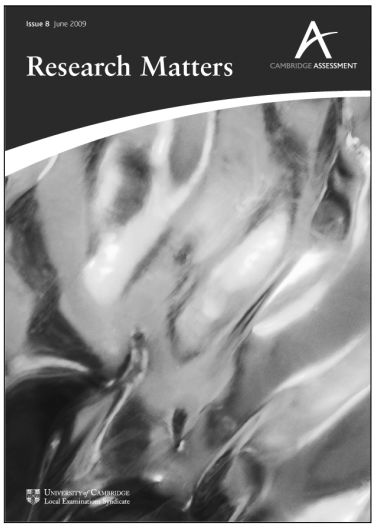
Issue 8 June 2009

Research Matters

CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate



If you would like to comment on any of the articles in this issue, please contact Sylvia Green.
Email:
researchprogrammes@cambridgeassessment.org.uk

The full issue and previous issues are available on our website:
www.cambridgeassessment.org.uk/ca/Our_Services/Research

Research Matters : 8

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

One key question keeps recurring: '…do you do fundamental work or is it all instrumental…?'. My own answer: '…both… at different times and sometimes entirely wrapped up together…'. An instance of fundamental work not tightly linked to operational work in the Group is the literature review on Birthdate Effect reported here under the title 'Happy Birthday to you – but not if it's summertime'. A seemingly-humble and increasingly unfashionable literature review, it cuts through the restrictions of cross-sectional studies – which look at individual phases of education and training – to illuminate the shocking persistence of the effect across the system as a whole. But much of the work in this volume exemplifies fundamental and operational work which is 'wrapped up together'. All too often, there exists an assumption that there is a contradiction between fundamental research and operational research. But in the most parochial of matters in assessment there lurk fundamental matters. And fundamental work can – and should – be used to drive improved 'evidence-based' practice. Perhaps we should take a lead from medical research – clinical practice of the most applied kind (ranging from surgical techniques to genetic counselling) – which is progressed by sound fundamental work. I characterise it as the 'janus-headed' nature of assessment research – looking both to enhance the canon of fundamental, generalisable knowledge and simultaneously improve the operation of complex public systems. Whilst tensions might arise in terms of issues such as '…onto the next project or do I disseminate well the outcomes of my existing work…?', the parallels with medical research (and with aeronautical engineering, meteorology, etc) suggest that those that preach an absolute distinction between fundamental and applied work will starve assessment systems of assessment of knowledge essential to their improvement. Look to the studies in this volume as examples of being 'janus-headed' in the best possible way.

Tim Oates *Group Director, Assessment Research and Development*

Editorial

Most of the articles in this issue relate to how judgements are made and the factors that impact on those judgements. In the first article Johnson and Nádas consider how on-screen essay marking is affected by mode. Reliability is also a focus in the article from Sykes *et al*. which reports on an investigation into the effectiveness of potential procedures for providing fast and automated feedback to examiners. Bramley's article concentrates on marker agreement at item level rather than at candidate level. He reports on a study that explored the features of question papers and mark schemes associated with higher and lower levels of marker agreement. Suto, Greatorex and Nádas consider the benefits of, and variations in, training procedures for examiners, drawing together research on examiner training and on the nature of the judgements entailed in the marking process. In the article from King, Novaković and Suto we learn how judgements are made using rank ordering, traditional awarding, and Thurstone pairs. The article focuses on the perspective of an examiner who took part in the research and his views are extremely valuable in informing the design of future research. Peter Bird, a member of OCR's operational research team, compares two methods of estimating missing marks and highlights issues and differences in the accuracy of the process. The article on the effect of birthdate on performance by Oates *et al*. outlines the findings of a research review which provides robust evidence from around the world that, on average, the youngest children in their year group at school perform at a lower level than their classmates. The review detailed in this article was released to the press in February 2009. It was widely reported in England and has also received attention in other countries, including China. At the same time it was submitted as evidence to the Rose Review of Primary Education which, as part of its interim report, had recommended that all children should start formal schooling at the age of four (rather than five, as is currently the case). Sylvia Green then reports on the latest in the Cambridge Assessment Parliamentary Research Seminar Series, hosted by Barry Sheerman MP, Chair of the Children, Schools and Families Select Committee.

Sylvia Green *Director of Research*

An investigation into marker reliability and some qualitative aspects of on-screen essay marking

Martin Johnson and Rita Nádas Research Division

A more detailed analysis of the reliability findings reported here will appear in 'Marking essays on screen: an investigation into the reliability of marketing extended subjective texts' to be published in the British Journal of Educational Technology *by the British Educational Communications and Technology Agency and Blackwell Publishing.*

Introduction

There is a growing body of research literature that considers how the mode of assessment, either computer- or paper-based, might affect candidates' performances (Paek, 2005). Despite this, there is a fairly narrow literature that shifts the focus of attention to those making assessment judgements and which considers issues of assessor consistency when dealing with extended textual answers in different modes.

This article argues that multidisciplinary links with research from domains such as ergonomics, the psychology of reading, human factors and human-computer interaction could be fruitful for assessment research. Some of the literature suggests that the mode in which longer texts are read might be expected to influence the way that readers access and comprehend such texts (Dillon, 1994; Hansen and Haas, 1988; Kurniawan and Zaphiris, 2001; Mills and Weldon, 1987; O'Hara and Sellen, 1997; Piolat, Roussey and Thunin, 1997; Wästlund, Reinikka, Norlander and Archer, 2005). This might be important since these factors would also be expected to influence assessors' text comprehension whilst judging extended textual responses.

Literature review

Gathering reliability measures is a significant practical step towards demonstrating the validity of computer-based testing during the transitional phase where assessments exist in both paper- and computer-based modes. In her review of comparability studies Paek (2005) notes that the transition from paper- to computer-based testing cannot be taken for granted and that comparability between the two testing modes needs to be established through carefully designed empirical work.

Paek suggests that one of the primary issues for such research is whether the computer introduces something unintended into the test-taking situation. In the context of assessing essays on screen this might demand enquiry into construct validity; exploring whether the same qualitative features of essay performance are being attended to by assessors in different modes.

Whilst Paek reports evidence that screen and paper versions of traditional multiple-choice tests are generally comparable across grades and academic subjects, she notes in her conclusion that 'tests with extended reading passages remain more difficult on computer than on paper' (p.18), and suggests that such differences might relate to computers inhibiting students' reading comprehension strategies.

Johnson and Greatorex (2008) extend this focus on comprehension to call for studies which explore the cognitive aspects of how judgements might be influenced when assessors read longer texts on screen. This concern appears to be important given a recent study which reports correlations between re-marked essays significantly lower when scripts are re-marked on screen compared with paper re-marking (Fowles, 2008).

There are a variety of cognitive aspects of reading whilst assessing. Just and Carpenter (1987) argue that working memory is directly linked to reading a text and that this involves an expectancy effect that relies on working memory to retain the words just read in order to allow the next words to be linked together in a meaningful way. They go on to suggest that increasing the complexity of the task or the number of component elements of the reading activity can also affect reading performance. Mayes, Sims and Koonce (2001) reiterate this point, reporting a study which found that increased reader workload related significantly to their reduced comprehension scores.

Another cognitive aspect of reading relates to the role of spatial encoding. Johnson-Laird (1983) suggests that the linear nature of the reading process leads to the gradual construction of a mental representation of a text in the head of the reader. This mental representation also accommodates the location of textual information with readers spatially encoding text during the reading process (Piolat, Roussey and Thunin, 1997). Spatial encoding hypothesis claims that positional information is processed during reading activity; the hypothesis is based on evidence that readers can regress to find a location within a visual text very efficiently.

Research suggests that the cognitive effort of reading can be augmented by other activities such as annotating and note taking, with these 'active reading' practices often operating concurrently with reading activity (O'Hara and Sellen, 1997; Piolat, Olive and Kellogg, 2005). Literature suggests that active reading can enhance reading comprehension by supporting working memory (Crisp and Johnson, 2007; Hsieh, Wood and Sellen, 2006; Marshall, 1997) and facilitate critical thinking (Schilit, Golovchinsky and Price, 1998). Schilit *et al*. (1998) observe that active reading is challenged by the screen environment due to difficulties in free-form ink annotation, landscape page orientation (leading to the loss of a full page view), and reduced tangibility.

Recent shifts in Human Factors research have been increasingly concerned with the cognitive demands related to reading across modes. Much of this work has focussed on the inherent features of computer displays and navigation issues. Since it has been found that protracted essay reading (and by inference essay assessment) can involve navigating a text in both linear and non-linear ways (O'Hara, 1996; Hornbæk and Frøkjær, 2001), on-screen navigation might exert an additional cognitive load on the reader. This is important given the suggestion that increased reading task complexity can adversely affect reading comprehension processes.

The literature has led to a model of the interactions that might influence mental workload whilst reading to comprehend. In the model, physical process factors such as navigation and active reading strategies are thought to support assessors' cognitive processing (e.g. spatial encoding) which could in turn affect their comprehension whilst they judge extended texts. Theory suggests that readers employ these physical processes differently according to mode and that this can affect reader comprehension. Studying physical reading processes might therefore help to explain any divergent assessment outcomes across modes. The model suggests that research might usefully include a number of quantitative and qualitative factors. Assessors' marking reliability across modes, their attention to different constructs, and their cognitive workloads could be quantitative areas of focus. These findings could be supplemented with qualitative data about factors such as navigation and annotation behaviours in order to explore influences on assessors' spatial encoding processes whilst comprehension building.

Research questions and methodology

The plan for this project considered 6 questions:

1. Does mode affect marker reliability?
2. Construct validity – do examiners consider different features of the essays when marking in different modes?
3. Is mental workload greater for marking on screen?
4. Is spatial encoding influenced by mode?
5. Is navigation influenced by mode?
6. Is 'active reading' influenced by mode?

One hundred and eighty GCSE English Literature examination essays were selected and divided into two matched samples. Each stratified sample contained 90 scripts spread as evenly as possible across the seven bands of the 30-point mark scheme.

The scripts were then blind marked for a second time by the subject Principal Examiner (PE) and Assistant Principal Examiner (APE) to establish a reference mark for each script. In this project the reference mark is therefore defined as the consensual paper mark awarded by the PE and the APE for each answer.

Twelve examiners were recruited for the study from those who marked the unit 'live' in January 2008. Examiner selection was based on the high quality of their past marking. In order to control the order of sample marking and marking mode, the examiners were allocated to one of four marking groups. Examiner groups 1 and 4 marked Sample 1 on paper and Sample 2 on screen; groups 2 and 3 marked Sample 1 on screen and Sample 2 on paper. Groups 1 and 3 marked Sample 1 first, and groups 1 and 2 marked on paper first. This design allowed subsequent analyses to separate out any purely mode related marking effects (i.e. direct comparisons of the marking outcomes of groups 1 and 4 with groups 2 and 3) from any marking order effects.

In order to replicate the normal marking experience as much as possible the examiners completed their marking at home. Before starting their on-screen marking all examiners attended a group training session to acquaint them with the marking software along with administrative instructions.

Marker reliability was investigated first by looking at the mean marks for each examiner in each mode. Overall comparisons of the mark distribution by mode and against the reference marks were also made. Statistical models were then used to investigate the interaction between each examiner and mode.

To investigate construct validity, the textual features that were perceived to characterise the qualities of each essay response were elicited through the use of a Kelly's Repertory Grid (KRG) exercise (Kelly, 1955; Jankowicz, 2004). This process involved the Principal Examiner (PE) and the Assistant Principal Examiner (APE) separately comparing essays that were judged to be worth different marks, resulting in 21 elicited constructs. The PE and APE then separately rated 106 scripts according to each individual construct on a 5-point scale. These construct ratings were added into the statistical models to investigate whether each construct influenced marking reliability in either or both modes.

To investigate mental workload in both marking modes, a subjective measure of cognitive workload was gathered for each examiner. The National Aeronautics and Space Administration Task Load Index (NASA-TLX) (Hart and Staveland, 1988) is one of the most commonly used multidimensional scales (Stanton *et al*., 2005). It is considered to be a robust measure of subjective workload (Moroney *et al*., 1995); demonstrating comparatively high factor validity; usability; workload representation (Hill *et al*., 1992); and test-retest reliability (Battiste and Bortolussi, 1988). This has led it to be used in a variety of studies comparing mode-related cognitive workload (e.g. Emerson and MacKay, 2006; Mayes *et al*., 2001).

For this study the NASA-TLX measure of mental workload was completed twice by each examiner, midway through their marking sessions in both modes. This enabled a statistical comparison of each marker across modes to explore whether screen marking was more demanding than paper marking.

The influence of mode on examiners' spatial encoding was investigated through their completion of a content memory task. After marking a randomly selected script in both modes, five of the examiners were asked to recall the page and the location within the page where they had made their first two annotations. A measure of spatial recall accuracy was constructed and used as a basis for comparison across modes.

To investigate how navigation was influenced by mode, information about reading navigation flow was gathered through observations of six examiners marking in both modes. This involved recording the directional flow of examiners' navigating behaviour as they worked through eight scripts.

Examiners' annotation behaviour was collected to explore how this aspect of 'active reading' was influenced by mode. Examiners' annotation behaviours were analysed through coding the annotations used on 30 paper and screen scripts from each of the examiners. This analysis of 720 scripts represented one-third of all the scripts marked.

Finally, concurrent information was gathered by the examiners in the form of an informal diary where they could note any issues that arose during marking. Alongside the marking observation data, this diary evidence provided a framework for a set of semi-structured interviews that were conducted with each examiner after the marking period had finished. This allowed the researchers to probe and check their understanding of the data.

Findings

Does mode affect marker reliability?

Initial analyses showed that neither mode order nor sample order had significant effects on examiners' reliability. Analyses of examiners' mean marks and standard deviations in both modes suggested no evidence of

any substantive mode-related differences (paper mean mark: 21.62 [*s.d. 3.89*]; screen mean mark: 21.73 [*s.d. 3.91*]). Five examiners tended to award higher marks on paper and seven awarded higher marks on screen. However, such analyses might mask the true level of examiner marking variation because they do not take into account the mark-disagreements between examiners at the individual script level.

To allow for this, further analysis considered the differences between examiners' marks and the reference marks awarded for the scripts. For the purposes of this analysis the chosen dependent variable was the script-level difference between the examiners' mark and the reference mark, known as the Mean Actual Difference, with negative values indicating that an examiner was severe and a positive value indicating that an examiner was lenient in relation to the reference mark.

Box plots for the distribution of the mark difference for scripts marked in both modes suggest little mode-related difference (Figure 1). These indicate that about half of the examiners showed a two–mark difference from the reference marks in both modes, with paper marking tending to be slightly more 'accurate'.

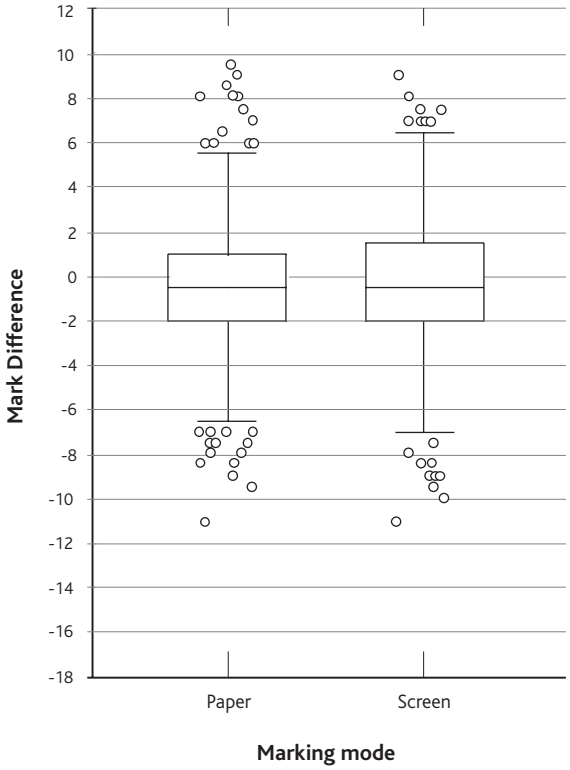


Figure 1: Box plots of the distribution of mark difference from the reference mark by marking mode[1]

To investigate the interaction between individual examiners and mode, least square means from an ANCOVA are plotted in Figure 2.

Figure 2 shows that the confidence intervals overlap for all examiners except for Examiner 4, suggesting no significant mode-related marking difference for 11 examiners. Where an examiner was severe or lenient in one mode they were also similarly severe or lenient in the other mode. Examiner 4 differed from the other examiners because his screen marking differed significantly from his paper marking with the screen marking being closer to the reference marks.

1 For ease of interpretation, the box includes 50% of the data, and each whisker represents 25% of the data. The horizontal line within the box is the median, below and above which lie 50% of the data.

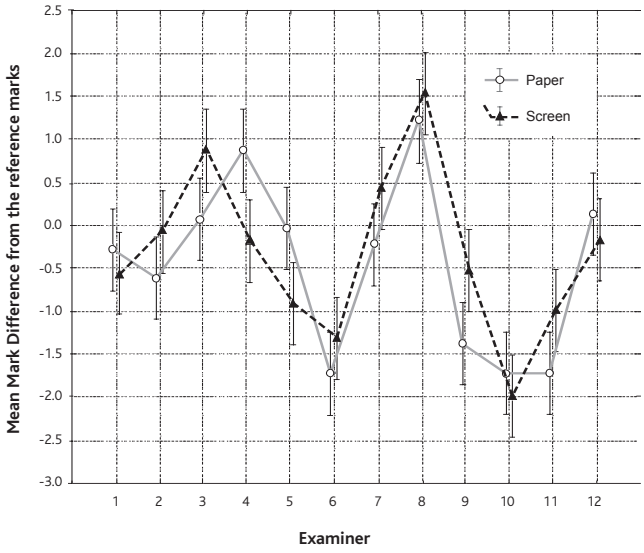


Figure 2: Least Square means for mark difference by examiner and mode

Construct validity – do examiners consider different features of the essays when marking in different modes?

21 sets of construct ratings were added in turn into the statistical reliability models in order to investigate whether each construct influenced marking in either or both modes. Data revealed that mode did not have a significant effect on the constructs examiners paid attention to while marking. However, some constructs did explain the difference between some individual examiners' marks and the reference marks; for example, 'points developed precisely and consistently'; 'insight into characters' motivation and interaction' or 'attention to both strands of the question'. Further research is currently underway on the relationship between examiners' use of constructs and essay marking performance.

Is mental workload greater for marking on screen?

Data suggest that overall cognitive load was greater for screen than paper marking ($t(11) = -2.95, p < 0.05$). Figure 3 shows the variations in the extent to which the subscales differed according to mode. The ***frustration*** subscale showed a large and statistically significant mode-related influence ($t(11) = -3.69, p < 0.01$), suggesting a greater factor in on-screen marking. A slight tendency was also found on the ***performance*** subscale ($t(11)=2.19, p=0.051$), suggesting that examiners were comparatively more satisfied with their marking on paper than on screen.

On all other dimensions marking mode did not have a significant effect on the cognitive load of the task, suggesting that the ***frustration*** experienced during screen marking contributed to the examiners' elevated overall cognitive load ratings for this marking mode.

Although overall cognitive workload ratings were higher for screen marking, significant variations were found between some of the total cognitive load ratings reported by examiners ($t(11) = 28.37, p < 0.001$). Furthermore, although all examiners reported concern for ***performance*** dimensions in both modes, there was variation among examiners concerning the rest of the dimensions.

In order to tease out which aspects of marking contributed to the above findings, and to explain the wide variation found in participants' profiles, follow-up semi-structured interviews were conducted with all participants.

On-screen marking was associated with significantly more ***frustration*** than traditional paper-based marking. Most examiners mentioned the

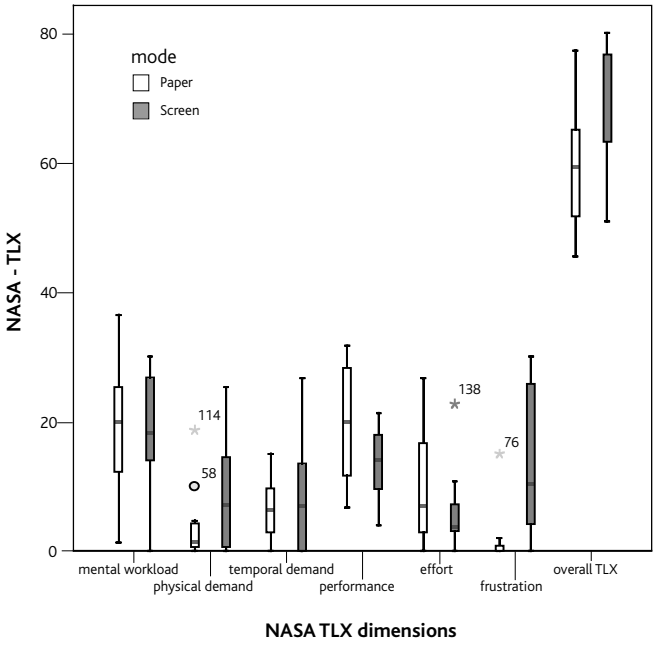


Figure 3: Mode-related differences in examiners' ratings on NASA-TLX dimensions and overall cognitive load

novelty of on-screen marking or specific elements of the software environment as causes for their initial frustration. However, once technical problems were resolved, examiners generally grew more comfortable with on-screen marking and frustration levels decreased.

