process. Meadows and Billington (2005) found some evidence in the literature suggesting that feedback increased marking reliability, but other studies found no positive effect. Kim (2010) found that repeated training and feedback sessions were necessary to achieve internal consistency, even for experienced examiners, when marking a second language speaking assessment. Baird *et al*. (2011) argue that the training meetings provided by awarding bodies may not be long enough to provide suitable practical exercises. Thus, ongoing training and feedback

have been suggested as essential for producing internally consistent marking behaviour (Kim and Wilson, 2009). The continuous monitoring processes employed in on-screen marking go some way to providing this.

4.5 Cognitive processes used in marking

Suto *et al.* (2008) conducted various inter-related studies investigating the process of marking GCSE and A level examinations. Greatorex and Suto (2006) identified five distinct cognitive marking strategies: *matching*, *scanning*, *evaluating*, *scrutinising*, and *no response*.

Matching	The examiner simply compares the letter(s)/number(s)/single word/part of diagram written by the candidate on the short answer line/ pre-determined spot in the answer space with those given in the mark scheme. [7]
Scanning	The examiner scans the whole of the space allocated to the question to find a key detail, which may either be visually recognisable, for example a number, letter or word, or more complex, for example a phrase, or calculation.
Evaluating	The examiner considers the truth/accuracy/meaning of what the candidate has written, evaluating the response using knowledge and information from a combination of sources.
Scrutinising	When a response is unexpected and/or wrong and/or not the same as any of those given in the mark scheme, the examiner may need to establish precisely where the problem lies, or whether the response is actually a correct and valid alternative. The examiner's overall aim is to reconstruct the candidate's line of reasoning or establish what the candidate has attempted to do.
No response	When the candidate appears to have written nothing at all, the examiner looks over the space allocated to the question more thoroughly to confirm this.

The same five strategies were used by examiners across three different subjects at two different levels (GCSE Mathematics, GCSE Business Studies and A-Level Physics) and for on-screen and paper-based marking (Greatorex and Suto, 2006). Greatorex and Suto (2006) showed that markers with an undergraduate degree in the subject being marked or a related subject, but with no experience of teaching or examining, used the same marking strategies as 'expert' markers and so could potentially become experts. Analysis of strategy use showed that, although there were some differences between individuals, the biggest differences occurred between subjects and questions (Suto *et al.*, 2008; Greatorex and Suto, 2006). Question type and mark scheme approach are key factors in determining cognitive

[7] Definitions in this table are from Greatorex and Suto, 2006, pp.8–12.

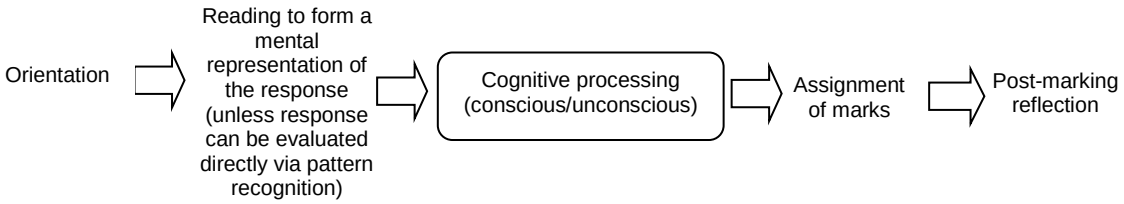
marking strategy complexity, but the strategy actually used will depend, in part, on what the candidate has written (Black *et al.*, in submission).

For individual items, the apparent marking strategy complexity (AMSC) that a question requires is strongly associated with marker accuracy (Greatorex *et al.*, 2007; Suto *et al*., 2011; Black, 2010).

Suto *et al.* (2008) suggest that marking accuracy for a particular question is determined by both the marker's expertise and the demands of the marking task. Suto *et al.* (2011) found empirically for IGCSE Biology that marker experience was associated with greater accuracy, and that this trend was more pronounced for questions requiring more complex marking strategies than for those requiring simple marking strategies.

On the basis of transcripts of examiners 'thinking aloud' as they marked, Crisp (2010) proposed a five stage model of judgment and decision-making underpinning A level marking for short and medium length responses and essays. She suggested that the model could provide a framework for investigating reasons for lower marking accuracy and consistency, and could provide a basis for mark scheme writing and examiner training in terms of anticipating marker judgment behaviours (Crisp, 2010).

Core process of marking (Crisp, 2010, diagram adapted from Suto *et al*., 2008)



This process will be influenced at different points by a number of variables, including teaching experience, subject knowledge, candidate response features, the complexity of the marking strategies needed to mark the question and whether the examiner knows which strategy to apply (Suto *et al*., 2008).

4.6 Technological advances in marking processes

4.6.1 Item-level marking

Pinot de Moira (2011) argues that item-level marking will always be more reliable than whole script marking, because it reduces the effects of individual examiner idiosyncrasies. When a single examiner marks a whole script, it is likely that there will be some relationship between the degree of leniency (or severity) shown on one item with that shown on another item. However, if several examiners contribute to the final mark of a complete paper, then there will probably be little or no relationship between items in terms of deviation from the true mark. That is, for each mark that is too high, there is likely to be one that is too low, and thus many of the marking errors will cancel one another out. The more examiners who contribute to the final mark, the more reliable the marking will be.

Item-level marking removes a number of biases that have been shown to affect marking accuracy at whole script level (Pinot de Moira, 2011). When marking a whole script, an examiner may carry forward preconceived, and possibly incorrect, ideas about a candidate's understanding based on answers to previous, unconnected items. Investigations into the thought processes of examiners show that they sometimes reflect upon the leniency or severity of their marking and may explicitly use compensation (e.g. "I was generous before so...") (Crisp, 2010, p. 15). This is known as the halo effect, and it is eliminated by the use of item-level marking.

There is still the potential for contrast effects, where the marking of a response is affected by the quality of the immediately preceding responses. Research has shown that a particular response is marked more severely if the responses preceding it are of high quality than if they are of low quality, and vice versa (see Meadows and Billington, 2005). In item-level marking, the responses are presented in a random order, and so any bias stemming from contrast effects will not be systematic. That is, the bias affecting one item from a script is likely to be different from that affecting another item on the same script (Pinot de Moira, 2011).

When an examination is marked at item-level, individuals with differing backgrounds can mark different items. For instance, items with a range of acceptable answers that can be fully defined could be marked by individuals who do not necessarily have the experience to mark more complex items. This is known as 'general' or 'clerical' marking. Some short answer items could be marked by individuals with subject knowledge but little or no teaching experience (Meadows and Billington, 2007; Suto *et al.,* 2011). In contrast, those items that elicit longer answers and require more complex marking strategies could be marked by experienced teachers and examiners (Pinot de Moira, 2011; Meadows and Billington, 2007; Suto *et al.*, 2008; Suto *et al.,* 2011).

4.6.2 On-screen marking

On-screen marking (or e-marking) is the process by which candidate scripts are scanned into digital format and sent to examiners for marking on-screen, via a secure system. This process has great potential for improving the marking of high-stakes examinations, not least because it facilitates item-level marking.

A number of studies have shown that characteristics of the candidate, such as gender, ethnicity and even first or given name, can affect the marks awarded (reviewed by Meadows and Billington, 2005). When item-level marking is conducted on-screen, candidate details are not provided and so biases related to the examiner knowing the candidate's name or school, and making inferences from these, are reduced. Some authors (e.g. Baird and Bridle, 2000) have argued, however, that concealing candidates' names from examiners does not completely solve these problems because handwriting style, content and style of language reveal personal characteristics of the candidates. In addition, the presence of crossings-out and parts of the responses being outside the designated area (although still visible on the screen) were both associated with reduced marking accuracy for a range of GCE and GCSE examinations (Black, 2010).

Delivering scripts electronically removes the economic and logistic burdens of sending paper scripts to and from examiners. Scripts remain in the possession of the awarding body and cannot be lost or damaged in transit. Online delivery also allows for more flexibility in marking allocations; examiners are sent scripts/items when they are ready for them and, if an examiner cannot continue marking, the scripts do not have to be retrieved and sent to someone else.

Examiner monitoring in the on-screen marking process is superior to the method used in paper-based marking for a number of reasons. First, problems can be identified earlier in on-screen marking and items that have been marked by unreliable examiners can be quickly re-allocated, removing the need for mark adjustment. Second, the continuous monitoring process in on-screen marking can identify examiners whose marking standards drift over time (Myford and Wolfe, 2009). Third, in the paper-based system, senior examiners are re-marking items with full knowledge of the marks/comments of the first examiner; this can lead to overestimation of marker accuracy (Billington, 2012). In on-screen marking the examiner's marks are compared with pre-determined definitive scores for the items, which produces a better estimate of marker accuracy.

Examiners marking on screen input their marks directly into the system. This eliminates errors associated with incorrect addition or transcription of marks, and prevents items being left unmarked.

However, all of these advantages would be irrelevant if something about the process of on-screen marking made it inherently less reliable than paper-based marking. Studies of the cognitive processes of reading, understanding and assessing texts suggest that reading mode (i.e. on screen versus on paper) might affect marking accuracy (see Johnson *et al*., 2012 for a summary of these arguments). Encouragingly, empirical studies investigating whether the mode in which short essays (150-600 words) are marked found that mode had negligible effects on marking accuracy (see references in Johnson *et al.,* 2012). Johnson and Nádas (2009b) found that examiners' reading behaviours were different when reading on screen than when reading on paper, and that their cognitive workload was heightened while marking on screen (Johnson and Nádas, 2009a). These findings led Johnson *et al.* (2012) to question whether marking accuracy would remain unaffected by marking mode for essays that were longer than 600 words. Using 180 GCE A-level American History scripts, the researchers showed that, despite the increased cognitive demands, examiners were able to mark extended essays with equal accuracy on screen as they do on paper (Johnson *et al.*, 2012).

On-screen marking has the potential to make the single biggest improvement to marking reliability. However, it is not without practical constraints. On-screen marking requires examiners to have a home computer and suitable internet access (Pinot de Moira, 2011). Examiners need to be trained to use the system and will require some time becoming familiar with the new processes and practices, but this issue should be short-term (Pinot de Moira, 2011).

4.6.3 Computer-based automatic marking

The first automated scoring system was introduced in 1938 for multiple choice tests and led to an increase in the use of this form of testing because marking could be conducted quickly and accurately (Blood, 2011). Since then, attempts have been made to produce systems that can mark longer answers. Blood (2011) provides a useful review of the use of automated systems for marking essays.

The first system for automatically scoring essays was developed in the 1960s and was known as Project Essay Grade (PEG). PEG assigned scores by analysing linguistic surface features such as number of words, word length, number of paragraphs and number of apostrophes. It took no account of the content of the essay, and yet the system scores agreed with those of a group of human markers as well as, or better, than the scores of individual human markers (Page, 1968). A more sophisticated version of this type of scoring system is E-Rater, which is used by ETS for high and low stakes testing situations. E-Rater can take account of grammar and perform topical analyses. Again, this system performs very well, producing high correlations between its scores and the combined scores of human markers (Attali and Burstein, 2006).

A further development is a latent semantic analysis-based approach (LSA), which ignores superficial surface features entirely and instead analyses the statistical relationships between meaningful units (i.e., words, sentences, paragraphs, and texts), thereby 'learning' what knowledge or concepts are contained in a piece of writing. These systems are designed to attend to essay content rather than style, and measure the similarity between a piece of writing of known quality and a candidate essay. The Intelligent Essay Assessor (IEA) is an example of this type of system and shows levels of reliability comparable to human markers and within generally accepted ranges for the testing purpose (Foltz *et al.,* 1999).

Another method of automatic scoring is text categorisation. This method divides previously scored essays into categories of quality and an algorithm is used to predict which quality category a newly scored essay would most likely be assigned to by a human marker. Larkey (1998) developed an automatic essay marking system that combined text categorisation technology with identification of linguistic surface features, with good results.

Raikes (2006) provides a description of a system that uses computational linguistics techniques for automatic marking of short, free response questions in GCSE Biology. The system was found to mark with high accuracy in many cases (judged as the agreement between the system mark and the 'correct' mark for the question as determined by a senior examiner). These results led the author to conclude that automatic marking is promising for 1-mark items requiring a short, textual response, but that more work is needed if these findings are to be generalised.

There are some obstacles to the introduction of automatic scoring systems. First, there has been much debate about whether automated systems are a valid way to mark essays. Supporters of the systems argue that the scoring rules are appropriate (and therefore valid) if either the marks awarded reflect the *product* of human ratings

or if the systems employ similar *processes* as those used by human markers (see Blood, 2011).

A second problem is that automatic scoring systems may be open to abuse. It is possible that, if students are aware that their essays will be marked by a machine, they will focus on those features (such as essay length and vocabulary choice) that most strongly affect the scoring. Although some systems can flag essays that are highly creative, off topic, in violation of standard formatting or too similar to another essay, and may therefore be able to identify essays written in '"bad faith"', this is still a valid concern (Blood, 2011, p. 49). However, automated systems may still be useful in high stakes situations as a second (or multiple) marker, especially as they are highly consistent (i.e. will always give the same score to the same piece of work) and fast.

4.7 Multiple marking

The first study of 'multiple marking' was conducted by Wiseman in 1949 (see also Meadows and Billington, 2005, for a thorough review). Wiseman combined the independent scores of four markers to produce a final mark for each script. This method differed markedly from previous methods because individual markers were not required to agree with one another. Wiseman argued that, provided the markers are self-consistent, differences in the marking provide a "truer 'all-round' picture" of the candidate's ability (Wiseman, 1949). Thus, diversity of opinion between markers becomes a virtue of the system rather than a problem. Another key feature of Wiseman's method is that markers were trained to use general impression marking. Wiseman claimed that this would speed up the process so that multiple markings could be achieved for the same amount of time/effort as a single analytic marking.

Further empirical and theoretical studies supported Wiseman's claim that multiple marking is more reliable than a single mark (e.g. Britton *et al.*, 1966; Pilliner, 1969; Head, 1966; Lucas, 1971; Wood and Quinn, 1976). Pilliner (1969) showed statistically that, provided there was reasonable agreement between the markers, the aggregation of multiple marks is a valid expression of the marking team's consensus and that reliability will increase as team size increases. Lucas (1971) investigated the relative gains to be made by scaling up the number of markers from one to two to three to four and found that the greatest increase in reliability came from increasing the size of the marking team from one to two markers. Additional benefits of using larger teams were statistically significant, but of smaller magnitude. Other authors confirm the finding that 'double-marking' can increase reliability (see references in Brooks, 2004). More recently Kim and Wilson (2009) used data from written compostitions and Generalizability theory to illustrate that increasing the number of markers beyond two had little effect on reliability.

Fearnley (2005) investigated different models of double marking in one component from each of GCSE English and GCE Business Studies (AS level). For each component, 100 scripts from the Summer 2004 examination were marked by two groups of examiners. The first group, comprising one senior examiner and 16 assistant examiners marked all of the scripts independently. The assistant examiners were then paired (either randomly or strategically) and examiner agreement was

calculated as the difference between the mean marks of the examiner pair and the senior examiner's mark. The second group comprised seven pre-selected examiner pairs. The first examiner of each pair marked each of the scripts, annotating them with comments and marks. The second examiner then re-marked the scripts with full knowledge of the original annotations (non-blind re-marking). Marker agreement was calculated as the difference between the mean mark of each examiner pair and the independent senior examiner's mark.

The results showed a lot of variation in the mean marks awarded by the examiners on the same 100 scripts. There was a small, but significant, increase in marker agreement when paired marks were used. That is, the difference between the independent senior examiner's mark and the mean of the paired marks was smaller, on average, than the difference between the senior examiner's mark and the individual examiner's marks. This increase in marker agreement was largest for randomly allocated pairs of examiners marking independently and smallest for pre-selected pairs with non-blind re-marking, with no gain at all for non-blind double-marking in the Business Studies unit. Fearnley (2005) suggested that the much lower increase in agreement for non-blind double-marking is likely to be the result of the second examiner being influenced by the annotations of the first examiner. This view is supported by relatively high correlations between the marks in each examiner pair in the non-blind exercise. Other studies comparing blind and non-blind re-marking have shown similar increases in correlation between examiner marks when the second marker could see the annotations of the first (e.g. Vidal Rodeiro, 2007; and references in Bramley and Dhawan, 2010).

Double marking should be targeted at examinations where genuine benefit can be demonstrated (Brooks, 2004). In subjects such as mathematics, where high levels of inter-marker reliability already exist, double-marking would serve little purpose (Brooks, 2004). In Fearnley's study, random pairing significantly improved the agreement of the marks of a quarter of the examiners, at best. In addition, the marks of one examiner actually agreed less with the senior examiner after pairing. This led Fearnley (2005) to question whether the gains in marker agreement found in the study would justify the introduction of double-marking. There are a number of obstacles to the introduction of multiple marking, including recruitment of enough examiners, cost implications, time constraints and logistical problems (Brooks, 2004; Fearnely, 2005; Vidal Rodeiro, 2007). The logistical problems of transporting the scripts from one examiner to the next and the additional systems and paperwork required to keep track of this process are mostly eliminated by on-screen marking (Brooks, 2004). Similarly, on-screen marking allows the same script to be independently marked by multiple examiners simultaneously, thus removing the problem of the additional time it would take to have the script marked by two or more examiners one after the other (Brooks, 2004).

