

2019

No. 2

✱INSIDE ASSESSMENT

Research news & comment

Quality of marking

Examiner decision making

Exploring cognitive strategies

Monitoring marking

Taking corrective action

Against the grain

Different approaches to
evaluating marking reliability

Engaging expertise

Considering the impact
of examiner specialism

Centre for
Education Research
and Practice (CERP)

cerp.org.uk

AQA 
Realising potential

✚ INSIDE ASSESSMENT

Research news & comment

Issue No. 2 Quality of marking

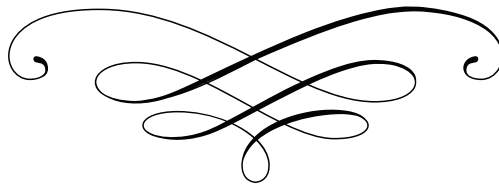
News and findings from AQA's Centre for Education
Research and Practice (CERP)

Edited by

Lena Gray, Will Pointer and Lindsay Simmonds

Contributors

Yaw Bimpeh, Sarah Hack, Liz Harrison, Zeek Sweiry and Claire Whitehouse



Inside assessment

Published by AQA

Except as permitted under current legislation, no part of this work may be photocopied, stored in a retrieval system, published, performed in public, adapted, broadcast, transmitted, recorded or reproduced in any form or by any means without the prior permission of the publisher.

Copyright © 2019 AQA and its licensors. All rights reserved.

AQA Education (AQA) is a registered charity (registered charity number 1073334) and a company limited by guarantee registered in England and Wales (company number 3644723). Registered address: AQA, Devas Street, Manchester M15 6EX.

Further articles and research papers are available to download at cerp.org.uk
Please send editorial correspondence to cerp@aqa.org.uk

Edited, designed and produced by Claire Jackson

Images © Shutterstock unless otherwise stated

Printed by Optichrome

CONTENTS

NO. 2 2019

- 6 Welcome to *Inside assessment*
- 8 Marking quality in context
- 10 Exploring the cognitive processes behind marking
extended written responses
- 14 Identifying inconsistent markers in real time
- 20 The impact of examiner characteristics
- 28 AQA's marking in numbers
- 30 Beyond classical statistics: three approaches to analysing
marking reliability
- 36 From the archives: features of a levels-based mark
scheme and their effect on marking reliability
- 46 Contextual notes
- 48 Meet the researchers – CERP's marking
reliability experts

From Lena Gray, Director of Research at AQA's Centre for Education Research and Practice (CERP)

EVERY ELEMENT OF AN assessment must be of a high quality in order to ensure that the measurement of a student's ability is reliable and valid. There are many different features of an assessment, including test design, marking and standard setting. Research produced by AQA's Centre for Education Research and Practice (CERP) broadly focuses on these three aspects. While the first instalment of *Inside assessment* showcased past and present work into standard setting, our second volume focuses on quality of marking.

AQA sets and marks around half of all GCSEs and A-levels taken in the UK every year (see contextual notes, p. 46). Marking is closely monitored at all stages of the process, providing rich datasets for CERP's researchers to analyse. Using this

data, CERP research has shown the impact of mark schemes on effective marking and has been able to identify which sorts of questions are complex to mark. Our findings inform AQA's work on assessment design as part of our commitment to continuous improvement.

The examiners who mark the papers also play a major role in the process. AQA's examiners must meet particular criteria (see p. 46). Certain characteristics – such as teaching experience, examining experience and level of education – have an impact on marking reliability, and CERP undertakes considerable research into this area. An example of a recent project can be found on p. 20.

It's also important to consider the cognitive strategies used by examiners in their marking; techniques for exploring

these strategies are expanding, and there are multiple lines of enquiry open to the assessment researcher (p.10).

Marking data can be analysed in a wide variety of ways, and CERP's work includes both qualitative and quantitative research. Researchers are always keen to push the boundaries and try new

methods: marking research offers an opportunity to look beyond classical statistics (p. 30).

This journal provides an overview of AQA's current activity, with a nod to our past achievements (p. 36). ■



Dr Lena Gray is Director of Research at AQA's Centre for Education Research and Practice, and is an Honorary Norham Fellow of the Department of Education, University of Oxford. Prior to this, Lena held a number of positions in the Scottish Qualifications Authority and its predecessor organisations, chiefly in relation to qualifications and assessment reform work. She also has experience as a teacher and tutor at the University of Strathclyde.

Quality of marking in context

Marking is a critical component of the exam life cycle, and, as such, it is high on the assessment research agenda. **William Pointer** explains how marking fits into broader aspects of quality assurance, and outlines some of the key issues within this area

IN ENGLAND, STUDENTS TAKE THE General Certificate of Secondary Education (GCSE) at the age of 16, and a further subject-specific qualification – the A-level – thereafter.

These exams are used to apply for further education and employment, and, because they affect life chances and even social mobility, they are referred to as ‘high-stakes’ exams. Quality of marking in high-stakes assessments is critical, and awarding organisations – including AQA – go to considerable lengths to build in quality assurance at every level of the process. As such, this theme underpins a great deal of CERP’s output.

A key concept in assessment is reliability, described as the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure

and hence are inferred to be dependable and repeatable for an individual test taker (Berkowitz, Wolkowitz, Fitch, & Kopriva, 2000). Reliability can be affected by the test, the candidates or the marking. There are a number of elements surrounding marking reliability, including:

- **Mark-scheme design**
Good quality assessments are predicated on good assessment design. The design of the mark scheme is influential in ensuring quality of marking
- **Standardisation**
This is the process of standardising markers so that they apply the mark scheme in a consistent manner
- **Quality assurance**
Monitoring markers to ensure they

continue to mark to the right standard and providing ongoing feedback

- **Markers**


What type of characteristics are linked with people who are able to mark consistently.

In addition to reliability, high-stakes assessment providers must also consider the importance of validity. Validity centres on ensuring that the test measures what it is intended to measure. Reliability is a prerequisite for valid assessment, but a test could be reliable while not valid.

To ensure that assessments are valid, we use a variety of question formats, ranging from multiple choice to essays. The marking of multiple choice questions is straightforward: generally, one of the responses is correct and gains one mark, and the others are incorrect and gain no marks. There is no room for subjectivity on the part of the marker and therefore these questions have very high marking reliability. However, not all skills can be tested using multiple choice questions. With questions that require longer and more open responses, markers will have some influence on the candidate's score. CERP's research examines ways to mitigate this risk to ensure that all assessments are fair.

One major innovation in assessment has been the move to on-screen marking, which involves the scanning and electronic distribution of students' exam papers. There are numerous benefits associated with this method, including:

- it has facilitated item-level marking, which allows markers to focus on certain items and improve reliability
- it is completely anonymous, which removes potential sources of bias from marking
- it allows for real-time quality assurance, removing the need for post-hoc adjustments or re-marking.

Technology continues to change many aspects of the marking process. For example, artificial intelligence has the potential to be used to monitor the quality of marking or to automate the marking process, reducing the reliance on human judgement. Alongside technological innovations come alternative ways of marking, such as comparative judgement, where responses are compared with one another rather than given a mark according to a mark scheme; some see this method as a way of increasing reliability. This all points to a rich area for further research in order to understand the potential impact any changes would have on our assessments. 

Understanding hidden processes

Marking essay-based answers requires considerable cognitive effort on the part of the examiner. We can explore the decision-making processes using cognitive strategies, but, as **Sarah Hack** writes, detailed analysis calls for multiple lines of enquiry

RESearch into decision making in assessment has tended to focus on an exploration of the cognitive marking strategies used by examiners in their marking. Five strategies have been identified: matching, scanning, evaluating, scrutinising and no response (Suto & Greatorex, 2008a, 2008b). However, these strategies were identified from a sample of examiners who were marking GCSE Business Studies and Mathematics, and, although subsequent research found further support for the five strategies in other subjects, including Biology GCSE (Suto, Nádas, & Bell, 2011), Physics A-level (Greatorex & Suto, 2006), and Geography A-level (Crisp, 2008), research to date has focused predominantly on subject areas where there is little scope for extended writing. It is therefore still not clear how applicable the cognitive marking strategies previously identified are applicable to higher-level examination questions involving extended writing – this was the starting point for the research outlined below.