Examiners were slightly less satisfied with their on-screen marking ***performance***. Some of the reasons for this related to the novelty of technology; the lack of a standardisation session; examiners' own personality traits, and the inherent responsibility of the marking process. Generally, it seemed that examiners perceived two types of performance: the satisfaction of completion and the professional accomplishment of performing high quality work.

Most of the sources of ***mental workload*** reported (e.g. cognitive processes, responsibility, unfamiliarity with the process, etc) are inherent characteristics of any marking process, and perhaps explain why mode did not have a significant effect on this subscale. Although causing a heightened initial mental workload, unfamiliarity with the situation eased as markers got used to the technology.

Mode had no significant effect on the ***physical demand*** of marking. A variety of activities, as well as the unfamiliarity and constraints of the physical environment, contributed to physical strain, which originated from inadequate working conditions characteristic of both marking modes.

Temporal demand was not significantly affected by marking mode, and was generally reported to be very low in the project overall. However, a live marking session with tight deadlines might result in heightened amounts of long-term temporal demand on examiners.

Data showed only a slight statistical tendency for on-screen marking to require more ***effort***. Participants listed a variety of elements which contributed to fatigue, for example, novelty and initial struggles with technology; sticking to and applying standards; physical strain and looming deadlines; mental fatigue; and administrative tasks/recording marks on paper. Others felt energised by some particular aspects of on-screen marking, for example, the ability to read poor handwriting; the lack of administrative requirements; and 'seeing the scripts off by a click'.

Is spatial encoding influenced by mode?

Whilst marking a randomly selected script in both modes, five of the examiners were asked to recall the page and the location within the page where they had made their first two annotations.

Although the number of examiners involved in this activity was limited it appears that the ability to recall not only the page but the location of a detail within that page was more precise on paper than on screen. On paper all five examiners could recall the page on which they made at least one of their first two annotations. Three of these annotations were located in the correct geographical ninth of the page and two were within the correct geographical third of the page. On screen only two of the examiners were able to locate the page of any of their annotations, and these were only positioned in the correct third of a page. The three remaining examiners could not remember the page where they made either of their first two annotations.

This suggests that the examiners' spatial encoding was better on paper and that this led to a better mental representation of the text read; as one examiner put it:

> *I do tend to have that sort of memory where I… know that it's at the top, middle or bottom of the page that I saw something. That sort of short term stays there, but with the zooming and scrolling it isn't quite as easy because on the paper you just turned, there it is and you've found it.* (Examiner: 10: Interview)

Theory suggests that readers spatially encode the location of features in a text when they construct a mental representation of it. It appears that the use of iterative navigational strategies can facilitate this process by affording readers the opportunity to efficiently locate and remember qualities within a text. At least two factors might influence this navigating activity: (i) reader annotation activity, and (ii) the characteristics of visual field and resolution levels in the reading environment.

Observations suggest that visual reading fields tend to be larger on paper and offer higher resolution levels, which in turn might influence navigation behaviour. Indeed, a number of examiners indicated that their marking practice involved them getting an overview of the script, reinforcing their mental image of it.

Is navigation influenced by mode?

Paired samples t-tests showed that examiners' paper navigation tended to be more iterative, using both linear and non-linear reading approaches, whilst on-screen navigation tended to be overwhelmingly linear ($t(5) = 2.84$, $p = 0.04$).

Iterative reading behaviours appeared to involve examiners establishing an overview of the script and it seems that the ability to gain an overview of the script positively influenced examiner confidence. Three examiners suggested that having an overview of the script made them feel more confident in the consistency of their marking. The reason for this perception seems to relate to the way that looking back over a script allowed examiners to confirm or question their previous reflections.

Three of the examiners suggested that navigational ease in the paper mode helped to support their working memory whilst building a sense of textual meaning. Another key mode-related factor appeared to be that the paper environment afforded fluid annotation across multiple pages. It appeared that not being able to navigate as freely around a script on screen led to some examiner frustration and their adoption of consciously different reading styles.

Interviews, observations and examiner diary evidence suggested that navigation away from the script was also related to mode. Examiners commonly described a reduced tendency to move their attention between different scripts on screen. Comparing the qualities of different scripts appears to be a key feature of the examiners' usual practice, with cross-referencing between scripts helping them to compare the qualities of different performances, establish or confirm a standard, and reinforce their confidence in the consistency of their own judgements.

It was very common for examiners to suggest that comparing the qualities of different scripts was less effective on screen. It is possible that such mode-related difference relates to how the tangibility of a text might support examiners' mental workload. One of the key links between tangibility and thinking might be the way that the paper environment can afford speedy comparisons to be made. One examiner noted that the process of identifying and accessing other potentially relevant scripts for comparative purposes is a rapid activity supported by speedy and targeted navigation:

> *When marking on paper, it's easy enough to look back at an earlier script. It's in a pile to one side and even if one does not remember the mark given, or the candidate's name or number, looking at the first sentence, paragraph, identifies the script wanted. With computer marking, 'flicking through the pile' is neither quick nor easy.* (Examiner 11: Diary)

Is annotation influenced by mode?

In order to compare examiners' annotation behaviours, 30 paper and screen scripts from each examiner were analysed and their annotation use coded. This analysis of 720 scripts represented one-third of all the scripts marked.

Examiners were able to use a wider variety of annotations on paper than on screen since the screen environment allowed only 10 annotation types. These annotations were built into the marking software following consultation with the examination's Principal Examiner.

Analysis showed that examiners used a wider variety of annotation types on paper (on average 7.58 annotation types per examiner) compared with on screen (6.75 annotation types per examiner). Written comments on paper accounted for most of the difference between the types of annotations used on screen and on paper. This type of annotation was used on average nearly 4 times per paper script and generally included sets of phrases directly linked to evidence found in the text to bring together subtle reflections (e.g. "possibly"), holistic and/or tentative judgements (e.g. "could be clearer"; "this page rather better"), to represent internal dialogue or dialogue with the candidate (e.g. "why?"), or to make note of particular features or qualities found in the text (e.g. "context"; "clear").

When comparing the use of the same ten annotations across modes, 8 of the 12 examiners annotated more on paper. Also, the mean number of annotations made on each paper script (19.48) was higher than on each screen script (18.62), although ANOVA analysis showed that this was not a statistically significant difference ($F(1, 22) = 0.13$, $p = 0.72$).

Despite this, ANOVA analyses showed significant mode-related differences between the mean number of paper and screen annotations for four specific annotation categories. "Underlining" ($F(1, 22) = 7.87$, $p = 0.01$) was used more heavily on paper whilst "Very Good" ($F(1, 22) = 4.78$, $p = 0.04$), "Excellent" ($F(1, 22) = 4.68$, $p = 0.04$) and "Support" ($F(1, 22) = 5.28$, $p = 0.03$) annotations were used significantly more frequently on screen. T-test analyses showed that examiners were also significantly more likely to use ideographic annotations to link text on paper such as circling and sidelining ($t(5) = 2.66$, $p < 0.05$), whereas screen annotations only allowed examiners to label discrete qualities found in the text.

It was usual for the examiners to write a final summative comment on the scripts in both modes. Analysis showed that summative comments were made on more than 99% of the paper script sample and more than 97% of the screen script sample. The importance of the summative comment was highlighted by two examiners who suggested that it factored into their final judgement about the quality of each script:

> *What I couldn't write in the margin, because the system didn't let me, I wanted to store up for the final comment. It seems to me that because you can't annotate, the final comment is more important on screen than it is on paper.* (Examiner 5: Interview)

> *In both cases it's in composing the comment that I harden up on exactly what mark I'm going to award.* (Examiner 8: Interview)

Discussion

It is important to acknowledge that this research project had a number of limitations relating to examiner sample, marking load and script distribution that could challenge the generalisability of the findings. First, the study involved only 12 examiners who were pre-selected for participation based on their high performance profiles, and thus their behaviour might not be representative of all examiners. Secondly, the examiners had a comparatively light marking load with a generous time allowance compared with live marking. Finally, the balance of the script sample characteristics did not necessarily reflect the balance of qualities that examiners might face during a live marking session.

This study was motivated by concerns that screen marking might interfere with examiners' reading processes and lead to marking variances when examiners assess longer texts on screen and on paper. This study found the variance across modes to be non-significant for all but one examiner, suggesting that the marking of these essays is feasible using this particular screen technology. Whilst this in itself is an interesting finding, it is only partial since the real issue concerns construct validity and whether marks were given for the same essay features in the different modes. Again, the Kelly's Repertory Grid analysis suggested that there were no significant relationships between specific essay constructs and differences between examiners' marks across modes. Most interestingly, some of these elicited constructs did explain the differences found between the marks given by different examiners regardless of mode, allowing an insight into the variances that are sometimes found between different examiners and providing obvious scope for further research.

Considering the research literature, these quantitative findings appear to sit uncomfortably with the qualitative study findings, and this requires some degree of exploration. The qualitative data suggest that the examiners in this study were able to assess equally well in both modes but that attaining this level of performance on screen exacted a greater cognitive workload on them. This finding mirrors those of other screen reading studies which suggest that reading on screen is cognitively more demanding than reading on paper (e.g. Wästlund *et al*., 2005). It also appears that the examiners were less able to spatially encode the

information accessed on screen compared with paper and that this contributed to them having a weaker mental representation of the text. Again, literature can be found which suggests this to be an unsurprising finding (e.g. Dillon, 1994; O'Hara and Sellen, 1997; Piolat *et al*., 1997). Most importantly, the qualitative analyses in this study help to explain the basis of this modal difference. Examiners' reading navigation styles and elements of their annotating behaviours differed substantially across modes and theory suggests that these differences are important because navigation and annotating can support readers in the process of building stronger mental representations.

Although this study suggests that examiners appeared to work harder on screen to achieve similar outcomes to paper marking, there are two key elements which might help to illuminate this relationship further. First, it is possible that the examiners attained similar levels of consistency across modes because they had enough spare cognitive capacity to accommodate the additional cognitive load exacted by the marking task in the screen environment. This suggests that in this study the examiners were still working below the threshold at which the cognitive effort was manageable enough to maintain currently acceptable levels of consistency. Secondly, the major factor which contributed to this heightened cognitive load in the screen marking environment related to frustration, with the novelty of the screen marking experience factoring heavily into this. Importantly, this factor had a transient quality, becoming clearly less important throughout the marking period as the examiners became more familiar with the experience.

A recommendation of this project is that future research should continue to explore how the characteristics of on-screen marking environments might affect examiner cognitive load and to explore whether there exists a point beyond which additional cognitive load might lead to unacceptable levels of marking consistency. Such a study might consider whether any mode-related marking effects exist when more examiners (with differing levels of expertise) mark a greater number of scripts which are lengthier, and include a wider diversity of characteristics.

References

Battiste, V. & Bortolussi, M. (1988). *Transport pilot workload: a comparison of two objective techniques*. Proceedings of the Human Factors Society 32nd Annual Meeting, 24–28 October, Anaheim, CA, 150–154.

Crisp, V. & Johnson, M. (2007). The use of annotations in examination marking: opening a window into markers' minds. *British Educational Research Journal*, **33**, 6, 943–961.

Dillon, A. (1994). *Designing Usable Electronic Text*. London: Taylor & Francis.

Emerson, L. & MacKay, B. R. (2006). Subjective cognitive workload, interactivity and feedback in a web-based writing program. *The Journal of University Teaching and Learning Practice*, **3**, 1, 1–14.

Fowles, D. (2008). *Does marking images of essays on screen retain marker confidence and reliability?* Paper presented at the International Association for Educational Assessment Annual Conference, 7–12 September, Cambridge, UK.

Hansen, W. J. & Haas, C. (1988). Reading and writing with computers: a framework for explaining differences in performance. *Comm. ACM*, **31**, 9, 1080–1089.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam: North Holland Press, 239–250.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L. & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors,* **34**, 429–439.

Hornbæk, K. & Frøkjær, E. (2001). *Reading electronic documents: the usability of linear, fisheye, and overview+detail interfaces*. In: Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI Letters, **3**, 1, 293–300.

Hsieh, G., Wood, K. R. & Sellen, A. (2006). *Peripheral display of digital handwritten notes*. CHI 2006 Proceedings, Montreal, Quebec.

Jankowicz, D. (2004). *The Easy Guide To Repertory Grids*. Chichester: John Wiley & Sons.

Johnson, M. & Greatorex, J. (2008). Judging text presented on screen: implications for validity. *E-Learning*, **5**, 1, 40–50.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge MA: Harvard University Press.

Just, M. A. & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn and Bacon.

Kelly, G.A. (1955). *The Psychology of Personal Constructs*. New York: Norton.

Kurniawan, S. H. & Zaphiris, P. (2001). *Reading online or on paper: Which is faster?* Proceedings of HCI International 2001. Mahwah, NJ: Lawrence Erbaum Associates.

Marshall, C. C. (1997). *Annotation: from paper books to the digital library*. Proceedings of the Second ACM International Conference on Digital Libraries; Philadelphia, Pennsylvania.

Mayes, D. K., Sims, V. K. & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics*, **28**, 367–378.

Mills, C. B. & Weldon, L. J. (1987). Reading text from computer screens. *ACM Comput. Surv*. **19**, 4, 329–357.

Moroney, W. F., Biers, D. W. & Eggemeier, F. T. (1995). Some measurement and methodological considerations in the application of subjective workload and measurement techniques. *International Journal of Aviation Psychology*, **5**, 87–106.

O'Hara, K. (1996). *Towards a Typology of Reading Goals. Rank Xerox Research Centre Affordances of Paper Project Technical Report EPC-1996–107*. Cambridge: Rank Xerox Research Centre.

O'Hara, K. & Sellen, A. (1997). *A comparison of reading paper and on-line documents*. In: S. Pemberton (Ed.), Proceedings of the ACM Conference on Human Factors in Computing Systems, Atlanta, Georgia. ACM Press: New York, 335–342.

Paek, P. (2005). *Recent Trends in Comparability Studies*. PEM Research Report 05–05.

Piolat, A., Olive, T. & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology*, **19**, 291–312.

Piolat, A., Roussey, J-Y. & Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, **47**, 565–589.

Schilit, B. N., Golovchinsky, G. & Price, M. N. (1998). *Beyond paper: supporting active reading with free form digital ink annotations*. Proceedings of CHI 98, Los Angeles, CA.

Stanton, N. A., Salmon, P.M., Walker, G. H., Baber, C. & Jenkins, D. P. (2005). *Human Factors Methods: A Practical Guide for Engineering Design*. Aldershot: Ashgate Publishing.

Wästlund, E., Reinikka, H., Norlander, T. & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior*, **21**, 377–394.

How effective is fast and automated feedback to examiners in tackling the size of marking errors?

Dr Elizabeth Sykes Independent Consultant in Cognitive Assessment, **Dr Nadežda Novaković, Dr Jackie Greatorex, John Bell, Rita Nádas and Tim Gill** Research Division

Introduction

Reliability is important in national assessment systems. Therefore there is a good deal of research about examiners' marking reliability. However, some questions remain unanswered due to the changing context of e-marking[1], particularly the opportunity for fast and automated feedback to examiners on their marking. Some of these questions are:

- will iterative feedback result in greater marking accuracy than only one feedback session?
- will encouraging examiners to be consistent (rather than more accurate) result in greater marking accuracy?
- will encouraging examiners to be more accurate (rather than more consistent) result in greater marking accuracy?

Thirty three examiners were matched into 4 experimental groups based on severity of their marking. All examiners marked the same 100 candidate responses, in the same short time scale. Group 1 received one session of feedback about their accuracy. Group 2 received three iterative sessions of feedback about the accuracy of their marking. Group 3 received one session of feedback about their consistency. Group 4 received three iterative sessions of feedback about the consistency of their marking. Absolute differences between examiners' marking and a reference mark were analysed using a general linear model. The results of the present analysis pointed towards the answer to all the research questions being "no". **The results presented in this article are not intended to be used to evaluate current marking practices. Rather the article is intended to contribute to answering the research questions, and developing an evidence base for the principles that should be used to design and improve marking practices.**

Background

It is imperative that General Certificates of Secondary Education (GCSE) examinations are marked validly, reliably and accurately. In this article the effectiveness of potential procedures for providing fast and automated feedback to examiners about their marking is evaluated.

For many years a great deal of research resource has focused on the reliability of marking and factors which influence the reliability of marking. The literature covers marking of academic, professional and vocational examinations, as well as marking the work of candidates of varied ages. Examples of research in the field are: Greatorex and Bell (2004; 2008), Akeju (2007), McManus *et al*. (2006), Baird (1998), Richards and Chambers (1996), Williams *et al*. (1991), Laming (1990), Braun (1988), Murphy (1979; 1982). Some, but not a great deal, of this literature focuses on GCSE marking, for example, Suto and Nádas (2007) and Vidal Rodeiro (2007). There are still some unanswered research questions about the effectiveness of different types of examiner training or feedback to examiners in the GCSE context. One such area is the effectiveness of fast and automated feedback to examiners about their marking. With this in mind, the research literature and current practice were used here to develop different approaches to providing feedback to examiners. Subsequently, the effect of each approach on marking accuracy was investigated.

Before setting out the context and the basis of the experimental approaches to feedback, some current pertinent GCSE examining practices need to be noted. For conventional paper-based marking at the beginning of the marking session, examiners normally attend a standardisation meeting. The aim of the meeting is to smooth the progress of high quality marking. In the meeting, scripts and the mark scheme are discussed. After the meeting, examiners submit some of their marked scripts to a senior examiner who reviews their marking and provides individualised feedback to each examiner. Usually the medium of communication is a standard paper form with hand written entries. The form includes marks given by the examiner, the marks given by the senior examiner for the same candidates, and any discrepancies. In some cases the hand written entries provide advice about how to improve marking. Sometimes other supplementary means of communication such as a telephone conversation are used as necessary. If the marking is sufficiently in line with the senior examiner's marking, the senior examiner allows the examiner to continue to mark as they have done so far. If the marking is not sufficiently in line with the senior examiner's, then the process outlined above is repeated. Depending upon the quality of marking, the examiner might not be allowed to mark any further scripts in that examination session. During the marking session further scripts marked by the examiners are sampled and the marking is checked by Team Leaders or the Principal Examiner[2], but feedback is not provided to the examiners. There are also other processes in place for quality control purposes, such as checking marking of scripts near to grade boundaries once grade boundaries have been set; see QCA (2008) for full details.

For each examination there is a range of marks around the Principal Examiner's (PE) or Team Leader's (TL) marking known as 'tolerance'. For many examinations, if an examiner does not mark within tolerance, then they are not an acceptable examiner. However, for some examinations, particularly those including essays, if the examiner's marking is outside

1 E-marking is used here to mean the marking of digital images of examination responses by examiners working at computers.

2 Principal Examiners generally write question papers and are responsible for leading the marking; Team Leaders also oversee some marking.

tolerance but is highly consistent, then the examiner's marking can be scaled. Scaling is the process of adding or subtracting a number of marks from the examiner's marking to bring it in line with the senior examiner's marking. When an examiner is scaled they might be scaled for the whole of the mark range or on part of the mark range. For instance, an examiner might be generous at the top of the mark range and accurate for the rest of the mark range. In which case the marks they gave for most of the mark range would remain unchanged but marks they gave at the top of the mark range would have some marks deducted. During the scaling process the rank ordering of the marks is preserved. One of the few research articles about scaling is Adams and Wilmutt (1982).