Proponents of multiple marking argue that two or three markers making a holistic judgment takes the same number of person-hours as a single marker using an analytic mark scheme and so multiple marking may not take longer, cost more or require more examiners than single analytic marking (Brooks, 2004). Meadows and Billington (2005) discuss some literature about the pros and cons of holistic and

analytic mark schemes. However, our review found no new information about the relative reliability of holistic marking (multiple or otherwise) compared with analytic marking. The authors who did touch upon this issue suggest that holistic marking is less valid (Blood, 2011), and possibly less reliable (Ahmed and Pollitt, 2011), than analytic marking. For example, Blood (2011) describes a study by Shi (2001) that showed that, while markers using a holistic scheme did not differ significantly in the scores they assigned, they did differ greatly in the justifications for their scores. This suggests that, even though the assigned scores were similar, the markers did not share a common understanding of what it means to be 'good'. The author concludes that this finding illustrates that reliability does not necessarily equate to construct-validity, and argues for the use of analytic mark schemes that encourage more thorough and balanced attention to the construct.

Ahmed and Pollitt (2011) make a similar argument in their production of a mark scheme taxonomy for unconstrained items. They argue that a "holistic implicit levels" (p. 273) mark scheme, which requires markers to make an overall assessment of the student's complete performance without explicitly weighting individual aspects, will be less reliable than an analytic levels mark schemes. The authors argue that the implicitness of the holistic mark scheme is the main source of marker unreliability, stating that it is "appropriate when a child must be classified as wholly belonging to one 'best fit' category, but is seldom appropriate for a single question in an exam" (p. 274). However, no empirical evidence is presented to support these assertions.

In contrast, Baker *et al*. (2008), in their study of international transferability of national curriculum key stage 3 English marking, found that Australian markers, who were used to holistic mark schemes, expressed concerns about the difficulties of using an analytic mark scheme. Many markers found the different strands of the mark scheme and their associated multiple criteria disconcerting, and expressed concerns that the marks they had awarded to one strand influenced their marking of other strands.

Fearnley (2005) defined reliability as the agreement with the independent mark of a senior examiner. However, some authors argue that double marking provides a more valid estimate of the candidate's true score than any single mark, even that of a senior examiner (see Brooks, 2004). This is especially true where a correct answer cannot be defined unambiguously and, therefore, legitimate differences of opinion can exist about what constitutes a 'good' answer. If double/multiple marking were to be introduced there would need to be further consideration of the best way to combine the scores to produce the final mark (Vidal Rodeiro, 2007). In addition, the combination of independent marks represents an acknowledgement that differences of opinion are legitimate and is a philosophical shift from the stance that there is one 'correct' mark for a piece of work and that it can be assigned by a single examiner.

4.8 Teacher assessment

Externally set and assessed tests are by their nature limited in scope. Aims in the curriculum that require performance assessment, for example musical skills or speaking and listening in languages, cannot be assessed by externally set and marked written tests. Current systems of high-stakes testing rely on teachers to

assess these skills using tasks that are set or constrained by the awarding bodies, to a greater or lesser extent depending on the subject (Johnson, 2011; Baird *et al*., 2011). However, there are few empirical studies that provide robust estimates of the levels of reliability of internally assessed components (for reviews, see Stanley *et al.*, 2009; Harlen, 2005). The results of the studies that do exist vary depending on the nature of the evidence being assessed (scripts, portfolios, oral interactions, practical performances) and the tightness of the criteria used for assessment (mark schemes, level descriptors, etc) (Stanley *et al.*, 2009; Harlen, 2005; Johnson, 2011), and so do not allow any general conclusions to be drawn. The Ofqual GCSE English 2012 Report also suggests that teacher marking is susceptible to pressures that encourage teachers to strive for the best outcomes for their students and their schools (Ofqual, 2012b).

The reliability of portfolio assessment has received some attention. Johnson (2011) identified two studies that compared the judgments of a number of experienced assessors but the candidate evidence assessed was just one portfolio in one case (Johnson, 2008) and two portfolios in the other (Greatorex, 2005). Thus, the findings cannot be generalised.

Johnson (2011) argues that for internally assessed components where marker variation is known or suspected, the best strategy to maximise reliability is to employ multiple markers. The author acknowledges that this strategy would be expensive, time consuming and difficult to implement but argues that it may be the only way to ensure candidates receive the correct mark.

In high stakes examinations a moderator is often employed by the awarding body to re-mark a sample of teacher-marked work. In some cases more than one moderator may be used. Where inter-moderator differences are found the awarded mark should be the average of the moderator marks and the original teacher's mark could also be included in this averaging process (Johnson, 2011). Johnson (2011) goes on to argue that the internally assessed work of all candidates should be double marked in this way.

To be a valid assessment the teacher assessment must be reliable. At present, the data does not exist to judge this in high-stakes examinations in England.

5 Detecting and correcting unreliable marking

5.1 What does this section cover?

The information gained from monitoring an examiner's marking is used to determine the accuracy and consistency of that marking. If systematic tendencies toward leniency or severity are discovered, the examiner's marks are all adjusted. Inconsistent marking (i.e. unsystematic errors) cannot be adjusted for and the items would have to be remarked.

In this section we discuss the processes used for monitoring traditional paper-based marking and on-screen marking. We also consider the methods by which inaccurate marking may be corrected.

5.2 Key findings

Our key findings about the evidence on detecting and correcting unreliability are:

- The process for monitoring paper-based marking has two potential flaws: the non-independent re-mark might overestimate marker agreement and the isolated sampling process might fail to identify changes in marking behaviour over time;
- On-screen marking facilitates continuous monitoring and removes the need to adjust scores;
- Marking is considered unreliable if it falls outside a given tolerance. Estimates of marking accuracy for different question types can be used to set the tolerance at the right level to ensure that marker monitoring is effective and fair.

5.3 What is the evidence base?

Ten of the papers that we reviewed contribute to the discussion on marker monitoring processes: Al-Bayatti and Jones (2005); Baird *et al*. (2012); Billington (2012); Black (2010); Fearnley (2005); Johnson (2011); Kim and Wilson (2009); Myford and Wolfe (2009); Suto *et al*. (2011) and Val Rodeiro (2007).

Black (2010), Dhawan and Bramley (2012), Johnson (2011) and Meadows and Billington (2005) all provide evidence for use in determining tolerance levels. Our review did not include any primary research on adjusting candidate scores, but the reviews by Meadows and Billington (2005) and Johnson (2011) both touch on this subject.

5.4 Externally marked components

Immediately after their training, examiners must mark a sample of items (standardisation sample) and submit the items to a more senior examiner for checking. The senior examiner ensures that the examiner is applying the mark

scheme accurately and consistently, and gives feedback to the examiner. If necessary, examiners must re-mark some of their standardisation sample, or mark a second sample until they reach a pre-determined level of accuracy. Examiners should not complete their marking until they have received clearance from the senior examiner. In a study of Biology IGCSE marking, Suto *et al.* (2011) found that experienced markers reached acceptable levels of marking accuracy after less standardisation than less experienced markers. The authors also showed that markers who failed to reach suitable levels of accuracy within two standardisation samples went on to become unreliable markers.

Further monitoring takes place during the marking period; paper-based marking and on-screen marking have different marker monitoring processes, described below.

5.4.1 Monitoring paper-based marking

After standardisation, a second sample (and sometimes a third sample) is taken during the marking process to check that the examiner's marking is accurate and consistent. Awarding bodies differ in the exact details of the sampling process, but in all cases, the sample is supposed to cover a good range of candidate performance and answer types. In general, the senior examiner will re-mark 15 of the sample scripts and, if the marking is outside the tolerance allowed for that particular paper or shows a pattern of consistent severity/leniency, will re-mark an additional 10 scripts. This sample of 25 re-marked scripts is used to make decisions about whether the examiner's marks need to be adjusted or included in review processes, or whether the examiner's allocation needs partial or total re-marking.

It is important to note that, when checking the marking, the senior examiner can see all the marks and annotations made by the first examiner. That is, the re-mark is not independent of the first mark. Various authors have demonstrated that marking consistency is higher when the second examiner can see the marks of the first than when re-marking is performed 'blind', that is, when the first examiner's marks/comments are not available to the second examiner (e.g. Murphy, 1979; Wilmut, 1984; Massey and Foulkes, 1994; Vidal Rodeiro, 2007; Fearnley, 2005).

Billington (2012) used GCE Sport and Physical Education papers to investigate different marker monitoring processes. She took as her starting point the conventional method of marker monitoring employed by AQA (known as second phase sampling or SPS) – in which examiners select and mark a sample of paper scripts, which are then sent, complete with marks and annotations, to be re-marked by a team leader. This conventional method was compared with two alternative second phase samples - in which the examiners marked scripts that had been selected, and previously marked, by the Principal Examiner. In the latter cases, one sample was marked on-screen and a second sample was marked on paper, but in both cases marking was conducted 'blind'.

Billington (2012) showed that marking appeared to be more accurate when judged using the conventional SPS method than when examiners were re-marking cleaned scripts (either on paper or on screen) for which a definitive score had already been determined. However, the blind re-marking, either on paper or on screen, produced a better estimate of real time marking than did the conventional paper SPS. Billington

suggested that the conventional SPS method tended to overestimate reliability because the second examiner's marks are biased by the presence of the annotations.

Billington (2012) argued that the conventional SPS may not be the most suitable method on which to base decisions about mark adjustments. She went on to suggest that pre-selected, common scripts should be used as SPSs because they provide a common basis for comparing examiners and remove the need for senior examiners to re-mark lots of scripts.

Al-Bayatti and Jones (2005) investigated the effect of sample size on the estimation of marker reliability. Unsurprisingly, more precise estimates of the differences between the senior and the assistant examiner's marks were achieved with larger sample sizes. However, there were diminishing returns and beyond a certain sample size the gains in precision were small. The authors calculated the minimum number of scripts that should be marked to detect marker differences of various sizes (in Key Stage 3 English written component with a maximum mark of 30). For example, for an experienced examiner a sample of 13 scripts would be required to detect a two-mark difference between the assistant examiner and the senior examiner, but only seven scripts are required to detect a three-mark difference.

The sample sizes necessary to achieve a particular level of precision decreased with examiner experience. Thus, examiners with more experience could be required to submit smaller samples for re-marking (Al-Bayatti and Jones, 2005). The sample size required is, however, sensitive to the values of the standard deviations of the mark difference between the senior and assistant examiner. Therefore, a lot more data would be required before any firm recommendations could be made regarding appropriate sample sizes for marker monitoring (Al-Bayatti and Jones, 2005).

Another issue for consideration is that a single sampling occasion cannot identify changes over time in marking accuracy. Myford and Wolfe (2009) developed a form of MFRM (Multi-Faceted Partial Credit Rasch Model) that allows analysis of changes in marking behaviour over time. The authors also present various statistics that can be used to indicate changes over time in the accuracy of marking or scale usage by an examiner (e.g. whether the examiner becomes more, or less, likely over time to use the extreme ends of the marking scale). Myford and Wolfe (2009) used these techniques to examine data from the College Board's 2002 Advanced Placement English Literature and Composition Examination. They found that most markers showed little evidence of change over time in their accuracy and marking scale usage. However, some markers did show statistically significant changes in their accuracy as marking progressed, while a larger number of others appeared to change the way in which they used the mark scheme.

Baird *et al.* (2012) and Meadows and Billington (2005) describe a number of other studies that show changes in severity of individual markers over time. While no pattern has emerged from the various studies, it is becoming apparent that instability exists in marking behaviour. This has implications on the estimation of marking reliability as a whole and how the findings should be interpreted or generalised (Baird

et al., 2012). Kim and Wilson (2009) argue that regular monitoring and feedback are required to reduce any effects of changing marking behaviour.

5.4.2 Monitoring on-screen marking

In contrast to paper-based marking, examinations that are marked on-screen are continually monitored by the routine inclusion into the marker's allocation of 'seed scripts/items' for which a definitive mark has already been determined. Examiners who fail to mark the seed items sufficiently accurately are suspended from marking until they have received feedback from a senior examiner. If an examiner continues to mark a particular item inaccurately, they are stopped from marking that item and responses to that item that they have already marked are sent for re-marking. If the problem is more widespread, the marker is removed from marking completely and all the responses marked by that examiner are sent for re-marking. This process therefore removes the need for mark adjustments.

In addition, senior examiners may access their team's completed marking to check it, or may have marked scripts forwarded to them by the system. In these instances, the senior examiner is able to view the marks and annotations of the markers. This provides another means of monitoring marking but is not a good method for estimating reliability because the re-marking is not blind (Baird *et al.,* 2012).

Seed items provide the main mechanism for checking marking accuracy and, therefore, the quality of the seed items is clearly important. Black (2010) argues that, while relatively straightforward and uncontroversial items may be sufficient for checking marking, they may not provide the best opportunities for feedback. She suggests that, for the purposes of providing feedback, it might be useful to include responses with crossings-out, or that are outside the designated answer space (two features that Black found to have influence on marking reliability in GCSE and GCE examinations). Non-standard responses are marked less reliably than standard responses because they are usually outside the scope of the mark scheme (Black, 2010), but Black questions whether they should be routinely included as seed items because, by definition, they are unlikely to be encountered.

5.5 Internally assessed components

For components that are assessed by teachers, external moderators evaluate samples of candidates' work in each subject from each centre. On the basis of this sampling, the centre's marks may be accepted without change or accepted with an adjustment, or a total remark may be requested.

The awarding bodies vary in the details of how the samples are drawn but all the sampling procedures are designed to include a spread of candidate achievement within the centre. Centres with 10 or fewer candidates submit work for all their candidates. For larger centres, work from between 10 and 20 candidates is submitted, depending on entry size. Wilmut (2005) and Johnson (2011) expressed concern about the small sample sizes used for moderation purposes. Where the outcome of the work is ephemeral, for example a musical performance, the moderator visits the centre during the teacher assessment process. The moderation sample in these cases will be opportunistic (see Johnson, 2011, for more details).

Moderators have full knowledge of the marks and annotations of the teacher-markers, which may affect their marking decisions in the same way as paper-based marking (Johnson, 2011). The exception to this is the case of opportunistic moderation where the teacher and moderator are assessing the work simultaneously and independently (Johnson, 2011).

Little is known about inter-moderator consistency at present. Even though research studies of inter-moderator consistency will be costly and complex to organise, they are necessary to ensure that the current system is assessing candidates fairly (Johnson, 2011).

5.6 Detection of errors after marking is complete

Awarding bodies employ a number of additional checks to ensure that candidates receive the correct grade, such as checking candidates on grade boundaries, comparing awarded grades with predicted grades, checking for large discrepancies between grades awarded to different components, and identifying examiners for which there is a 'lingering' doubt about the quality of their marking. Meadows and Billington (2005) provided a detailed summary of these processes and their effectiveness. No new evidence was found in the literature surveyed for this review.

5.7 Tolerance levels

A tolerance level is set for each paper/item and only marking that falls outside of this tolerance is adjusted. The use of tolerance recognises that there may be legitimate differences in professional judgment. In addition, small adjustments are hard to justify on the basis of re-marking only a small sample of scripts: a different sample may have resulted in a different adjustment (Meadows and Billington, 2005).

Tolerance levels are set differently for different papers/items. Clearly, the level at which the tolerance is set will affect the sensitivity of the detection of unreliability. In-depth investigations of question and mark scheme features that affect marker agreement (Black, 2010; Massey and Raikes, 2006; Suto *et al*., 2008; and references in Curcin, 2010) have provided some empirical evidence for the levels of marking reliability that can realistically be expected. In general, higher maximum mark, less constrained mark schemes and the requirement for more complex marking strategies all reduce the accuracy with which an item can be marked (Black, 2010; Massey and Raikes, 2006; Suto *et al*., 2008; review by Curcin, 2010). These features should clearly influence the tolerance levels that are set for seed items. For instance, items with a maximum mark of less than three probably should not have any tolerance, and neither should objective and constrained items. Short answer and extended response questions, on the other hand, may well need some level of tolerance (Black, 2010).