The first study

The initial exploratory study saw five experienced examiners of A-level Psychology interviewed about their own marking strategies. The examiners were also asked to consider the marking strategies of less experienced colleagues. In addition, three of the five examiners completed a task that required them to mark an extended written response to an A-level Psychology question worth 24 marks, while ‘thinking aloud’ – i.e. vocally expressing their thoughts during the marking process. Although this was a small study, thematic qualitative analysis of the interviews and the ‘think aloud’ transcripts suggested that while the cognitive marking strategies previously identified were both recognised and used by the A-level examiners, the marking of extended written responses is a more complex and cognitively demanding process than previously identified. Ultimately, it is an iterative process, involving references to the question, the mark scheme and the written response, and involves an ongoing process of

judgement. The predominant strategy is 'evaluating'.

The second study

This involved 43 participants who were asked to complete a marking task while thinking aloud, then complete an online questionnaire that further explored the marking process. In this study, the participants marked a range of AS-level Psychology questions while thinking aloud, as described above. The questions marked represented a range of question types: a multiple-choice question, a short-answer question and two questions requiring extended written responses – one worth six marks and one worth 12 marks. The last two questions had levels-based mark schemes, the application of which has been linked to lower marking reliability (Massey & Raikes, 2006; Raikes & Massey, 2007). Marking reliability was investigated by comparing the marking decisions made by the participants with the marks originally awarded by the principal examiner.

Senior examiners made reference to the development of an 'internalised marking schema' over the marking period, which enabled them to mark more quickly

Statistical analysis found that the marking of the questions requiring extended responses was associated with a higher likelihood that the marker would

change his/her mind as to the final mark to award, a greater number of readings of the response and lower confidence ratings regarding the marking decision reached. Further, although marking reliability was generally very good (91% of markers were within tolerance for the total mark awarded to the four questions), there was a significant association between the question type and the proportion of markers who were within tolerance, with 98% of the short-answer question marks being within tolerance compared to 72% of the marks awarded to the 12-mark 'Discuss...' question requiring an extended written response. With regard to cognitive marking strategies, as before they were all recognised as being used in the marking process, but it was clear that the evaluating strategy was key to the marking of extended written responses.

Further qualitative analysis led to the development of a model of marking with evaluating at its core. Markers were found to make informal evaluations concurrent-

ly with a careful, initial reading of an extended written response. These informal evaluations were seemingly more

subjective judgements on the quality of the response. However, the marking decision was made with a subsequently more objective, formal evaluation against

the mark scheme, generally associated with a second lighter-touch reading of the response. Further evaluation often occurred when the marker reflected on their marking decision and this was accompanied by another light-touch rereading of the response, seemingly to confirm the marking decision. Reference was also made in the questionnaire data to further evaluation occurring sometime later, after the marking of subsequent responses or following feedback from a team leader, when a marker might question their earlier judgement and return to responses to reconsider their marking decisions.

Study 2 reflected marking at the very start of the marking period, so two further studies were conducted in June and July of this year, once again using A-level Psychology examiners to explore changes in the marking process over the intensive three-week A-level examining period (see contextual notes; p.46). Study 2 identified a clear two-stage process where formal evaluation against the mark scheme led to the marking decision. However, the senior examiners interviewed in the first study made reference to the development of an 'internalised marking schema' over the marking period, which enabled them to mark more quickly. They also raised the issue of having an experience-informed 'gut feeling' or 'professional expectation' as to the quality of the response. These two ideas suggest that the two-stage model identified in study 2 may change over the marking period, as examiners internalise the mark scheme and are perhaps able to

'recognise' a response as being a particular level and/or mark.

The third study

In order to explore what happens to the marking process as examiners become increasingly familiar with the mark scheme, A-level Psychology examiners were emailed four online questionnaires to complete, sent at regular intervals throughout the three-week marking period. The examiners were asked to reflect on their marking at the four time points. The analysis of this data, which is in process, will allow an investigation of how the model of marking developed from study 2 changes over the marking period.

The fourth study

Another study ran concurrently; it is hoped this work will provide further insight into changes in the marking process. The study involved a small sample of examiners ($n = 5$) whose eye movements were tracked as they carried out a marking activity. The examiners marked an extended written response worth 16 marks from this summer's Psychology A-level, once at the start of the marking period and then again, three weeks later, at the end of the marking period.

Of particular interest in this study are the reading behaviours of markers and their reliance on the formal mark scheme at the two time periods data was collected. In both studies, it will be possible to explore the marking process in relation to marking reliability, as the response marked in the eye-tracking study was one of the seeds, and data from the seeds will also be

integrated into the data analysis of study 3. Analysis of both studies is in process, but it is hoped that in addition to providing insight into how the marking process changes as examiners become experienced over the marking period, a final study might lead to the development and testing of an intervention to improve the speed and accuracy of examiners' judgements, although the exact nature of the intervention will depend on the findings of the current analysis. ■

REFERENCES

- Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, 38, 247–264.
- Greatorex, J., & Suto, I. (2006, May). *An empirical exploration of human judgement in the marking of school examinations*. Paper presented at the International Association for Educational Assessment Conference, Singapore.
- Massey, A., & Raikes, N. (2006, September). *Item-level examiner agreement*. Paper presented at the British Educational Research Association (BERA) Annual Conference, University of Warwick
- Raikes, N., & Massey, A. (2007). Inter-level examiner agreement. *Research Matters: A Cambridge Assessment publication*, 4, 34–37.
- Suto, I., & Greatorex, J. (2008a). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policy & Practice*, 15, 73–89.
- Suto, I., & Greatorex, J. (2008b). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, 34, 213–233.
- Suto, I., Nádas, R., & Bell, J. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, 26, 21–51.



Sarah Hack is the recipient of a CASE PhD studentship, funded by the ESRC SEDTC and AQA, investigating decision making in educational assessment. Prior to starting the research in October 2016, she was a teacher and examiner of A-level Psychology. Her supervisors are Dr Naomi Winstone and Dr Adrian Banks from the University of Surrey and Dr Neil Stringer from AQA. The researchers are extremely grateful to all the A-level Psychology examiners who have participated in the four studies described above.

Identifying inconsistent markers

Monitoring marker reliability is a critical part of an awarding organisation's quality assurance; however, most observation methods require data that can only be captured once marking has ended. **Yaw Bimpeh** outlines an ongoing research project that explores a way of detecting inconsistent markers in real time, presenting an opportunity for corrective action

HIGH-STAKES EXAMS SUCH as GCSEs and A-levels (see contextual notes p. 46) must have high levels of marking reliability, and considerable resources are

in place to ensure that is the case.

However, most observation methods – such as generalisability theory and many-facet Rasch modelling (see p. 30) – require evidence that accumulates over time; therefore, they cannot be used to identify inconsistent markers and take corrective action during live marking. Recent CERP research has explored ways to provide real-time feedback on the quality of marking.

This work adapts the empirical Bernstein's concentration inequality (Boucheron, Lugosi, & Massart, 2013) to study the divergence of markers' scores from the definitive scores. It expresses

how the score awarded by the marker differs to its expectation. This method provides a valuable way of detecting potential inconsistent markers, and is more compatible with operational marking.

In England, the increase in on-screen marking allows clear monitoring of markers' work. Awarding organisations use seed scripts or responses and peer-pair double marking to monitor on-screen marking. Seed responses or scripts are pre-selected and are marked by the senior examiner, or senior examiner panel; they are then introduced randomly into a marker's allocation. The marker is not aware that a particular response or script is a seed and has no knowledge of the exact mark that has been awarded. If markers pass seeds, they can continue to mark their allocation of responses/scripts; if they fail a set number of seeds, they will

Seed responses or scripts are pre-selected and are marked by the senior examiner, or senior examiner panel; they are then introduced randomly into a marker's allocation. The marker is not aware that a particular response or script is a seed and has no knowledge of the exact mark that has been awarded. If markers pass seeds, they can continue to mark their allocation of responses/scripts; if they fail a set number of seeds, they will be stopped.

be stopped. In current practice, the tolerance of seed marks is often set according to the maximum mark for the item. The critical question is whether the current rules for stopping a marking can be enhanced.