For e-marking the process of examiner standardisation is somewhat different to that of conventional paper-based marking. Senior examiners meet to mark a minimum of 35 scripts and agree on what are known as 'definitive marks' for these scripts. The examiners mark a practice sample of scripts remotely. The definitive marks and associated annotations are available for the examiners to consult. Subsequently, the examiners mark ten scripts (standardisation scripts) and submit their marking. Once the marking has been submitted the software informs the examiner of the definitive item level marks for each script. They also receive feedback on their marking from a senior examiner. If an examiner's marking is acceptable they are allowed to go ahead and mark the rest of their allocation. If the marking is not acceptable they can revisit the original standardisation scripts; they also mark another ten scripts and receive feedback on their marking from a senior examiner. If after this second round of feedback the examiner's marking is acceptable, the examiner is cleared by the senior examiner to go ahead and mark the rest of the allocation. If their marking is not acceptable then they are not cleared to continue marking. Once the marking is underway examiners are provided with feedback and monitored. This is accomplished by every 20th script that the examiner marks being a 'seeded script', that is a script for which there is a definitive mark. The differences between definitive marks and examiners' marks can be monitored. If the marking of a seeded script is unacceptable then the Team Leader can review the marking of the last 20 scripts and ask the examiner to re-mark them. The e-marking procedure for standardising marking is different to the conventional paper-based approaches, as feedback can be provided throughout the e-marking session.

There is a wide ranging literature about training and feedback to examiners, much of which is about marking on paper. It is likely that much of the research about paper-based marking is relevant to e-marking. As already noted by Greatorex and Bell (2008), e-marking and linked innovations are associated with the prospect of Awarding Bodies up-dating their practices. In an automated environment, there is the possibility of introducing new training and feedback approaches. For instance, there is the possibility of providing feedback to examiners more quickly than relying on the post. What is more, there is the possibility for the feedback to be automated rather than involving a person-to-person aspect, for example, telephone calls or a face-to-face element, such as co-ordination/standardisation meetings. Bearing these possibilities in mind, our article is intended to investigate which would be the best approach to providing feedback to examiners in an automated environment, based on research evidence.

The traditional reasoning which underpins current paper marking practice is that after examiners have had one or, in some cases, two rounds of feedback and their marking is deemed acceptable, the examiners should continue to mark. It is argued that if they have further feedback then their marking behaviour might change part way through the marking session which makes scaling untenable. There is research that indicates that when conventional paper marking practices are followed some examiners still drift a little over time in terms of their leniency or severity (Pinot de Moira *et al*., 2002). This finding is consistent with other research from beyond the GCSE and A-level context; see Aslett (2006) for a summary. Another argument associated with this traditional line of reasoning is that if feedback is given part way through the marking session the examiners can overcompensate by swinging from severe to lenient or vice versa. This view is also supported by research from outside the GCSE or A-level context such as Shaw (2002), Hoskens and Wilson (2001), as well as Lumley and McNamara (1993). This would then make scaling untenable (unless Awarding Bodies know when responses are marked and are happy to apply different levels of scaling at different times as necessary). In e-marking it is possible to provide feedback iteratively during the marking session. However, this approach contradicts the traditional reasoning.

In some research about feedback to examiners the feedback has been provided shortly after the marking had taken place, perhaps within 24 hours, for example, Hoskens and Wilson (2001). This highlights a limitation of some of the other research in this area such as Shaw (2002) and Greatorex and Bell (2008) where the feedback was received by post and so there was some delay between the marking and receiving feedback.

Another line of traditional reasoning is that examiners should be encouraged either to replicate the marking of the senior examiner, or to be consistently more lenient or severe than the senior examiner. This latter practice is maintained so that examiners can be scaled. The research literature suggests that training or feedback aimed at getting the examiner to be self-consistent (increasing intra-examiner consistency) is likely to be more successful than feedback or training which encourages the examiners to replicate the senior examiner's marking (increasing examiner accuracy or "inter-examiner reliability") (Weigle, 1998; Lunz *et al*., 1991).

To our knowledge some of these issues have not been investigated in the GCSE context. With this in mind the following questions arise:

1) will iterative feedback result in greater marking accuracy than only one feedback session?
2) will encouraging examiners to be consistent (rather than more accurate) result in greater marking accuracy?
3) will encouraging examiners to be more accurate (rather than more consistent) result in greater marking accuracy?

Method

Design

Interventions

This marking experiment applied combinations of four types of interventions:

- examiners receiving one round of feedback
- examiners receiving iterative feedback
- examiners receiving 'accuracy feedback'
- examiners receiving 'consistency feedback'

Each type of intervention is explained in more detail below.

One round of feedback

Examiners received one round of feedback on their marking near the beginning of the marking session.

Iterative feedback

Examiners received feedback on their marking at regular intervals during the marking session.

'Accuracy feedback'

This type of feedback drew examiners' attention to differences between their marks and the reference marks (the reference marks were taken to be the true score for this experiment, more details are given below). The differences between the reference marks and the examiners' marks were provided as *actual differences*. That is, the examiners could see whether the differences were positive or negative and therefore whether they were more lenient or severe than the reference mark. The feedback was presented in graph form so that examiners could see how accurate they were across the entire mark range.

'Consistency feedback'

Examiners received feedback that drew their attention to those responses where the mark they had given deviated in some way from their usual marking level (for example, if they showed a tendency to be in line with the PE or lenient, their attention was drawn to those responses where they marked more harshly). The feedback was presented in graph form so that examiners could see how consistent they were across the entire mark range. In this way, drawing their attention to differences between their marks and the reference marks was avoided, as this could potentially sway the examiners in their marking.

For both the 'accuracy feedback' and 'consistency feedback' interventions, the examiners received written detailed instructions on how to interpret the graphs, before marking began (see Appendix 1). As far as possible the instructions were the same for all groups. The examiners were also given ample opportunity to get in touch with the research team both before and during the marking to raise any queries about the feedback they received. This process was intended to simulate an automated system of providing feedback to examiners on their marking.

During the marking phase, each examiner marked a total of 100 paper responses to one question. The examiners were asked to mark at the item level rather than at the script level because this approach reflects an e-marking environment, where examiners might mark assigned questions rather than assigned scripts (whole question papers).

The four groups marked the same batch of scripts in the same order. There were 5 batches, each consisting of 20 responses covering a wide range of marks. Each batch included the same number of responses in order to avoid a practice effect influencing the accuracy of the marking at each stage in the experiment. Examiners marked one batch of responses per day. The examiners marked the first batch on day one and repeated this exercise with the consecutive batches over each of the following 4 marking days. They received the feedback on their marking (as appropriate) the following morning. Table 1 illustrates the experimental design used in the study.

The first set of 20 responses constituted a practice sample which served as a 'warm-up' exercise to help the examiners remind themselves of the mark scheme and prepare them for marking the remaining four sets of responses. Thus, no group received any feedback after marking the first batch.

Table 1: Experimental design

Day	*Accuracy feedback*		*Consistency feedback*	
	Group 1	*Group 2*	*Group 3*	*Group 4*
1	Batch 1	Batch 1	Batch 1	Batch 1
2	Batch 2	Batch 2	Batch 2	Batch 2
3	*Feedback on batch 2*	*Feedback on batch 2*	*Feedback on batch 2*	*Feedback on batch 2*
3	Batch 3	Batch 3	Batch 3	Batch 3
4		*Feedback on batch 3*		*Feedback on batch 3*
4	Batch 4	Batch 4	Batch 4	Batch 4
5		*Feedback on batch 4*		*Feedback on batch 4*
5	Batch 5	Batch 5	Batch 5	Batch 5

Procedure

Only one question was selected to be used in the research. After live marking responses to that one question in some OCR scripts were copied. All the copies were cleaned of marks. Thus multiple copies of the same responses could be marked by many examiners.

Two PEs were asked to give their own reference marks for candidates' responses. The PEs then compared their marks and agreed on a reference mark for each response. This approach reflects the procedures used to determine definitive marks in an e-marking context/ environment.

Each experimental group (1 to 4) experienced the interventions as described above. All the marking was undertaken remotely. Examiners were expected to spend around 120 minutes marking each batch (it takes approximately five minutes or less to mark the question). The 5 batches of responses were sent out to examiners by post. After marking a batch the examiners sent their marks back to the research team by e-mail, and received the feedback by e-mail.

Script samples

A GCSE English Higher Tier examination question was used in the experiment. Candidates could score 30 or fewer marks on the question. A total of 100 responses with reference marks were divided into 5 batches of 20 responses. Each batch of 20 responses was intended to include a similar range of achievement. The resulting frequency of reference marks by batch is given in the Table 2 below.

Participants

In addition to the two PEs a total of 33 examiners took part in the study. All the examiners were experienced examiners who had marked the GCSE English Higher Tier examination in live marking. Other reasons for recruiting these particular examiners included that they were all contactable by email and available to mark at the scheduled times. The examiners were divided into four experimental groups: two groups consisted of nine examiners, one group consisted of eight examiners and one group of seven examiners. The differences in numbers in groups were due to issues like availability and dropout.

To form the groups, the examiners were matched in terms of their

Table 2: Reference marks (agreed by 2 PEs) and the frequency of each reference mark in each batch

Reference mark	*batch 1*	*batch 2*	*batch 3*	*batch 4*	*batch 5*
13	0	0	0	0	1
14	0	0	1	1	0
15	2	2	1	1	1
16	0	0	1	1	2
17	2	2	2	1	0
18	2	3	1	2	2
19	3	2	3	3	3
20	3	1	2	0	2
21	0	2	1	3	1
22	1	2	3	2	3
23	3	1	1	1	2
24	1	2	1	2	0
25	1	2	2	1	0
26	0	0	0	0	1
27	1	0	0	1	1
28	0	1	1	0	1
29	0	0	0	1	0
30	1	0	0	0	0

Table 3: Number of participants who were lenient or severe in a previous session of live marking of the examination

Group	*Neither lenient or severe*	*Lenient*	*Severe*	*Total*
1	3	5	1	9
2	4	5	0	9
3	3	4	1	8
4	2	4	1	7

severity in the previous live marking session; the intention was that there would be a similar variety of marking severity in each group to avoid group differences.

Table 3 provides a summary of the final distribution of the historical severity and leniency of examiners who went on to complete all aspects of the study. For the purposes of Table 3 examiners were classified according to their live marking of the examination in the previous live marking session. The classifications were 'neither lenient nor severe' if they were not scaled, 'lenient' if their scaling resulted in marks being deducted and 'severe' if their scaling resulted in marks being added.

Results

A statistical analysis of the absolute differences between the examiner's mark and the reference mark was conducted. When we discuss the analysis and results from our data in this article we refer to absolute differences as a measure of accuracy[3] or marking error. To report *absolute differences* all negative differences were changed to positive values. This has the advantage that the overall size of the marking error can be seen, regardless of the levels of the severity or leniency of marking. Reporting absolute differences also has the advantage that a lower mean absolute difference is an improvement in accuracy, whereas this information is lost when reporting actual differences (where positive and negative differences can negate each other).

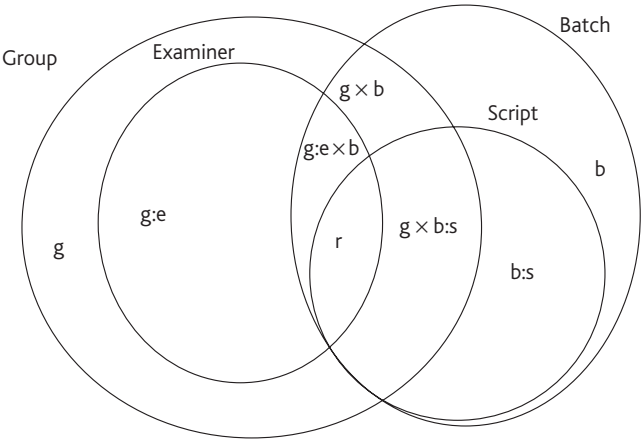


Figure 1: Diagram representing the structure of the experiment

Figure 1 above represents the structure of the experiment. In the diagram 'g' represents experimental group, 'e' represents examiner, 'b' represents batch, 's' represents response and 'r' represents residual error. Examiners are nested within groups (g:e) crossed with responses nested within batches (b:s). This indicates that it is possible to estimate two main effects, batch and group, and five interaction effects (examiner within group, response within batch, group crossed with batch, examiner within group crossed with batch and a group crossed with response within batch). Finally, there is a confounded residual error. Ideally there should be no differences between groups. Examiners within groups and responses within batches are expected to be different. Batch and group crossed with batch are effects that the experiment was designed to estimate.

The above model can be represented as an equation:

$$y_{gebs} = \mu + \tilde{\mu_g} + \tilde{\mu_b} + \tilde{\mu_{g\times b}} + \tilde{\mu_{g:e}} + \tilde{\mu_{b:s}} + \tilde{\mu_{g:e\times b}} + \tilde{\mu_{g\times b:s}} + \tilde{r_{gebs}}$$

where y_{gebs} is the marks difference for examiner e in group g marking response s in batch b,

- μ is the grand mean,
- $\tilde{\mu_g}$ is the overall effect of group g,
- $\tilde{\mu_b}$ is the overall effect of batch b,
- $\tilde{\mu_{g\times b}}$ is the effect of the interaction between batch b and group g,
- $\tilde{\mu_{g:e}}$ is the effect of examiner e in group g,
- $\tilde{\mu_{b:s}}$ is the effect of response s in batch b,
- $\tilde{\mu_{g:e\times b}}$ is the effect of the interaction between examiner e within group g crossed with batch b,
- $\tilde{\mu_{g\times b:s}}$ is the effect of the interaction between group g crossed with response s within batch b,
- $\tilde{r_{gebs}}$ is the error term.

The foci of this study are the batch effect, the interaction between group and batch and the interaction of examiner within group crossed with batch.

3 The reader might notice that when we discuss accuracy in the context of accuracy feedback we are concerned with actual differences, when we are discussing the analysis of our data we are discussing absolute differences and when we are discussing the research literature we might be referring to actual or absolute differences. These uses of the term accuracy are in keeping with much of the research literature.

Table 4: The General Linear Model

Source	*df*	*Type III SS*	*Mean Square*	*F Value*	*Pr > F*
Group	3	81.84	27.28	5.96	0.0005
Batch	4	166.96	41.74	9.11	<.0001
Examiner (group)	29	2553.45	88.05	19.22	<.0001
Response (batch)	95	7282.17	76.65	16.73	<.0001
Group*batch	12	104.03	8.67	1.89	0.0308
Examiner (group) * batch	114	1993.45	17.49	3.82	<.0001
Group*response (batch)	285	1105.77	3.88	0.85	0.9656
Error	2717	12446.63	4.58		

A general linear model was applied to the absolute differences between the examiner's marks and the reference mark.

The results in Table 4 indicate that most of the effects were significant ([Pr>F]<0.05). The results can be taken to mean that:

- in general the marking of each group was different;
- in general each examiner's marking changed over batches;
- individual examiners within a particular group had different levels of marking accuracy;
- the accuracy of marking was different for different responses;
- each group's marking accuracy changed from batch to batch (generally accuracy was improved over time until batch 5 when marking became more inaccurate);
- the examiners in different groups marked the different batches differently;
- the experimental groups of examiners did not generally mark the same response differently, i.e. the experimental groups tended to have similar accuracy levels for the same response.

Figure 2 illustrates that the marking accuracy of all groups generally increased with each batch except for the final batch of marking. In this analysis least square (LS) means can be used in the way that an arithmetic mean would be used in other situations.

Multiple comparisons procedures, like the general linear model, are used to control for the familywise error rate. For example, suppose that we have four groups and we want to carry out all pairwise comparisons of the group means. There are six such comparisons: 1 with 2, 1 with 3, 1 with 4, 2 with 3, 2 with 4 and 3 with 4. Such a set of comparisons is called a family. If we use, for example, a t-test to compare each pair at a certain significance level $\propto$, then the probability of Type I error (incorrect rejection of the null hypothesis of equality of means) can be guaranteed not to exceed $\propto$ only individually, for each pairwise comparison separately, and not for the whole family. To ensure that the probability of incorrectly rejecting the null hypothesis for any of the pairwise comparisons in the family does not exceed $\propto$, multiple comparisons methods that control the familywise error rate (FWE) need to be used (Westfall *et al*., 1999).

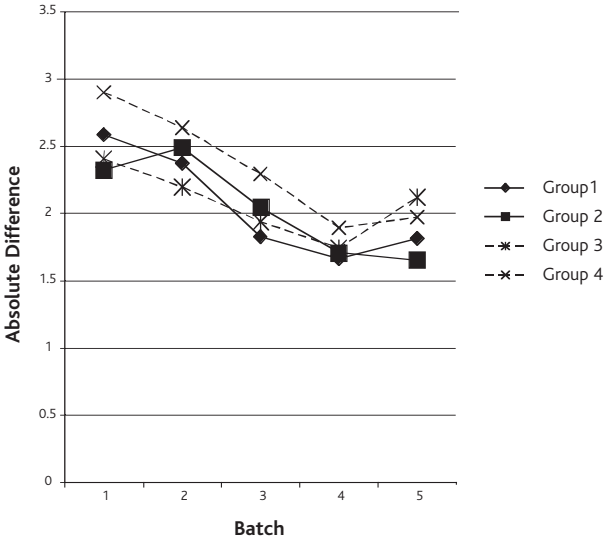


Figure 2: LS means of group by batch

The LS means for the effect of batch are shown in Table 5 and illustrated in Figure 3. Table 6 shows whether the means of each pair of batches are statistically significantly different.

Table 5: Adjustment for multiple comparisons: absolute differences

Batch	*LS mean*	*95% Confidence Limits*	
1	2.55	2.47	2.68
2	2.43	2.30	2.56
3	2.02	1.89	2.15
4	1.75	1.62	1.88
5	1.88	1.74	2.01

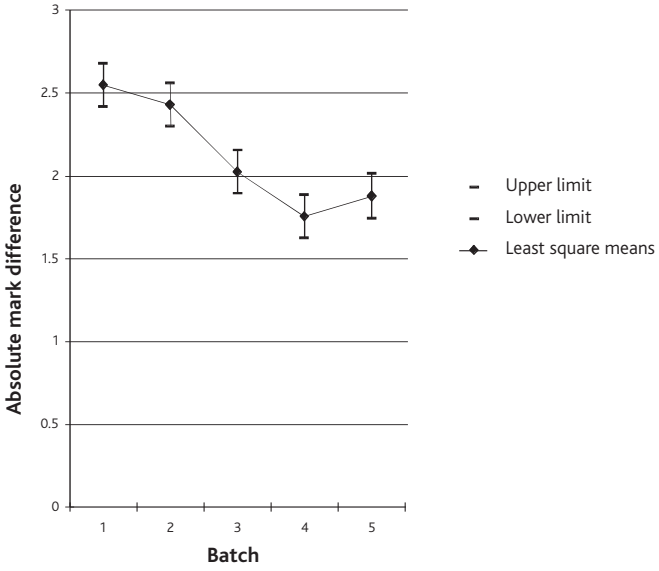


Figure 3: LS means by batch (with confidence intervals) for all groups

Table 6: LS means for the effect of batch: absolute differences

Batches	*t value*	*Pr > (t)*
1–2	1.27	0.71
1–3	5.57	< 0.01
1–4	8.43	< 0.01
1–5	6.99	< 0.01
2–3	4.23	< 0.01
2–4	7.16	< 0.001
2–5	5.74	< 0.001
3–4	2.87	0.03
3–5	1.53	0.55
4–5	1.28	0.71

Overall, the changes in accuracy as measured by mean absolute differences were as follows:

(1) **Batch 1–Batch 2:** there was no feedback provided but there was increased familiarity with the Mark Scheme. There was a slight non-significant improvement in accuracy.

(2) **Batch 2–Batch 3:** all groups received feedback. This comparison showed that there was a significant improvement in accuracy and this was the largest improvement between consecutive batches.

(3) **Batch 3–Batch 4:** Groups 1 and 3 had no feedback, Groups 2 and 4 had feedback. This comparison showed that there was a significant improvement in accuracy and that all groups continued to improve in accuracy.

(4) **Batch 4–Batch 5:** Groups 1 and 3 had no feedback, Groups 2 and 4 had feedback. There was no improvement in accuracy for any group. In fact there was a slight non-significant decline.

Accuracy improved for all groups of examiners after they had the first round of feedback. The improvement was sustained for another round of marking for all groups whether they received continued feedback or not. Performance then levelled off on the last round of marking for all groups.

Thus, in terms of LS means, the findings showed that the first round of feedback (accuracy and consistency) was effective in bringing the examiners' marking nearer to the reference mark and that the difference in the mean marks between examiners and the reference mark was reduced from 2.55 marks to 2.02 marks. There was continued improvement for one more round of marking, reducing the difference in the mean marks from 2.02 to 1.75. The mean mark for every group was within 2 marks of the reference mark by the fourth batch. Improvement appeared to level off at this point although the mean mark difference between examiners and the reference mark for the fifth batch remained below two marks. The pattern was the same for all of the groups, suggesting that initial feedback per se was effective in reducing marking error, but that neither the type nor the amount of feedback were important in contributing to improved accuracy.

It is worth noting that in this analysis the main comparison is between the marking trajectories of the different groups rather than a direct comparison between each group's marking at each stage of the experiment.