There will also be significant and demonstrable differences in marking reliability between question papers as a result of different profiles of item types (Dhawan and Bramley, 2012). Dhawan and Bramley (2012) investigated marker agreement on seed items during on-screen marking of eight live GCSE and GCE examination. The components were chosen to form four pairs matched by features such as maximum mark, number of seed scripts, grade bandwidth and raw score distribution. In each

pair, one component (the 'Long' component) had at least one question worth eight or more marks, and the other component (the 'Short' component) contained only questions worth less than eight marks. In all the pairs, the 'Short' components were marked more reliably than the 'Long' components. In addition, the marking of all the 'Short' components was within the tolerance levels for the paper, but the same was not true for 'Long' components. The tolerance level was higher for each of the 'Long' components, but in three out of four of the pairs the difference was only one mark. Dhawan and Bramley (2012) argue that this is a very narrow range of 'extra' tolerance in a 'Long' component as compared to a 'Short' component with the same paper total, and suggest that "setting the tolerance value at a slightly higher percentage of the paper total in the Long components of these three pairs might have given more fair marker-monitoring results." (p. 24).

The simplest method for setting a tolerance level is to calculate it as a percentage of the paper total, but this does not take into account question type. A slightly more sophisticated method would be to calculate tolerance levels for each item and then combine these to produce a script-level tolerance (this does not necessarily imply item-level marking) (Black, 2010; Dhawan and Bramley, 2012). In addition, extra weighting could be given to factors like complexity of mark scheme and length of expected answer (Dhawan and Bramley, 2012). Dhawan and Bramley (2012) conclude that "setting tolerance at the right level, particularly for essay-type questions, would be an important step in effective and fair monitoring of markers." (p. 33).

Tolerance levels for internally assessed components are set at six per cent of the paper total (see Johnson, 2011). Our review did not find any studies that suggested whether this amount was reasonable or not.

5.8 Methods for adjusting scores

It is sometimes the case, in both internally and externally assessed components, that a marker is found to be consistently severe or lenient. Systematic marking errors of this sort can be corrected quickly and inexpensively (compared with a total re-mark) by applying an adjustment to all the marks awarded by the marker in question. Adjustments can be positive or negative and can be different for different mark ranges. For example, if a marker has been more severe at the top end of the scale, it would be appropriate to make a larger adjustment to the higher scores than to the lower scores. Adjustments must not, however, change the rank order of the candidates (GCSE, GCE, Principal Learning and Project Code of Practice, Ofqual, 2011). A number of methods have been suggested for evaluating whether an adjustment should be applied. There are also a number of different adjustments possible. See Meadows and Billington (2005) for an extensive review of mark adjustment, and Johnson (2011) for a description of mark adjustment in internally assessed components.

Johnson (2011) raises concerns about the process of adjustment of marks for internally assessed components because the decision to adjust the marks is based on the evaluation of a small sample of work by a single moderator. Taylor (1992) argued that it may be unreasonable to assume that an individual moderator can

‘carry’ the standard of the assessment and that their mark is any more ‘correct’ than that of the teacher. The author demonstrated, across three GCSE and one GCE component, that between 15 per cent and 40 per cent of candidates would have been awarded a different grade had a different moderator evaluated their work. The same concern could be extended to the practice of a single senior examiner basing adjustment decisions on a small sample of another examiner’s marking (Johnson, 2011). The basis of these concerns is whether or not the senior examiner’s/moderator’s mark is ‘truer’ than that of the marker/teacher.

It should be emphasised that mark adjustment only works in cases where there are systematic errors; it cannot correct for inconsistent marking or marking that changes in severity/leniency over time (e.g. see Myford and Wolfe, 2009). In addition, it has been shown that, while many candidates can benefit from adjustments, some have their marks moved further from that which a senior examiner would have awarded (see Meadows and Billington, 2005). Thus, adjustments should be used with caution and methods of making the initial marks more accurate should be pursued.

6 Conclusions

The literature that we reviewed covered a wide range of factors pertaining to marking reliability focusing on advances made since the Meadows and Billington work. Our review has highlighted new ideas about methods for quantifying marking reliability, factors which affect marking reliability and corresponding methods that could be used to correct unreliability.

In terms of quantifying marking reliability a number of studies have been conducted since 2004, adopting a range of different statistical techniques, such as Item Response Theory, Generalizability Theory and multi-level modelling. A number of reports were also produced as part of the Ofqual Reliability Programme reviewing measurement theories and models used in the study of reliability. It is clear that there are pros and cons for the different methods, and that in some cases the different methods lead to different results. However, the literature does not suggest we are nearing a consensus in terms of a shared approach to quantifying reliability. This remains a target that must be reached if we are to be able to directly compare reliability measures across different studies. As there is not a clear solution to this issue, it may be an area where an organisation, such as Ofqual, could take a lead and lay down some clear guidelines about what methods could be used and what statistics should be produced. In fact, in the Final Report of the Reliability Programme, Opposs and He (2011) make the following recommendation for further research: "Use of multiple reliability indices for a range of assessment types should be explored to assess the practical applications of specific estimation techniques and the differences in estimation between different techniques" (p. 54).

There appears to be greater consensus in terms of the factors that can be modified as a way of improving the reliability of marking. These factors include:

- Modifying features of the items and mark schemes, for example: increasing item constraint; using highly specified mark schemes; having lower maximum marks; or clarifying the cognitive demands placed on markers;
- Selecting markers with particular characteristics for different item types, for example using more experienced examiners on items that are complex to mark;
- Using item-level marking so that more than one marker contributes to a candidate's overall mark rather than a single marker assessing a whole script, because it reduces the effects of random error, removes the halo effect and eliminates biases relating to candidate characteristics such as gender, ethnicity, school and name;
- Using on-screen marking as it allows continuous marker monitoring, which enables inaccuracy to be detected early and corrected. On-screen marking also eliminates errors resulting from incorrect addition or transcription of marks and prevents items being left unmarked;
- Using blind double-marking to detect unreliability because re-marking with the comments and marks visible on the script is likely to underestimate unreliability.

A number of the published studies that we reviewed were conducted by awarding bodies, which demonstrates that they are actively working to improve the reliability of their examinations.

The increase in the use of on-screen marking over recent years has enabled a number of improvements to reliability, not least because of the availability of item level data. The move to on-screen marking is also thought to enable unreliability to be better detected throughout the marking process. However, more work is needed into the tolerances that should be considered acceptable for different modes of assessment.

7 Glossary

Analytic mark scheme Analytic levels-based mark schemes separate the aspects of interest and provide level descriptors, and associated mark bands, for each aspect. That is, they explicitly weigh the different features of response.

Average An 'average' value purports to represent or to summarise the relevant features of a set of values. There are three types of average: mean, median and mode. For their definitions, please refer to the *mean, median and mode* sections of this Glossary.

Classical Test Theory A statistical model used to estimate the reliability of assessments. According to this theory, the mark awarded to a candidate in an examination contains an amount of error. Also known as True Score Theory.

Correlation A measure of association between two measurements, e.g. between size of school and the mean number of GCSE passes at grades A, B and C obtained by each student. A positive correlation would occur if the number of passes increased with the size of the school. If the number of passes decreased with size of school there would be a negative correlation. Correlations range from -1 to +1 (perfect negative to perfect positive correlations); a value of zero indicates no linear association between the two measures.

Cronbach's alpha A measure of the consistency of test scores, also a measure of internal consistency. Notionally, this approach splits the test questions into two halves and looks at how candidates do on each half. This is then repeated for every possible combination of "halves", and an average correlation between the two halves is calculated. This measure is called Cronbach's alpha and is the starting point for much reliability work. It is a widely used form of Kuder-Richardson formula 20 (KR 20), it can be used for test items that have more than two answers which KR 20 cannot be. Like other reliability coefficients, Cronbach's alpha ranges from 0 to 1. Scores towards the high end of the range suggest that the items in a test are measuring the same thing.

Holistic mark scheme In a holistic levels-based scheme, markers make an overall judgment of the candidate's performance. Each level may include a number of different response features but no explicit weightings are given to the different features.

Inter-marker reliability Reproducibility of the marks assigned to an examination script by different markers.

Interaction It is sometimes the case in regression models that the relationship between one of the variables and the outcome measure is different for different groups – for example the relationship between achievement and prior attainment may be different for boys and girls. This is modelled using an *interaction term*, which takes account of this possibility. If statistically significant, it implies that the strength of the underlying relationship is not the same for all groups.

Intra-marker reliability Reproducibility of the marks assigned to an examination script by an individual marker.

Item An item is the smallest separately identified question or task within an Assessment, accompanied by its mark scheme. Often but not always a single question.

Item Response Theory (IRT) IRT is a statistical approach to the design, analysis, and scoring of assessments. IRT is a modern test theory (as opposed to classical test theory). IRT attempts to model the interaction between the test taker and each individual question. It is the branch of psychometrics that is concerned with the probability of success when someone attempts to answer a test item (widely used in tests of intelligence, aptitude, ability, achievement and knowledge). Different types of model are used, with one, two or three parameters. The one-parameter model is sometimes known as the Rasch model.

Levels-based mark scheme Levels-based mark schemes divide the mark range into several bands, each representing a distinguishable level of quality of response. The level descriptors may include features of language, content or both.

Mean The conventional way of calculating the 'average' of a set of data values, by adding them up and dividing by the number of data values. Can be seriously affected by a few extreme data values (see *median)*.

Measurement error Measurement error is the difference between a measured value and its true value. In statistics, variability is an inherent part of the measurement process, and so error in this sense is not a "mistake".

Median is the central value in a set of data, such that half the cases lie below and half above that value. It is less affected by extreme values than the *mean* as a measure of the 'average' of a dataset.

Mode The most common response in a set of data.

Multilevel modelling Multilevel modelling is a recent development of linear regression which takes account of data which is grouped into similar clusters at different levels. For example, individual students are grouped into year groups or cohorts, and those cohorts are grouped within schools. There may be more in common between students within the same cohort than with other cohorts, and there may be elements of similarity between different cohorts in the same school. Multilevel modelling allows us to take account of this hierarchical structure of the data and produce more accurate predictions, as well as estimates of the differences between students, between cohorts, and between schools. (Multilevel modelling is also known as hierarchical linear modelling).

Points-based mark scheme Points-based mark schemes list objectively identifiable words, statements or ideas. Marks are awarded one at a time for each creditworthy point in the candidate's response.

Reliability refers to the consistency of a measure. A test is considered reliable if we get the same result repeatedly.

Standardisation is a process which awarding organisations carry out to ensure that assessment criteria for an assessment are applied consistently.

Standard deviation Standard deviation is a measure of the spread of some quantity within a group of individuals. If the quantity is distributed approximately Normally, we would expect about 95% of the individuals to be within 2 standard deviations either side of the mean value.

Standard error A measure of the uncertainty in the estimation of a statistical parameter. It is expressed as the *standard deviation* of the errors in the estimate, so that there is roughly a 95% chance that the 'true' value lies within 2 standard errors either side of the estimate.

Standard Error of Measurement (SEM) A measure of the uncertainty in individuals' test scores resulting from factors unrelated to the purpose of the test.

Validity Whether what is being measured is what the researchers intended.

Variance In statistics, the variance is used as a measure of how spread out a set of numbers are. A low variance indicates similar values and high variance indicates diverse values.

8 References

Ahmed, A. and Pollitt, A. (2011). 'Improving marking quality through a taxonomy of mark schemes', *Assessment in Education: Principles, Policy & Practice*, **18**, 3, 259–278.

Al-Bayatti, M. and Jones, B. (2005). *NAA Enhancing the Quality of Marking Project: the Effect of Sample Size on Increased Precision in Detecting Errant Marking.* London: QCA [online]. Available: http://dera.ioe.ac.uk/9451/1/The_effect_of_sample_size_on_increased_precision_in_detecting_errant_marking.pdf [12 November, 2012].

Attali, Y. and Burstein, J. (2006). 'Automated essay scoring with e-rater V.2', *The Journal of Technology, Learning, and Assessment,* **4**, 3, 1–31. Cited in: Blood, I. (2011). 'Automated Essay Scoring: a Literature Review', *Working Papers in TESOL & Applied Linguistics*, **11**, 2, 40–64 [online]. Available: http://journals.tc-library.org/index.php/tesol/article/download/745/470 [29 November, 2012].

Baird, J.A., Beguin, A., Black, P., Pollitt, A. and Stanley, G. (2011). *The Reliability Programme: Final Report of the Technical Advisory Group.* In: Ofqual *Reliability Compendium* (Chapter 20). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/standards/reliability [12 November, 2012].

Baird, J. and Bridle, N. (2000) *A Feasibility Study on Anonymised Marking in Large-scale Public Examinations* (AQA Research Report RC/91). Cited in: Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [12 November, 2012].

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

Baird, J.A., Hayes, M., Johnson, R., Johnson, S. and Lamprianou, L. (2012). *Marker Effects and Examination Reliability: a Comparative Exploration from the Perspectives of Generalizability Theory, Rasch Modelling and Multilevel Modelling*. Coventry: Ofqual.

Baker, E., Ayres, P., O'Neil, H.F., Chli, K., Sawyer, W., Sylvester, R.M. and Carroll, B. (2008). *KS3 English Test Marker Study in Australia: Final Report to the National Assessment Agency of England.* Sherman Oaks, CA: University of Southern California.

Bew, P. (2011). *Independent Review of Key Stage 2 Testing, Assessment and Accountability: Final Report*. London: DfE [online]. Available: https://media.education.gov.uk/MediaFiles/C/C/0/%7BCC021195-3870-40B7-AC0B-66004C329F1F%7DIndependent%20review%20of%20KS2%20testing,%20final%20report.pdf [29 November, 2012].

Billington, L. (2012). *Exploring Second Phase Samples: What is the Most Appropriate Basis for Examiner Adjustments?* Manchester: AQA, Centre for Education Research and Policy.

Black, B. (2010). 'Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks.' Paper presented at International Association for Educational Assessment 36th Annual Conference, Bangkok, Thailand, 22-27 August.

Black, B., Curcin, M. and Dhawan, V. (2010). 'Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks.' Paper presented at the 36th Annual Conference of the International Association for Educational Assessment , Bangkok, Thailand, August. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Black, B., Suto, W.M.I. and Bramley, T. (in submission). *The Interrelations of Features of Questions, Mark Schemes and Examinee Responses and their Impact Upon Marker Agreement*. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Blood, I. (2011). 'Automated essay scoring: a literature review', *Working Papers in TESOL & Applied Linguistics*, **11**, 2, 40–64 [online]. Available: http://journals.tc-library.org/index.php/tesol/article/download/745/470 [29 November, 2012].

Bramley, T. (2007). 'Quantifying marker agreement: terminology, statistics and issues', *Research Matters*, **4**, 22–27.

Bramley, T. (2009). 'The effect of manipulating features of examinees' scripts on their perceived quality.' Paper presented at the Association for Educational Assessment – Europe Annual Conference, Malta, November [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/186235_TB_Script_features_AEA_Europe09.pdf [12 November, 2012].

Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Britton, J.N., Martin, N.C. and Rosen, H. (1966). *Multiple Marking of English Compositions: an Account of an Experiment* (Schools Council Examinations Bulletin, 12). London: HMSO. Cited in: Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.

Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.

Burslem, S. (2011). *The Reliability Programme: Final Report of the Policy Advisory Group.* In: Ofqual *Reliability Compendium* (Chapter 21). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=782 [12 November, 2012].

Crisp, V. (2007). 'Do assessors pay attention to appropriate features of student work when making assessment judgements?' Paper presented at the 33rd International Association for Educational Assessment Annual Conference, Baku, September. Cited in: Curcin, M. (2010). 'A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice', *Research Matters*, **10**, 27–32 [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/186572_Rearch_Matters_10_2010.pdf . [29 November, 2012].

Crisp, V. (2010). 'Towards a model of the judgement processes involved in examination marking', *Oxford Review of Education*, **36**, 1, 1–21.

Cronbach, L. J. and Shavelson, R.J. (2004). *My Current Thoughts on Coefficient Alpha and Successors. Procedures* (CSE Report 643). Los Angeles, CA: Centre for the Study of Evaluation. Cited in: Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [12 November, 2012].

Curcin, M. (2010). 'A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice', *Research Matters*, **10**, 27–32 [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/186572_Rearch_Matters_10_2010.pdf . [29 November, 2012].

Department for Education (2012). *Reforming Key Stage 4 Qualifications: Consultation Document* [online]. Available: http://www.education.gov.uk/aboutdfe/departmentalinformation/consultations/a00213902/reforming-key-stage-4-qualifications [12 November, 2012].

Dhawan, V. and Bramley, T. (2012). *Estimation of Inter-rater Reliability*. Coventry: Ofqual.

Evers, A., Lucassen, W., Meijer, R. and Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie)* [COTAN assessment sytem for the quality of tests (revised version)]. Amsterdam: NIP. Cited in: Baird, J.A., Beguin, A., Black, P., Pollitt, A. and Stanley, G. (2011). *The Reliability Programme: Final Report of the Technical Advisory Group.* In: Ofqual *Reliability Compendium* (Chapter 20). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/standards/reliability [12 November, 2012].