Recent CERP research, which follows on from Pinot de Moira (2010), addressed the essential question of whether the quality assurance system in on-screen marking can be further improved. Specifically, it answers the following questions: (1) how many seeds does an

examiner need to mark for quality assurance? (2) how can we infer the likelihood of the marker marking within tolerance? (3) how do we determine optimal error tolerance for an item or script to ensure marking quality? and (4) how can we detect markers who are not marking to the required standard with some level of assurance?

The proposed method can tell us how many seeds are required for quality assurance. Table 1 illustrates how the minimum 'cost' in terms of samples for

Table 1: Minimum sample size required to ensure that the standard tolerances applied to marking of items guarantee 80%, 90%, 95% and 99% assurance in detecting marking error

Maximum mark	Tolerance	Sample size assurance probability 80%	Sample size assurance probability 90%	Sample size assurance probability 95%	Sample size assurance probability 99%
1-3	0	13-116	17-150	21-185	29-265
4-7	1	19-57	24-74	30-91	42-130
8-12	2	19-42	24-54	30-67	42-96
13-20	3	22-52	29-67	35-82	50-118
21-30	4	32-65	42-85	51-104	74-150
31+	5	44	58	71	102

quality control relates with tolerance and assurance probability using Bernstein’s concentration inequality. As shown in Table 1, if we would like a smaller tolerance, sample size for an assurance probability of 80% should increase accordingly. Assurance probability as high as 95% to 99% would demand impossibly high sample sizes. Note that Table 1 has a range of values because each row represents a range of maximum marks; for example, the required sample size is 19 when the maximum mark is four, but 57 when the maximum mark is seven.

The concentration inequality also provides detailed information on how we should

reframe our thinking about the examiners’ performance in the marking of examinees’ work. For example, Table 2 provides empirical quality of marking statistics by different items and markers.

Performance statistics for a sample of 10 markers are presented in Table 2. For each item it shows how many seeds were marked out of tolerance. Our proposed method for detecting inconsistent markers is applied to the sample. With the new method, the processes of monitoring the quality of marking can be posed as evaluating whether we believe the marker is consistent or not.

Table 2: Monitoring of markers using multiple marking data

Number of items out of tolerance by marker/Total number of seeds marked

Item	Max. mark	Tol	Total seeds	Av. seeds marked	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
a	2	0	42	19.45	3/15	3/19	2/15	1/14	1/15	1/25	4/22	6/25	2/23	1/21
b	2	0	42	20.31	6/21	9/26	3/18	4/21	6/25	1/19	3/26	1/25	0/22	0/23
c	5	1	40	21.10	1/15	5/12	1/19	3/21	0/21	1/25	2/31	0/35	0/29	1/32
d	6	1	44	18.42	1/14	3/15	3/17	1/17	0/12	1/18	1/20	4/29	0/25	2/31
e	5	1	39	21.36	3/20	2/18	5/28	2/25	3/27	1/23	8/34	0/25	3/34	0/33
f	10	2	38	21.91	2/13	1/17	1/25	1/28	2/31	0/25	2/33	2/34	1/33	0/33

Table 3: Monitoring of marking using concentration inequality

Marking within tolerance (yes (Y)/no (N)) by marker

Item	Max. mark	Tol	Total seeds	Av. seeds marked	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
a	2	0	42	19.45	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
b	2	0	42	20.31	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
c	5	1	40	21.10	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
d	6	1	44	18.42	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
e	5	1	39	21.36	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
f	10	2	38	21.91	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

Table 4: Marking within tolerance (yes/no) by marker

Assurance probability by marker

Item	Max. mark	Tol	Total seeds	Av. seeds marked	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
a	2	0	42	19.45	<u>76.69</u>	80.43	84.01	89.05	91.46	94.44	94.44	95.29	97.98	98.19
b	2	0	42	20.31	<u>33.03</u>	<u>32.22</u>	<u>46.7</u>	<u>47.59</u>	<u>52.98</u>	<u>75.27</u>	<u>78.55</u>	89.57	93.08	94.06
c	5	1	40	21.10	<u>76.41</u>	<u>78.09</u>	<u>79.76</u>	87.64	93.41	95.11	96.11	98.69	99.45	99.56
d	6	1	44	18.42	91.58	92.59	93.67	94.86	95.81	99.55	99.34	99.84	99.9	99.97
e	5	1	39	21.36	86.68	92.44	97.34	97.72	98.08	98.36	98.81	99.53	99.69	99.97
f	10	2	38	21.91	<u>61.96</u>	89.17	96.76	97.17	97.28	98.13	98.77	98.8	99.46	99.72

The results of applying this method are summarised in Table 3, which shows that for all cases considered, the marker did not exceed the tolerance threshold significantly. So any variation in marking may be due to chance; there is not strong evidence that the marker is applying the mark scheme incorrectly.

Table 4 shows the assurance analysis of the quality of marking data using empirical Bernstein’s concentration inequality. The results show that in general the frequently used sample size for seeding provides adequate assurance probability at the conventional level of 80%. However, in 12 cases (underlined), the assurance probability was below the nominal level of

80%. This means that in these 12 cases there were too few seeds to detect potential inconsistent markers with a high level of certainty. This method offers a quick and practical way of detecting evidence of marking that is out of tolerance, with as few false alarms as possible. An interesting feature of our method is that, unlike the classical method (e.g. paired t-test) based on the central limit theorem, it holds for all fixed sample sizes as opposed to sample sizes approaching infinity.

The proposed method can also be applied automatically to find the optimal number of seeds or instances of double marking needed for quality assurance,

and the optimal error tolerance for an item response or script to ensure marking quality. ■

REFERENCES

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration Inequalities: A Non-asymptotic Theory of Independence*. Oxford University Press.

Lamprianou, I. (2004) *Marking quality assurance procedures: Identifying good practice internationally*. Report produced for the National Assessment Agency.

Pinot de Moira, A. (2010). *Identifying errant markers: quality assurance systems in an e-marking environment*. Manchester: AQA Centre for Education Research and Policy.



Yaw joined CERP in 2014. He holds a PhD in Statistics, an MSc in Mathematical Sciences and a BSc (Hons) in Mathematics. His current areas of research include marking reliability, application of the Bayesian method to standard setting and test equating, and construct validity of assessment designs. Yaw has experience of analysing and modelling data in a variety of fields, and is skilled in the research and application of statistical methods.

Examiner characteristics

The decision makers involved in marking exam papers play a vital role within assessment. **Claire Whitehouse** describes how particular examiner characteristics impact on marking accuracy, and how research findings are contributing to improved reliability

AS EXPRESSED throughout this publication, marking is a critical part of the examination process. Along with the demands of the marking task (this includes the properties of the questions and the mark schemes, and the nature of the students' responses), examiner characteristics exert an influence on marking accuracy, as indicated in a model developed by Suto, Nadas and Bell (2008). Naturally, awarding organisations have stringent criteria when recruiting examiners, in order to ensure the integrity of assessments. For example, potential AQA examiners should have recent teaching experience in the subject and at the level for which they're applying to mark, as well as having appropriate academic qualifications. Experienced examiners with recent examining experience, who no longer teach in the classroom, are also encouraged to continue examining.

CERP research considers how examiner characteristics – primarily, teaching experience, examining experience and level of education – impact on marking reliability. We have also looked at whether different types of question or assessment require other human factors to be taken into consideration, and whether it is an effective use of examiners' expertise to direct certain types of response to individual examiners. The findings are then used to inform examiner recruitment and retention strategies to improve the marking process.

A number of projects investigated the three selection criteria for examiners. These were initiated with a view to widening the pool of potential examiners. Most of the experiments focused on low-tariff items in assessments (questions that require short answers and are worth a small number of marks). Responses to these items are usually one or two words, phrases or sentences, or a short

paragraph. These require examiners to use simple cognitive marking strategies such as matching a student's response to the correct answer provided in the mark scheme (see p. 10 for more about cognitive marking strategies).

Two key findings came out of the designed experiments that used short response items. First, examiners with no teaching experience are likely to mark as accurately as experienced teachers who've taught for at least three years. This finding holds across national curriculum tests in English that were aimed at 14 year olds and selected items from curriculum-embedded high-stakes exams in English, biology, physics and mathematics taken by 16 year olds. The second finding was that the highest level of general education is the best predictor of marking accuracy, better even than the highest level of relevant subject education and marking experience.