Discussion

Awarding Bodies have indicated a keen interest in examiner training in the GCSE context. Advances in computerised technology have provided the opportunity to consider their impact on the possibility of updating methods for providing training and feedback to examiners during the marking sessions. Being able to mark responses on screen and receive feedback by email shortly after each marking session rather than by post might both be expected to impact on the reliability of examiner marking.

The aim of this study was to investigate how feedback might affect levels of reliability of examiners' marking in the GCSE context and to consider the results in the context of an automated environment. The administration of two different amounts of feedback (once and three times) and of two different types of feedback (accuracy and consistency) were investigated.

The accuracy of examiners' marking was investigated by measuring the absolute differences between the examiners' marks and the reference mark. There were significant differences between the four groups of examiners and the five batches of responses. However, all of the groups performed similarly across batches. The marking of all groups improved in accuracy over the course of the study, with the greatest improvement being evident after the first round of feedback. The improvement was sustained for another round of marking for all groups whether they received continued feedback or not. Performance then levelled off on the last round of marking. The mean mark for each group was approximately 2 marks off the reference mark by the fourth batch and remained at this level to the end of the study. The mean mark for all groups together was within 2 marks of the reference mark by the fourth batch and remained so to the end of the study. Thus initial feedback per se was effective in reducing marking error, but neither the type nor the amount of feedback was important in contributing to improved accuracy. In other words our analysis of absolute differences indicated that the answer to all three research questions is 'no'.

Similarly, Shaw (2002) noted increases in accuracy up to batches 3 and 4, although these were not maintained in the fifth batch of marking. By the end of his study, accuracy levels had returned to the level they were at the start of the study. The tailing off in increases in accuracy may have been the result of 'participation fatigue' (Shaw, p. 17). Shaw suggested that the increases in accuracy were the result of feedback but there was no control group to test this theory. Likewise Greatorex and Bell (2008) suggested that feedback could have led to an increase in marking accuracy, but these researchers recognised that, as in Shaw's study, the research design did not include a no-feedback control condition in order to clarify this suggestion. Furthermore, Greatorex and Bell found no clear pattern to suggest which kind of feedback might account for the rise in accuracy. The current study had the benefit of a control group to make identification of an effect (or non-effect) of iterative feedback more discernible.

In Shaw (2002), and Greatorex and Bell (2008), the feedback was not given immediately after the marking had taken place, but it was provided a little later due to providing the feedback by post. Although this reflects some current practice, examiners might benefit from more immediate feedback. In both studies feedback on the previously marked batch was provided just before the next batch was marked. One of the aims of the current study was to provide feedback within 24 hours of marking, as in Hoskens and Wilson (2001).

A limitation of the present study is that not all possible forms of feedback were researched. Arguably, a further limitation of the research concerns the allocation of participants to groups, which was based on the severity of previous live marking at the examination level. The marking in this study is at the item level. It is possible that the severity of live marking at the examination level is not linearly related to severity of experimental marking at the item level, and it is beyond the scope of this article to investigate this relationship. However, the size of the mean marking error for different groups in batches 1 and 2 differs by less than a mark (see Figure 2). This suggests that the groups were fairly well matched at the beginning of the study in terms of the size of the marking error.

There is a caveat for using the results presented in this article, as follows. We analysed only absolute differences and not actual differences between the examiner's mark and the reference mark. For *absolute differences* all negative differences were changed to positive values. This has the advantage that the overall size of the marking error can be seen, regardless of the levels of severity or leniency. Analysis of *actual*

differences between the examiner's mark and the reference mark (negative differences remain negative) provides information regarding levels of severity or leniency. Sometimes the analysis of actual and absolute differences can lead to different research outcomes, one such case in a marking study is one of the experiments reported in Baird *et al*. (2004). However, for this article we were concerned with the accuracy of the marking or the size of marking errors, which is estimated using absolute differences.

It should also be noted that the results presented in this article cannot be used alone to evaluate the utility of current live marking practices. To use the results presented here it is advisable to:

- **investigate how different types of feedback affect severity and leniency which are not considered in this article;**
- **note that the experiment intended to simulate potential procedures for an automated environment and answer research questions, and not to evaluate the utility of live marking practices, which are different to the procedures in the experiment.**

One line of traditional reasoning that underpins current practice is that after examiners have had one (or in some cases two) round(s) of feedback and their marking is acceptable, the examiners should be left to mark. Some research indicates that some examiners drift a little over time in terms of their leniency or severity even with the initial feedback (Pinot de Moira *et al*., 2002). Other research shows that iterative feedback can lead to examiners swinging from leniency to severity or vice versa (Shaw, 2002; Hoskens and Wilson, 2001; Lumley and McNamara, 1993). It was beyond the scope of this article to investigate whether examiners' marking swung from severe to lenient. The analysis of absolute differences in the present article indicated that marking accuracy tended to increase throughout the study (except for the final batch) but that the iterative feedback was no better than one-off feedback in tackling marking errors. Indeed the initial feedback was the most effective; this might be partly because at the beginning of the study there was a greater marking error to rectify. This suggests that there would be no apparent benefit in providing feedback (of the types used in this study) throughout an e-marking session based on absolute differences between examiners' marking and the reference marks.

The other lines of traditional reasoning are that examiners should be encouraged either to replicate the marking of the senior examiner, or to be consistently more lenient or severe than the senior examiner. Previous research suggests that training or feedback aimed at getting the examiner to be consistently severe or lenient in comparison to the senior marker is likely to be more successful than feedback or training to encourage the examiners to replicate the senior examiner's marking (Weigle, 1998; Lunz *et al*., 1991). The analysis of absolute differences did not indicate that one approach was more beneficial than the other.

References

Adams, R.M. & Wilmut, J. (1982). A measure of the weights of examinations components, and scaling to adjust them. *The Statistician*, **30**, 263–9.

Akeju, S. A. (2007). The reliability of General Certificate of Education Examination English composition papers in West Africa. *Journal of Educational Measurement*, **9**, 3, 175–180.

Aslett, H. J. (2006). Reducing variability, increasing reliability: exploring the psychology of intra- and inter-rater reliability. *Investigations in University Teaching and Learning*, **4**, 1, 86–91.

Baird, J. (1998). What's in a name? Experiments with blind marking in A-level examinations. *Educational Research*, **40**, 2, 191– 202.

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

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational and Behavioural Statistics*, **13**, 1, 1–18.

Greatorex, J. & Bell, J. F. (2004). Does the gender of examiners influence their marking? *Research in Education*, **71**, 25–36.

Greatorex, J. & Bell, J.F. (2008). What makes AS Marking Reliable? An Experiment with some stages from the Standardisation Process. *Research Papers in Education*, **23**, 3, 333–355.

Hoskens, M., & Wilson, M. (2001). Real-Time Feedback on Rater Drift in Constructed-response Items: An Example from the Golden State Examination. *Journal of Educational Measurement*, **38**, 2, 121–145.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgements. *The Quarterly Journal of Experimental Psychology Section A*, **42**, 2, 239–254.

Lumley, T. & McNamara, T.F. (1993). Rater Characteristics and Rater Bias: Implications for Training. *Language Testing*, **12**, 54–71.

Lunz, M.E., Stahl, J.A. & Wright, B.D. (1991). *The invariance of judge severity calibrations*. Paper presented at the annual meeting of the American Research Association, Chicago, IL. Quoted in Weigle, S.C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, **15**, 2, 263–287.

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, **6**, 42 http://www.biomedcentral.com/1472–6920/6/42

Murphy, R. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, **52**, 1, 58–63.

Murphy, R. (1979). Removing the marks from examination scripts before re-marking them: does it make a difference? *British Journal of Educational Psychology*, **49**, 1, 73–78.

Pinot de Moira, A., Massey, C., Baird, J. & Morrissy, M. (2002). Marking consistency over time. *Research in Education*, **67**, 79–87.

Qualifications and Curriculum Authority (2008). *GCSE, GCE and AEA code of practice 2008*, (London, Qualifications and Curriculum Authority) http://ofqual.gov.uk/files/Code_of_practice_April_2008.pdf

Richards, B. & Chambers, F. (1996). Reliability and validity in the GCSE oral examination. *Language Learning Journal*, **14**, 28–34.

Shaw, S (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes*, **8**, 13–17.

Suto, I. & Nádas, R. (2007). The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters: a Cambridge Assessment Publication*, **4**, 2–5.

Vidal Rodeiro, C. L. (2007). Agreement between outcomes from different double marking models. *Research Matters: a Cambridge Assessment Publication*, **4**, 28–34.

Weigle, S.C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, **15**, 2, 263–287.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D. & Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*, SAS Institute.

Williams, R., Sanford, J., Stratford, P. W. & Newman, A. (1991). Grading written essays: a reliability study. *Physical Therapy*, **71**, 9, 679–686.

APPENDIX 1:
INSTRUCTIONS FOR INTERPRETING CONSISTENCY FEEDBACK FOR GROUP 3

Dear examiner,

This document is intended to prepare you for the feedback you will receive after marking Batch 2. It contains explanations as to what the feedback will look like and how to interpret it. Please read this before you start marking. If you have any questions or are unclear about anything related to the feedback you will receive, please do not hesitate to contact us as soon as possible.

The feedback you will receive will be different from the feedback you receive in live marking (after standardisation sample). In live marking, the feedback you receive shows the difference between your marks and Principal Examiner's marks. However, the feedback you will receive here will show the extent to which the marks you have given to certain responses differ from your average marking for that specific mark range. In other words, the feedback will not focus on how different your marking is from that of the PE, but it will focus on the consistency of your marking.

You will receive feedback on all the marks you have given to responses within a batch. The feedback you receive will be in the form of a graph similar to the graph presented below (these are made-up data).

As you can see, the graph consists of two axes. The X-axis is a thick horizontal line running through the middle of the graph. The ticks on this line represent marks, from 0 to 30, that can be given to a candidate's work.

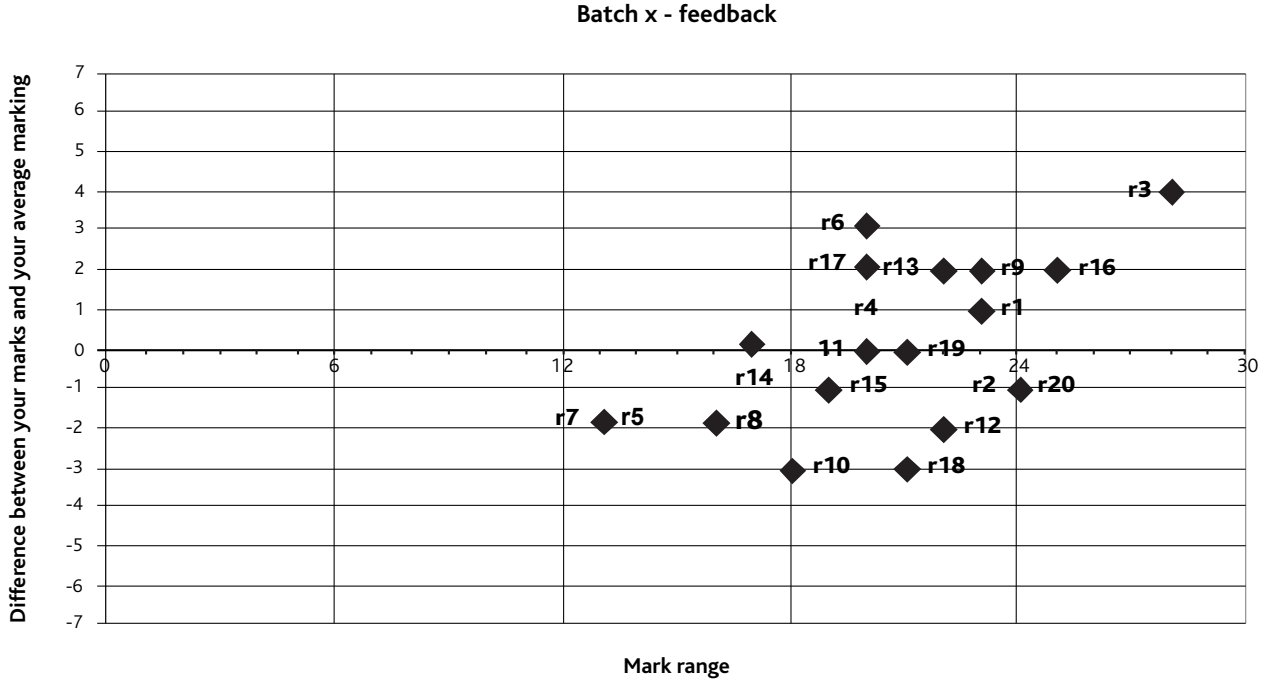
The Y-axis is the leftmost vertical line and it shows to what extent your marks differ from your average marking. If this difference is above 0, this means that you have marked a candidate's work more generously than would be expected if your average marking is taken into account. If the difference is negative, i.e. below 0, this means that you were harsher than would be expected if your average marking is taken into account.

The "diamonds" scattered over the graph plot area represent candidates' responses from the batch. These are marked as r1 (response 1), r2 (response 2) etc. and refer to the number on the first page of each candidate's response, which is also the number in the mark recording sheet which we will send you to record your marks.

Let us take, for example, responses number r6 and r14. If you traced an imaginary line from the "diamond" representing script r6 onto the horizontal X-axis, it would cross it at 20, showing that you have given this response a mark of 20. If you traced an imaginary line onto the vertical Y-axis, it would cross it at close to +3, indicating that the mark you gave to this candidates' work was about three marks higher than your average marking. In other words, if your average marking is taken into account, we would have expected you to have given this response r6 a mark of 17, rather than 20. On the other hand, the mark you gave to candidate response number 14 (r14) is consistent with your average marking for this mark range.

The more clustered your marks are around the X-axis, the more consistent you are in your marking for that specific mark range. The more spread out your marks are, the more inconsistent you are in your marking. Furthermore, by taking a look at the graph as a whole you can get an overall impression as to the overall spread of your marks.

We will email you feedback as part of an attached Microsoft Excel sheet.



Mark scheme features associated with different levels of marker agreement

Tom Bramley Research Division

This is a shortened version of a paper presented at BERA in 2008. It does not include the statistical modelling of the results. See Bramley (2008) for full details.

Introduction

Most of the marker agreement analysis reported in research on examinations in England has been at the level of the whole question paper, rather than at the individual item level. The general finding has been that higher correlations among examiners occur on exams containing structured, analytically marked questions than on exams containing essays, and that the less subjective the mark scheme, the greater the reliability (e.g. Murphy, 1978, 1982; Newton, 1996; Massey and Raikes, 2006). The purpose of the research reported here was to concentrate on agreement at the item level (rather than the candidate level) and to dig deeper into the features of the question papers and mark schemes associated with higher and lower levels of marker agreement.

Recent and ongoing research (Suto and Nádas, 2008a, b, *in press*) at Cambridge Assessment is investigating the factors contributing to accurate marking of examinations. These factors can usefully be grouped according to whether they reside in the marker (e.g. factors contributing to marker expertise, such as subject knowledge, level of education, amount of training, etc); or whether they reside in the task (e.g. clarity of mark scheme, nature of candidate response, complexity of marking strategy needed, etc.). For a brief summary of some of this work, see Suto and Nádas (2007).

The study reported here is about the second group of factors, that is, those residing in the task. However, the approach taken contrasted somewhat with that of Suto and Nádas, whose work involved detailed subject-specific analysis in only two subjects (GCSE Maths and Physics). The present study was broader-brush, aiming to identify relatively coarse features of question papers and mark schemes that could apply across a wide range of subjects and be objectively coded by someone without particular subject expertise or examining experience. The aims were to discover which features were most strongly related to marker agreement, to discuss any possible implications for question paper (QP) and mark scheme (MS) design, and to relate the findings to the theoretical framework described above.

The data came from 38 public examinations in mainstream subjects taken at age 16–18 from OCR (GCSE, AS and A-level in June 2006), and CIE (IGCSE, O-level and A-level in November 2006). In contrast to the research cited above, these data were collected from the process of marker monitoring in the live examinations, as opposed to a research exercise taking place later.

In general, marker monitoring is achieved by a hierarchical system where a Team Leader (TL) is responsible for monitoring the quality of the marking by the Assistant Examiners (AEs) in their team. This monitoring is achieved by the TL re-marking a sample of each of their team's allocation of scripts, at one or more points in the marking process. The data used in this study came from sampling from these re-marked scripts across each team (panel) of examiners. Once the scripts had been obtained, the marks awarded by AE and TL at item level were keyed into a database.

The final data set contained over 114000 records, with each record containing a mark from an AE and their TL on a single item. (38 units[1] × an average of 100 candidates per unit × an average of 30 items per unit = 114000).

The coding framework for categorising QP/MS features

The coding framework was developed iteratively – an initial set of features and coding categories was produced after a 'brainstorming' discussion with colleagues, and this framework was gradually modified in the light of experience with applying it to some specific QP/MS combinations.

Hypothesised effects of coding features on marking accuracy

The features to be coded, and the coding categories for each feature, were selected to meet the criteria of being easy to code in a relatively objective way (i.e. not to require specialist subject expertise) and because they were hypothesised to be relevant to marking accuracy, as described below. See the Appendix for some examples of how the coding framework was applied.

Maximum mark [item_max][2]

The maximum mark is an easily codable indicator of the length and weight given to the response. We might expect it to be related to the number (or complexity) of cognitive processing tasks the marker needs to accomplish in marking it. We would probably expect less agreement between markers on questions worth more marks.

Item type [item_type]

This feature was coded using the same definitions of item type as used by Massey and Raikes (2006):

1 Here a 'unit' means a single examination paper – usually just one component of several in the complete assessment.

2 The abbreviation for each category given in square brackets is the variable name which appears in some of the tables and graphs elsewhere in the report.

An **Objective** item was here considered to be one where the mark scheme precisely gives the *only* accepted answer (e.g. a single number or word, or a multiple choice item, or an item where a candidate has to rank given information, etc.). Objective items require only very brief, heavily constrained responses from candidates.

A **Points-based** item is one which is marked against a "points" mark scheme. These items generally require brief responses ranging in length from a few words to one or two paragraphs, or a diagram or graph, etc. The key feature is that the salient points of all or most credit-worthy responses may be pre-determined to form a largely prescriptive mark scheme, but one that leaves markers to locate the relevant elements and identify all variations that deserve credit. There is generally a one-to-one correspondence between salient points and marks.

A **Levels** item is one which is marked against a "levels" mark scheme. Often these items require longer answers, ranging from one or two paragraphs to multi-page essays or other extended responses. The mark scheme describes a number of levels of response, each of which is associated with a band of one or more marks. Examiners apply a principle of best fit when deciding the mark for a response.

Massey and Raikes (op. cit.) found that there was more agreement on objective items than on points-based and levels-based items. This coding feature in effect records the amount of constraint in the acceptable answers. We would expect less agreement on the less constrained responses, but then these are often worth more marks (see above) and require more writing (see below) so we might expect these effects to be confounded. Suto and Nádas (b, *in press*) found that 'Mark scheme flexibility' and 'Single letter answer' were related to marking accuracy in GCSE Physics (in the expected direction).

Answer space [ans_space]

This feature is likely to be strongly related to the maximum mark and the amount of writing required, but it is conceivable that it might have an effect on marker agreement over and above those two features. For example, it might be that the larger the area the marker has to scan visually to locate the correct response, the greater the opportunity for a cognitive processing error, hence lowering the marker agreement.

Writing [writing]

The greater the amount of writing required, the more opportunity there is for candidates to express their answer (correct or incorrect) in a way which is different from what appears on the mark scheme, and thus to require an increasing degree of understanding and interpretation on the part of the marker. We might therefore expect the task of marking questions requiring more writing to be more cognitively demanding, and hence for there to be less marker agreement. Suto and Nádas (b, *in press*) found it to be related to marking accuracy (in the expected direction). For the longer written responses with levels-based mark schemes we might expect differences between the markers in their internalisation of the construct being assessed, and hence differences in marks awarded.

Points to marks ratio [PM_ratio]

We hoped that this feature might be able to distinguish among points-based items worth equal numbers of marks. It seems plausible that where the marker has a wider range of acceptable responses against which to compare the actual responses, the marking task is more complex and we might expect less agreement. As seen in Table 1, this was not always an

Table 1: Coding framework used to code different features of the question papers and mark schemes[3]

QP/MS feature	*Valid values*	*Notes*
Maximum mark	1,2, etc.	Use QP/MS to decide what the sub-questions are. Usually square brackets e.g. [2].
Item type	O (objective) P (points-based) L (levels-based)	Use definitions from Massey & Raikes (2006).
Answer space	N/A '1' up to and including 1 line '2' more than 1 line but less than ½ page '3' ½ page or more	The N/A category is for answers in separate booklets. The 'answer space' does not include the question stem – it is the (maximum) amount of physical space the marker has to scan to locate the answer. This feature can be coded just by looking at the QP.
Writing	N/A '1' one word or simple numerical answer '2' few words / single sentence '3' two or more sentences	The N/A category is for diagrams, sketches, formulas, equations, arrows etc. This feature can be coded by looking at the QP/MS combination.