Fearnley, A. (2005). *An Investigation of Targeted Double Marking for GCSE and GCE.* London: QCA [online]. Available: http://dera.ioe.ac.uk/9450/1/QCDA104979_an_investigation_of_targeted_double_marking_for_GCSE_and_GCE.pdf [29 November, 2012].

Foltz, P.W., Landauer, T.K. and Laham, D. (1999). 'Automated essay scoring: Applications to educational technology.' In: Proceedings of EdMedia '99. Retrieved from http://www.psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html. Cited in: Blood, I. (2011). 'Automated essay scoring: a literature review', *Working Papers in TESOL & Applied Linguistics*, **11**, 2, 40–64 [online]. Available: http://journals.tc-library.org/index.php/tesol/article/download/745/470 [29 November, 2012].

Fowles, D. (2009). 'How reliable is marking in GCSE English?' *English in Education*, **43**, 1, 49–67.

Greatorex, J. (2005). 'Assessing the evidence: different types of NVQ evidence and their impact on reliability and fairness', *Journal of Vocational Education and Training,* **57**, 2, 149–164. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Greatorex, J. and Bell, J.F. (2008). 'What makes AS marking reliable? An experiment with some stages from the standardisation process', *Research Papers in Education,* **23,** 3, 233–255. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Greatorex J., Nádas R., Suto W.M.I. and Bell J.F. (2007). 'Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training.' Paper presented at the European Conference on Educational

Research, , Ghent, Belgium, 19–22 September. Cited in: Suto, I., Crisp, V. and Greatorex, J. (2008). 'Investigating the judgemental marking process: an overview of our recent research'**,** *Research Matters*, **5**, 6–8 [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/136163_Research_Matters_5_web_.pdf [12 November, 2012].

Greatorex, J. and Suto, I (2006). 'An empirical exploration of human judgement in the marking of school examinations.' Paper presented at the 32nd International Association for Educational Assessment Conference, Singapore, 21–26 May.

Harlen, W. (2004). A systematic review of the evidence of the reliability and validity of assessment by teachers for summative purposes. In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November , 2012].

Harlen, W. (2005). 'Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes', *Research Papers in Education,* **20**, 3, 245–270

He, Q. (2009). *Estimating the Reliability of Composite Scores*. In: Ofqual *Reliability Compendium* (Chapter 12). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/files/2010-02-01-composite-reliability.pdf [12 November, 2012].

Head, J.J. (1966). 'Multiple marking of an essay item in experimental O-level Nuffield biology examinations, *Educational Review'*,**19**, 1, 65–71. Cited in: Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.

Johnson, M. (2008). 'Assessing at the borderline: judging a vocationally related portfolio holistically', *Issues in Educational Research*, **18**, 1, 26–43.

Johnson, M., Hopkin, R., Shiell, H. and Bell, J.F. (2012). 'Extended essay marking on screen: is examiner marking accuracy influenced by marking mode?' *Educational Research and Evaluation*, **18**, 2, 107–124.

Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Johnson, M. and Nádas, R. (2009a). 'An investigation into marker reliability and some qualitative aspects of on-screen essay marking', *Research Matters*, **8**, 2–7. Cited in: Johnson, M., Hopkin, R., Shiell, H. and Bell, J.F. (2012). 'Extended essay marking on screen: is examiner marking accuracy influenced by marking mode?' *Educational Research and Evaluation*, **18**, 2, 107–124.

Johnson, M. and Nádas, R. (2009b). 'Marginalised behaviour: digital annotations, spatial encoding and the implications for reading comprehension', Learning*, Media and Technology*, 34, 323–336. Cited in: Johnson, M., Hopkin, R., Shiell, H. and Bell, J.F. (2012). 'Extended essay marking on screen: is examiner marking accuracy influenced by marking mode?' *Educational Research and Evaluation*, **18**, 2, 107–124.

Kim, H. J. (2010). Investigating Raters' Ddevelopment of Rating Ability on a Second Language Speaking Assessment. Unpublished doctoral dissertation. Columbia University, Teachers College. Cited in: Blood, I. (2011). 'Automated Essay Scoring: A Literature Review', *Working Papers in TESOL & Applied Linguistics*, **11**, 2, 40–64 [online]. Available: http://journals.tc-library.org/index.php/tesol/article/download/745/470 [29 November, 2012].

Kim, S.C. and Wilson, M. (2009). 'A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model', *Journal of Applied Measurement*, **10**, 4, 408–423.

Larkey, L.S. (1998). 'Automatic essay grading using text categorization techniques.' Paper presented at 21st International Conference of the Association for Computing Machinery-Special Interest Group on Information Retrieval (ACM-SIGIR), Melbourne, Australia. Retrieved from http://ciir.cs.umass.edu/pubfiles/ir-121.pdf . Cited in: Blood, I. (2011). 'Automated Essay Scoring: A Literature Review', *Working Papers in*

TESOL & Applied Linguistics, **11**, 2, 40–64 [online]. Available: http://journals.tc-library.org/index.php/tesol/article/download/745/470 [29 November, 2012].

Lucas, A.M. (1971). 'Multiple marking of a matriculation biology essay question', *British Journal of Educational Psychology,* **41**, 1, 78–84. Cited in: Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.

Massey, A. (1983) 'The effects of handwriting and other incidental variables on GCE 'A' level marks in English Literature', *Educational Review*, **35**, 1, 45–50. Cited in: Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [12 November, 2012].

Massey, A. and Foulkes, J. (1994). 'Audit of the 1993 KS3 acience national test pilot and the concept of quasi-reconciliation', *Evaluation and Research in Education*, **8**, 119–132. Cited in: Curcin, M. (2010). 'A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice', *Research Matters*, **10**, 27–32 [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/186572_Rearch_Matters_10_2010.pdf . [29 November, 2012].

Massey, A.J. and Raikes, N. (2006). *Item-level Examiner Agreement*. Cambridge: Cambridge Assessment [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/171646_BERA06_Massey_and_Raikes.pdf [12 November, 2012].

Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [12 November, 2012].

Meadows, M. and Billington, L. (2007). *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection.* London: QCA [online]. Available: http://archive.teachfind.com/qcda/orderline.qcda.gov.uk/gempdf/184962531X/QCDA104980_marker_selection.pdf [12 November, 2012].

Murphy, R.J. (1978). 'Reliability of marking in eight GCE examinations', *British Journal of Educational Psychology*, **48**, 2, 196–200. Cited in: Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [12 November, 2012].

Murphy, R.J.L. (1979). 'Removing the marks from examination scripts before re-marking them: does it make any difference?' *British Journal of Educational Psychology,* **49**, 73–78. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available:

http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Myford, C.M. and Wolfe, E.W. (2009). 'Monitoring rater performance over time: a framework for detecting differential accuracy and differential scale category use', *Journal of Educational Measurement*, **46**, 4, 371–389.

Newton, P.E. (1996). 'The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English', *British Educational Research Journal*, **22**, 405-420. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Newton, P.E. (2009). 'The reliability of results from national curriculum testing in England', testing in England, *Educational Research*, **51**, 2, 181–212.

Ofqual (2011). *GCSE, GCE , Principal Learning and Project Code of Practice.* Coventry: Ofqual [online]. Available: http://dera.ioe.ac.uk/3648/1/53-codes-of-practice%3Fdownload%3D680%3Acode-of-practice [29 November, 2012].

Ofqual (2012a). *Corporate Plan 2012-15*. Coventry: Ofqual [online]. Available: http://www.google.co.uk/url?q=http://www2.ofqual.gov.uk/downloads/category/139-information%3Fdownload%3D1404%253Acorporate-plan&sa=U&ei=QkYKUcvGEPGR0QX66YC4Cw&ved=0CBUQFjAA&usg=AFQjCNFz0PI5sQGo7nSYDutB9DaZeA1t1w [30 January, 2013].

Ofqual (2012b). *GCSE English Awards 2012: a Regulatory Report.* London: Ofqual [online]. Available: http://www.ofqual.gov.uk/files/2012-08-31-gcse-english-awards-2012-a-regulatory-report.pdf [29 November, 2012].

Opposs, D. and He, Q. (2011). *The Reliability Programme Final Report.* In: Ofqual *Reliability Compendium* (Chapter 22). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=783 [12 November, 2012].

Page, E. B. (1968). 'The use of the computer in analyzing student essays', *International Review of Education,* **14**, 210–225. Cited in: Blood, I. (2011). 'Automated Essay Scoring: A Literature Review', *Working Papers in TESOL & Applied Linguistics*, **11**, 2, 40–64 [online]. Available: http://journals.tc-library.org/index.php/tesol/article/download/745/470 [29 November, 2012].

Pilliner, A.E.G. (1969). 'Multiple marking: Wiseman or Cox?' *British Journal of Educational Psychology*', **39**, 3, 313–315. Cited in: Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.

Pinot de Moira, A. (2011). *Why Item Mark? The Advantages and Disadvantages of E-Marking.* Manchester: AQA, Centre for Education Research and Policy.

Pollitt, A. (2012). 'The method of adaptive comparative judgement', *Assessment in Education: Principles, Policy & Practice*, **19**, 3, 281–300.

Raikes, N. (2006). 'The Cambridge Assessment/Oxford University automatic marking system: does it work?' *Research Matters*, **2**, 17–20.

Raikes, N., Fidler, J. and Gill, T. (2010). 'Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology', *Research Matters*, **10**, 21-27

Royal-Dawson, L. (2005). *Is Teaching Experience a Necessary Condition for Markers of Key Stage 3 English?* London: QCA. Cited in: Newton, P.E. (2009). 'The reliability of results from national curriculum testing in England', testing in England, *Educational Research*, **51**, 2, 181–212.

Shi, L. (2001). 'Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing', *Language Testing,* **18,** 303–325. Cited in: Blood, I. (2011). 'Automated essay scoring: a literature review', *Working Papers in TESOL & Applied Linguistics*, **11**, 2, 40–64 [online]. Available: http://journals.tc-library.org/index.php/tesol/article/download/745/470 [29 November, 2012].

Stanley, G., MacCann, R., Gardner, J., Reynolds, L. and Wild, I. (2009). *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development.* London: QCA [online]. Available: http://oucea.education.ox.ac.uk/wordpress/wp-content/uploads/2011/01/2009_03-Review_of_teacher_assessment-QCA.pdf [29 November, 2012].

Stemler, S.E. (2004). 'A comparison of consensus, consistency and measurement approaches to estimating interrater reliability', Practical Assessment, Research & Evaluation, 9, 4 [online]. Available: http://PAREonline.net/getvn.asp?v=9&n=4 [12 November, 2012].

Suto, I., Crisp, V. and Greatorex, J. (2008). 'Investigating the judgemental marking process: an overview of our recent research'**,** *Research Matters*, **5**, 6–8 [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/136163_Research_Matters_5_web_.pdf [12 November, 2012].

Suto, W.M.I. and 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. Cited in: Suto, I., Nadas, R. and Bell, J. (2011). 'Who should mark what? A study of factors affecting marking accuracy in a biology examination', *Research Papers in Education*, **26**, 1, 21–52.

Suto, I., Nadas, R. and Bell, J. (2011). 'Who should mark what? A study of factors affecting marking accuracy in a biology examination', *Research Papers in Education*, **26**, 1, 21–52.

Taylor, M. (1992). *The Reliability of Judgements Made by Coursework Assessors.* Guildford: Associated Examining Board. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Taylor, R. (2011). *A Qualitative Exploration of Key Stakeholders' Perceptions and Opinions of Awarding Body Marking Procedures*. Manchester: AQA, Centre for Education Research and Policy [online]. Available: https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP-RP-RT-01052007.pdf [12 November, 2012].

Vidal Rodeiro, C.L. (2007). 'Agreement between outcomes from different double marking models', *Research Matters*, **4**, 28–33 [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/136145_Research_Matters_4_Jun_2007.pdf [29 November, 2012].

Wiliam, D. (2000). 'Reliability, validity, and all that jazz', *Education*, **29**, 3, 9–13. Cited in: Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability*. Manchester: AQA [online]. Available: https://orderline.education.gov.uk/gempdf/1849625344/QCDA104983_review_of_the_literature_on_marking_reliability.pdf [12 November, 2012].

Wilmut,J.(1984). *A Pilot Study of the Effects of Complete or Partial Removal of Marks and Comments from Scripts Before Re-marking Them* (AEB Research Report RAC315). Cited in: Curcin, M. (2010). 'A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice', *Research Matters*, **10**, 27–32 [online]. Available: http://www.cambridgeassessment.org.uk/ca/digitalAssets/186572_Rearch_Matters_10_2010.pdf . [29 November, 2012].

Wilmut, J. (2005). *Experiences of Summative Assessment in the UK*. London: QCA. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Wiseman, S. (1949). 'The marking of English composition in grammar school selection', *British Journal of Educational Psychology*', **19**, 3, 200–209. Cited in: Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.

Wood, R. and Quinn, B. (1976). ' Double impression marking of English language essay and summary questions', *Educational Review*, **28**, 3, 229–246. Cited in: Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.

Further reading

Altman, D.G. and Bland, J.M. (1983). 'Measurement in medicine: the analysis of method comparison studies', *The Statistician,* **32**, 307–317. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Bell, J. F., Bramley, T., Claessen, M., and Raikes, N. (2006). 'Quality control of marking: some models and simulations.' Paper presented at the Annual Conference of the British Educational Research Association, University of Warwick, September. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Bland, J.M. and Altman, D.G.1. (1986). 'Statistical methods for assessing agreement between two methods of clinical measurement', *Lancet*, **i,** 307–310. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Fowles, D. (2002). *Evaluation of an E-marking Pilot in GCE Chemistry: Effects on Marking and Examiners' Views* (AQA Research Report RC/190). Cited in: Meadows, M. and Billington, L. (2007). *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection.* London: QCA [online]. Available: http://archive.teachfind.com/qcda/orderline.qcda.gov.uk/gempdf/184962531X/QCDA104980_marker_selection.pdf [12 November, 2012].

Harth, H. and Hemker, B. (2011). *On the Reliability of Results in Vocational Assessment: the Case of Work-based Certification*. Coventry: Ofqual [online]. Available: www.ofqual.gov.uk/research-and-statistics/research-reports/92-articles/20-reliability. Cited in: Opposs, D. and He, Q. (2011). *The Reliability Programme Final Report.* In: Ofqual *Reliability Compendium* (Chapter 22). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=783 [12 November, 2012].

Harth, H. and Van Rijn, P. (2011). *On the reliability of results in vocational assessment: the case of work-based certification*. Coventry: Office of the Qualifications and Examinations Regulator. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Hoge, R.D. and Coladarci, T. (1989). 'Teacher-based judgements of academic achievement: a review of literature', *Review of Educational Research,* **59**, 297-313. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Johnson, M. (2006**). '**A review of vocational research in the UK 2002-2006: measurement and accessibility issues', *International Journal of Training Research*, **4**, 2, 48–71. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in*

GCSE and GCE. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Martinez, J.F., Stecher, B. and Borko, H. (2009). 'Classroom assessment practices, teacher judgments, and student achievement in mathematics: evidence from the ECLS', *Educational Assessment*, **14**, 78–102. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Murphy, R.J.L. (1982). 'A further report of investigations into the reliability of marking of GCE examinations', *British Journal of Educational Psychology*, **52**, 58–63. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Suto, W.M.I. and Nadas, R. (2009). 'Why are some GCSE questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features', *Research Papers in Education,* **24**, 3, 335-377. Cited in: Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. In: Ofqual *Reliability Compendium* (Chapter 7). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=768 [12 November, 2012].

Taylor, M. (2009). *Sample Sizes for Moderation from Summer 2009* (Draft Paper for JCQ Adoption). Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Wilmut, J., Wood, R. and Murphy, R. (1996). *A Review of Research into the Reliability of Examinations*. Nottingham: University of Nottingham. Cited in: Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. In: Ofqual *Reliability Compendium* (Chapter 9). Coventry: Ofqual [online]. Available: http://www2.ofqual.gov.uk/index.php?option=com_content&view=article&id=770 [12 November, 2012].

Appendix 1

A summary of the key findings of Meadows & Billington (2005) A Review of the Literature on Marking Reliability

Meadows and Billington (2005) produced an extensive review of the literature on marking reliability spanning nearly 100 years. They covered "the levels of marking reliability achieved in different forms of assessment and research into methods of improving marking reliability" with a focus on "the marking of externally assessed examination scripts, rather than on the assessment of coursework, performance or of competence..." (p. 4).