The basic skills needed to mark low-tariff, short-response questions consistently and accurately tend to be generic. An examiner needs to understand and interpret a mark scheme as intended and to be able to follow a set of marking instructions. These skills are likely to be found in the graduate population, regardless of subject studied. Broadly speaking, those examiners educated

to degree level and above are likely to experience the least difficulty in marking. Using these findings, awarding organisations started to recruit and train general markers to mark low-tariff, short-response questions with highly constrained correct answers that don't require in-depth subject knowledge.

But what about questions that *do* require in-depth subject knowledge, such as extended responses that most often reward students' work using a levels of response mark scheme (see p. 36)? Here, examiners need to use evaluation skills and complex marking strategies that involve multiple readings to understand the intention behind the student's response.

A study that used selected questions from GCSE English found that some of the questions were marked more accurately by trained teachers with examining experience. That is, that some questions *do* require in-depth subject knowledge. A similar conclusion was made in the context of the marking of a national curriculum test of English. Here, reading and writing tasks based on Shakespeare texts were marked more reliably by examiners with teaching experience. Perhaps teachers' curriculum knowledge of how Shakespeare was taught

QUALITY OF MARKING



in the classroom gave them an advantage when it came to marking responses.

The influence of the examining experience on marking accuracy has proved to be difficult to describe fully. Some early work using multi-level modelling suggested that increasing years of examining led to greater consistency in marking. However, in an examiner

management system that is working well, good examiners will be retained. This makes disentangling the effects of examiner longevity from marking reliability very difficult.

Another way of looking at the question of the effect of examining experience is to compare the marking quality of novice examiners with that of examiners with some experience. Previous experimental

projects have found that when the general education was of a sufficient level (degree and above), novice examiners' marking was of a similar quality to the marking of experienced examiners. Within those participating in the experiments, there were some novice examiners whose marking was better than that of some experienced examiners and vice versa, but on the whole there was little difference between the two groups. What did prove to be an important observation was that there is a training effect. With the same quality of training, novices are able to mark low-tariff, short-response questions with the same accuracy and consistency as experienced examiners.

to be small scale. For example, one of the largest studies involved 359 markers marking the same 200 responses to five items from a GCSE English question paper. Focusing on a few questions or one or two assessments weakens the generalisability of a designed study. Nevertheless, such studies have drawn similar conclusions about the effects of different examiner characteristics and demands of the marking task.

It is fairly straightforward to identify questions that are suitable for marking by a general marker who has a 'common sense' knowledge of a subject. Tariff often acts as a key identifier. However, tariff

Environmental factors may influence the quality of marking for different types of examiner

The training effect has less influence when it comes to marking higher-tariff questions and those that require extended responses. With these question types, examiners who have examining and teaching experience in the relevant subject are least likely to experience difficulties with marking. While designed experiments have proved to be useful in identifying key examiner characteristics, they are costly. Designed studies tend

alone is not sufficient, which is why the allocation of questions to examiners based on their characteristics has not moved beyond the current dichotomy of general marker and expert marker.

Quality monitoring data gathered during live marking, whether whole script or item-based or from on-screen or paper-based marking can be modelled. With these data coming from many

questions and components across a number of subjects, the demands of the marking task and examiner characteristics can be explored. Statistical models are built to facilitate these explorations. Multi-level modelling can be useful because it recognises the hierarchical structure of marking and teams of markers. For example, a four-level model was used to explore the marking reliability of novice examiners with that of experienced examiners. Marking events (level 1) were nested within questions (level 2) which were nested within examiners (level 3) who were nested within the team working on a question paper (level 4).

Evidence, though weak, from this four-level model suggests that examiners' marking was influenced by the nature of the marking teams they found themselves in. In a team with a high proportion of novice examiners (here defined as just under two thirds of examiners), novices' marking is likely to suffer. This may be because team leaders have insufficient time to ensure all novice examiners are marking to the required standard during the entire marking period. This finding suggests that environmental factors may influence the quality of marking for different types of examiner.

So far the focus has been on low-tariff items requiring simple cognitive marking strategies. But what about higher-tariff, extended response items such as essays for which examiners use complex marking strategies? Examiner characteristics combine with the demands of the marking task to influence marking accuracy. A marking task may be described by a number of features related to the question, the mark scheme and the candidates' responses. Models can be set up that control for these demands, while exploring the influence of more nuanced examiner characteristics on that same marking reliability.

Operationalisation of examiner characteristics to create new independent variables can be completed by combining a number of data sets. This was undertaken to explore the influence of examiner specialisms on marking accuracy in an examination in an English literature specification. Candidates responded to two essay prompts worth 40 marks each. In section A of the question paper there was a choice of one essay question from 18. Each question addressed a specific text, for example, *Macbeth* from the Gothic genre and *As You Like It* from the Pastoral genre. In section B the optionality was one essay from six.

Two sets of data were combined to create a variable, 'examiner specialism'. This gave an estimate of the number of texts that an examiner was likely to be involved in teaching. The findings suggested that most examiners (92.5%) specialised in teaching four or fewer texts. The number of texts that an examiner specialised in teaching ranged from one to nine, with the most frequent number being three. Just over a third of students' responses (37.8%) were marked by an examiner who specialised in the text associated with the optional question in section A. Using the new variable of 'examiner specialism' probably underestimated the number of texts per examiner but, nonetheless, gave an indication of the relative coverage of texts taught by the panel of examiners.

model. This model also accounted for question difficulty, quality of response and a training effect. Key features of the marking task were kept constant in this analysis because it focused on one question paper. The optional essay questions had the same maximum mark and format. Marking accuracy was operationalised as the mean absolute mark difference between a senior examiner and a junior examiner. A key finding from this investigation was that it was the number of texts in which an examiner specialises that exerted an influence on the accuracy of marking, not whether the examiner was a specialist in the text on which the student chose to offer a response. For every additional text an examiner was involved in teaching, the mean absolute mark difference decreased

A key finding from this investigation was that it was the number of texts in which an examiner specialises that exerted an influence on the accuracy of marking, not whether the examiner was a specialist in the text on which the student chose to offer a response

Examiner specialism and a second binary variable indicating whether or not a student's response was based on a text that was the examiner's specialism were included in a two-level hierarchical linear

by 0.27 marks. So an examiner who specialised in five texts could be expected to have a mark difference some 1.35 marks lower than an examiner who specialised in one text.

One interpretation of this finding is that experience of teaching a number of texts enables an examiner to appreciate the set of skills necessary to analyse, interpret and evaluate a range of texts. Examiners who teach a limited number of texts may focus their attention on favoured interpretations, rather than skills that can be applied consistently across all texts. Operationally, the finding provided evidence to suggest that matching students' responses to examiners' specialist texts would not show an improvement in the quality of marking that justified the cost.

Using examiner specialism as an independent variable assumed that an examiner is involved to a greater or lesser extent in the teaching of the texts on

which candidates chose to answer questions. He or she might teach all of those texts or they might teach some of the texts and be involved in curriculum decisions around how all of them were taught. This is another aspect of examiners' teaching environments that is worthy of further investigation in future studies.

Many characteristics of examiners and their working and examining environments have yet to be operationalised. Further investigation could yield fruitful insights into how best to recruit and train examiners across the range of question types and subjects offered in curriculum-embedded high-stakes assessments. ■

REFERENCES

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

Pinot de Moira, A. (2005). *Do examiner characteristics affect marking reliability?* Manchester: AQA Centre for Education Research and Practice.

Royal-Dawson, L. and J.-A. Baird (2006). *Is teaching experience necessary for reliable marking?* Manchester: AQA Centre for Education Research and Practice.

Suto, I., et al. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, 26, 21–51.

Whitehouse, C. (2014). *Who is the specialist? The effect of specialisms on the marking reliability of an English Literature examination*. Manchester: AQA Centre for Education Research and Policy.

Whitehouse, C. (2015). *Does examining experience count?* Manchester: AQA Centre for Education Research and Practice.



Claire joined AQA in 2003 working in subject management before taking up a post in the Centre for Education Research and Practice. Her current research work has two strands. The first is the application of e-processes to aspects of the assessment procedure, such as marking long-form responses on screen. The second strand is the teaching and learning processes applicable to younger learners. Claire gained a BSc(Eng) from University College London and a PhD from the University of Cambridge, both in chemical engineering.