Points to marks ratio	N/A
	S (same)
	M (more)
		N/A category is for levels-based mark schemes, calculations, QoWC.
		Same = # correct possible answers equals the number of marks available.
		More = # correct possible answers exceeds the number of marks available.
		N.B. Aim to distinguish separate points, not relatively trivial variations in acceptable wording within the same point.

Qualifications, restrictions and variants	N/A N (No) Y (Yes)	N/A is for levels-based mark schemes. This is to capture where the mark scheme **explicitly** says (for example) 'allow xxx' or 'also accept yyy' etc; or where a qualification/restriction is given e.g. 'only if…' or 'must also have…'. It also applies to mark schemes where there is 'error carried forward' (ecf).
Wrong answers specified	N/A N (No) Y (Yes)	N/A is for levels-based mark schemes. This is to capture where the mark scheme **explicitly** specifies an incorrect or unacceptable response, (for example) 'do not accept xxx' or 'NOT yyy' etc.

3 More features than this were coded, but only those features referred to later are listed. See Bramley (2008) for full details.

easy feature to code, because when deciding on the ratio of points to marks the coder has to distinguish between relatively trivial variations in acceptable wording for what is substantively the same point, and substantively different points. Suto and Nádas (b, *in press*) found that a similar feature of 'alternative answers' was related to marking accuracy (in the expected direction).

Qualifications, restrictions and variants [QRV]

It was difficult to predict what the effect of this feature might be on marker agreement. On the one hand, the purpose of adding qualifications, restrictions and variants to the mark scheme is presumably to clarify to the marker exactly what is worthy of credit. Thus it should make it easier to apply the MS accurately, and therefore items with QRV might have higher levels of agreement. On the other hand, the need to bear in mind all the extra information when considering a response might increase the complexity of the marking task and increase the likelihood of a marker error, decreasing the levels of agreement. It is also possible that these two opposing effects might be different for items with different maximum marks. The QRVs might be a help for the larger questions, but a hindrance for the shorter questions. One particular example is where the mark scheme allows 'error carried forward' (ecf)[4]. Suto and Nádas (b, *in press*) found that questions with ecf were marked less accurately.

Wrong answers specified [wrong]

This is where the mark scheme explicitly mentions a possible response which is not worthy of credit. We decided to code this feature separately from the other QRVs because it might be expected in some cases to 'interfere' with the marking strategy. For example, a strategy of matching text in the answer to text in the mark scheme might result in a marker awarding a mark to a wrong answer which has been explicitly specified on the mark scheme, thus lowering agreement levels. On the other hand, as described above, by clarifying what is not worthy of credit, items with wrong answers specified in the mark scheme might be marked more accurately and hence with higher levels of agreement.

Results

The index of marker agreement chosen was the percentage of exact agreement (P_0) between the AE and the TL. This statistic has the great advantages of simplicity and transparency (Bramley, 2007). It does not indicate the direction of any differences (severity or leniency), but these are arguably of less interest here given that they are likely to pertain to individual markers.

The P_0 statistic was calculated for each item in each unit for which there were more than 10 data points. It seemed sensible to compare 'like with like' as much as possible, and to this end we chose to group items by maximum mark. The most natural grouping, based on the numbers of items in the data, is shown in Table 2 below.

Table 2: Distribution of items by maximum mark category

Max. mark	1	2	3	4	5–9	10–20	21–60	Total
No. of items	329	267	139	87	98	50	42	1012

4 Ecf is where a candidate is not penalised for using an incorrect answer obtained in an earlier part of the question as part of their working for a later part of the question. It is most often seen in questions involving calculations.

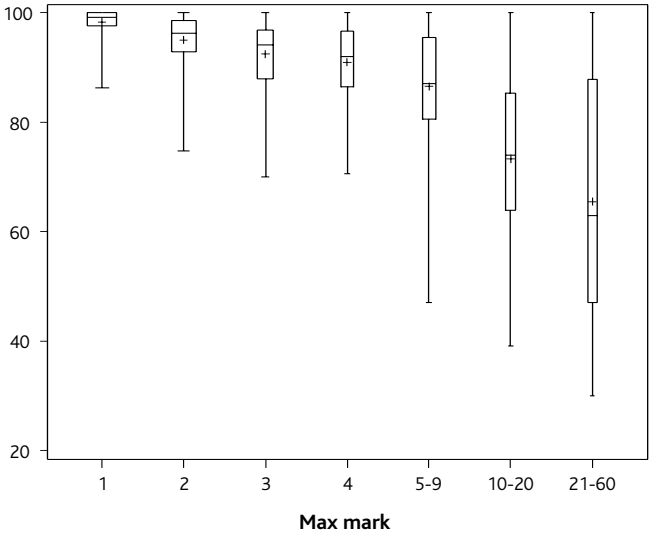


Figure 1: Distribution of P_0 values by item maximum mark. (Width of box is proportional to number of items in each mark category)

We would expect the level of exact agreement between AE and TL to be higher on the lower-mark questions. Figure 1 shows that there was a high level of agreement for the 1-mark items. The median value was around 99% which means that half the 1-mark items had a P_0 value higher than 99%. The vertical length of the box (the interquartile range, IQR) shows that the middle 50% of the 1-mark items had a P0 value in the range ≈97% to 100%. Figure 1 shows that as the maximum mark increased, the average (median or mean) value of P_0 decreased, and that the spread (IQR) of P_0 values tended to increase.

The following graphs show, for each maximum mark category, the median P_0 value for the items with a given feature coding. Many of the coded features were only applicable to objective and points-based items. These items tended to be worth 9 marks or fewer.

Item type

Figure 2 clearly shows that for items with a given maximum mark, there was a higher average level of agreement for 'objective' items than for 'points-based' items. The average difference was about 3 percentage points for 1-mark items, growing to about 10 percentage points for 5–9 mark items. This finding fits the expectation that the amount of constraint in the mark scheme (the essential difference between objective and points-based items) affects the marking accuracy, and agrees with the results of Massey and Raikes (2006).

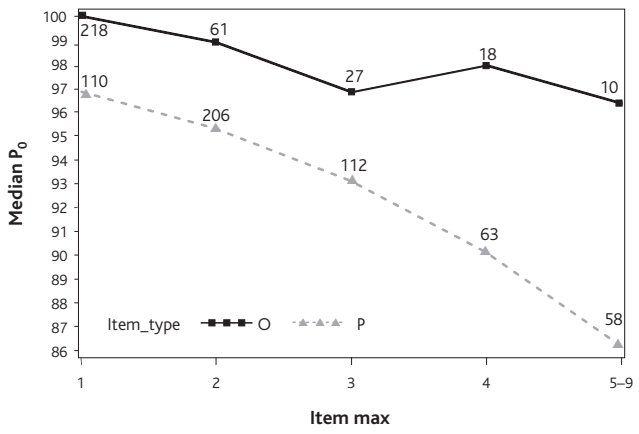


Figure 2: Median P_0 values for objective (O) and points-based (P) items

Points-to-marks ratio (PM_ratio)

Figure 3 shows that for points-based items with a given maximum mark, there was higher agreement for the 'S' items where the number of points equals the number of marks than for the 'M' items where the number of valid points exceeds the number of marks. The differences were around 4 percentage points for 1 and 2 mark items, but larger for the larger items.

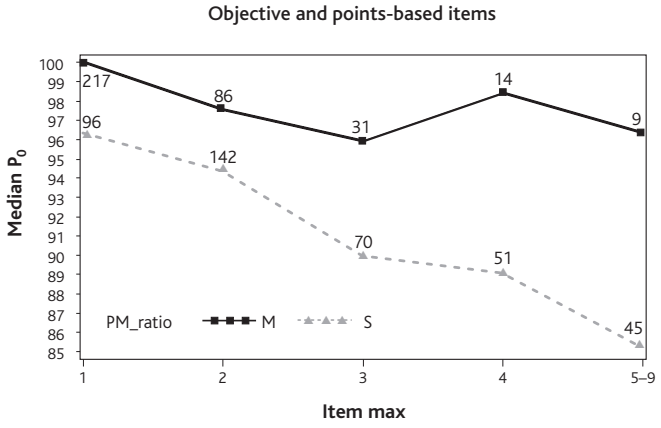


Figure 3: Median P_0 values for objective and points-based items with the same (S) and more (M) points than marks

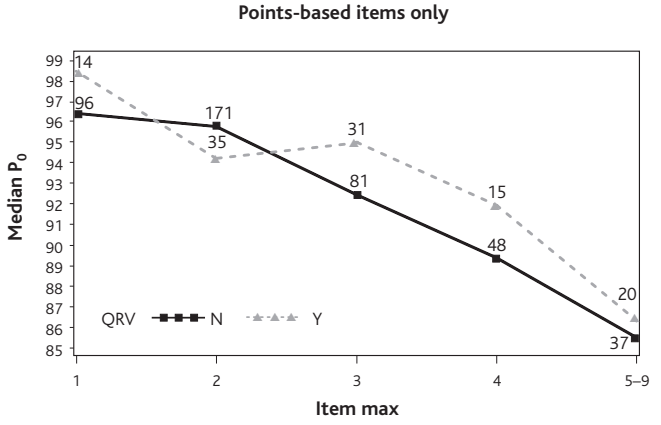


Figure 4: Median P_0 values for points-based items with (Y) and without (N) any QRVs

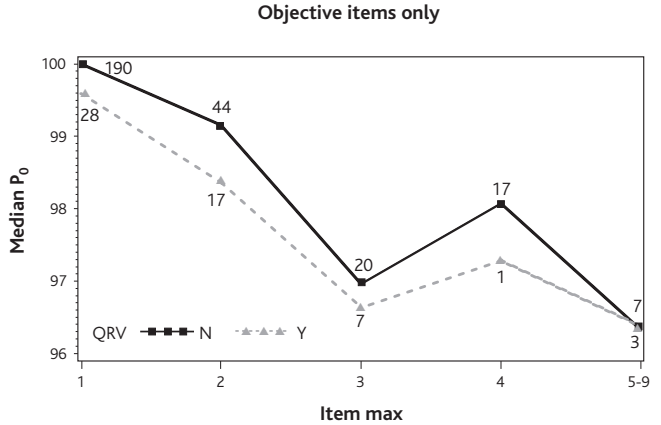


Figure 5: Median P_0 values for objective items with (Y) and without (N) any QRVs

Qualifications, restrictions and variants (QRV)

Figures 4 and 5 show an interesting interaction between item type and the presence of QRVs in the mark scheme. For the points-based items Figure 4, the presence of qualifications, restrictions and variants seemed to increase the level of agreement very slightly. The pattern is spoiled by the 2-mark items, but for the other marks there seemed to be a difference of around 2–3 percentage points. For the objective items, on the other hand, the presence of qualifications, restrictions and variants seemed to reduce the level of agreement very slightly (note the change of scale on the y-axis), as shown in Figure 5. See the discussion for a possible explanation of this result.

Wrong answer specified (wrong)

As with the QRV, it is interesting to separate the objective items from the points-based items, shown in Figures 6 and 7. The presence of a specific wrong answer in the mark scheme appeared to be associated with lower marker agreement for objective items, and also for the 1 and 2-mark points-based items.

The features of 'answer space' and 'amount of writing required' were applicable to all items (that is, not just objective and points-based items up to 9 marks), although obviously in many places there was little overlap between the different cross-categorisations according to maximum mark and item type.

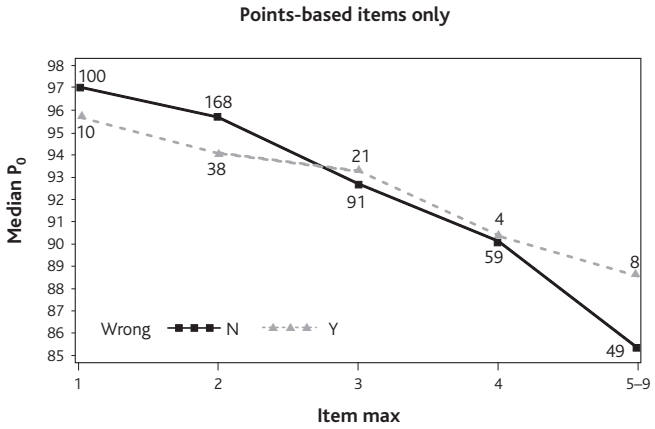


Figure 6: Median P_0 values for points-based items with (Y) and without (N) any wrong answers specified in the mark scheme

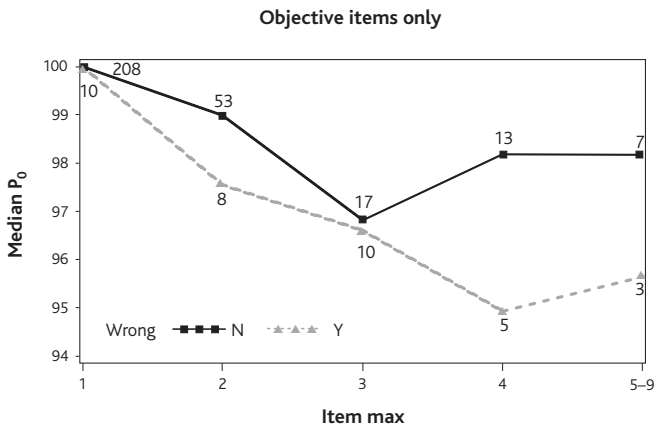


Figure 7: Median P_0 values for objective items with (Y) and without (N) any wrong answers specified in the mark scheme

Answer space (ans_space)

Figure 8 shows that there was a small effect of the amount of answer space for a given maximum mark, in the expected direction – that is, slightly higher agreement corresponding to less physical space for the marker to examine to locate the answer. Perhaps the most interesting feature of Figure 8 is the lack of difference between the values for '2' (answer spaces of more than one line but less than half a page) and 'N/A' (the category for responses in a separate answer booklet). This suggests that although there may be reasons for favouring combined question-answer booklets over separate answer booklets (or vice versa) in terms of the quality and quantity of the candidate's response (Crisp, 2008), the effect on marker agreement is not one of them.

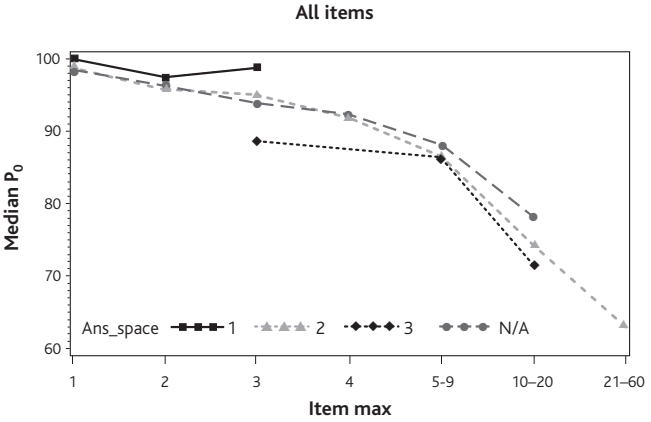


Figure 8: Median P_0 values (all items) for different amounts of answer space

Writing (writing)

In Figure 9 the comparisons based on meaningful numbers of items across the mark range mainly come from items coded '3' or 'N/A' for Writing in the range 2–9 marks. The graph shows that there was much higher agreement (about 6 percentage points) for the 'N/A' items than for items coded '3'. The former were items requiring diagrams, sketches, formulas, equations, arrows, circles, ticks etc. The latter were items requiring two or more sentences.

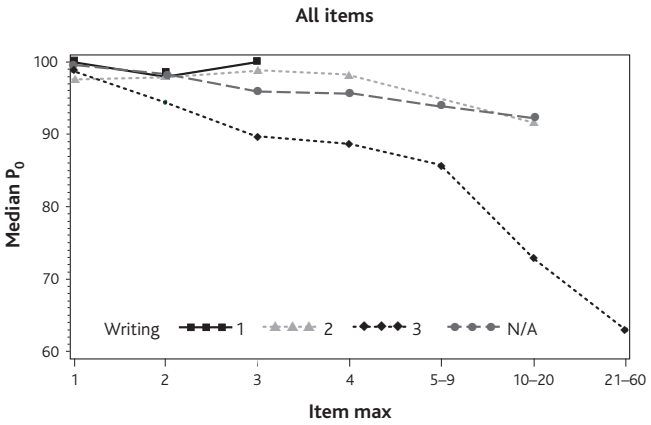


Figure 9: Median P_0 values (all items) for different amounts of writing expected in the response

Points v levels

It is interesting to compare the P_0 values for points-based and levels-based items in the mark ranges where they overlap. Figure 10 shows that the median P_0 value was slightly higher for points-based items worth 4 marks, but that the median values were the same for items worth 5–9 marks, and the levels-based items had higher P_0 values for items worth 10 or more marks. This shows that it is not necessarily the case that a more 'subjective' mark scheme will lead to less accurate marking. This finding should be treated with some caution however, because the high-mark levels-based items were strongly clustered in particular units (subjects).

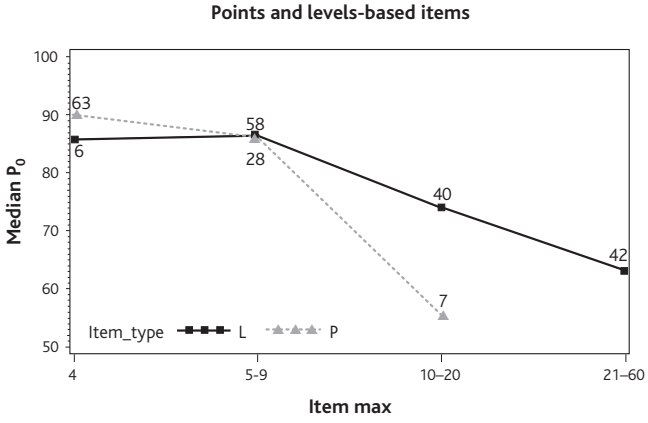


Figure 10: Median P_0 values for points and levels-based items against maximum mark category

Discussion

The qualitative features we coded were all shown to be associated with marker agreement to a greater or lesser extent. Are there any implications for question or mark scheme design? This question cannot be answered without considering validity. As Newton (1996) and many others have pointed out, changing the format of questions or mark schemes to increase the reliability of marking may change what is being assessed. In altering a mark scheme to improve the level of marker agreement it would be very easy to reduce the validity.

A (grossly unrealistic) example would be to decide only to accept one answer in a situation where several valid answers are possible – clearly this would greatly reduce the validity of the question even if it did improve marker agreement. Or imagine a 2-mark question that asked candidates to name two types of rock. The mark scheme might say 'Any two from: igneous, sedimentary, metamorphic'. This question has a points/marks ratio greater than one, which we have shown is associated with lower levels of marker agreement. The question could be changed to ask candidates to name two types of rock other than igneous. The mark scheme would then be constrained to 'sedimentary' and 'metamorphic'. Alternatively, the question could ask for three kinds of rock, changing the mark allocation to 3 and awarding one mark for each type of rock. Either of these would bring the points/marks ratio to one, which would be expected to increase marker agreement (although other things being equal questions worth more marks have lower marker agreement). However, the first might be objected to on the grounds that it is 'unfair' on pupils who only know two out of three rocks, one of them being igneous. The second might in some contexts give too much weight to the question.

To make predictions about marker agreement at this very fine level requires understanding of what causes variation in marker agreement, rather than what is merely associated with it, which is likely to require further experimental work systematically manipulating different features

of questions and mark schemes. The following paragraphs contain some speculative suggestions of how marker agreement on objective and points-based items might be considered in terms of the probability of an 'execution error' in a cognitive processing task.

If the decision to award each mark reflects a single process with a constant probability of error, then the proportion of exact agreement on an *n*-mark question should be equal to the proportion of exact agreement on a 1-mark question raised to the power *n*. Table 4 shows these expected proportions for objective and points-based items separately.

Table 4: Observed and expected proportions of agreement for objective and points-based items

		Item maximum mark					
		1	*2*	*3*	*4*	*5*	*6*
Objective	# items	218	61	27	18		
	observed	0.994	0.983	0.969	0.970		
	expected	0.994	0.988	0.982	0.976		
Points	# items	110	206	112	63	21	17
	observed	0.967	0.944	0.920	0.897	0.857	0.850
	expected	0.967	0.934	0.903	0.873	0.843	0.815

The agreement between the observed and expected proportions is quite close, especially for the objective items. This suggests that considering the award of each mark as an independent process with a constant probability of incorrect execution is a reasonable 'baseline' model. The fact that the agreement for points-based items is slightly higher for an *n*-mark task than for *n* 1-mark tasks is interesting. It seems plausible to assume that there is less of a shift of 'task set' (e.g. Allport *et al*., 1994; Rogers and Monsell, 1995) when carrying out multiple tasks in the same semantic context than when carrying them out across contexts, and this could be related to the probability of an execution error occurring.