Summary of the key findings

- An estimate of the reliability of a test is necessary to interpret its results fully. There are strong arguments that examination results should be reported with their associated coefficient of reliability and/or estimates of the errors associated with the scores.
- Marker unreliability can stem from many sources including: contrast effects, the text of the script itself, characteristics of the candidate, and characteristics of the examiner. Some means of controlling these effects are suggested but their effectiveness is debatable.
- Interrater reliability is strongly associated with question format. Tightly prescribed questions with definite answers are marked much more reliably than essays.
- If essays continue to be viewed as a valuable question format, the associated unreliability of marking may simply have to be accepted.
- Improving the mark scheme, training examiners, fostering a community of practice and providing exemplar material may all help to improve marking reliability.
- Mark adjustments can help to correct systematic errors caused by consistent leniency/severity, but are no help when examiners are inconsistent.
- More research should be done on alternatives to traditional marking.

Definition and estimation of Reliability

Meadows and Billington (2005) begin by discussing what is meant by *reliability*, in relation to assessment in general, and more specifically in relation to marking. Different researchers define reliability in slightly different ways (p. 7):

1. "Reliability is the extent to which the test measurements are the result of properties of those individuals being measured." If this definition is satisfied, the results of repeated applications of the procedure should be repeatable and dependable; if not, they will vary unsystematically.

2. Reliability can be defined as how “consistent or error free measurements are. When random error is minimal, one can expect scores to be accurate, reproducible and generalisable...”
3. *“*A theoretical definition of reliability is the proportion of score variance caused by systematic variation in the population of test takers. This definition is population specific and sees reliability as a joint characteristic of a test and an examinee group…*”*

An estimate of the reliability of a test is necessary to interpret its results fully. As we cannot know the variation of the true abilities of the population, it is not possible to calculate a reliability statistic based on the third definition. However, there are several ways to estimate the stability of a set of test scores for a group of candidates: test-retest reliability, split-half reliability, internal consistency, and alternate form reliability.

Test-retest reliability – A test-retest reliability coefficient is obtained by administering the same test twice and correlating the scores. Theoretically, it is useful because it allows direct measurement of consistency from one occasion to the next. However, there are problems with using it in practice because it requires the same set of students to take the same test twice. If the testing sessions are too close together consistency may be artificially inflated because the students remember some of the questions and their responses. If the sessions are too far apart the results will be confounded by learning and maturation: that is, changes in the students themselves.

Split-half reliability – This coefficient is calculated by splitting the test in half, correlating the scores on each part and correcting for length. This method requires only one administration of the test but the coefficient will vary depending on how the test was split.

Internal consistency – These statistics reflect how the individual items are correlated with each other and include Cronbach’s alpha, the Kuder Richardson Formula 20 (KR-20) and Richardson Formula 21 (KR-21). Only one administration of the test is required to calculate these statistics, but they are only valid if all the items measure the same skill. If a test measures a set of different skills, internal consistency would not be expected.

Alternate-form reliability – The correlation between the scores on alternate forms of a test that are matched in content and difficulty provides another measure of consistency. The results will be affected by the choice of items in each test, and therefore slight differences in content and difficulty, as well as changes in the students between the tests.

All of these statistics are based on the correlation coefficient, but Meadows and Billington (2005) found a number of studies that highlight the shortcomings of this approach. Notably, that the correlation coefficient does not convey much information about the distribution of the two variables that are being correlated, and even a perfect correlation may ignore systematic differences between sets of scores. Also, correlation is affected by the spread of performance in a sample.

Thus, alternatives to the correlation coefficient have been sought. One such alternative (and complementary) measure is the standard error of measurement (SEM) from classical test theory. When the errors are small in relation to the actual scores, the test is relatively reliable, and vice versa.

SEM has the advantage that it does not depend on the spread of performance in the sample and is more directly related to the likely error on an individual candidate's mark. In addition, it is argued that defining reliability as the uncertainty associated with a score is easier for those who use the test scores to understand, particularly if they have no statistical knowledge. Meadows and Billington (2005) found a number of authors supporting the use of SEM as a measure of reliability and some argue that it is the "most important single piece of information to report" (p.16).

Other authors argue that, where grades or levels are reported, it is useful to include the expected percentage of misclassifications. Even tests with high reliability coefficients can misclassify a substantial proportion of students, with the problem worsening as the number of levels increases. This is particularly important in criterion referenced tests, where true-score variance may be small, or the distribution of errors unusual (pp.10-11). Similar considerations need to be made when using tests to predict future performance (p.12) or select individuals (p.13).

Meadows and Billington (2005) note that reliability is also a pre-requisite for validity; with estimates of validity being meaningless without an estimation of the error of measurement. However, attempts to increase reliability (such as using a stricter mark scheme or limiting the amount of the curriculum that is tested) may result in reduced validity (p.13).

Sources of unreliability

Meadows and Billington (2005) discuss three major sources of assessment error (based on Wiliam (2000)): the test itself; the candidates taking the test; and test scoring.

First, error can stem from multiple factors of the test itself. These include the effectiveness of the distractors in multiple choice tests, partially correct distractors, multiple correct answers, and difficulty of items relative to candidate ability, with the largest source of unreliability usually being the choice of items included in the test (pp.7-8). Second, error will be introduced as a result of changes in candidates' concentration, attitudes, health, fatigue, etc. which can affect their test-taking consistency. Finally, there are multiple factors affecting the reliability of the scoring (or marking) of a test and these are the main focus of the review by Meadows and Billington (2005).

According to classical test theory, the reliability of a test can be improved either by reducing the error variance or by increasing the true-score variance. Error variance can be reduced by improving the items selected, and by making the marking more consistent, although it has been argued that the effect of such changes is small and that "the most effective ways of increasing reliability of a test are to make the scope of the test narrower, or make the test longer" (p. 11). Increasing the length of the examination, or the number of component parts, increases the chance that the random effects of marking error will cancel each other out. However, other researchers have argued that techniques for making marking more systematic and objective should not be overlooked, particularly in an environment where increasing the amount or length of examinations is unlikely to be popular.

Types of interrater reliability

Stemler (2004) identifies three classes of statistical methods for reporting interrater reliability: consensus estimates; consistency estimates; and measurement estimates. He emphasizes that it is important to specify which type is being discussed.

Consensus estimates

Consensus estimates of interrater reliability assume that observers should be able to come to exact agreement about how to apply the various levels of a mark scheme. These estimates are often reported as a per cent agreement or using Cohen's kappa statistic, but both measures can be affected by the distribution of the candidates across categories.

Consistency estimates

In this case, it is not necessary for examiners to have the same understanding of the rating scale, provided each examiner is consistent in their own classifications. Consistency estimates may be high whilst the average scores awarded by the different examiners may be very different. The Pearson correlation coefficient can be used if the data are normally distributed and the Spearman rank coefficient should be used if they are not. Cronbach's alpha can also be used to give a single consistency estimate of interrater reliability across multiple examiners, but each examiner must give a score for every test. Consistency estimates are also sensitive to the distribution of the observed data.

Measurement estimates of reliability

Measurement estimates use all the information available from all examiners to create a summary score for each candidate. It is not necessary for examiners to come to a consensus, providing it is possible to estimate and account for examiner severity when creating the final score. Measurement estimates can be calculated using generalizability theory, the many-facets Rasch model or principal component analysis.

This approach has several advantages. First, errors are taken into account at the level of individual examiners so the summary score should be a more accurate measurement of the true score than the raw scores from the examiners. Second, ratings from multiple examiners can be handled effectively by simultaneously calculating estimates across all the items that were rated, rather than separately for each item and pair of examiners. Third, examiners are not required to mark every item. However, measurement estimates require the use of specialized software and certain methods can only handle ordinal data.

Studies of the reliability of marking

Meadows and Billington (2005) state that: "the reliability of marking has been studied at all levels of education across various subjects and assessment methods" (p. 17). Their examples include a number of studies on the reliability of marking in high stakes examinations, often conducted by the awarding bodies themselves, and also studies of marking across various subjects at Higher Education level, and in other contexts. These studies show extremely variable levels of reliability.

By the 1970s it was clear that marking reliability varied across subjects and with examination type. Interrater reliability appears to be best in mathematics and the physical sciences, and poorest in the arts and social sciences. The most reliably marked examinations tended to be

those made up of highly structured, analytically marked questions; while those examinations that used essay-type questions were least reliably marked, regardless of subject.

Despite the prevalence of descriptive studies of marking reliability, Meadows and Billington (2005) found that "it is often difficult to draw conclusions about the factors that influence reliability. … because the studies often vary in so many important respects (the training of the markers, the type of assessment, the mark scheme, the subject assessed and so on)" (p. 20). Instead, systematic research is needed in which these variables are manipulated and the resultant effect on reliability is measured.

Changes in the consistency and severity of marking over time

Meadows and Billington (2005) discuss the research on changes in examiner severity/leniency during the marking of a particular batch of scripts, over an entire a marking session, and over more extended periods of time (pp. 20-23).

In the short-term, variations may occur in the way examiners mark because:

- an examiner may be more accurate at the beginning of marking when training is fresh;
- the pressure towards the end of the examination period and/or examiner fatigue may reduce the accuracy of marking;
- examiners may improve at marking with practice;
- an examiner may respond to feedback by overcompensating.

Meadows and Billington (2005) found that studies of the extent to which the severity/leniency of examiners' marking varied over time were contradictory. Some studies showed a relative stability in examiner severity, with neither position of the script within the allocation nor chief examiner feedback affecting the way in which the examiner marked. However, there were many studies that suggested otherwise and showed changes in examiner severity within and between examination sessions.

Meadows and Billington (2005) note that "it is common practice that candidates' marks are adjusted to account for any inconsistencies in examiner severity, but this is undermined if examiner severity varies across the marking period" (p. 22). Fortunately, statistical methods exist that can help to detect the effects of changes in examiner severity and eliminate them from the final marks. Changes in examiner severity/leniency over longer time periods have implications for maintaining standards, and must be monitored.

Sources of bias in marking

When an examiner marks a script, bias can be introduced from several sources: the standard of the script relative to others in the allocation (contrast/context effects); the text of the script itself; the candidate; or the examiner. Although it is hard to assess the extent of these biases, Meadows and Billington (2005) found a large body of research pertaining to these areas (pp. 23–25).

Contrast/context effects

Meadows and Billington (2005) describe various studies that show that the mark awarded to a script can be influenced by the standard of the immediately preceding scripts (pp. 23–25). Good work is assessed more favourably when it follows work of lower quality than when it precedes such work. Poor quality work is assessed more severely when it follows work of higher quality. The contrast effect occurs even when the target script is preceded by only two contrasting scripts. Thus, some authors suggest that reading through several pieces of work before starting to mark may be insufficient to prevent contrast effects biasing the marks awarded to the first few pieces of work.

Research has found that both analytic and holistic marking are equally susceptible to contrast effects.

Some research exists into ways to eliminate contrast/context effects. In one study markers were explicitly warned about the effect and asked to categorize the essays qualitatively before re-reading them and awarding final grades. Another study provided the examiners with model essays. But in both cases there was no difference in the extent to which the examiners were influenced by context. It is still possible that providing model essays may reduce the influence of context on marking in subject areas where factual accuracy rather than written communication is being assessed.

The text of the script

Meadows and Billington (2005) found many studies that showed that the text of the script itself affected the marks awarded (pp. 25–-27). Handwriting had a major effect; with higher grades being awarded to scripts with good handwriting. A similar effect was found in a study of recording quality in a spoken test; poor quality recordings were marked more harshly than good quality recordings.

A more complex study showed that the effect of handwriting quality is not uniform. Examiners' marks were influenced by handwriting and the attractiveness of the alleged author when the student was female, but no such effects were found if the student was male.

Meadows and Billington found substantial evidence that other factors of written work, such as essay length, reading level, spelling and grammar, influence the marks awarded.

Encouragingly, two studies that investigated the marking of A-level scripts by experienced examiners found no evidence of bias related to handwriting. Meadows and Billington suggest that "the well-defined marking schemes and good community practice brought about by well-managed standardisation meetings … might reduce the effects of presentational style*"* (p. 26).

An obvious measure to remove the influence of handwriting and presentation is to have candidates type their work. There is evidence, however, that typed scripts are marked more harshly than handwritten answers.

The candidate

Research has shown that examiners' marking can be influenced by characteristics of the candidate, including gender, race, social class, physical attractiveness, and attractiveness of the first given name (pp. 27–-30).

Meadows and Billington (2005) found that the largest body of literature on this topic related to gender bias. Gender biases appear to be subject specific but there were no clear patterns (pp. 27–28).

A simple way to reduce gender bias and also the effect of name stereotypes is 'blind marking', that is, not providing the candidate's name on the script. However, the effectiveness of blind marking might be limited because there is evidence that the candidate's gender can be determined from the handwriting, content and style of language used.

Meadows and Billington also note that no gender biases were found in the few studies that investigated it in the marking of public examinations by experienced examiners. This may be because the "tightly defined marking schemes used … leave little room for sex bias …" (p. 29). One study directly tested this hypothesis and found it to hold true.

Only two studies investigating ethnic bias were cited (p. 30). Meadows and Billington question whether the results could be generalized to experienced examiners using tightly defined mark schemes and suggest this as an area for further research.

The examiner

Meadows and Billington (2005) found evidence in the literature for a number of biases stemming from the examiner him/herself (pp.30–35).

Ideological bias – Caused by examiners having fundamental disagreements about what constitutes the ideal in their subject. However, "it is likely that the tightly defined mark schemes and standardisation of examiners removes [this] effect in GCSE and A level marking" (p. 30).

Examiner background – Investigation of the influence of marker background on marking reliability is important for establishing examiner recruitment criteria.

Meadows and Billington found a number of studies that suggested that inexperienced markers tend to mark more severely than experienced ones, and that training eliminates these differences. But some other studies found no such effect (p.30-31).

The evidence of a relationship between marker experience and marking consistency is more inconclusive (pp. 31–34). Some studies found no effect of examiner experience while others found that experienced examiners were slightly more accurate but that this effect could be negated by item choice or training. Many authors argued that examiner selection criteria could be relaxed if the correct training were provided or if unskilled/semi-skilled examiners were only marking certain items (clerical marking).

Examiner traits – Attempts to link personality traits with marking performance have been made. However, the small scale of these studies, and rather ambiguous personality measures, did not allow "sensible interpretation of the effect … on marking reliability" (p. 34). Similarly, transient aspects of the marker, such as fatigue and mood, may have important effects on marking reliability but the studies on this effect are too few and varied to draw any meaningful conclusions.

The effect of question format, subject and choice of essay topic on marking reliability

Question format

Meadows and Billington (2005) found that "numerous studies [show that] closely defined questions, which demand definite answers, are associated with higher reliability" (pp. 35–37). Question type and subject are intrinsically connected. Examinations in subjects that are predominately mathematically based require tightly prescribed questions with definite answers, which in turn result in high interrater correlations. Whereas, subjects that placed most dependence on essay-type questions, such as English, are least reliably marked.

Objective tests, by definition, can be scored with perfect reliability. Meadows and Billington (2005) note that objective testing is used extensively in the United States of America and discuss whether it should be used more widely in the UK. The main argument against its use is that reliability may be achieved at the expense of validity and that "where the nature of the domain [examined] calls for extended writing, the attendant difficulties of marking consistently have to be accepted" (Meadows and Billington, p. 36). Other authors have supported this view that unreliability is inevitable for some subjects if essay-type questions are valued. However, researchers have found a correlation between holistic ratings of essays and objective test scores, and have shown objective tests to be a more valid predictor of the quality of essays than other essay tests.

Meadows and Billington (2005) describe various studies that show that both interrater and intrarater reliability is poor when marking essays (pp. 37–38). In one case "the level of agreement between marks awarded to essays by the same examiner over time was no better than the level of agreement between two different examiners" (p. 37). It has, however, been found that agreement tends to be better at the extreme ends of the performance range.

Candidates' choice of essay topic

Meadows and Billington found many studies showing that the problem of low reliability in the marking of essays is exacerbated by the candidates' choice of essay topic (p. 38).

Reliability is lower if the subject matter is discursive and inexact. One study showed that essay topics that were considered more difficult tended to get higher scores, suggesting that raters "may be unconsciously rewarding test takers who choose the more difficult prompt or may have lower expectations for that topic" (p. 38). Offering a choice allows candidates with different strengths to choose a topic that suits them, and research backs this up by showing that although the marks awarded were affected by the question answered, this was mostly accounted for by differences in quality of the answers.

Studies of the processes by which examiners rate essays

Meadows and Billington (2005) found a number of studies that examine the process by which examiners make their decisions. They suggest that "an understanding of the processes by which examiners rate essays is needed to inform techniques to improve essay marking and reliability" (pp. 38–41). These studies found that different examiners use different approaches to decide what score to allocate to a script, and that examiners develop their own individual method regardless of mark schemes and training. Research into the thought process behind holistic marking showed that examiners are influenced by factors

such as handwriting, writing style, and grammar, and that when reading one essay after another the examiners naturally begin to make comparative statements about the work, rather than considering each piece individually.

Other work suggests that reliability is reduced when an essay is not ‘conventional’ and does not fall into the pattern expected by the examiner.