Marking in numbers


AQA ensures that its exams are informed by expertise from design until delivery. But how many people are involved in the process? **Will Pointer** crunches the numbers

Around **27,000** academics, teachers, lecturers and subject experts helped us write, set and mark our exams last year

We marked more than **10,000,000** scripts during 2018

There were approximately **12,000,000** seed responses marked

We set more than **12,000** different questions this year



More than **1,100,000**
students sat our qualifications

We delivered **2,853,430**
GCSEs in the 2018
summer series

Over **500,000**
students took our
most popular GCSE
– English Language

For further information,
see contextual notes
on p. 46 ►

Beyond classical statistics

Like other awarding bodies, AQA closely monitors the exam marking process to ensure that mark schemes are applied consistently and fairly. The resulting data is then analysed within CERP; there are several approaches open to the research team, as **Yaw Bimpeh**, **Liz Harrison** and **William Pointer** explain

MARKING IS CLOSELY monitored as part of AQA's quality assurances. This process generates data that can be used to analyse, among other things, marking reliability. As is often the case, there is no single way to carry out this analysis. A variety of approaches exist, and each rely on different underlying assumptions. The chosen approach depends on the purpose and specific focus of the task. The following outlines three of the approaches trialled within current CERP projects.

The Many-Facets Rasch Measurement Model (MFRM)

The Rasch model (Rasch, 1960) was developed by Georg Rasch in 1960 to analyse intelligence tests. It is a probabilistic model that can be applied to tests with dichotomous items (i.e. questions that are scored 0 or 1). The model

has two parameters, item difficulty and the person ability. It is a very popular model within the field of educational measurement: it is used, among other things, to equate tests to allow meaningful comparisons between students who sit different exams. The model has been extended to deal with polytomous items (i.e. questions that have a maximum score greater than 1).

The Many-Facets Rasch Measurement Model (Linacre, 1989) is an extension of the Rasch model. The model has an extra parameter, which is the severity of the marker. This model allows us to investigate and evaluate marker behaviour.

Missing data

Our data is often incomplete. This is because not all students will answer every item – and not all markers will mark every student. This is not a problem for Rasch models; estimates are computed from the data that is observed, and no extra

The Many-Facets Rasch Measurement Model is an extension of the Rasch model. The model has an extra parameter, which is the severity of the marker

assumptions or imputations are needed. The only requirement is that there are no disconnected subsets – this means that, to use the model, we can't have one set of students marked by one set of markers and another by a different set of markers.

Types of marker effects

There are different types of marker behaviour that we may want to identify. The first is leniency; a marker is lenient if they tend to assign marks higher than the rest of the markers. Conversely, a marker is severe if they assign marks lower than the average. If a marker is lenient (or severe) then the model can factor this in when calculating a student's ability. For example, we may rank one student higher than another, despite the fact that they have achieved the same score. Previously, AQA would apply adjustments to markers' marks to counteract marker leniency or severity.

A second issue relates to randomness in marking, this could be characterised as

erratic marking, i.e. awarding marks in a manner inconsistent with the other markers. It is not possible to apply adjustments to markers in this case. The model allows inconsistent markers to be identified; if this can be done during live marking these markers can be given extra guidance to improve the consistency of their marking.

A third issue is central tendency; this is where markers tend to avoid awarding marks at the extremes, which will restrict the range of marks awarded. For example, if a question has a maximum mark of 5, the marker may only tend to award marks of 2, 3 or 4. This will compress the spread of marks. The MFRM allows the user to see if markers are failing to award low or high marks.

Another potential source of bias that can be diagnosed using the MFRM is the halo effect. Tests are composed of multiple different items which may be testing different skills (traits). A halo effect

is when a marker assigns similar marks to a student on different items. A number of AQA's tests are marked at item level, this means that the tests are split into individual items and different markers will mark different items. This protects against a halo effect.

An extension of leniency is to look for evidence of differential leniency. This is where a marker shows leniency to a particular sub-group, e.g. a tendency to award higher marks for male students. MFRM provides a framework to investigate this. Most high-stakes exams are marked blind i.e. the student's details are anonymised, which should protect against this potential source of bias.

Summary

In summary the model is really powerful as it allows us to analyse students, items and markers in detail. It can be used to diagnose various different types of marker effects. The model performs well even with lots of missing data, which is invaluable when working with operational data.

Generalisability theory (G-theory)

G-theory can be viewed as an extension of classical test theory (CTT). CTT assumes that for each student taking an exam, there is a 'true' mark – the one that accurately reflects their ability in the subject, unaffected by any factor irrelevant to their ability, such as the occasion, the paper or marker. CTT

then proposes that the mark the student receives is their true mark plus some random measurement error. This random error is undifferentiated, in that the various possible sources of error are not individually evaluated. For example, across all the students, we do not know what proportion of the overall error is due to markers or any other source. Cronbach and others developed G-theory to enable assessment researchers to disentangle the error term into different components and then use this information to design more reliable tests. This is a two-stage process, including:

The G-study: where data are collected through a carefully designed study in order to estimate the error due to as many sources as possible, including that due to markers.

The Decision-study (D-study): using these error estimates, the theoretical impact of using, say, more items or more markers can be assessed. This can be used to design future, higher-quality tests.

Using G-theory

At AQA, we have applied G-theory post-hoc to our marker monitoring data. We have a sample of student responses that have been marked multiple times, gathered across many markers, items or papers. We have analysed these data using techniques that handle their sparse and unbalanced nature. As part of the G- and D-study analysis we have some useful

metrics: the estimate of the marker error in an exam – this can be compared with the maximum mark for the paper and the number of marks between grade boundaries – and a measure of inter-marker reliability. Ideally, marker error is small and inter-marker reliability is high. The marker error is made up of two components: the main effect, which summarises how consistently lenient or severe markers are, and an interaction term, which can be viewed as describing how much markers have disagreed in their marking of certain responses. If the interaction is relatively high, it might indicate a mark scheme that was difficult to interpret or apply. With the marker error estimate, we can also assess how confident we are that students were awarded the appropriate grades for the assessment.

Summary

Each metric can be used to summarise how marking has performed on average, and to identify problem items or papers, which is then used to inform future item writing as part of a process of continuous improvement. However, they cannot be used to identify individual inconsistent markers, and the analysis does require a good sample of student responses that have been marked multiple times.

Confirmatory factor analysis

While MFRM provides micro-level analysis

of marker performance and the G-theory approach provides macro-level analysis of marker performance, the confirmatory factor approach (CFA) provides both micro and macro-level perspectives on marker performance. It allows the researcher to see both the detail and the bigger picture. The method provides marker performance information as well as aggregate-level information that is not specific to individual markers and inter-marker reliability for multiple marking data. It relates the true item score to scores awarded by markers, taking into account the measurement error that reflects non-systematic error on the part of the marker in assigning a score.

The purpose of CFA (Bollen, 2002) is to understand the underlying structure that produced relationships among scores assigned by multiple markers.

The confirmatory factor analysis has several advantages over the alternative approaches. For example, a confirmatory factor-analytic approach does not impose any stringent assumptions regarding the relationship among the markers' scorings. It attempts to determine the marker model that best describes the sample data.

The CFA approach investigates unobserved or unmeasured factors that are thought to cause scores assigned by markers to covary. The idea behind a CFA approach is that there are a small number of latent factors that influence

QUALITY OF MARKING

TABLE 1

Marker	Item 1e (tarriff = 6)	Item 2a (tarrif = 10)	Type of examiner (N = normal; S = senior)
A1	0.849	0.536	S
A2	0.774	0.630	S
A3	0.905	0.336	S
A4	0.655	0.661	N
A5	0.528	0.522	N
A6	0.582	0.558	S
A7	0.668	0.400	N
A8	0.359	0.531	N
A9	0.407	0.528	N
A10	0.787	0.499	N
A11	0.673	0.366	N
A12	0.704	0.704	N
A13	0.707	0.427	N
A14	0.814	0.941	S
A15	0.506	0.694	N
A16	0.370	0.409	N
A17	0.508	0.417	N

each of scores assigned by markers. The confirmatory factor analysis has been successfully used with reliability theory by Joreskog in his classic paper on congeneric measurement (Joreskog, 1971).

Table 1 (left) illustrates the potential information to assess reliability among the markers for each response item, as well as determining the invariance of the reliability estimate across the various items, using a CFA approach.