The difference between 'objective' and 'points-based' items as defined here is based on constraint. This is likely to affect the marking strategy used. The simpler strategies of 'matching' and 'scanning for simple items' (Suto and Greatorex, 2008) are more likely in general to be applicable to items with highly constrained mark schemes. The greater automaticity of these strategies presumably implies that they are more likely to be executed without error, and hence that the agreement will be higher, even once the number of marks has been taken into account.

A points/marks ratio greater than one can also be seen as increasing the complexity of a given processing task. In the 'types of rock' example above, we might tentatively assume that: i) 'matching' is an appropriate marking strategy; and ii) that it is a serial process rather than a parallel one. Then for the original question ('name two types of rock') the first response from the candidate has to be matched against 'igneous', 'sedimentary' and 'metamorphic', and the second response has to be matched against either all three (if the first response was not one of the three correct types) or whichever two remained (if the first response was one of the three types). For the modified question ('name two types of rock other than igneous') the number of correct answers to match the candidate response against has been reduced. If there is a finite probability of an execution error at each matching step then this would lead to higher marker agreement in the second case.

Qualifications, restrictions and variants in the mark scheme (here including wrong answers specifically mentioned) could help when applying the more complex marking strategies such as 'evaluating' or 'scrutinising' by increasing the information available to the AE and ensuring that their decision matches the (assumed correct) decision of the TL. However, it might be that this extra information interferes with the more simple strategies of 'matching' and 'scanning'. One possibility is that the presence of variant responses forces the marker to use a different cognitive strategy (e.g. 'matching' as opposed to 'scanning') and that this switch carries with it an increased probability of error. For example, if the marker had got into an automatic routine of 'scanning' for the most common correct response and then did not notice when a correct response was different from the one being scanned for, yet nevertheless matched a QRV in the mark scheme, they would wrongly mark it as incorrect. This would fit with the finding that QRVs were associated with higher agreement on points-based items, but lower agreement on objective items.

It is more difficult to relate marker agreement on levels-based questions to the probability of an execution error in a cognitive strategy because it is more difficult to argue that the TL mark (or any one person's mark) is correct. Overall patterns of marker variation are better handled statistically within a many-facet IRT model, or a generalisability theory model, which separate out leniency/severity and erraticism (Bramley, 2007). These models do not say anything, however, about the processes within an individual which lead to the award of a mark. Presumably some kind of matching process is going on in some instances (e.g. those with 'best fit' judgements), but this is not the same kind of 'matching' referred to above. Also, it is plausible that the TL monitoring role is somewhat different when second-marking essays with a levels-based mark scheme, as opposed to shorter points-based items. In the latter, it might be clear to them that their AE has applied the mark scheme incorrectly, whereas in the former they might be prepared to tolerate differences within a certain range and not award a different mark from the AE unless they seriously disagreed.

We can speculate that the lower marker agreement for items requiring a longer written response might be due to the greater interpretation required by the marker to form a representation of the response which can be compared to the mark scheme. In other words, the marker is likely to encounter more ways of expressing the same concepts and thought processes in writing than in (for example) formulas and equations.

Two caveats in interpreting these results should be mentioned: i) when carrying out the qualitative coding of the question papers and mark schemes we were working from the final version of the question papers, and the latest version of the mark scheme that we were able to obtain. There was some inconsistency across different units in what mark scheme was available. In some cases, it is likely that changes made to the mark scheme at the standardisation meeting[5] would not have appeared on the versions we coded. This is likely to have affected some of the coding categories more than others – for example, it is plausible that more items would have been coded positively for QRV and Wrong if we had had access to the final definitive mark scheme used by the markers; and ii) the live setting gave the advantage of no possible artefacts (e.g. time lags, the need for extra or special training, the use of photocopied scripts) which might be introduced in a specialised 'research' setting.

5 The point in the process when final clarifications and amendments are made to the mark scheme, in the light of the PE's marking of a sample of actual candidate responses.

On the other hand, it removed the opportunity for experimental control of the different features of question papers and mark schemes that were coded. We relied on the fact that the sample of units was large and representative of written papers in general qualifications.

In conclusion, this research has shown that some general features of examination question papers and mark schemes, which can be relatively objectively coded across a wide range of subjects, are related to the level of agreement between two markers (or marking accuracy, if one of the marks can be taken as the 'correct' mark). This could be useful in deciding how to allocate resources where there is the option to assign different types of marker to different types of question. In terms of understanding the underlying causes of variation in marker accuracy, these findings fit into a framework that looks to relate question features to cognitive task complexity and to cognitive marking strategies.

References

Allport, D.A., Styles, E.A. & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In: C. Umiltà and M. Moscovitch (Eds.), *Attention and performance Vol. XV*. Cambridge: Bradford, 421–452.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement*. Paper presented at the annual conference of the British Educational Research Association (BERA), Heriot-Watt University, Edinburgh. Available at http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers . Accessed 06/03/09.

Crisp, V. (2008). Improving students' capacity to show their knowledge, understanding and skills in exams by using combined question and answer papers. *Research Papers in Education*, **23**, 1, 69–84.

Massey, A.J. & Raikes, N. (2006). *Item level examiner agreement*. Paper presented at the annual conference of the British Educational Research Association (BERA), University of Warwick, UK. Available at http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers Accessed 09/03/09.

Murphy, R.J.L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*, **48**, 196–200.

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.

Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405–420.

Rogers, R.D. & Monsell, S. (1995). Costs of a predictable switch between simple cognitive tasks. *Journal of Experimental Psychology: General*, **124**, 207–231.

Suto, W.M.I. & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, **34**, 2, 213–233.

Suto, W.M.I. & Nádas, R. (2007). 'The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters: A Cambridge Assessment Publication*, **4**, 2–5.

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

Suto, W.M.I. & Nádas, R. (b, *in press*). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*.

APPENDIX – EXAMPLES OF HOW SOME OF THE CODING CATEGORIES WERE APPLIED

1: Points to marks ratio

The question below was coded as **M** (More) because there were more distinct acceptable points than marks available.

Question:

1 **(a)** Study Fig. 1, a scatter graph which shows the birth and death rates of seven countries in 2004.

(iv) Suggest reasons why Botswana has a higher death rate than the USA. **[3]**

Mark Scheme:

(iv) Ideally answer should be comparative, however be prepared to link points from separate accounts.

Ideas such as:
better quality health care in USA;
more likely to be preventative measures in USA/vaccination;
better diet/food supply in USA/less likelihood of starvation;
better sanitation in USA;
cleaner water supply in USA;
healthier lifestyle in USA;
AIDS is more of a problem in Botswana;
Education re: health care, etc.

3 @ 1 mark or development **[3]**

The following question was coded as **S** (same) because the number of substantive valid points (ignoring slight variations in wording) was equal to the number of marks available. It also contains an example of a wrong answer specifically mentioned.

Question:

Q3 **(c)** Explain in detail how carbon monoxide, produced in this reaction, is poisonous. **[2]**

Mark Scheme:

(c) (CO is poisonous...)

due to complexing / ligand exchange with (Fe of) haemoglobin **[1]**
(NOT redox involving Fe^{2+}/Fe^{3+})

stopping O2 being transported around body/in blood/to tissues/from lungs (1) **[2]**

2: Qualifications, Restrictions and Variants (QRV)

The following two questions were coded **Y** (Yes) for the presence of QRVs. The first one also contains an example of an explicit wrong answer (A stands for 'accept' and R stands for 'reject'), so would also have been coded **Y** for Wrong. The second example allows 'error carried forward' (ecf).

Question:

2 Fig. 2.1 shows a transverse section of a root nodule of a legume. Fig. 2.2 is a drawing of a cell from the centre of the nodule made from an electron micrograph.

(a) Name three structures that are present in cells in the cortex of the root that are not present in bacterial cells. **[3]**

Mark Scheme:

2(a) nucleus/nuclear membrane/nuclear envelope/nucleolus;
ER/SER/RER;
Golgi (body/apparatus) / lysosomes;
larger ribosomes/80S ribosomes;
linear DNA/chromosomes/protein + DNA (in chromosomes);
mitochondrion/mitochondria;
cell wall made of cellulose;
R cell wall unqualified microtubules;
A spindle fibres/centriole large vacuole/tonoplast;
plasmodesmata. **[max 3]**

Question:

4(b)(iv) Calculate the total energy transformed by the three lamps in kilowatt hours when operated for 12 hours.

Mark Scheme:

4(b)(iv) energy = 0.018 × 12 × 3 C1
energy = 0.648 = 0.65 (kW h) (Possible ecf) A1

(0.22 (kW h) scores ½)
(648 (kW h) scores ½)
(2.3 × 106 (J) scores ½)

3: Wrong *(a wrong answer specified in the mark scheme)*

The following question was coded **Y** (Yes) for the 'Wrong' category:

Question:

2 Repondez:
À quelle occasion a-t-elle envoyé les fleurs? **[1]**

Mark Scheme:

Q2
pour son anniversaire de marriage **[1]**
Reject: *anniversaire* t.c.
Reject: *anni versaire* – two words

ASSESSMENT JUDGEMENTS

Thinking about making the right mark: Using cognitive strategy research to explore examiner training

Dr Irenka Suto, Dr Jackie Greatorex, and Rita Nádas Research Division

This article is based on a presentation, "Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training", given by Jackie Greatorex, Rita Nádas, Irenka Suto and John F. Bell at the European Educational Research conference, September 2007, Ghent, Belgium.

Introduction

In England, school-leavers' achievements are assessed through a system of public examinations, taken primarily at ages 16 and 18 (Broadfoot, 1996). High stakes examinations for General Certificate in Secondary Education (GCSE) and Advanced (A) level qualifications are administered by three independent awarding bodies, and are marked externally by professional examiners rather than within schools (Ofqual, 2008). Since employers and higher education institutions use GCSE and A-level grades in their selection procedures (Lamprianou, 2008), it is imperative to ensure that examination marking is valid and reliable. This is a considerable task, given the wide variety of question structures and response formats entailed (Eckstein and Noah, 1993). Awarding Bodies therefore conduct rigorous checks on their marking processes and organise highly specialised examiner training, for example in the form of 'standardisation' or 'co-ordination' meetings (National Assessment Agency, 2008). In this article, we investigate the benefits of, and some possible variations in, these training procedures.

GCSE and A-level assessments are in a period of transition. In this context and beyond there has been particular interest in new developments such as on-screen marking (Hamilton, Reddel and Spratt, 2001; Whetton and Newton, 2002; Leacock and Chodorow, 2003; Raikes and Harding, 2003; Sturman and Kispal, 2003; Sukkarieh, Pulman and Raikes, 2005; Knoch, Read and von Randow, 2007; Raikes and Massey, 2007) and the employment of examiners with differing levels of teaching and examining experience (Powers, Kubota, Bentley, Farnum, Swartz and Willard, 1998; Royal-Dawson, 2005; Raikes, Greatorex and Shaw, 2004; Meadows and Wheadon, 2007; Suto and Nádas, 2007a). The focus on examiners with potentially varying expertise has arisen in part because the UK has recently faced shortages of experienced examiners (usually experienced schoolteachers) in some subjects. Moreover, on-screen

marking enables a single candidate's script to be divided up so that individual questions can be assigned to different examiners according to marking demands and personal examiner expertise (Suto and Nádas, 2007). Alongside the need to ensure that new systems enhance valid and reliable marking, for example through anonymising candidates' responses, lies the growing requirement for effective and optimal forms of training for examiners of varying expertise.

In this article we draw together research on examiner training and on the nature of the judgements entailed in the marking process. We report new analyses of data from two recent empirical studies, Greatorex and Bell (2008) and Suto and Nádas (2008a), exploring possible relationships between the efficacy of training and the complexity of the cognitive marking strategies apparently needed to mark the examination questions under consideration. In the first study reported in this article, we consider the benefits of three different training procedures for experienced examiners marking AS-level biology questions. In the second study reported here, we explore the effects of a single training procedure on experienced and inexperienced (graduate) examiners marking GCSE mathematics and physics questions.

Current practice in examiner training in England

As some GCSE and A-level examinations are taken by several thousands of candidates at a time (Broadfoot, 1996), many examiners may be needed to ensure that all candidates' scripts for a single examination are marked within a reasonable time period. Since a marking team may comprise over a hundred examiners, training plays an essential role in ensuring that mark schemes are applied consistently, so that responses are marked to identical criteria.

The practices and procedures of the awarding bodies are regulated by Ofqual, the Office of the Qualifications and Examinations Regulator, who issue a code of practice and associated guidance (Ofqual, 2008). (Until recently, this was the responsibilty of the Qualifications and Curriculum Authority (QCA), who now focus on national curriculum development). Generally, newly recruited examiners take a subject-specific induction course to learn about relevant marking principles. Next they undertake training together with experienced examiners in their subject. This training often includes attending a standardisation (coordination) meeting prior to marking candidates' responses for a particular examination (usually after marking a small 'practice' sample of responses). The purpose of the meeting, led by a Principal Examiner[1], is to establish common standards of marking that are to be maintained throughout the marking period. During a typical meeting, examiners are briefed on the mark scheme, undertake some closely supervised marking, and discuss questions and candidates' responses with each other.

Personalised feedback is a further aspect of examiner training, and is usually given on marking undertaken soon after the standardisation meeting. Examiners submit some of their marked scripts (a 'standardisation' sample) to a Team Leader or other senior examiner who reviews the marking and provides written feedback on a structured form. This written feedback is supported with telephone and/or e-mail contact where necessary. If an examiner's marking of the standardisation sample is not sufficiently reliable, then he or she is required to provide a further sample for review, and will receive further feedback. An examiner can only go ahead and mark their allocation of candidates' responses once the senior examiner is confident that their marking will be valid and reliable.

The training procedures described above are the traditional GCSE and A-level approach which is widely used. However, some school examinations are now marked on screen, and sometimes this goes hand in hand with new training procedures. Although the GCSE and A-level training procedures outlined above may differ from those used in other assessment contexts, such as the marking of high stakes tests in the USA, they combine several features purported to benefit marking reliability. These include: feedback to individuals (Shaw, 2002; Greatorex and Bell, 2008); marking practice and experience; the generation and propagation of communities of practice (Baird, Greatorex and Bell, 2004; Wenger 1998); a common understanding of the mark scheme (Baird *et al*., 2004), which might also serve as a common reference point (Laming, 2004); and opportunities to boost confidence (Greatorex, Baird and Bell, 2002).

Efficacy of training procedures

The efficacy of training has been investigated widely, and while it is not possible to provide an exhaustive review of the literature here, we describe some of the most significant studies within the context of educational assessment. Unsurprisingly, broadly beneficial effects of various forms of training on inter-marker agreement have been reported in studies of diverse examinations, ranging from Key Stage 3 English tests in England to graduate business school admissions tests in the US (Shohamy, Gordon and Kraemer, 1992; Wigglesworth, 1993; Stahl and Lunz, 1996; Powers *et al*., 1998; Hoskens and Wilson, 2001; Elder, Knoch, Barkhuizen and von Randow, 2005; Royal-Dawson, 2005).

Several studies have focussed on some of the differential effects of training. In the context of examining English as a Second Language (ESL) in the US, Weigle (1998, 1999) investigated differences between experienced and inexperienced examiners. She found that prior to training, inexperienced examiners marked more severely than experienced examiners did. However, the effects of training ('norming sessions' – a form of standardisation meeting) included eliminating this group difference, as well as reducing the overall spread of examiner severity. The findings of Elder *et al*. (2005), who explored the writing component of a diagnostic English language needs assessment in New Zealand, are in line with those of Weigle (1998, 1999). Elder *et al*. (2005) found that following feedback on their marking (in the form of individualised statistical reports explicated at a group briefing session) inexperienced examiners were more likely to make changes to their marking than experienced examiners were.

Another notable study focussing on examiners' backgrounds is that of Shohamy, Gordon and Kraemer (1992). Working within the context of English for Speakers of Other Languages (ESOL) examinations, Shohamy *et al*. (1992) used a 2×2 design to compare four marker or 'rater' groups marking a writing task: two groups had an EFL (teaching) qualification whereas two did not, and two groups received training (broadly akin to a standardisation meeting) whereas two did not. It was found that:

> *Raters are capable of rating reliably, regardless of background and training, however, reliability [marker agreement] can be improved when raters receive intensive procedural training.* (p. 31)

1 In the 'live' marking of syllabuses with large candidatures, a Principal Examiner leads a group of Team Leaders, each of whom leads a team of Assistant Examiners.

Drawing together the findings of the above studies, it seems reasonable to conclude that at least in some cases, training can result in inexperienced examiners achieving a quality of marking akin to that of experienced examiners.

In another strand of research, Baird *et al*. (2004) investigated whether variations in the style of standardisation meetings affected examiner agreement; they found minimal differences in the marking of examiners in hierarchically-led and consensually-led meetings. The possibility of self-training has also been examined. Kenyon and Stansfield (1993) reported that in the USA, examiners trained themselves successfully in the holistic scoring of an oral proficiency test. However, the efficacy of this self-training depended considerably upon examiners' background characteristics, including familiarity with the assessment, motivation, and teaching experience.

In a recent empirical study, Greatorex and Bell (2008) explored the relative efficacies of three different examiner training procedures in the context of experimental AS-level biology marking. (AS-level examinations are usually taken after the first year of two-year A-level courses, but are also stand-alone qualifications.) The study involved a traditional standardisation meeting (as described previously), personal feedback using a standard form with telephone or e-mail support, and pre-written feedback from a Principal Examiner. There were four groups of experienced examiners in the study, and each group undertook a different combination of two of the three training procedures. (In professional or 'live' AS-level marking, each examiner receives two forms of training.) When the total marks awarded to whole scripts were analysed, it was found that no particular combination of procedures was significantly more beneficial than any other.

Overall, the relative merits of different training procedures as reported in the research literature are far from clear-cut. One possible explanation for this may lie in the level of detail of the analyses conducted to date. Arguably, accuracy measures that stem from comparisons of the total marks awarded to candidates by examiners are likely to conceal differences in the marks awarded to individual questions. Examination questions and their mark schemes are known to have varied structural and stylistic features, which contribute differently to the demands of the marking task and therefore to marking accuracy (Suto and Nádas, 2008b, *in press*). It is plausible that this occurs partly because questions are affected by training procedures differently. For example, for some questions, accuracy levels may benefit most from an oral discussion engendering clarifications of mark scheme ambiguities that affect the majority of examiners. For other questions, however, personalised feedback in the form of precisely written instructions relating to individual marking errors or highly unusual candidate responses may be more fruitful.

In Greatorex and Bell (2008) accuracy data were analysed at the *whole script* level. For the first study reported in this article, we re-analysed marking accuracy data from Greatorex and Bell (2008) at the *question* level. We also investigated potential relationships between the benefits of the three training procedures and the cognitive strategies needed to mark the questions (discussed below).

Cognition in marking

A major strand of recent research addresses the judgements that marking entails (Sanderson, 2001; Crisp, 2007; Suto and Greatorex, 2008a, b). However, it has yet to be related to training procedures. Thus far, there is evidence that for a variety of GCSE and A-level examinations, both experienced (with both teaching and marking experience) and inexperienced (with neither teaching nor marking experience) graduate examiners use five cognitive strategies to mark short and medium-length responses to questions (Greatorex and Suto, 2006; Greatorex, 2007; Suto and Greatorex, 2006, 2008a). The strategies have been named *matching*, *scanning*, *evaluating*, *scrutinising* and *no response* and are described fully by Suto and Greatorex (2008a). For brief descriptions, see Appendix 1.

Suto and Nádas (2008a) classified the five marking strategies according to the sophistication and depth of cognitive processing demanded, and in a study of experimental GCSE mathematics and physics marking, judged questions as falling into two categories:

- **apparently simple:** appears to require the use of only the matching and/or simple scanning marking strategies;
- **apparently more complex:** appears to require the use of more complex marking strategies such as evaluating, scrutinising, and complex scanning, in addition to, or instead of, simple strategies.

Experienced examiners (with both teaching and marking experience), and inexperienced graduate examiners (with neither teaching nor marking experience) participated in the study, which entailed question-by-question marking. They marked identical *pre-training* samples of candidates' responses to selections of GCSE questions, received training in the form of a single standardisation meeting led by a Principal Examiner, then marked identical *post-training* response samples. An analysis of post-training marking accuracy revealed very few differences between experienced and inexperienced markers. However, all examiners marked *apparently simple* questions more accurately than they marked *apparently more complex* questions.