Improving the reliability of essay marking

Meadows and Billington (2005) found in the literature a number of suggestions to improve the reliability of essay marking, including: all candidates writing on the same topic; removing the names from the scripts; examiner training; double-marking; averaging marks from two samples of writing and encouraging examiners to read quickly and score their first impression (p.41). Empirical tests of methods to improve reliability, such as matching scripts to exemplars or producing a mark from several separate assessments of the same piece of work by the same examiner, had no effect on reliability.

The effect of mark scheme/rating system on marking reliability

Meadows and Billington (2005) found that ”research has revealed that an unsatisfactory mark scheme can be the principal source of unreliable marking” and that ”with some exceptions, the introduction of detailed assessment criteria leads to improvements in marking consistency” (p. 42). The more subjective the marking, the less reliable the final mark set is likely to be. Further improvements in the understanding of the mark scheme can be made by providing exemplars, piloting the mark scheme, joint development of the criteria by those assessing the work, and periodical review of the criteria (p. 42).

However, experiments involving manipulation of the mark scheme did not appear to increase marking reliability, and there is evidence to suggest that agreement between markers can be obtained even in the complete absence of assessment criteria (construct referencing or general impression marking) (p.43).

Meadows and Billington (2005) discuss the use of holistic and analytic marking of essays (pp.44–47). Holistic scoring is rapid but only a single score is reported, thus the same score assigned to two separate pieces of work may represent two entirely distinct sets of characteristics. In contrast, in analytic scoring a mark is awarded to each of a number of different aspects of the task. It is, therefore, much slower than holistic marking, but provides more information about the candidate’s ability.

Comparative studies of the reliability of the different marking methods show that analytic marking is more reliable than holistic marking in some cases, but that there is no difference in other cases. Analytic marking is more labour intensive so, in terms of time and cost, several holistic markers are equivalent to one analytic marker, and there is some evidence that the pooled results of a set of holistic markers are more reliable than that of one analytic marker.

Meadows and Billington (2005) found a number of criticisms of analytic marking. One problem is that error could be compounded when a single marker makes multiple judgments. Evidence showed that the reliability of analytic marking decreased as the level of

sophistication of the essay increased. Other studies showed that experienced examiners can have difficulty assigning a score based on certain descriptors. There is also concern that ”the analytic method of scoring may fragment effects that remain intact in global reading” (p. 46). The process of concentrating on individual aspects of a piece of writing may divert attention away from the overall effect of its whole, and may therefore not be a valid means of assessment.

The validity of holistic marking has also been questioned. It is suggested that agreement between holistic scores may be because examiners depend on ”characteristics in the essays which are easy to pick out but which are irrelevant to ‘true’ writing ability” (p. 46) such as handwriting, vocabulary, spelling and length of essay.

Procedural influences on marking reliability

Consensus versus hierarchical approaches to achieving marking reliability

Meadows and Billington (2005) discuss the hierarchical approach to standardisation of marking that is employed by examination boards in the UK (pp. 47–48). One of the aspects of the system is that assistant examiners have samples of their work re-marked by more senior examiners. Evidence shows that the marks allocated by the second examiner are influenced by those awarded by the first examiner, but not by the latter’s written comments. However, in one case, removing the initial marks and/or the comments made no difference to the second set of marks.

Removing the first set of marks appears to be important when measuring marker reliability. However, if the scores of the two judges are to be combined in some way to determine the final mark it may not be necessary to have independent judgments. In fact, processes of reconciling differences, rather than averaging independent scores, may be a better way to determine final score. Many authors believe that marker agreement does not, necessarily, equate to marking quality and that individual self-consistency is more important than differences between markers (pp. 49–50). Meadows and Billington note, however, that “in public examinations the grades … have great currency so consistency between examiners is crucial*”* (p. 50).

Training and feedback

Training is often cited as essential for compensating for different examiner backgrounds and expectations, and familiarizing examiners with the mark scheme. However, Meadows and Billington (2005) found little empirical research to assess which parts of training are effective and why. Of the few studies they found, some showed that training was successful while others showed that it had no lasting effect (p.50-51). Similarly, when examiners received feedback on their marking reliability was increased in some cases, but in other cases no effect was found. Many authors argue that training should focus on making examiners more self-consistent and that it is most effective on new examiners. Finally, if the mark scheme is explicit, training may not be needed at all (p. 52).

Community of practice

Meadows and Billington (2005) found a large body of recent work that considered whether “reliable marking [is] the product of an effective community of practice” (p. 53). This theory assumes that “standards do not solely reside in explicit assessment criteria or mark

schemes, some knowledge cannot be committed to paper. The latter tacit knowledge is instinctive and commonly held" (p. 53). They found considerable evidence to support the argument that discussion between examiners is needed for reliability. It is likely that these effects explain findings where examiner meetings, rather than mark schemes, are crucial to reliability (pp. 53–54).

There have been suggestions that 'ownership' and shared decision making about the mark scheme would improve reliability. Meadows and Billington found one study that aimed to test this empirically, but the data did not support the idea that consensus improved reliability (pp. 53–54). The authors argued that the mark scheme had a strong standardizing effect in itself. However, responses to questionnaires showed that examiners valued the co-ordination meetings and appreciated the opportunity for discussion.

Exemplar material

Meadows and Billington (2005) briefly discuss the use of exemplar scripts in marking (pp. 55–56). They note that while exemplars can be useful there are some drawbacks. Exemplars of the same standard can differ dramatically from one another and can become quickly outdated. It is also important to provide exemplars that illustrate the range of achievement associated with each mark band.

Double and multiple marking

Meadows and Billington (2005) discuss the large body of literature on double and multiple marking. The research shows large gains in reliability from double marking (p.56-59). Markers are not required to agree with one another, and many authors suggest that this is a merit of the system, allowing a "truer, all-round picture" (p. 56) to be established. However, if there is too much disagreement, aggregating the marks would lead to bunching around the mean, which in turn would reduce discrimination.

The way in which the marks of multiple raters should be aggregated has received much discussion, with suggestions ranging from simple addition and averaging to complex formulae. Also, the exact method of re-marking has been debated (pp. 59–60).

During the 1960s and 1970s double marking was used by awarding bodies in examinations with subjective assessment. It has now mostly disappeared in this context, mainly due to the difficultly in recruiting enough examiners, although it is still common in Higher Education. Suggestions have been made that each script could be marked by a human and by a computer, with a second blind marking by a human in the event of large disagreement.

Remedial measures to detect/correct unreliable marking

Meadows and Billington (2005) found little information about how unreliable marking is detected or corrected, with the exception of the code of practice of UK awarding bodies. They did, however, find a large body of literature discussing the various methods that could be used to adjust marks and their relative merits (p.60-64).

Awarding bodies have a tolerance limit for each paper, and only marking falling outside this tolerance is adjusted. The use of tolerance recognizes that there may be legitimate differences in professional judgment. In addition, small adjustments are hard to justify on the basis of re-marking only a small sample of scripts: a different sample may have resulted in a different adjustment.

Various factors can be used to determine whether an adjustment should be made: percentage of marks that lie outside tolerance; average absolute mark difference; confidence intervals; background information on the reliability of the examiner's marking; and direction of adjustment.

There are also a number of different ways of making adjustments: the mean difference between the assistant examiner's marks and those of the senior examiner is applied to all the assistant examiner's marks; the median difference between the marks is applied to all the marks; different adjustments are applied to different mark ranges; a line of best fit between the senior and assistant examiner's marks is calculated (regression adjustment).

One researcher showed that the estimate of marker reliability increased with the number of scripts re-marked, but that there was little to be gained beyond a certain number of scripts. He could not, however, draw any firm conclusions about the exact number of scripts that should be sampled.

Another study argued that even if adjustment is small it can significantly affect candidates who were unlucky enough to be marked by an especially severe examiner on most of their work. However, candidates with exceptionally good answers may end up being unfairly downgraded if they were marked by a lenient examiner. Compared to double marking, adjustment is quick and inexpensive. The authors warn, however, that if an examiner knows their marks have been adjusted they may become inconsistent in their marking.

Meadows and Billington (2005) also found a number of studies investigating the effectiveness of mark adjustment. The research shows that for many students adjustment is effective, but for a reasonable minority the final mark awarded is actually further from that awarded by the senior examiner than the original mark. Adjustment only works for consistent severity/leniency and cannot overcome all the inconsistencies in marking, especially if the examination contains different tasks. Thus adjustments must be applied with caution.

Methods for detecting unreliable examiners used by UK awarding bodies

Meadows and Billington (2005) review the methods used by UK awarding bodies to detect unreliable examiners including: enquiries on results, comparison of predicted and achieved grades, office review, borderline review, identification of 'lingering doubt' examiners using regression analysis, and checking for clerical errors (pp.64–67). They identify various pieces of research that investigate the effectiveness of these measures.

A study on grade comparison showed that, of the examiners selected for re-marking, the percentage of marks adjusted varied widely with subject and, although it had a substantial effect on grade distribution, there was no difference in the number of result enquiries.

Two studies on office review showed that examiners whose marks were adjusted had a relatively low level of changes post-results. However, there were contradictory findings concerning examiners who were referred to the office review but whose marks were not adjusted. One study found that these examiners had a higher proportion of upgrades than examiners who were not referred to the review, while another study found no such effect. Neither study was able to determine how many upgrades were avoided by the process of office review.

Studies on borderline review suggest that it can identify and correct a number of marking errors, but argue that, to be fairer, the process should cover all grade boundaries and include a larger mark range around the boundary. One author argued that grades should move down as well as up, so that the assessment more accurately reflects achievement, while another showed a subject bias in the number of mark increases.

The reliability of e-marking

Meadows and Billington (2005) note the benefits of e-marking in terms of the monitoring of examiner reliability, the early identification of problems and the elimination of clerical error. However, they were able to find only a few studies on the reliability of e-marking and these show *"*small and inconsistent differences in [its] reliability*"* (p. 67). E-marking often involves examiners marking individual items rather than whole scripts. There are some theoretical advantages to this approach, but there is little research into the effects of part versus whole script marking.

Conclusions

"The literature reviewed has made clear the inherent unreliability associated with assessment in general, and associated with marking in particular. The extent of this unreliability may vary across subjects and assessment formats, and may be improved through marker training, attention to marking schemes and so on. Nonetheless while particular assessment formats, for example essays, are valued by those involved in education there has to be an acceptance that the marks or grades that candidates receive will not be perfectly reliable. There are two possible responses to that acceptance, report the level of reliability associated with marks/grades, or find alternatives to marking." (p. 68)

The need to routinely report reliability statistics alongside grades

Even when the reliability coefficient is high, the number of candidates who are wrongly graded can be large. Thus, many authors have called for awarding bodies to publish the reliability coefficient and/or the possible margin of error associated with a result so that users of the results can be better informed as to the limitations of the examination. In fact, a number of examination bodies in the USA report a range of marks for each candidate, based on the standard error. Another way to report reliability is the number of candidates getting the 'correct' grade.

Meadows and Billington (2005) note that "to not routinely report the levels of unreliability associated with examinations leave awarding bodies open to suspicion and criticism*"* (p. 70). However, "there would need to be further empirical and conceptual groundwork aimed at reaching consensus on the degree of reliability that is acceptable and unacceptable for the uses to which test results are put" (p. 69).

Alternatives to marking

Meadows and Billington discuss two alternatives to marking: Thurstone paired comparison of scripts; and computer marking (p. 70). The limited studies show that these methods can be as reliable as human marking, but more research is needed.

Computer marking of closed questions is used routinely. Methods to extend computer marking to open questions are being investigated. Some research has looked at essay marking, using a computer to analyze features of the answer such as number of characters, number of sentences, sentence length, number of low frequency words used, and so on. One study found that “the correlation between the number of characters keyed by the candidate and the scores given by human markers are as high as the correlation between scores given by human markers” (p. 71). However, there are serious concerns about the validity of a scoring system such as this.

Computer marking has also been investigated for questions where a range of acceptable responses can be compiled, such as short answer science questions. The computer will be completely reliable, in that the same marks will be produced if the responses are re-marked, but different marks might have been allocated if a different examiner had to provided the marking rules that the computer followed.

Appendix 2

Search strategy and the review process

Search strategy

This appendix contains details of the search strategy, which used five different types of source to ensure thorough coverage of the evidence base:

- A range of general bibliographic databases
- Websites of key organisations
- Reference harvesting of key documents
- Contact with UK awarding bodies and assessment organisations for unpublished studies
- Contact with individual experts to identify additional unpublished sources.

The first stage in the process was for the NFER's information specialists to match database keywords to the research questions and agree the search strategy with Ofqual.

Searching was next carried out across the specified databases and web resources. These websites were searched on main keywords and/or the publications/research/policy sections of each website were browsed as appropriate. In addition, the journal "Research Matters" was hand searched. All searches were limited to publication years 2004-2012, in English language only.

Individual subject experts were also contacted and references were harvested from key documents.

A brief description of each of the databases searched, together with the keywords used, is outlined below. The search strategy for each database reflects the differences in database structure and vocabulary. Smaller sets of keywords were used in the more specialist databases. Throughout, the abbreviation 'ft' denotes that a free-text search term was used, the symbol * denotes truncation of terms and '?' denotes a wildcard used to replace any single character.

British Education Index (BEI)

(searched via Dialog Datastar 15/10/2012)

BEI provides information on research, policy and practice in education and training in the UK. Sources include over 300 journals, mostly published in the UK, plus other material including reports, series and conference papers.

#1 Analytical based marking (ft)
#2 Assessment criteria (ft)
#3 Assessment objectives (ft)
#4 Automated marking (ft)
#5 Awarding (ft)
#6 Awarding bod* (ft)
#7 Blind marking (ft)
#8 Classical Test Theory (ft)
#9 Computer assisted testing (ft)
#10 Computer based marking (ft)
#11 Construct validity
#12 Criteria based marking (ft)
#13 Cultural bias
#14 Data interpretation
#15 Double marking (ft)
#16 E marking (ft)
#17 Ethnic bias
#18 Examiner selection (ft)
#19 Examiner training (ft)
#20 Examiner* (ft)
#21 Generalisability Theory
#22 Grade descriptors (ft)
#23 Grades (scholastic) (ft)
#24 Grading (ft)
#25 Grading criteria (ft)
#26 Holistic assessment
#27 Interrater reliability
#28 Item analysis
#29 Item based marking (ft)
#30 Item Response Theory
#31 Item/ question/ script seeding (ft)
#32 Latent Trait Theory
#33 Level descriptors (ft)
#34 Many Facets Rasch Model (ft)
#35 Mark adjustments (ft)
#36 Mark schemes (ft)
#37 Marker agreement (ft)
#38 Marker judgement (ft)
#39 Marker selection (ft)
#40 Marker training (ft)
#41 Markers (ft)
#42 Marking (ft)
#43 Marking (scholastic)
#44 Mark* bias (ft)
#45 Marking reliability (ft)
#46 Marking tolerance (ft)
#47 Measurement techniques
#48 Moderation (marking
#49 Multilevel modelling
#50 Multiple marking (ft)
#51 Online marking (ft)
#52 On-screen marking (ft)
#53 Paper based marking (ft)
#54 Predictive validity
#55 Quality control
#56 Question formats (ft)
#57 Ranking systems (ft)
#58 Rater agreement (ft)
#59 Rater reliability (ft)
#60 Rating scales
#61 Rating systems (ft)
#62 Reliability
#63 Re marking (ft)
#64 Sample size
#65 Scores
#66 Scoring
#67 Script based marking (ft)
#68 Sex bias
#69 Social bias
#70 Standardisation (ft)
#71 Standards
#72 Test bias

#73 Test format

#74 Test items

#75 Test questions (ft)

#76 Test reliability

#77 Test results

#78 Test validity

#79 Testing

#80 #1 or #2 or #3 or … #77 or #78 or #79

#81 A level examinations (ft)

#82 A level examinations (AS)

#83 A levels (ft)

#84 Examination papers

#85 Examination results

#86 Examination scripts (ft)

#87 Examinations

#88 General Certificate of Educational Achievement (ft)

#89 General Certificate of Secondary Education

#90 I GCSEs (ft)

#91 International GCSE Level 1/2 certificates (ft)

#92 National Curriculum

#93 Standardised tests

#94 Mode 3 examinations

#95 Scottish Certificate of Education

#96 #81 or #82 or #83 or … #93 or #94 or #95

#97 #80 and #94

Education-*line* (searched 15/10/12)

Educati*on-line* represents the collection of documents submitted directly to the BEI by their authors, with any newer content typically resulting from annual conferences of the British Educational Research Association (BERA).

#1 Mark*

#2 Examiner*

#3 Awarding bod*

#4 Examinations

#5 Examination paper*

#6 Examination script*

#7 A levels

#8 GCSEs

Education Resources Information Center (ERIC)

(searched via Dialog Datastar 12/10/12)

ERIC is sponsored by the United States Department of Education and is the largest education database in the world. Coverage includes research documents, journal articles, technical reports, program descriptions and evaluations and curricula material.