Summary

There are several advantages for using a CFA model to evaluate maker reliability. First, a CFA approach does not impose an implicit stringent assumption regarding the correlations among the markers' scores. CFA attempts to identify the marker model that best describes the sample marking data. Second, various statistics useful in examining the fit of a given model may be computed to help

determine a most suitable model when there are competing models. Third, it allows for determining the several potential threats to the reliability of a group of markers, and how different markers may apply the mark scheme differentially across candidates. Finally, the model can be used reasonably when there are four or more markers and there is a fairly large number of students. ■

REFERENCES

- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests.
- Linacre, M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.
- Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag
- Summary

This topic was presented as a symposium at Association for Educational Assessment – Europe's 18th annual conference (2017) and the slides can be found at <https://tinyurl.com/y455eq74>.

We are always interested in finding new ways of examining data to gain insights. The work we have been done is being operationalised to ensure that there is a feedback loop to improve the quality of our assessments.

From the archives

In each issue of *Inside assessment*, AQA's research team trawls the archives to present significant historical papers. With our current focus on marking, in this edition we look at levels-based mark schemes and examine their effect on marking reliability. Below is an abridged version of **Anne Pinot de Moira's** 2013 paper, which brings together evidence based on operationally collected mark remark data and contemporary literature to present a guide for best practice in mark scheme design

Introduction by Zeek Sweiry

Extended constructed response items (and particularly those which are scored holistically through levels-based mark schemes) are central to the make-up of UK-based high-stakes assessments. Despite the many justifications for their use, they also constitute a significant challenge to marking reliability. Until relatively recently, most research on marking reliability in constructed response

items focused on marker characteristics (e.g. subject expertise, level of education and marking experience). In recent years, however, attention has increasingly turned to characteristics of the task, and particularly to features of mark schemes and their relationship with marking agreement levels.

Research to investigate the possible relationship between features of mark schemes and marking reliability has broadly taken two forms. Empirical

studies have involved the manipulation of one or more properties or features of mark schemes and a comparison of the marking agreement levels achieved when marking the original and amended versions.

The second form involves the coding of selected mark scheme features and an analysis of the relationship between these features and marking agreement levels across larger data sets in multiple subjects.

The following paper by Anne Pinot de Moira constitutes one of the seminal research papers of this second form. The paper focuses specifically on levels-based mark schemes, which have historically received inadequate attention, particularly given that previous research has tended to associate them with lower levels of marking reliability than other mark scheme types such as points-based mark schemes.

Perhaps the most influential aspect of the paper is the list of variable levels-based mark scheme design features, based on an analysis of 300 level-based mark schemes. The importance of this

classification is two-fold. First, it has provided a foundation for future research in the area by identifying features of levels-based mark schemes that may warrant further attention. Second, it has increased the focus of item writers and assessment designers on these features, raising debate about whether or not variability in these features between subjects and specifications reflects genuine differences in practice, or whether in fact there is no subject or specification-based rationale for this variability.

Although the classified mark scheme features were found to explain only limited variation in the data, the recommendations made in relation to them have been highly influential in levels-based mark scheme design at GCSE and A level. The findings also highlighted the importance of response factors (and in particular, response quality) in contributing to the difficulty of marking, and helped draw attention to cognitive demand as a key variable that is moderated by the different mark scheme features addressed in the study.

Features of a levels-based mark scheme and their effect on marking reliability
By Anne Pinot de Moira

LEVELS-BASED MARK schemes are used predominantly for questions with a high mark tariff where there is an extended written response. As such questions have scope for multiple valid approaches, a point-based marking system or exemplar answers are impractical. An examiner is expected to make an initial assessment of a response and, once the response is classified into a single defined level, refine this judgement to award a

mark (see Figure 1 for an example of a levels-based mark scheme). While there may be a common understanding of the philosophy behind levels-based mark schemes, there is little commonality in their design and formulation. There are considerable differences in the look and feel, and these differences present varying cognitive challenges.

Key features

From the many levels-based mark schemes used to mark items on summer 2011 A-level question papers, a sample of over 300 were scrutinised to establish a list of the variable design features. These mark schemes were selected on the basis that the features could be quantified for future modelling and, while not an exhaustive list, they represented a high

FIG. 1 Excerpt from GCE Citizenship Studies Unit 1 generic mark scheme for items 1 and 5 (summer 2011)

Level	Assessment objective: knowledge and understanding
Level 3	(4-5 marks) Answers demonstrate a range of citizenship knowledge and an accurate understanding of relevant citizenship concepts and theories. A range of examples is used to relate knowledge and understanding to citizenship issues.
Level 2	(2-3 marks) Answers are characterised by a good level of citizenship knowledge and an understanding of relevant citizenship concepts and theories. Examples are used to relate knowledge and understanding to citizenship issues.

proportion of all AQA large-entry A-levels with long-form answer questions at the time of writing. Areas of difference are listed below, along with a brief description of the implication of these differences as understood from the current research literature.

1. The number of levels in the mark scheme.

Clearly, the number of levels is inextricably linked to the maximum marks for an item and therefore to the intended weight of that item (and area of specification content) within the assessment. However, there is an extensive literature, succinctly summarised in Peterson (2000, p. 63), which discusses the optimal number of categories or levels for a rating scale.

2. The number of marks within a level.

As with the number of levels, decisions regarding the number of marks within a level are linked to the limits of cognitive discrimination and to the desired content weight within the specification.

3. The distribution of marks between levels.

Theoretical evidence suggests that the number of marks should be equal across all levels described in the mark scheme for an item (Pinot de Moira, 2012).

4. The evaluation, or otherwise, of quality of written communication in the mark levels.

Quality of written communication (QWC)

is most often assessed and evaluated in open-ended response items. In the past, a judgement has been made across an entire script but, with the introduction of item level marking, QWC marks are often assigned on the basis of one item alone. In many subjects, QWC has low correlation with the subject specific construct being measured (see for example, Massey & Dexter, 2002). Effective design of a levels-based mark scheme for items where QWC is integrated into the assessment would, therefore, appear to require its separate evaluation outside the levels.

5. The presentation of levels in a grid-like format to separate the evaluation of assessment objectives.

For some items, the mark scheme makes a distinction between performances on the different assessment objectives tested within. It is assumed that the correlation between these performances may be low and it would, therefore, be invalid to use a single levels-based model. Where this is the case, the mark scheme is often presented as a grid with levels forming the rows and assessment objectives forming the columns. The drawback of such a design is that, as the number of cells in the grid increases, so the mark scheme tends towards a points-based system where the award of every mark is specified in detail. It would contradict evidence which suggests that levels-based mark schemes are better, in terms of marking reliability, for items with

a maximum tariff of 10 and above (Bramley, 2008).

6. The inclusion, or otherwise, of a mark of zero in the bottom level.

In some mark schemes the mark of zero (nothing creditworthy) is included in the lowest level and in others it is identified separately outside the levels.

7. The inclusion, or otherwise, of indicative content within the levels.

In some levels-based mark schemes the level descriptions are generic, while in some they contain indicative content. While there is no research evidence to suggest which design is preferable, the cognitive load would undoubtedly differ dependent upon the wordiness of the mark scheme.

8. The order of presentation of levels: lowest first or highest first.

Perhaps surprisingly, there is variation in mark schemes as to whether the highest or lowest level is described first. This may introduce a tendency towards positive or negative reward, which differentially influences examiners, especially as they are entreated to be open-minded and positive when marking scripts; crediting what a candidate knows.

9. The documentation, or lack thereof, to describe the application of the levels-based mark scheme.

Very few mark schemes include any instructions to examiners on how to use

levels-based mark schemes. While there are undoubtedly pragmatic reasons for variations in mark scheme design, in the interests of improving marking reliability, a clearer understanding of the impact of the varying features would be desirable. This understanding would also facilitate the assembly of a coherent evidence-based guidance for use in assessment development to eliminate arbitrary, and potentially damaging, variations. Matching mark remark data to a sample of levels-marked items gives the opportunity to consider whether marking reliability is influenced to any measureable extent by the mark scheme.

Modelling Marking Reliability

To improve understanding of mark remark reliability, data from 16 units and 133 items were explored using a series of multilevel models. The data were taken from long-form answers that were double-marked in summer 2011. They represented an opportunity sample of responses which were remarked for quality control purposes during the marking period. All the items included in the analysis were on-screen marked and this accounts, in part, for attrition from the 300 items originally scrutinised to identify features of levels-based mark schemes.