While Suto and Nádas (2008a) addressed important questions surrounding post-training accuracy, they did not explore the process by which it was achieved. Pre-training accuracy was not considered, and the effects of the training on the two examiner groups may have been different. From the literature reviewed earlier (Elder *et al*., 2005; Weigle, 1998, 1999), we hypothesise that inexperienced examiners benefited more from the training than did experienced examiners. Moreover, it can be hypothesised that in the studies of both Suto and Nádas (2008a) and Greatorex and Bell (2008), training was more beneficial for the marking of *apparently more complex* strategy questions than for *apparently simple* strategy questions. If this were indeed the case, then there may be implications for the focussing and emphasis of training procedures. For instance, perhaps training of all examiners should emphasise the marking of *apparently more complex* strategy questions. For the second study in this article, we re-analysed data from Suto and Nádas (2008a), in order to test the above hypotheses.

Study 1

Many of the following method details are available in Greatorex and Bell (2008). However, the exceptions are the information about coding questions according to the complexity of the cognitive marking strategies apparently needed, as well as the analysis and results of question level marking accuracy.

Examination paper

A question paper from a mainstream biology AS-level syllabus, administered by Oxford, Cambridge and RSA examinations (OCR) in

2005, was selected for use in the study. It entailed a traditional points-based mark scheme and candidates' scripts comprised individual booklets containing subdivided questions with answer spaces either beneath each printed question part or very nearby. The paper was one of four assessments needed to obtain this particular AS-level qualification.

The paper was to be marked on a script-by-script basis rather than assigning different questions to different examiners. For each question in it, the complexity of the cognitive marking strategies apparently needed was considered: two researchers independently studied each question and its accompanying mark scheme and coded it as either *apparently simple* ('appears to require the use of only the matching and/or simple scanning marking strategies') or *apparently more complex* ('appears to require the use of more complex marking strategies such as evaluating, scrutinising, and complex scanning, in addition to or instead of simple strategies'). (For a full discussion of GCSE examination marking strategies, see Suto and Greatorex, 2008a.) The coding was undertaken with reference to a small number of scripts, the question paper and the mark scheme, but no statistics. There was agreement between the researchers on over 90% of codes, but where disagreements arose, they were discussed and resolved. The paper was judged to comprise 5 *apparently simple* strategy questions and 13 *apparently more complex* strategy questions.

Script samples

A limited number of candidates' scripts were made available by OCR for use in the study. From these scripts, four samples were drawn:

- Sample 1 (23 scripts): used to obtain a pre-training measure of accuracy for each marker.
- Sample T (10 scripts): used in training.
- Sample 2 (10 scripts): marked in between two training procedures.
- Sample 3 (23 scripts): used to obtain a post-training measure of accuracy for each marker.

Samples 1 (*pre-training*) and 3 (*post-training*) were matched samples, selected by the researchers to cover a majority of the available mark range and drawn from a variety of candidate centres. The scripts in these samples were checked by the acting PE (see 'Participants' section) to ensure that they were not atypical. Script samples T and 2 were selected by the acting PE. All scripts were photocopied, and marks and annotations were removed from the copies. Multiple copies of these 'cleaned' scripts were then made.

Participants

As the Principal Examiner for the professional or 'live' marking of the examination paper (the 'live PE') was unable to take a major role in the study, a Team Leader from the live marking was recruited to lead the experimental marking (the 'acting PE'). The acting PE led a total of 29 paid participants, all of whom were experienced examiners. (An 'experienced marker' was defined as someone who had marked AS Biology examinations from the specification under consideration, but not the particular examination paper used in the study). The examiners were assigned to experimental groups 1 to 4, each of which comprised at least six examiners.

Procedure

Initially the acting PE marked all scripts, and some of her marking was checked by the live PE. As the acting PE's marking was deemed acceptable by the live PE, the acting PE's marks were used as reference marks in the study.

All other examiners marked script sample 1. Each experimental group then underwent two of the following three training procedures, interspersed with the marking of sample 2:

1. *Standardisation meeting*, in which script sample T was available for use.
2. *Personal feedback*, as described above.
3. *Pre-written feedback*, which is not a form of training currently used in live examining practices in England and Wales. It is similar to a type of training that has been included in previous studies (Shaw, 2002). After marking some scripts (sample A), the examiner received a copy of the same scripts marked by the acting PE accompanied by some notes (also from the acting PE) explaining why the marks had been credited to the candidate. The examiner was asked to check whether his or her marking was sufficiently close to that of the acting PE, and if not, then to take this information into account in subsequent marking.

The standardisation meeting and the personal feedback were as similar as possible to the training undertaken in usual live examining practices in England and Wales, but within the confines of the research setting.

Sample 3 was marked by all examiners once all training had taken place. The combinations of training procedures experienced by the four experimental groups are given in Table 1.

Table 1: Summary of procedures experienced by experimental groups 1 to 4

Experimental group of examiners	*Pre-training marking (sample 1)*	*First training session*		*Further marking (sample 2)*	*Second training session*		*Post -training marking (sample 3)*
		Standardisation meeting *(sample T available)*	Pre-written feedback on *marking of sample T*		Personal feedback *on marking of sample 2*	Pre-written feedback *on marking of sample 2*	
1	✓	✓	✗	✓	✓	✗	✓
2	✓	✓	✗	✓	✗	✓	✓
3	✓	✗	✓	✓	✓	✗	✓
4	✓	✗	✓	✓	✗	✓	✓

Notes: The marking and training experience of group 1 was most similar to current examining practices. The sequence of events in the study reads from left to right, and each experimental group is represented by one row. For example, examiners in Group 3 marked sample 1 then sample T. They then received pre-written feedback on their sample T marking. Next, they marked sample 2 and received personal feedback on that marking. Finally, they marked sample 3.

Table 2: Mean P_0 (and s.d.) values for the four experimental groups pre- and post- training (i.e. on the first and third candidate response samples)

Experimental group	*Pre-training (sample 1)*			*Post-training (sample 3)*		
	All questions	Apparently simple *strategy questions*	Apparently more complex *strategy*	*All questions questions*	Apparently simple *strategy questions*	Apparently more complex *strategy questions*
1	0.74 (0.02)	0.92 (0.02)	0.67 (0.02)	0.80 (0.02)	0.98 (0.01)	0.73 (0.02)
2	0.74 (0.02)	0.94 (0.01)	0.66 (0.02)	0.79 (0.01)	0.96 (0.01)	0.72 (0.02)
3	0.74 (0.02)	0.93 (0.02)	0.66 (0.02)	0.79 (0.01)	0.97 (0.01)	0.72 (0.01)
4	0.74 (0.02)	0.94 (0.01)	0.65 (0.02)	0.77 (0.01)	0.96 (0.01)	0.70 (0.02)

Samples 1, 2 and 3 were identical for all examiners, and overall, examiners were given just over 4 weeks to complete their marking and training (including time for the post).

Analysis and results

The marking data were analysed to yield P_0 values for each examiner on each question for the pre- and post-training samples. P_0 is the proportion of exact agreement between a marker and the PE; values range from 0 to 1, and the measure indicates how frequently a marker differs from the PE in his or her marking. (See Bramley, 2007, for a full discussion of some common accuracy measures.) Mean P_0 values are displayed in Table 2 above, which indicates that questions of all types were marked more accurately after training than beforehand. Table 2 also indicates that, in line with previous findings (Suto and Nádas, 2008a), apparently simple strategy questions were generally marked more accurately than apparently more complex strategy questions were, on both the pre-training and the post-training samples.

Wilcoxon tests comparing accuracy on all questions revealed that the improvement in P_0 from the pre-training to post-training condition was significant for all four experimental groups ($Z = 1.65$, $p < 0.001$ for group 1; $Z = 1.78$, $p < 0.001$ for group 2; $Z = 1.85$, $p < 0.001$ for group 3; and $Z = 1.48$, $p < 0.05$ for group 4). Therefore, all four combinations of training procedures were beneficial for marking accuracy.

To investigate the *relative* benefits of the training procedures, changes in accuracy for each examiner on each question were calculated for use as the dependent variable in a Kruskal-Wallis test (the non-parametric equivalent of one-way ANOVA with independent measures). This analysis revealed no significant effect of experimental group ($X^2 = 1.64$, d. f. = 3, $p > 0.05$), indicating that no one combination of training procedures was more beneficial than any other. To confirm that the analysis had not masked any differential effects of individual training procedures, Mann-Whitney U tests were conducted with all combinations of pairs of experimental groups. Again, no significant differences in change in accuracy were found; this suggests that the three types of training procedures in the study were all equally effective in improving accuracy.

The relative benefits of training on *apparently simple* strategy questions and *apparently more complex* strategy questions were also explored. For each experimental group, a Mann-Whitney U-test was conducted to investigate possible differences in accuracy changes for the two question types. However, these tests revealed no significant differences between *apparently simple* strategy questions and *apparently more complex* strategy questions ($Z = -0.69$, $p > 0.05$ for group 1; $Z = -1.40$, $p > 0.05$ for group 2; $Z = -0.68$, $p > 0.05$ for group 3 and $Z = -0.09$, $p > 0.05$ for group 4). This indicates that the training procedures in the study were equally beneficial for the two question types.

Although marking strategy complexity was found not to be related to how beneficial training was, a Kruskal-Wallis test was conducted to analyse differences in the effects of training among individual questions. A significant main effect was found ($X^2 = 124.96$, d.f. = 17, $p < 0.001$), indicating that training had different effects on different questions, as illustrated in Figure 1. For example, training greatly improved accuracy on question 5, whereas on question 14, accuracy levels either remained constant or decreased after training. Overall, it appears that for AS-level biology, question features other than those that contribute to marking strategy complexity must therefore play a role in determining how beneficial training will be.

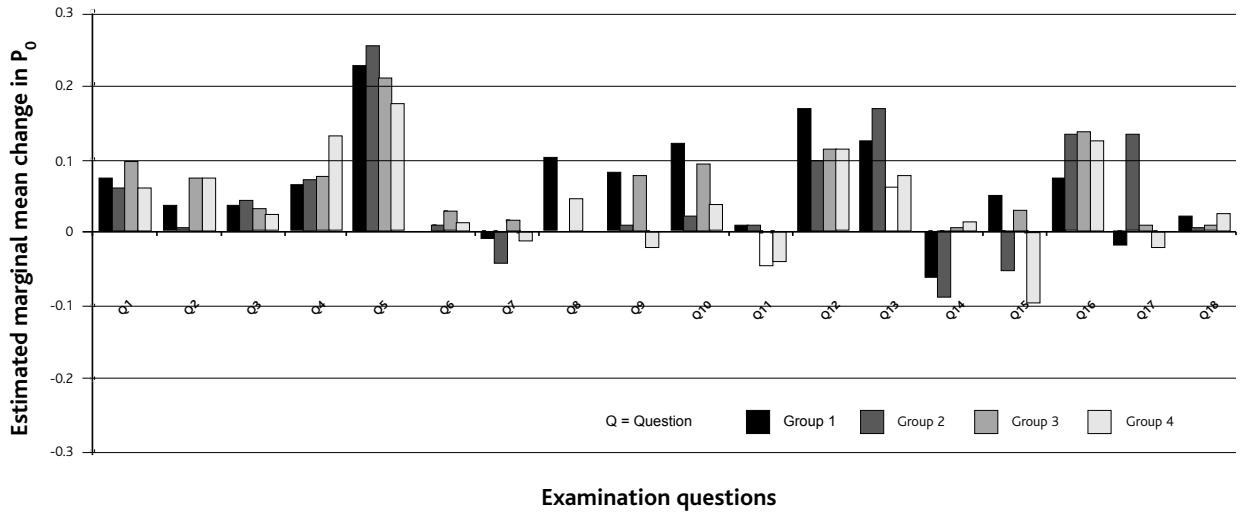


Figure 1: Graph showing changes in accuracy after training for individual AS-level biology questions

Study 2

Examination questions

Questions were selected from end-of-course examination papers from mainstream mathematics and physics syllabuses, administered by OCR in 2005. All entailed points-based mark schemes and candidates' scripts comprised individual booklets containing subdivided questions with answer spaces beneath each question part. For each subject, the question selection was intended to cover the full range of difficulties for candidates (grades A* to D) and be approximately equivalent to one examination paper in length and in the total marks available.

As with Study 1, the complexity of the cognitive marking strategies apparently needed to mark each question was also considered: two researchers independently studied each question and its accompanying mark scheme and coded it as either *apparently simple* ('appears to require the use of only the matching and/or simple scanning marking strategies') or *apparently more complex* ('appears to require the use of more complex marking strategies such as evaluating, scrutinising, and complex scanning, in addition to or instead of simple strategies'). There was agreement between the researchers on over 90% of codes, but where disagreements arose, they were discussed and resolved. The maths question selection comprised 7 *apparently simple* strategy questions and 13 *apparently more complex* strategy questions. The physics selection comprised 4 *apparently simple* strategy questions and 9 *apparently more complex* strategy questions.

Response samples

For both subjects, stratified sampling methods were used to draw two representative samples of candidates' responses to the selected questions: the *pre-training* sample comprised 15 different responses to each question and was to be marked before training (a standardisation meeting); and the *post-training* sample comprised 50 responses to each question and was to be marked after training. The selected responses were photocopied, 'cleaned' of all previous marks and annotations, copied again, and collated into identical response samples, to be marked on a question-by-question basis. This arrangement ensured that each examiner would be able to mark exactly the same candidates' responses.

Participants

For each subject, a highly experienced PE (who had been the PE in the live marking of at least half of the questions) led the marking of twelve examiners: six 'experts' had experience of GCSE teaching and first-hand professional experience of marking at least one tier of the selected examination paper; six 'graduates' had a relevant Bachelor's degree but neither professional marking experience nor teaching experience.

Procedure

The procedure was the same for each subject. Initially, the PE marked all of the selected candidate responses; these marks were to be used as reference marks in the subsequent analysis. All other examiners then marked the *pre-training* sample of 15 responses. Training then took the form of a single standardisation meeting for all examiners in the subject, which lasted 5–6 hours and was led by the PE. Each question was discussed in turn, and issues and difficulties arising on the *pre-training* sample were addressed. The examiners then marked the *post-training* sample of 50 responses.

Analysis and results

The marking data were analysed to yield P_0 values for each examiner on each question for each sample. Mean P_0 values are displayed in Table 3.

Table 3 indicates that maths marking was generally more accurate than physics marking, that *apparently simple* strategy questions were generally marked more accurately than *apparently more complex* strategy questions, and that, after training, there were very few differences in marking accuracy between expert and graduate examiners. These findings are considered in depth elsewhere (Suto and Nádas, 2008a). What is of most interest in the present article however, are the *changes* that occurred in marking accuracies before and after training. These changes were explored using ANOVA. For each subject, two full-factorial models were constructed:

- Model 1 explored the effects of examiner type and individual questions on change in accuracy after training,
- Model 2 explored the effects of examiner type and apparent marking strategy complexity on change in accuracy after training.

For maths, Model 1 revealed significant main effects of both examiner type ($F(1) = 14.25$, $p < 0.001$) and individual question ($F(19) = 7.13$, $p < 0.001$) on change in accuracy. There was no interaction between examiner type and individual question. These findings indicate that training affected experts and graduates differently, and affected accuracy on individual questions differently. When Model 2 was run, it again revealed a significant main effect of examiner type on change in

Table 3: Mean P_0 (and s.d.) values for maths and physics examiners pre- and post- training (i.e. on the practice and main response samples)

	Pre-training			*Post-training*		
	All questions	Apparently simple *strategy questions*	Apparently more complex *strategy questions*	*All questions*	Apparently simple *strategy questions*	Apparently more complex *strategy questions*
All maths markers	0.87 (0.13)	0.93 (0.07)	0.83 (0.14)	0.89 (0.11)	0.92 (0.10)	0.87 (0.10)
Maths experts	0.90 (0.11)	0.95 (0.06)	0.88 (0.12)	0.89 (0.10)	0.93 (0.04)	0.87 (0.11)
Maths graduates	0.84 (0.15)	0.92 (0.08)	0.79 (0.15)	0.88 (0.11)	0.91 (0.14)	0.87 (0.10)
All physics markers	0.80 (0.19)	0.98 (0.04)	0.71 (0.17)	0.84 (0.16)	0.99 (0.02)	0.78 (0.14)
Physics experts	0.83 (0.17)	1.00 (0.01)	0.76 (0.15)	0.85 (0.16)	0.99 (0.02)	0.79 (0.15)
Physics graduates	0.76 (0.20)	0.96 (0.06)	0.67 (0.17)	0.84 (0.16)	0.99 (0.03)	0.77 (0.14)

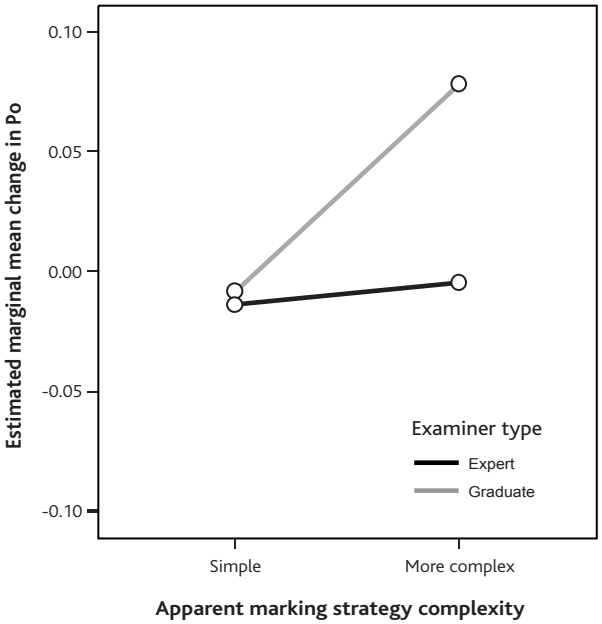


Figure 2: Graph showing estimated marginal mean changes in P_o values for expert and graduate maths examiners for questions with different apparent marking strategy complexities.

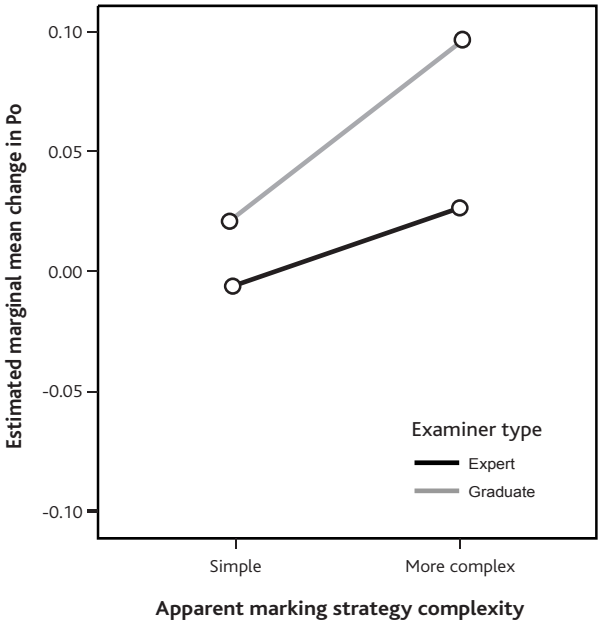


Figure 3: Graph showing estimated marginal mean changes in P_o values for expert and graduate physics examiners for questions with different apparent marking strategy complexities.

accuracy ($F(1) = 5.45$, $p < 0.05$), and also indicated a significant main effect of apparent marking strategy on change in accuracy ($F(1) = 6.53$, $p < 0.05$). Again, there was no interaction between examiner type and apparent marking strategy complexity. Figure 2 illustrates these results.

As Figure 2 shows, for questions requiring *apparently simple* marking strategies, the training appears in general to have had little effect on either experts or graduates on their marking accuracy. That is, the frequency with which maths examiners agreed with their PE decreased very slightly. For questions requiring *apparently more complex* marking strategies, however, there was a sizeable improvement in accuracy for graduates but not for experts.

For physics, Model 1 revealed a significant main effect of examiner type on change in accuracy ($F(1) = 12.92$, $p < 0.001$). There was also a significant main effect of individual question on change in accuracy ($F(12) = 9.40$, $p < 0.001$). In contrast with maths, there was a significant interaction between examiner type and individual question on change in accuracy ($F(1,12) = 2.22$, $p < 0.05$). These findings indicate that: (i) the training affected experts and graduates differently; (ii) training affected accuracy on individual questions differently; and (iii) experts and graduates were affected differently on different questions.

When Model 2 was run for physics, there was a significant main effect of examiner type on change in accuracy ($F(1) = 4.82$, $p < 0.05$), and there was a significant main effect of apparent marking strategy on change in accuracy. There were no significant interactions between examiner type and apparent marking strategy complexity. Figure 3 illustrates these results.