#1 Analytical based marking (ft)
#2 Assessment criteria (ft)
#3 Assessment objectives (ft)
#4 Automated marking (ft)
#5 Awarding (ft)
#6 Awarding bodies
#7 Blind marking (ft)
#8 Classical Test Theory (ft)
#9 Computer assisted testing
#10 Computer based marking (ft)
#11 Construct validity
#12 Criteria based marking (ft)
#13 Cultural bias
#14 Data interpretation
#15 Double marking (ft)
#16 E marking (ft)
#17 Ethnic bias
#18 Examiner selection (ft)
#19 Examiner training (ft)
#20 Examiner* (ft)
#21 Generalisability Theory
#22 Grade descriptors (ft)
#23 Grades (scholastic) (ft)
#24 Grading (ft)
#25 Grading criteria (ft)
#26 Holistic assessment
#27 Interrater reliability
#28 Item analysis
#29 Item based marking (ft)
#30 Item Response Theory
#31 Item seeding (ft)
#32 Question seeding (ft)
#33 Script seeding (ft)
#34 Latent Trait Theory
#35 Level descriptors (ft)
#36 Many Facets Rasch Model (ft)
#37 Mark adjustments (ft)
#38 Mark schemes (ft)
#39 Marker agreement (ft)
#40 Marker judgement (ft)
#41 Marker selection (ft)
#42 Marker training (ft)
#43 Markers (ft)
#44 Marking (ft)
#45 Marking (scholastic)
#46 Mark* bias (ft)
#47 Marking reliability (ft)
#48 Marking tolerance (ft)
#49 Measurement techniques
#50 Moderation (marking
#51 Multilevel modelling
#52 Multiple marking (ft)
#53 Online marking (ft)
#54 On-screen marking (ft)
#55 Paper based marking (ft)
#56 Predictive validity
#57 Quality control
#58 Question formats (ft)
#59 Ranking systems (ft)
#60 Rater agreement (ft)
#61 Rater reliability (ft)
#62 Rating scales
#63 Rating systems (ft)
#64 Reliability
#65 Re marking (ft)
#66 Sample size
#67 Scores
#68 Scoring
#69 Script based marking (ft)
#70 Sex bias
#71 Social bias
#72 Standardi?ation (ft)
#73 Standards
#74 Test bias

#75 Test format

#76 Test items

#77 Test questions (ft)

#78 Test reliability

#79 Test results

#80 Test validity

#81 Testing

#82 #1 or #2 or #3 or …#79 or #80 or #81

#83 A level examinations

#84 A level examinations (AS)

#85 A levels (ft)

#86 Examination papers

#87 Examination results

#88 Examination scripts (ft)

#89 Examinations

#90 General Certificate of Educational Achievement (ft)

#91 General Certificate of Secondary Education

#92….I GCSEs (ft)

#93 International GCSE Level 1/2 certificates (ft)

#94 National Curriculum

#95 Standardised tests

#96 Mode 3 examinations

#97 Scottish Certificate of Education

#98 #83 or #84 or #85…or #95 or #96 or #97

#99 #82 and #98

Idox (searched 16/10/12)

The IDOX Information Service covers all aspects of local government. Key areas of focus include public sector management, economic development, planning, housing, social services, regeneration, education, and environmental services.

#1 Mark*

#2 Examiner*

#3 Awarding bod*

#4 Examinations

#5 Examination paper*

#6 Examination script*

#7 A levels

#8 GCSEs

Websites

Website	URL	Number of results
AQA Centre for Educational research and policy	http://web.aqa.org.uk/ http://cerp.aqa.org.uk	18
Cambridge Assessment Research Division	http://www.cambridgeassessment.org.uk/ca/About_Us/Our_Structure/Research_and_Consultancy/Research_Department	
Cambridge International Examinations	http://www.cie.org.uk/	0
Edexcel	http://www.edexcel.com/Pages/Home.aspx	0
International Curriculum and Assessment Agency (ICAA)	http://www.icaa.com/	0
OCR	http://www.ocr.org.uk/	0
Ofqual	http://www.ofqual.gov.uk/	28
WJEC	http://www.wjec.co.uk/	0
Council for the Curriculum Examinations and Assessment (CCEA)	http://www.rewardinglearning.org.uk/	0
Chartered Institute of Educational Assessors	http://www.ciea.org.uk/	0
Centre for Evaluation and Monitoring	http://www.cemcentre.org/	1
Standards and Testing Agency	http://www.education.gov.uk/aboutdfe/armslengthbodies/b00198511/sta	0
Oxford University Centre for Educational Assessment	http://oucea.education.ox.ac.uk/	9
Institute of Education (IOE)	http://www.ioe.ac.uk/	0
Joint Council for Qualifications	http://www.jcq.org.uk/	0
Federation of Awarding Bodies	http://www.awarding.org.uk/	0
American Educational Research Association	http://www.aera.net/	0
Educational Testing	http://www.ets.org/	

Website	URL	Number of results
Service		
International association for educational assessment (IAEA)	http://www.iaea.info/	28
SQA	http://www.sqa.org.uk/sqa/CCC_FirstPage.jsp	3
College board	http://collegeboard.org/	
International Baccalaureate	http://www.ibo.org/	8

Appendix 3

The evidence base for the review

This appendix provides a brief description of the items of literature included in the main body of the review, together with the review team's rating of the quality and relevance of each item. Descriptions of the ratings appear below the table.

Item of literature	Brief description	Quality	Relevance
Ahmed, A. and Pollitt, A. (2011). 'Improving marking quality through a taxonomy of mark schemes', *Assessment in Education: Principles, Policy & Practice*, **18**, 3, 259–278.	This work aims to develop a taxonomy to show how mark schemes may be designed, or improved, to minimise any threats to valid interpretation of the results of an examination. It is based on the premise that a mark scheme should help markers decide how many marks to award each response, concentrating on responses that are close to a score boundary. In addition, the markers should award these marks based on a consensual view of the trait they want students to demonstrate (as described in the Importance Statement for the subject).	High	High
Al-Bayatti, M. and Jones, B. (2005). *NAA Enhancing the Quality of Marking Project: the Effect of Sample Size on Increased Precision in Detecting Errant Marking.* London: QCA	Secondary analysis and simulation based on real NCA data.	Modest	Medium
Baird, J., Greatorex, J. and Bell, J.F. (2004). 'What makes marking reliable? Experiments with UK examinations', Assessment in Education: Principles, Policy & Practice, 11, 3, 331-348.	This paper presents the results of two research studies that investigated aspects of examiner standardisation procedures. The first study looked at the effects on marking accuracy of different types of exemplar scripts. The second study looked at the effects of different types of standardisation meetings.	High	High

Item of literature	Brief description	Quality	Relevance
Baird, J.A., Hayes, M., Johnson, R., Johnson, S. and Lamprianou, L. (2012). *Marker Effects and Examination Reliability: a Comparative Exploration from the Perspectives of Generalizability Theory, Rasch Modelling And Multilevel Modelling*. Coventry: Ofqual.	Collaborative research project comprising a comparative study of the contributions that three different analysis methodologies could make to the exploration of rater effects on examination reliability.	Modest	High
Bramley, T. and Dhawan, V. (2010). *Estimates of Reliability of Qualifications*. Coventry: Ofqual	Investigating and reporting information about marker reliability in high-stakes external school examinations. The report also contains a useful review of findings in this area.	High/Strong	High
Bramley, T. (2009). 'The effect of manipulating features of examinees' scripts on their perceived quality.' Paper presented at the Association for Educational Assessment – Europe Annual Conference, Malta, November	Investigation of the effect of 'non-relevant' features of an exam script on the score given by an examiner.	Modest	Of some relevance
Bramley, T. (2007). 'Quantifying marker agreement: terminology, statistics and issues', *Research Matters*, **4**, 22–27. Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational*	Review of the terminology used to describe indicators of marker agreement and discussion of statistics which are used in analyses.	Modest	Medium

Item of literature	Brief description	Quality	Relevance
Studies, **52**, 1, 29–46.			
Johnson, M. (2008). 'Assessing at the borderline: judging a vocationally related portfolio holistically', *Issues in Educational Research*, **18**, 1, 26–43.	A small scale study which focused on how assessors holistically judged a portfolio of evidence. The study investigated the cognitive strategies that underpinned their judgments of a school-based vocationally-related assessment containing borderline pass and merit characteristics.	Modest/Impressionistic	Of some relevance
Johnson, M., Hopkin, R., Shiell, H. and Bell, J.F. (2012). 'Extended essay marking on screen: is examiner marking accuracy influenced by marking mode?' *Educational Research and Evaluation*, **18**, 2, 107–124.	Comparison of onscreen vs paper marking of extended essays. Part of a wider research project which looked broadly at the influence of marking mode on 12 examiners' marking outcomes and processes when assessing samples of extended essays.	High	Medium
Johnson, S. (2011). *A Focus on Teacher Assessment Reliability in GCSE and GCE*. Coventry: Ofqual [online].	Literature review on the reliability of teacher summative assessment in GCE and GCSE examinations.	High	High
Meadows, M. and Billington, L. (2007). *NAA Enhancing the Quality of Marking Project: Final Report for Research on Marker Selection.* London: QCA	Research project comparing quality of marking of four groups of possible markers.	High	Mostly relevant
Newton, P.E. (2009). 'The reliability of results from national curriculum testing in England', testing in England,	Assessment of the reliability of results from National Curriculum Assessment.	High	Highly relevant/ strong

Item of literature	Brief description	Quality	Relevance
Educational Research, **51**, 2, 181–212.			
Opposs, D. and He, Q. (2011). *The Reliability Programme Final Report.* Coventry: Ofqual	A two-year research programme, conducted by Ofqual to investigate the reliability of results from national tests, public examinations and other qualifications in order to develop regulatory policy on reliability.	High	Strong
Pollitt, A. (2012). 'The method of adaptive comparative judgement', *Assessment in Education: Principles, Policy & Practice*, **19**, 3, 281–300.	This paper describes the theoretical basis of Adaptive Comparative Judgment (ACJ), and illustrates it with outcomes from some trials.	High	Mostly relevant/ impressionistic
Baker, E., Ayres, P., O'Neil, H.F., Chli, K., Sawyer, W., Sylvester, R.M. and Carroll, B. (2008). *KS3 English Test Marker Study in Australia: Final Report to the National Assessment Agency of England.* Sherman Oaks, CA: University of Southern California.	Marker studies were conducted collaboratively by the National Assessment Agency (NAA) in London, the University of New South Wales in Sydney, and Advance Design Information in Los Angeles.	High	Mostly relevant/ strong
Billington, L. (2012). *Exploring Second Phase Samples: What is the Most Appropriate Basis for Examiner Adjustments?* Manchester: AQA, Centre for Education Research and Policy.	For examinations that are marked on paper, two samples of each examiner's marking are evaluated. The first phase sample (FPS) of 10 scripts is done immediately after training to check that standardisation has been successful. The second phase sample (SPS) is taken half way through marking and comprises 50 scripts, selected by the examiner. The Team Leader will re-mark 15 of these and, if the original marking is outside the tolerance, will re-mark an additional 10 scripts. This sample of 25 re-marked scripts is used to make decisions about examiner adjustments. Examiners thought to be consistently lenient (or severe) will have an adjustment applied to all the scripts in their	High	Mostly relevant/ modest

Item of literature	Brief description	Quality	Relevance
	allocation. On-screen monitoring involves the introduction of seed items for which 'true' scores have already been determined by the Principal Examiner. The procedural differences can be summarised as follows: **Paper**: • Sample is self-selected by examiner from their allocation • Re-marked by Team Leader on paper • Team Leader sees marks/annotations of first examiner. **Online**: • Pre-selected sample assigned a 'true' score by the Principal Examiner • Re-marked by examiners onscreen • No marks/annotations are present. Research suggests that a Team Leader's re-marking of paper SPSs is influenced by the marks/annotations of the first examiner, resulting in greater marking accuracy than would be found for cleaned scripts (Murphy, 1979; Baird and Meadows, under review).		
Black, B. (2010). 'Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks.' Paper presented at International Association for Educational Assessment 36th Annual Conference, Bangkok, Thailand, 22-27 August.	In on-screen marking, the quality of the seeding items that are used to monitor (and improve) marker accuracy is important. As is the use and interpretation of the data gathered. An understanding of how various features of seeding items influence marker agreement will have implications for the levels of agreement that might be realistically expected. Factors that affect marker agreement can be grouped into three categories: i) item features; ii) mark scheme features; iii) candidate response features. Previous research has shown that many features have an effect on marker accuracy. The ones with the strongest effect appear to be: maximum mark; whether the mark scheme is objective, points-based or levels-based; points/marks ratio (agreement was higher for items where the number of acceptable answers equals the number of marks than for those where this number exceeds the number of marks). That is, in general, the more	High	Highly relevant/ strong

Item of literature	Brief description	Quality	Relevance
	constrained the mark scheme, the higher the marking accuracy. Also, items which require markers to make simple intuitive judgements (matching or scanning) were associated with higher marking accuracy than those which require more complex reflective judgements (evaluation or scrutinising). There is mixed evidence of the influence of superficial candidate response features (e.g. neatness and legibility) on examiners' choice of marks. Some experimental studies involving teachers as markers (e.g. Shepherd, 1929; Briggs, 1970, 1980; Bull and Stevens 1979; Markham, 1976) have found that neater and more legible handwriting is associated with higher marks. However, studies which involve experienced exceeds the number of marks). That is, in general, the more constrained the mark scheme, the higher the marking accuracy. Also, items which require markers to make simple intuitive judgements (matching or scanning) were associated with higher marking accuracy than those which require more complex reflective judgements (evaluation or scrutinising).		
	There is mixed evidence of the influence of superficial candidate response features (e.g. neatness and legibility) on examiners' choice of marks. Some experimental studies involving teachers as markers (e.g. Shepherd, 1929; Briggs, 1970, 1980; Bull and Stevens 1979; Markham, 1976) have found that neater and more legible handwriting is associated with higher marks. However, studies which involve experienced examiners have not shown such effects (e.g. Massey, 1983; Crisp, 2007).		
Brooks, V. (2004). 'Double marking revisited', *British Journal of Educational Studies*, **52**, 1, 29–46.	A review of the "all but forgotten" literature on double marking and a consideration of its current (2004) relevance	Medium	Mostly relevant
Burslem, S. (2011). *The Reliability Programme: Final Report of the Policy Advisory Group.* Coventry: Ofqual	The Reliability Programme undertaken by Ofqual investigated the reliability of results from National Curriculum assessments, public examinations and vocational qualifications with the aim of developing regulatory policy on reliability. The Policy Advisory Group (PAG) was appointed to investigate public perceptions of reliability and develop regulatory policy on reliability. It was	High	Modest

Item of literature	Brief description	Quality	Relevance
	made up of representatives from various stakeholders, including assessment experts, assessment providers, employers, communications experts, teachers, students and parents. The group explored ways to improve public understanding of reliability concepts, communicate reliability evidence to the public and increase public confidence in the examinations system. They also considered the adequacy and appropriateness of the recommendations from the Technical Advisory Group to the Reliability Programme.		
Curcin, M. (2010). 'A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice', *Research Matters*, **10**, 27–32.	Literature review focussing mainly on inter-marker agreement in the context of on-screen marking. The increasing use of on-screen marking provides new possibilities for monitoring marking and ensuring higher agreement levels.	Medium	Mostly relevant
Dhawan, V. and Bramley, T. (2012). *Estimation of Inter-rater Reliability*. Coventry: Ofqual.	An analysis of data gathered from on-screen marking of 8 components of the June 2011 live OCR examination session. In particular, data from multiple markings of 'seed' scripts, for which a 'definitive' mark had been determined, was used to investigate marker accuracy. Four of the components comprised short-answer questions, where each item was worth less than eight marks. The other four, referred to as long components, had at least one item which was worth eight marks or more. Marker accuracy was compared between the short and the long components, with the expectation that the long components would be more difficult to mark reliably.	High	Mostly relevant/ modest
Fearnley, A. (2005). *An Investigation of Targeted Double Marking for GCSE and GCE.* London: QCA	A research study to investigate whether double marking can improve reliability. Scripts were used from a live examination session, but the study was not conducted at the same time as the live marking.	High	Highly relevant/ modest
Fowles, D. (2009). 'How reliable is marking in GCSE	Marking reliability was explored in two current AQA GCSE English specifications. Specification A differentiates mainly by outcome, while	Medium	Of some relevance/