The rules for determining the final mark for double marked responses state that, where there is agreement between two examiners within a predetermined tolerance, the original examiner's mark is

awarded. Where there is disagreement, defined as a difference greater than the predetermined tolerance, the response is sent for adjudication. The adjudicator, who is normally a senior examiner, will judge which mark of the two is correct and this mark will be chosen as the final mark. If the adjudicator is not happy with either of the marks, the final mark will be of his or her choosing.

Findings

Explained variation

The key finding, emerging from all three models, was that features of the mark scheme explained very little variation in the data. Even after all main effects were fitted, there remained some unexplained variation at response level and this variation dwarfed that at item, examiner and unit level. The individual responses given by students, therefore, appeared to be the limiting factor for reliable marking.

This suggests that, in terms of assessment design, a focus on the interaction between mark scheme and item, rather than the mark scheme alone, might prove more profitable in the quest for improved marking reliability.

Independent Variables

There were some independent variables which related to the dependent variable in the same way no matter which model was considered. These were not always significant in a statistical sense and therefore require cautious interpretation.

The probabilistic chance of all outcomes coinciding must be considered but, in most cases where there was coincidence, there was some existing research evidence to support the finding.

Whilst some of the existing literature suggests that it is the difficult items which are problematic to mark (Sweiry, 2012), the models used in this study appear to show that it is a higher quality of response rather than a greater item difficulty which lowers reliability; an observation also made by Pinot de Moira (2003).

The models all showed that there was an extremely small but statistically significant improvement in marking over the marking period. Examiners were on average a quarter of a mark closer to the final mark by the end of their marking. Evidence, perhaps, that rather than suffering from fatigue as time progressed, the examiner's increasing experience and regular feedback on performance served to hone skills. Another administrative feature that appeared to affect marking reliability was time at which marking was completed. Marking out of school hours (before 8am or after 4pm) tended to be less reliable.

The remaining three independent variables which suggested a consistent interpretation across all models described features of the mark scheme. The first related to the band descriptions. Marking reliability was higher, although not

statistically significantly, when the band descriptions were generic rather than including indicative content specific to the particular item. It might seem counterintuitive that mark schemes with more supporting information result in less reliable marking. However, visual comparison of the generic and specific levels-based mark schemes reveals what may well be the root of the problem. Generic mark schemes are often simple, neat and uncluttered. They are the same in format throughout the unit and, therefore, require less cognitive demand of the user. Levels-based mark schemes which include indicative content in the bands tend to be lengthier and, by definition, differ across the unit.

It might seem counterintuitive that mark schemes with more supporting information result in less reliable marking.

The second consistent mark scheme feature described, albeit non-significantly, by the models was the effect of the distribution of marks across bands. Previous theoretical research showed that, in order to reduce bias in the distribution of marks across the mark range for an item, the number of marks within each of the levels of a levels-based mark scheme should be the same (Pinot de Moira, 2012).

The models showed some support for this finding insofar as they indicated that marking was more reliable when each

band was composed of the same number of marks. The extent of this increased reliability appeared to be of the order of half a mark.

Finally, each of the three models suggested that marking was more reliable if the lowest level was described first on the mark scheme. Even allowing for the limitations of the model, the effect size was small and, unlike the other findings, was unsupported by independent literature or simple reasoning. On many mark schemes, among the general marking guidelines, are instructions to be positive in marking, to award marks which reflect the expected level of performance for the qualification, to use the whole mark range and not to deduct

marks for irrelevant or incorrect answers.

While

doubtless these instructions are not regularly reread, they reflect the philosophy for marking. The interaction of this philosophy with the design of the mark scheme might give clues to the better reliability for mark schemes where the marks are described in ascending order.

On using a mark scheme with the maximum mark at the top of the page and reading downwards, an examiner will be starting from the point of perfection. Thus, the examiner is required to deduct

rather than to award marks; undermining the established philosophy. Of the 12 units in which the mark scheme detailed the highest level first, two explicitly described the need for positive marking and four required a top down approach to arriving at a final mark. At best, marking philosophy, whether explicitly described or etched in folklore, sometimes seems to be at odds with mark scheme design.

This effect could be seen as analogous to Bramley's (2008) finding which suggested that the addition of qualifications, restrictions and variants (QRVs) to a mark scheme reduced marker agreement. Bramley argued that including QRVs led to examiners switching to more complex cognitive strategies to mark; leading to more errors. Even if QRVs are not included explicitly in levels-based mark schemes, they may still be implicit in examiners' thinking if the top band is presented first.

Conclusions

Whether the models described herein could be improved is a moot point. The number of units marked at item level within AQA has reached a plateau and there are still many high stakes subjects with high tariff items which are traditionally marked. To expand the models for greater generalisability would require the introduction of a new swathe of units, with levels-marked items, to the item level marking system. Maybe, on the other hand, improvements could be made if the

features of the mark scheme were described subjectively rather than analytically. Bramley (2008), for example, included variables such as the complexity of marking strategy in his model of marking reliability. Furthermore, and at the risk of overanalysing the data, it would be possible to revisit the inclusion of interactions into the models. This might, in particular, shed light on the reasons that good responses prove more difficult to mark.

Recommendations

Rather than providing unequivocal evidence to support effective design of levels-based mark schemes, this study serves to highlight the differences in practice that currently exist between specifications. Plainly there is an argument for flexibility in mark scheme design so that the mark scheme suits the subject being assessed. However, there is also an argument for greater commonality to improve reliability and to increase the transferability of skills. There seems to be no rationale for differing marking philosophies and guidelines.

Furthermore, it seems logical that we should strive to present mark schemes in a way which minimise the cognitive demand to the examiner. Returning to the areas of difference identified earlier, the following recommendations are made with a view to improving marking reliability:

QUALITY OF MARKING

- The number of levels and marks within levels in the mark scheme.
- The number of levels in a mark scheme should be determined by the intended weight of the item and by the extent to which the levels can be uniquely described. As with the number of levels, decisions regarding the number of marks within a level should be determined by the limits of cognitive discrimination and to the desired content weight within the specification
- The distribution of marks between levels.

As far as possible, the number of marks within each level of a levels-based mark scheme should be equal.

The evaluation, or otherwise, of quality of written communication in the mark levels. Quality of written communication (QWC) should be evaluated separately from the subject-based content and its evaluation should be independent of the levels-based mark scheme.

The presentation of levels in a grid-like format to separate the evaluation of assessment objectives. The inclusion, or otherwise, of indicative content within the levels.

Mark schemes should be designed with cognitive demand in mind. Clear, concise and simple mark schemes are likely to elicit more reliable marking.

The documentation, or lack of documentation, to describe the application of the levels-based mark scheme.

Mark schemes, and in particular levels-based mark schemes, should include clear and concise instructions for use. They should promote a consistent philosophy to marking which, in turn, should allow greater transferability of skills between units and specifications.

Above and beyond design of the mark schemes, it seems evident that marking might be improved if time is invested in providing support to examiners who manage their examining workload alongside a teaching schedule. Furthermore, given that marking reliability appears to have the greatest variation at the individual response level, careful item design might help alleviate marking difficulties. A schema such as that proposed by Pollitt et al (2008) might be used to limit item ambiguity and reduce the multiplicity of responses without compromising the validity of the assessment. At the same time, this focus on the assessment as a whole could be used to consider the effective design of items and mark schemes to discriminate accurately between the higher quality responses. ■

REFERENCES

- Bramley, T. (2008). Mark scheme features associated with different levels of marker agreement. Presented at the British Educational Research Association (BERA) Annual Conference, Heriot-Watt University, Edinburgh, UK.
- Long, J. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage Publications.
- Massey, A., & Dexter, T. (2002). An evaluation of Spelling, Punctuation and Grammar assessments in GCSE. Presented at the British Educational Research Association (BERA) Annual Conference, Exeter University, Exeter, UK.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63, 81-97.
- Peterson, R. A. (2000). *Constructing effective questionnaires*. Thousand Oaks, CA: Sage Publications.
- Pinot de Moira, A. (2003). *Examiner Background and the Effect on Marking Reliability*. Manchester: AQA Centre for Education Research and Policy.
- Pinot de Moira, A. (2012). *Levels-based mark schemes and mark bias*. Manchester: AQA Centre for Education Research and Policy.
- Pinot de Moira, A., Massey, C., Baird, J., & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, 67, 79-87.
- Pollitt, A., & Ahmed, A. (2008). Outcome Space Control and Assessment. In 9th Annual Conference of the Association for Educational Assessment – Europe. Presented at the 9th Annual Conference of the Association for Educational Assessment – Europe, Hissar, Bulgaria.
- Pollitt, A., Ahmed, A., Baird, J., Tognolini, J., & Davidson, M. (2008). Improving the quality of GCSE assessment. Qualifications and Curriculum Authority. Retrieved from <http://www.lifeinbits.org/camexam/htdocs/papers/2008ImprovingQualityofGCSE.pdf>
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis : an introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications.
- Sweiry, E. (2012). Conceptualising and minimising marking demand in selected and constructed response test questions. Presented at the Association for Educational Assessment (AEA) Europe Annual Conference, Berlin, Germany.