Item of literature	Brief description	Quality	Relevance
English?' *English in Education*, **43**, 1, 49–67.	Specification B differentiates by task. Both specifications comprise two written papers (each 30% of the total mark) and two coursework assessments (20% each), for each of two tiers of assessment, the Higher tier (targeted on grades A* to D) and the Foundation tier (targeted on grades C to G). In Specification A the questions are the same in both tiers, other than that for the Foundation tier a number of bullet points are provided to guide the candidates' responses. The mark scheme for the two tiers is, therefore, virtually the same. Specification B questions have no overlap in the two tiers.		modest
Massey, A.J. and Raikes, N. (2006). *Item-level Examiner Agreement*. Cambridge: Cambridge Assessment	This study considers the degree of inter-examiner agreement that should be expected at item level. It also considers surface features of the items and their mark schemes that might be expected to influence the reliability with which they are marked. Surface features considered are: 1. The subject 2. The level of examination 3. The maximum mark for the item 4. The implied time restriction (ITR) imposed on candidates. This is: Total time in minutes x (item max mark/total max mark) 5. Type of marking: objective, points based or levels based. Objective marking – items require very brief responses and greatly constrain how candidates may respond. E.g. candidates must make a selection, order information, match information according to criteria, locate or identify a piece of information, write a single word or give a single numerical answer. Credit-worthy responses can be sufficiently pre-determined to make a mark scheme that only requires superficial judgements by the marker. Points based marking – items require brief responses ranging from a few words to one or two paragraphs, or a diagram or graph. The salient points of all or most credit-worthy responses may be pre-determined so that the marker only has to locate the relevant elements and identify all variations that deserve credit. There is generally one-to-one correspondence between salient points and marks. Levels based marking – items require longer answers, from one to two	High	Mostly relevant/ modest

Item of literature	Brief description	Quality	Relevance
	paragraphs to multi-page essays or other extended responses. The mark scheme describes levels of response, each of which is associated with a band of one or more marks. Markers apply a principle of best fit when deciding the mark.		
Opposs, D. and He, Q. (2011). *The Reliability Programme Final Report.* Coventry: Ofqual	The Office of Qualifications and Examinations Regulation (Ofqual) in England conducted a two-year research programme, from 2008 to 2010, to investigate: the reliability of results from national tests, public examinations and other qualifications; and the public's understanding of and attitudes towards unreliability. The information produced would be used to develop regulatory policy on reliability of examinations. The Programme had three strands: **Strand 1:** generating evidence on the reliability of results from a selection of national qualifications, examinations and other assessment in England through empirical studies **Strand 2:** interpreting and communicating evidence of reliability **Strand 3:** Investigating public perceptions of reliability and developing regulatory policy on reliability. Two advisory groups were formed. The Technical Advisory Group, made up of educational assessment experts, advised on strands 1 and 2. The Policy Advisory Group, made up of representatives from a wide range of stakeholders, advised on Strand 3.	High	Mostly relevant
Pinot de Moira, A. (2011). *Why Item Mark? The Advantages and Disadvantages of E-Marking.* Manchester: AQA, Centre for Education Research and Policy	A short article on the advantages and disadvantages of splitting papers into items for use in e-marking.	Medium	Mostly relevant
Raikes, N. (2006). 'The Cambridge Assessment/Oxford	A three-year research project that investigated the application of computational linguistics techniques to the automatic marking of short, free text answers to examination questions.	High	Of some relevance

Item of literature	Brief description	Quality	Relevance
University automatic marking system: does it work?' *Research Matters*, **2**, 17–20.	The research focussed on GSCE Biology because the question papers contained large numbers of questions requiring short, factual, written answers. Two broad approaches to automatic marking were taken: 1. 'Information extraction' involved writing by hand 'machine marking schemes' for each item to be automatically marked. 2. 'Machine learning' involved trying various machine learning techniques to learn the marking scheme from a sample of human marked answers. A hybrid approach using semi-automatic methods to produce the machine marking scheme was also investigated. The machine learning and hybrid approach showed promising results in terms of reducing the amount of specialised work required to set up new items. For details see Pulman and Sukkarieh (2005). A complete prototype marking system was developed using Information Extraction techniques and it is this system that is the focus of this article. The system works by matching candidate's answers to pre-written patterns to extract pertinent information that has been judged creditworthy (or not) by human examiners. Essentially, the pattern covers the synonyms for each pertinent piece of information. The patterns are written by hand. In this investigation the pattern writers (based in Oxford) were provided with question papers, mark schemes and 200 sample answers that had been marked and annotated by two senior examiners independently to indicate exactly which parts of the answer gained (or forfeited) marks. Three sets of marks for each of the 200 sample answers were also made available to the pattern writers (the marks of each of the two senior examiners and the original live mark awarded to the answer).		
Raikes, N., Fidler, J. and Gill, T. (2010). 'Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology', Research	This paper presents the results of a research study into the effectiveness of face-to-face meetings for examiner standardisation. It investigates the effectiveness of examiner standardisation on new and experienced examiners for short-answer questions and structured essay questions.	High	High

Item of literature	Brief description	Quality	Relevance
Matters, 10, 21-27			
Suto, I., Crisp, V. and Greatorex, J. (2008). 'Investigating the judgemental marking process: an overview of our recent research', Research Matters, 5, 6–8	An extensive research programme that considers the process of marking GCSE and A level examinations from different angles. The projects explore the information people attend to and utilise and the sequences of mental operations involved in marking items.	High	Of some relevance
Taylor, R. (2011). *A Qualitative Exploration of Key Stakeholders' Perceptions and Opinions of Awarding Body Marking Procedures*. Manchester: AQA, Centre for Education Research and Policy.	Qualitative study of knowledge and perceptions of the examination marking process.	Medium	Of some relevance/ impressionistic

Descriptions of quality ratings

High: large scale quantitative study; or in-depth case studies that cover a range of institutions and a wide range of stakeholders, where views are triangulated; or a meta-analysis or systematic review.

Medium: quantitative or qualitative studies with smaller sample sizes, or covering only a small number of institutions. Qualitative studies that do not cover a full range of stakeholders. Non-systematic reviews.

Low: based on observation or opinion, or on one school case-study, or the views of one person, for example.

Descriptions of relevance ratings

High: very relevant to all or most questions

Medium: at least moderately relevant to most questions

Of some relevance: relevant to some questions

Low: at least slightly relevant to one question

What is the strength of the evidence base for this item?
Strong (e.g. large scale quantitative study with adequate sample sizes to allow scope for statistical analysis – ideally an RCT or a QED such as baseline/follow-up; or a comparison group design, or in-depth case studies that cover a range of institutions and a wide range of stakeholders, where views are triangulated)
Modest (quantitative or qualitative studies with smaller sample sizes, or covering only a small number of institutions. Qualitative studies that do not cover a full range of stakeholders)
Impressionistic (based on observation or opinion, or on one school case-study, or the views of one person, for example)

Appendix 4

Mark scheme types

This Appendix provides a brief description of the different types of mark scheme mentioned in this report.

Objective/constrained mark scheme

Items that are objectively marked require very brief responses and greatly constrain how candidates must respond. An unambiguous correct answer exists for the question which can be completely defined in the mark scheme. The distinction between right and wrong is completely transparent and the marker does not need to use any subjectivity. Examples include multiple choice questions, answers in the form of a single word or number, questions that require matching or sequencing of given information and questions that require the indication or identification of information on the question paper (e.g. indicating an area on a diagram). Objective mark schemes can be applied with a high degree of accuracy.

For example:

Name the capital city of Finland.

or

Write the chemical symbol for Sodium.

Points-based mark schemes

These items usually need responses ranging in length from a few words to one or two paragraphs, or a diagram or graph. Points-based mark schemes list objectively identifiable words, statements or ideas. Marks are awarded one at a time for each creditworthy point in the candidate's response. There is generally a one-to-one correspondence between the number of correct answers that the candidate gives and the number of marks that should be awarded (up to the maximum mark). All the creditworthy points are listed in the mark scheme but the marker still needs to find the relevant elements in the response.

One criticism of this type of mark scheme is that the relative importance of different statements is rarely addressed – every point is treated as equal in value. Therefore, if the maximum mark is lower than the number of creditworthy points, a candidate can achieve full marks even if they omit fundamental parts of the answer. Similarly, the tactic of simply writing down everything that comes to mind, even if it is not relevant, can achieve high marks without the candidate fully understanding what they are writing.

Marker agreement on points-based mark schemes decreases as the number of points increases (Black, 2010).

Levels-based mark schemes

These items usually require longer answers, ranging from one or two paragraphs to multiple page essays. Levels-based mark schemes divide the mark range into several bands, each representing a distinguishable level of quality of response. The level descriptors may include features of language, content or both.

In a holistic levels-based scheme, markers make an overall judgment of the performance. Each level may include a number of different response features but no explicit weightings are given to the different features. Therefore, if a response merits different levels for different aspects, the marker must use their judgment to decide the 'best fit' category, without explicit information about which aspects are most highly valued. The result is that different markers may award different marks because they have different understandings of what it means to be 'good'. Alternatively, markers may award the same mark for different reasons. These issues both undermine the construct-validity of the test: that is, the same marks may not mean the same thing in terms of the trait that the test is supposed to measure.

Analytic levels-based mark schemes separate the aspects of interest and provide level descriptors, and associated mark bands, for each aspect. That is, they explicitly weight the different features of response.

Appendix 5

Classical Test Theory

Classical test theory assumes that each person has a true score on the trait being measured, be it a body of knowledge, competence in a skill or prediction of future potential in work or further study. The theoretical definition of true score is the average score over infinite independent replications of the test. Clearly, it is impossible to perform infinite replications of a test and, therefore, we can never directly measure true score, only the observed score. Thus, it is assumed that:

Observed score (X) = True score (T) + measurement error (E)

Measurement error is assumed to be a random variable that is normally distributed with a mean of zero. If the standard deviation (spread) of the error is small then replications of the measurement will produce similar results, that is, the distribution of observed scores will be similar across testing occasions. The reproducibility of results, or the degree to which they are error-free, is known as *reliability*.

The reliability of the test results ρ_{XT}^2 is defined as the ratio of true score variance σ_T^2 to observed score variance σ_X^2

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$$

The variance of the observed scores can be shown to equal the sum of the variance of true scores and the variance of errors[8], so

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2}$$

This equation shows that reliability increases as the proportion of error in the test scores decreases, and vice versa. In addition, it shows that reliability is equivalent to the proportion of variance in the test scores that we could explain if we knew the true scores. However, we cannot know the true scores so reliability must be estimated using other methods.

One method of estimating reliability is to use parallel tests. It is assumed that the parallel forms produce the same true score for every individual, *i*, and the same distribution of errors on each test. Under these assumptions it can be shown that the correlation between the scores on the parallel tests is equal to reliability.

Where parallel tests are not available, a measure of internal consistency, known as Cronbach's α, can be used to measure reliability. For a test with *k* items u_j, *j* = 1, ... , *k*. The total test score for an individual, *i,* is defined as

$$X_i = \sum_{j=1}^{k} U_{ij}$$

[8] Assuming that the scores of any examinee are uncorrelated with any other examinee.

And Cronbach's alpha is

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma_{U_j}^2}{\sigma_X^2}\right)$$

Where $\sigma_{U_j}^2$ is the variance on the j^{th} item.

Cronbach's alpha is a measure of the covariance between items in a test. It treats any covariance between items as true score variance

Generalizability Theory

Generalizability theory generalises the assumptions of classical test theory by assuming that the items making up a test are a random sample from a larger 'universe' of items. A candidate's expected score in the universe is analogous to a true score. The Generalizability coefficient is analogous to reliability in classical test theory and is defined as the ratio of the variance in universe scores to the variance of observed scores.

A major difference between G-theory and classical test theory is that G-theory can separate out the relative effects of different sources of error (facets), whereas classical test theory only deals with one source of error at a time. A G-theory analysis will quantify the amount of measurement variance attributable to each facet under investigation (item, marker, occasion etc.) and to the interaction between the facets. Ideally, most of the variance in measurement should come from the object of measurement (i.e. individual candidates), with little variance resulting from the other facets (which all represent measurement error).

The results of a Generalizability study can also be used to design better assessments because they can be used to model what would happen if different aspects of the measurement were altered. For example, the effects of changing the number of items in a test or employing multiple markers can be investigated.

Item response theory

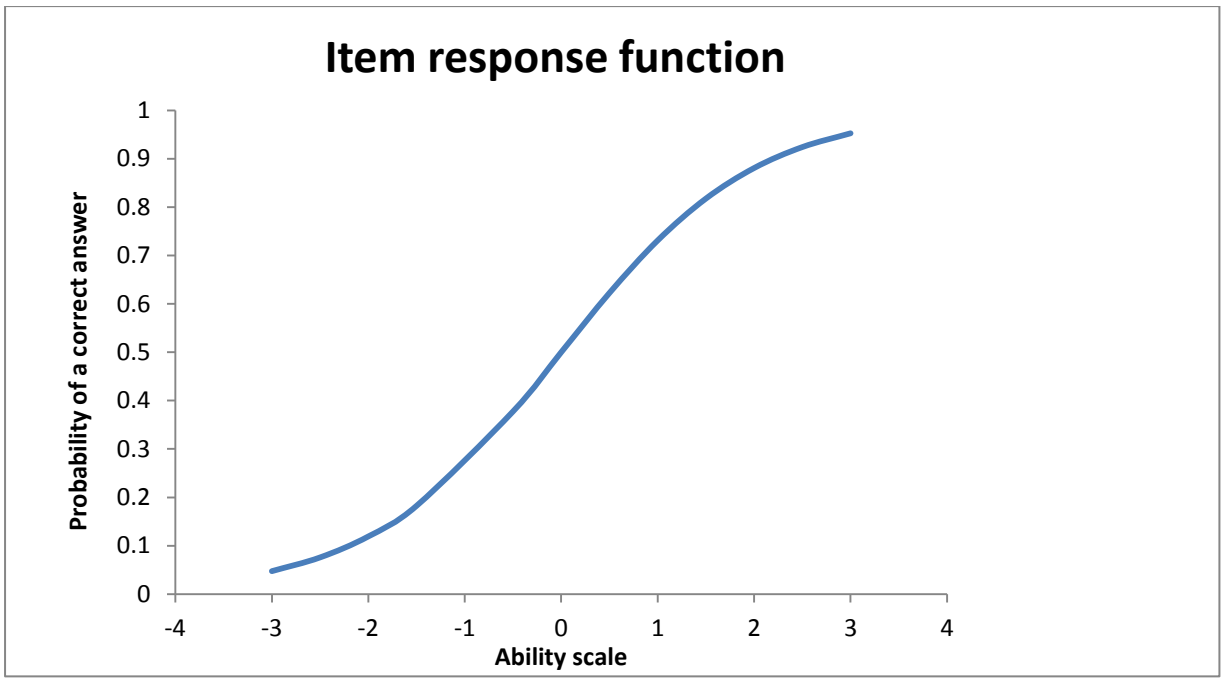
In item response theory an ability scale is created through statistical analysis and candidate ability, item difficulty and marker severity are all placed on the same scale. The ability measure represents performance in the trait of interest and measurement error is a function of ability. Where data obeys its assumptions, it makes it feasible to give comparable scores to candidates who may have taken different tests, provided there is some means of linking these results.

Traditional[9] item response theory makes three assumptions: first, that the trait of interest is unidimensional; second, that the items are unrelated apart from the fact that they measure the same trait, i.e. the items are locally independent; and third, that a candidate's response to an item can be modelled with an item response function (IRF).

The item response function gives the probability that a candidate of a given ability will answer an item correctly; the lower the candidate's ability the lower the probability of a correct answer, and vice versa. The exact probability will depend on the 'item parameters'

[9] Multi-dimensional IRT models do not assume a unidimensional trait.

which essentially determine the shape of the IRF. IRT models can also incorporate partial-credit scoring.



The most general model for dichotomous items has three item parameters:

- Difficulty – the position of the item on the ability scale. This is the point at which the probability of a correct answer is 0.5. In the example above, the item is medium difficulty because the probability of 0.5 coincides with the centre of the ability scale.
- Discrimination – the scale or slope of the IRF at the point on the ability scale where the probability of a correct answer is 0.5, which equates to how well an item distinguishes between candidates of varying ability.
- Guessing/chance – the asymptotic minimum of the function, i.e. the lowest probability of a correct answer for that item. For example, in a multiple choice question with four (equally plausible) answers, even the lowest ability candidates would have a probability of 0.25 of getting the answer correct by guessing.

If guessing is unlikely to occur or is irrelevant then the asymptotic minimum is zero. This is known as a two parameter model. In this case, a candidate whose ability is equal to the item difficulty will have a probability of 0.5 of answering correctly. If the candidate's ability is higher than the item difficulty the probability of a correct answer will be between 0.5 and 1. If the candidate's ability is lower than the item difficulty the probability of a correct answer will be between 0 and 0.5.

In some models discrimination is assumed to be the same for all items and so the only parameter included is item difficulty. This is described as a one parameter model, or sometimes as a Rasch model. It is also possible to include an asymptotic maximum into the model, but this is rarely done in practice.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2013

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346