Contextual notes

IN ENGLAND, THE MAIN ACADEMIC qualification is known as the General Certificate of Secondary Education (GCSE), which pupils take at the age of 16. AQA is one of the awarding bodies that offers these qualifications. Students take GCSEs in a number of subjects, some of which are compulsory – for example, English, Mathematics and Science. These are taken alongside optional subjects, including History, modern foreign languages, Music and Physical Education. The English exam system differs to many other countries in that GCSEs are awarded for each individual subject, rather than a diploma that summarises achievement across the curriculum.

After receiving their GCSEs, many pupils decide to continue their studies and select three or four subjects to study for another two years. If successful, they obtain further subject-specific qualifications, called (GCE) A-levels, which can be used to apply to higher

education institutions. AQA sets and marks around half of all GCSEs and A-levels taken in the UK every year.

Because both GCSEs and A-levels act as passports to the next stage for young people, they are often referred to within the research community as 'high-stakes' exams. The variety of subjects on offer means that these qualifications are assessed in several different ways. Written papers include questions in a variety of formats, from multiple choice through to extended response essays. This ensures validity of measurement, by testing a variety of skills, but means that marking is more complex.

AQA requires examiners – referred to throughout this publication by the research term 'marker' – to have recent teaching experience in the subject and level that they are applying to mark. Qualified teachers who have a number of years' examining experience with a recognised awarding body but no longer

work in a school or college are also encouraged to apply. The examining experience has to be within the last three years in the subject and at an appropriate level.

The markers' work is checked during the marking period – known as 'live marking' – by a senior examiner or a senior examiner panel. When exam papers – called 'scripts' – are marked using the traditional pen and paper method, samples of markers' work are re-marked by senior examiners. This is to make sure that marking is within an acceptable tolerance.

Most marking is now conducted on-screen and marking is monitored using 'seeds'. These are pre-selected answers that are marked by the senior examiner, or senior examining panel. They are then introduced randomly into a marker's allocation. The marker is not aware that a particular question or script is a seed and has no knowledge of the exact mark

that has been awarded. If the new mark matches that of the senior markers, the marker can continue to mark their allocation of questions; if they fail a set number of seeds, they will be stopped.

Research therefore needs to consider those who carry out the marking, although, as demonstrated in several of the pieces within this publication, marking accuracy is not dependent upon easily measurable features (p10). Examiner characteristics have a role to play (p20), as do training and standardisation (p14). It is also important to monitor marking (p30).

At AQA's Centre for Education Research and Practice (CERP), research teams analyse assessment data to inform the creation of question paper and mark schemes, as well as to deliver quality assurance. In practice, these findings have been used to introduce new technology and to allocate questions to markers more intelligently. ■

Meet the researchers

AQA's **Centre for Education Research and Practice (CERP)** comprises a range of experts, including statisticians, psychologists, educationalists and scientists. In this second instalment of our regular series of researcher profiles, we introduce you to the staff who specialise in marking reliability



William Pointer
Senior Researcher

William joined AQA in 2010, having graduated with a Masters in Mathematics

from the University of Bath. His work focuses on the quality of marking: how we measure and monitor marking reliability, and how we can improve it. He is also interested in research on standards and comparability. William initially worked as a Qualification Developer in the GCSE Science department, before joining CERP in August 2013 to pursue a career in research.

What sparked your interest in marking reliability?

During my time in the GCSE Science subject team, I saw the great efforts that

senior examiners go to in order to standardise markers to make marking as consistent and fair as possible. When I moved to CERP, I was surrounded by large amounts of data that came out of the marking process. I was interested in how we could better use this to support examiners and drive improvements in the quality of marking – and the quality of overall assessments.

Describe a notable research highlight.

I really enjoyed conducting an evaluation of mark schemes from the reformed AS specifications in 2016. While I focused mainly on the quantitative analysis, my colleague considered the qualitative aspects. Working collaboratively on a project really brought it home to me that while data can provide insights into assessment functioning, it doesn't provide *all* the answers, and we need to make time to carefully review our assessments.

Time spent in CERP: 5.5 years



Claire Whitehouse
Senior Researcher

Claire joined CERP in 2004, having spent two years in qualification

management within AQA. Her current research is focused on improving the validity and reliability of assessments. Claire is working with colleagues to investigate the interactions of students, examiners and teachers with assessment materials and the assessment process, and to explore how new technologies may influence the nature of assessment. Claire gained a BSc(Eng) from University College London and a PhD from the University of Cambridge, both in chemical engineering.

How did you come to specialise in marking reliability within your work as an assessment researcher?

Research colleagues at AQA had investigated the influence on marking reliability of a number of variables such as allocation size, demand of question and quality of response. There seemed to be a gap where examiner characteristics should sit.

So, I started to look at how to operationalise these characteristics using the available data – that’s how my exploration of examiner expertise in optional questions and team environments began.

Which project has been the most rewarding?

Probably the project I worked on with Qingping He (now of Ofqual) on semi-automated test construction from an item bank. Apart from the great team-working environment, the experimental work yielded interesting results on how a subject expert is able to construct a test using item functioning indices and other metadata.

What would you like to spend more time researching?

Understanding how examiners interact with mark schemes and candidates’ responses, and how students interact with question papers is fascinating. With equipment such as eye trackers, which are capable of logging real-time measurements, this is an area that could provide interesting descriptions of cognitive processes that have the potential to improve our assessment materials.

Time spent in CERP: 14 and a half years



Yaw Bimpeh
Senior
Researcher

Yaw joined CERP in 2014. He holds a PhD in Statistics, an MSc in Mathematical

Sciences and a BSc (Hons) in Mathematics. His current areas of research include marking reliability, application of the Bayesian method to standard setting and test equating, and construct validity of assessment designs. Yaw has experience of analysing and modelling data in a variety of fields, and is skilled in the research and application of statistical methods.

How have your degrees prepared you for your work in assessment research?

I believe my training in mathematics and statistics equipped me with analytical tools to deal with many aspects of the diverse educational assessment problems I have come across so far.

Tell us about a recent research achievement.

I am currently working on procedures for maintaining and adjusting standards that can potentially combine the judgements of subject-matter experts with the statistical predictions.

What would you like to investigate next, in terms of marking reliability?

I am interested in chance-corrected ways of measuring marking reliability, and also how best to choose a suitable model for evaluating human marking.

Time spent in CERP: Four and a half years.



Liz Harrison Researcher

Liz joined CERP in 2015. She previously worked as a data analysis manager at a

secondary school, where she supported staff in target setting, monitoring progress, and understanding performance measures. As well as working in schools, Liz has experience as a research assistant at Nottingham University, and as a statistician in the pharmaceutical industry. She has a BSc in Mathematics with Statistics and is currently studying for an MSc in Educational Assessment.

What do you enjoy most when it comes to researching marking reliability?

It is important that students can have confidence that their grades are fair. A lot of work is going into this and it's good to be part of the team.

Any career highlights so far?

I enjoy working with the awarding teams and assessment designers as they seek to provide high-quality and fair assessments to students. I find it rewarding to support colleagues in interpreting their data and providing specialist analyses.

What projects are you currently working on?

I am looking at how to maintain standards in small-entry subjects and the relationship between the number of grades and grading accuracy.

Time spent in CERP: Three and a half years.

Centre for
Education Research
and Practice (CERP)