Figure 3 shows that, for questions requiring *apparently simple* marking strategies, the training appears to have had little effect on expert examiners' P_0 values. For graduate examiners, however, it appears to have improved marking accuracy slightly: that is, the frequency with which physics graduates agreed with their PE increased slightly, and more so than with the maths graduates (Figure 2). For questions requiring *apparently more complex* marking strategies, there was a sizeable improvement in accuracy for physics graduates (even more than there was for maths graduates) and a small improvement for physics experts (again, more so than for maths experts). A comparison of Figures 2 and 3 would suggest that overall, the physics training improved the frequency of physics examiners' agreement with their PE more than the maths standardisation meeting improved the frequency of the maths examiners' agreement with their PE.

General discussion

In this article we presented further analyses of data from two recent empirical studies in which we explored possible relationships between the complexity of the cognitive marking strategies apparently needed to mark some AS-level and GCSE examination questions and the efficacy of some examiner training procedures. In both studies, it was found that: (i) marking accuracy was better after training than beforehand; and (ii) the effect of training on change in marking accuracy varied across all individual questions. Our hypothesis that training would be more beneficial for *apparently more complex* strategy questions than for *apparently simple* strategy questions was upheld for both subjects in Study 2, but not in Study 1. (However, as in Study 2, levels of marking accuracy per se were lower for *more complex* strategy questions in both subjects in Study 1.) The hypothesis that graduates would benefit more from training than expert examiners would, was supported in both subjects in Study 2.

Limitations

Our research had a number of limitations. First, the original studies had different aims to those of the analyses reported here, which did not warrant the inclusion of control groups receiving no training. Consequently, it is somewhat difficult to disentangle the effect of the training from the practice effect or any fluctuations in examiner accuracy over time. Whilst this might appear to be a limitation in both studies, general psychological research in many areas suggests that feedback

leads to more accurate judgements (Laming, 2004), and there is no clear reason for expecting marking to be an exception. Research by Awarding Bodies is somewhat constrained by the availability of resources and operational concerns. Arguably, it is more important for an Awarding Body to know which training is the most effective for which types of questions, than to know whether a particular type of training is better than no training, hence the lack of control groups.

Secondly, the studies represent a limited number of school disciplines, a non-exhaustive set of question or mark scheme characteristics, and have a limited number of participants and scripts in comparison with the live marking of some examinations. Despite these points, the studies are as similar to operational practice as it was possible to arrange within the constraints of an empirical setting.

Thirdly, we did not control the standardisation meetings and the feedback to examiners to ensure that the PEs put the same amount of effort into training examiners on each and every question. However, such controls might have resulted in communications between PEs and the examiners which were not necessarily geared towards the needs of the examiners, and as such would have low ecological validity. For instance, it could have been decided that an equal amount of time would be spent discussing each individual question in the standardisation meeting. This would guard against questions that received extensive attention in the standardisation meeting having larger changes in the accuracy of marking than questions which received less attention. However, such an experimental control might have resulted in time-wasting (explaining how to mark a question(s) not genuinely warranting much explanation).

Implications

Nevertheless, our findings have some important implications. First, the finding that the conventional training provided by a standardisation meeting and personal feedback is as effective as the alternatives trialled, confirms that current practice is sound, and is in line with the earlier findings of Greatorex and Bell (2008).

The finding that training is more effective for graduate examiners than for expert examiners is in line with the findings of Weigle (1998, 1999) and Elder *et al*. (2005), who found that inexperienced examiners benefited more from training than did experienced examiners. It indicates a need for more intensive training for graduate examiners, and Awarding Bodies need to be mindful of this finding if numbers of graduate examiners were to be increased. The expert examiners in Study 1 had not marked the examination under consideration before, and the expert examiners in Study 2 were new to approximately half the questions under consideration, yet we found that experts marked all questions accurately, even prior to training. It is possibly the case that less intensive training than is currently provided is sufficient for expert examiners. Our findings also raise the question of whether more effort should be put into retaining accurate expert examiners and using their skills as much as possible, or into ploughing resources into recruiting many new graduate examiners who might need more intensive and possibly more expensive training than the expert examiners. Clearly, comprehensive cost-benefit analyses may need to be undertaken.

Whilst training is more effective for graduates than for expert examiners, this does not mean that training is an irrelevant process for experts. It could be that training provides opportunities for experts to share their knowledge and thereby contribute to the improvements in graduates' marking accuracy. However, there are many other factors which could have facilitated changes in graduate examiners' accuracy. It can also be argued that training is valuable because it gives the expert examiners the confidence to mark. The latter is a view proposed by Greatorex *et al*. (2002).

As mentioned above, the expert examiners in Study 1 had not marked the examination under consideration before, and the expert examiners in Study 2 were new to approximately half the questions under consideration, yet we found that experts marked all questions fairly accurately, even prior to training. This finding is similar to that of Baird *et al*. (2004), who found that experienced examiners' marking was at a similar level of agreement, whether they had participated in a standardisation meeting or not. Perhaps then, expert examiners have more transferable skills within their subject domains than we have thus far anticipated. That is, at present expert examiners receive training on how to mark all of their questions, but training might only be necessary for some of these questions. However, if a 'partial' training approach were to be adopted, then it would be essential that this approach include checks on marking accuracy for *all* questions to be marked (as in current practice). The issue of the transferability of expert examiners' skills is the focus of research in progress.

The classification of questions into the categories *apparently simple* and *apparently more complex* marking strategies can sometimes account for differences in change in marking accuracy, as exemplified by Study 2. However, this was not found to be the case in Study 1. It appears that some other additional features of examination questions, and/or the candidates' answers are affecting changes in accuracy. Further research in this area is currently underway, following on from a recent study of question features associated with accuracy levels per se (Suto and Nádas, 2008b, *in press*). Additionally, our findings draw attention to the issue of how PEs and examiners decide which questions to spend most time and discussion on during meetings, personal feedback or other forms of training. This might be a source of the variation of change in marking accuracy that has yet to be investigated.

In summary, we found that the current training practices are as effective as the alternatives which we tested; training was sometimes more beneficial for questions which required apparently more complex rather than simple marking strategies; and graduates benefited more from training than expert examiners did.

References

Baird, J-A., Greatorex, J. & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education*, **11**, 3, 333–347.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication*, **4**, 22–28.

Broadfoot, P.M. (1996). *Education, assessment and society*. Buckingham: Open University Press.

Crisp, V. (2007). Researching the judgement processes involved in A-level marking. *Research Matters: A Cambridge Assessment Publication*, **4**, 13–17.

Eckstein, M.A. & Noah, A.J. (1993). *Secondary school examinations: International perspectives on policies and practice*. New Haven: Yale University.

Elder, C., Knoch, U., Barkhuizen, G. & Von Randow, J. (2005). Individual Feedback to Enhance Rater Training: Does it Work? *Language Assessment Quarterly*, **2**, 3, 175–196.

Greatorex, J. (2007). Did examiners' marking strategies change as they marked more scripts? *Research Matters: A Cambridge Assessment Publication*, **4**, 6–12.

Greatorex, J., Baird, J. & Bell, J.F. (2002, August) '*Tools for the trade': What makes GCSE marking reliable?* Paper presented at the conference Learning

communities and Assessment Cultures: connecting Research and Practice. The conference was jointly organised by the EARLI Special Interest group on assessment and Evaluation and the University of Northumbria.

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

Greatorex, J. & Suto, W.M.I. (2006, May). *An empirical exploration of human judgement in the marking of school examinations*. Paper presented at the 32nd annual conference of the International Association for Educational Assessment, Singapore.

Hamilton, J., Reddel. S. & Spratt, M. (2001). Teachers' perceptions of on-line examiner training and monitoring. *System*, **29**, 4, 505–520.

Hoskens, M. & Wilson, M. (2001). Real-Time Feedback on Rater Drift in Constructed-response Items: An Example from the Golden State Examination. *Journal of Educational Measurement*, **38**, 2, 121–145.

Kenyon, D. & Stansfield, C.W. (1993, August). *Evaluating the efficacy of examiner self-training*. Paper presented at the 15th annual Language testing research colloqium, University of Cambridge.

Knoch U., Read J. & von Randow, J. (2007). Re-training writing examiners online: How does it compare with face-to-face training? *Assessing Writing*, **12**, 1, 26–43.

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

Lamprianou, I. (2008). Comparability of examination standards between subjects: an international perspective. *Oxford Review of Education*.

Leacock, C. & Chodorow, M. (2003). C-examiner: Automated Scoring of Short-Answer Questions. *Computers and Humanities*, **37**, 4, 389–405.

Meadows, M. & Wheadon, C. (2007, September). *Selecting the conscientious marker – a study of marking reliability in GCSE*. Paper presented at the meeting of the International Association of Educational Assessment, Baku, Azerbaijan.

National Assessment Agency (2008). http://www.naa.org.uk/

Office of the Qualifications and Examinations Regulator (Ofqual) (2008). http://www.ofqual.gov.uk/

Powers, D., Kubota, M., Bentley, J., Farnum, M., Swartz, R. & Willard, A. E. (1998). Qualifying Essay Readers for an On-line Scoring Network (ETS RM -98 -20), Princeton, NJ , Educational Testing Service. In: Y. Zhang, D. E. Powers, W. Wright & R. Morgan (Eds.), (2003), *Applying the On-line Scoring Network (OSN) to Advanced Placement Program (AP) Tests*. (RR-03-12) Princeton, NJ: Educational Testing Service.

Qualifications and Curriculum Authority (2007). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice 2006/7*. London: Qualifications and Curriculum Authority.

Raikes, N. & Harding, R. (2003). The Horseless Carriage Stage: replacing conventional measures. *Assessment in Education, Principles, Policies and Practices*, **10**, 3, 267–277.

Raikes, N., Greatorex, J. & Shaw, S. (2004, June). *From paper to screen: some issues on the way*. Paper presented at the meeting of the International Association of Educational Assessment, Philadelphia, USA.

Raikes, N. & Massey, A. (2007). Item-level examiner agreement. *Research Matters: A Cambridge Assessment Publication*, **4**, 34–37.

Royal-Dawson, L. (2005). *Is Teaching Experience a Necessary Condition for Markers of Key Stage 3 English?* Assessment and Qualifications Alliance report, commissioned by the Qualification and Curriculum Authority.

Sanderson, P. J. (2001). *Language and Differentiation in Examining at A level*. Unpublished doctoral dissertation, University of Leeds.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes*, **8**, 13–17.

Shohamy, E., Gordon, C.M., & Kraemer, R. (1992). The Effects of Examiners' Background and Training on the Reliability of Direct Writing Tests. *The Modern Language Journal*, **76**, 27–33.

Stahl, J.A. & Lunz, M.E. (1996). Judge Performance Reports: Media and Message. In: J.R. Engelhard & M. Wilson (Eds.), *Objective Measurement. Theory into practice*. 113–116. Norwood, NJ: Ablex Publishing.

Sturman, L. & Kispal, A. (2003). *To e or not to e? A comparison of electronic marking and paper-based marking*. Paper presented at the 29th meeting of the International Association for Educational Assessment, Manchester, UK.

Sukkarieh, J. Z., Pulman, S. G. & Raikes, N. (2005). Automatic marking of short free text responses. *Research Matters: A Cambridge Assessment Publication*, **1**, 19–22.

Suto, W.M.I. & Greatorex, J. (2006). A cognitive psychological exploration of the GCSE marking process. *Research Matters: A Cambridge Assessment Publication*, **2**, 7–10.

Suto, W.M.I. & Greatorex, J. (2008a). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal*, **34**, 2, 213–233.

Suto, W.M.I & Greatorex, J. (2008b). A quantitative analysis of cognitive strategy usage in the marking of two GCSE examinations. *Assessment in Education: Principles, Policies and Practices*, **15**, 1, 73–90.

Suto, W.M.I. & Nádas, R. (2007). The 'Marking Expertise' projects: Empirical investigations of some popular assumptions. *Research Matters: A Cambridge Assessment Publication*, **4**, 2–5.

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

Suto, W.M. I. & Nádas, R. (2008b, *in press*). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education*.

Weigle, S.C. (1998). Using FACETS to Model Rater Training Effects. *Language Testing*, **15**, 2, 263–287.

Weigle, S.C. (1999). Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing*, **6**, 2, 145–178.

Wenger, E. (1998). *Communities of Practice learning, meaning and identity*. Cambridge: Cambridge University Press.

Whetton, C. & Newton, P. (2002, September). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong SAR, China.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving examiner consistency in assessing oral interaction. *Language Testing*, **10**, 3, 305–335.

APPENDIX 1: SUMMARY OF THE MARKING STRATEGIES IDENTIFIED BY SUTO AND GREATOREX (2008A AND B)

Strategy	*Usage and description*	*Complexity of judgement processes entailed**
Matching	When the response to a question is a visually recognisable pattern, for example, a letter, word, number, part of a diagram, the examiner looks at a fixed part of the answer space and contrasts the candidate's response with the right answer, making a judgement about whether they match.	Simple
Scanning	When an examiner scans the whole of the answer space, in order to discover whether a specific detail in the mark scheme is there or not. When the detail is simple (for example, a single number or letter), pattern recognition takes place. When the detail needs additional meaningful or semantic processing, for example, a stage of mathematical working, a supplementary marking strategy may also be utilised.	Both simple and complex depending on the complexity of the detail to be scanned for

Appendix 1 – *continued*

Strategy	*Usage and description*	*Complexity of judgement processes entailed**
Evaluating	When an examiner attends to either all or part of the answer space and must process the content semantically, considering the candidate's response for structure, clarity, and logic or other features the mark scheme deems creditworthy.	Complex
Scrutinising	Only when a candidate's response is unanticipated or wrong. The examiner endeavours to spot the route of the error, and whether a valid substitute to the mark scheme solution has been given. During the process, the examiner considers various aspects of the candidate's answer with the intention of recreating what the candidate was attempting. The examiner may have to deal with a lot of uncertainty and re-read the response several times.	Complex
No response	When there is nothing in the answer space, the examiner checks the answer space a couple of times to confirm there is no answer and then awards 0 marks.	Simple

*Note: when interpreted within the context of dual-processing theories of judgement, 'simple' strategies entail *System 1* (intuitive) judgements, whereas 'complex' strategies entail *System 2* (reflective) judgements.

Capturing expert judgement in grading: an examiner's perspective

Peter King, Cambridge Examiner, **Dr Nadežda Novaković and Dr Irenka Suto** Research Division

Introduction to the study

There exist several methods of capturing expert judgement which have been used, or could potentially be used, in the process of determining grade boundaries for examinations. In a recent study conducted within Cambridge Assessment's Research Division, we sought to explore the judgements entailed in three such methods: (i) rank ordering, (ii) traditional awarding, and (iii) Thurstone pairs. Rank ordering requires judges to make relative holistic judgements about each of a series of up to ten scripts, in order to place them in order of overall quality (Black and Bramley, 2008, Gill *et al*., 2007). Traditional awarding, which is England's current principal grading method (QCA, 2008), utilises limen referencing (Christie and Forrest, 1981; French *et al*., 1988; Greatorex, 2003). Recommendations for grade boundaries are made by a committee of senior examiners based upon absolute judgements of whether selected scripts are worthy or unworthy of particular grades. Finally, like rank ordering, the Thurstone pairs method (Thurstone, 1927a, b) requires judges to make relative holistic judgements about scripts. However, judgements are comparisons of pairs of scripts, rather than rankings of larger series of scripts.

The study was conducted in the context of two contrasting examinations from AS level biology and GCSE English. A key aim was to identify the features of candidates' scripts that affect the judgements made in each of the three methods. To achieve this, sixty experienced examiners were invited to participate in the study (thirty for each subject). Each examiner made judgements about overall script quality, using each method on a different batch of scripts. Additionally, each examiner completed a research task in which he or she was asked to rate a fourth batch of scripts for a series of features, using rating scales devised by the researchers. Subsequent data analysis entailed relating the judgemental data on script quality to the script feature data.

Obtaining an examiner's perspective

Immediately after taking part in the study, one examiner recorded and offered the Research Division his views and experiences of participation. His perspective is the focus of this article. While researchers have many opportunities to report their views, the first-hand experiences of research participants generally receive much less attention, yet perspectives of this nature can be immensely valuable. On some occasions, they can be used to triangulate research findings or provide greater depth and explanation of phenomena. At other times they may prove valuable in informing the design and direction of future research. Furthermore, recruitment of these crucial volunteers and their colleagues for further studies may depend upon research being perceived as meaningful and valid, and affecting policy and practice positively.

The examiner is one of Cambridge Assessment's most experienced examiners. He became an English teacher in 1957 and was appointed a Cambridge examiner for O-levels two years later. Over the past fifty years, he has also been involved in GCSE marking, the moderation of coursework, and the training of examiners, amongst other assessment activities. He has retired as Head of English at a comprehensive school in England, and wrote the following account of his participation as a judge in the study.

A first-hand account of participation

"Just as we are currently asking searching questions about our public examination system, so questions are now being asked about the best methods of assessing candidates' work. This may stem from a variety of reasons: the need to make assessment as economically viable as possible; awareness through research projects that there are valid alternatives to traditional marking and awarding; technological changes that make a reality of reliable on-screen assessment.

These are thoughts that were inspired by my recent involvement in one such project by the Research Division of Cambridge Assessment. The work was carried out entirely at home rather than at an award meeting in Cambridge. It involved four batches each of about twenty scripts from OCR English GCSE Unit 1900, Paper 2431/2 (Non-fiction, Media and Information) for the 2006 and 2007 summer examinations. Each batch required a different approach:

- Rank ordering of Batch 1 scripts.
- Traditional awarding exercise for Batch 2.
- Thurstone pairs (paired comparisons) for Batch 3.
- Rating scripts for individual features for Batch 4.

Such an all-embracing exercise proved thought-provoking, leading me to ask some searching questions after years of traditional assessing of English examination scripts. As someone whose experience included moderation of folders of coursework, where rank order is sacrosanct, Batch 1 posed few problems of placing the scripts in what I considered the correct descending order after reading but not re-marking. Whilst it brought home again the importance of comparison and discrimination between scripts, it seemed to have little advantage over the traditional assessment required for Batch 2 where it is essential to assess each script in relation to specified criteria, with clear descriptors for each band or level in which they are to be placed. The latter approach, however, is extremely time-consuming, requiring an initial close scrutiny of the mark scheme before one feels that one has a complete grasp of its complexity. It is also an approach where the ability to make concise, apt comments (based on the criteria) at the end of each task is at a premium. It should, however, be a highly reliable method of assessment, provided examiners put in this groundwork and don't try to work too quickly – something not easy to guarantee, especially where such work is done in the evening or at the weekend after a highly demanding day or week as a full-time teacher or lecturer. It is a distinct advantage to be retired!

The Batch 2 traditional approach highlighted another possible problem with the holistic approach required of Batch 1 (where the script is not re-marked but considered in its entirety). Holistic approaches still require complete familiarity with a complex mark scheme, something not easily acquired for Batch 2 assessment, before one can have complete confidence in one's judgement.

Thurstone Pairs was a new and attractive approach for me but poses the same problems as suggested for the first holistic exercise. Where it was of particular value was that it involved comparisons between scripts from 2006/2007, with valuable cross-checking of whether standards are comparable year on year. The scripts were cunningly paired, often involving reading the script a second time and comparing it with another new script. I suspect it has more advantages with extended writing papers/exercises than with my paper where different tasks require different approaches and criteria. I can see how it could act as a quick, valuable cross-check of standards where scripts have first been traditionally assessed.

Unlike the three methods requiring judgements about overall script quality, the research task of rating a fourth batch of scripts for a series of features (from mechanical aspects such as spelling or handwriting to questions of relevance, the length of the response or the degree of sophistication or understanding or coherence in the writing) proved to be a slightly unsatisfactory exercise. It was generally not as demanding, failing to involve one fully and leaving one wondering whether one had really done justice to the script by such a fragmentary approach. It was a salutary reminder of how assessors often fail to do justice to a piece of work when they focus on particular features rather than the overall quality.

I concluded it is good to be made to think in different ways about methods of assessment. However, in terms of justice to each candidate, I feel that there are no short cuts in English, and that of the three methods of judging overall script quality, the traditional approach is the fairest. Where such research and re-thinking could be an advantage, however, is if it brought home to hard-pressed English teachers that they need to use a variety of approaches when assessing day to day work (holistic, paired, traditional) rather than predominantly focussing on detailed 'correcting' of pupils' work."

Findings

The data analysis for this project has been complex and lengthy. It is intended that the findings will be disseminated in a subsequent report.

References

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357–373.

Christie, T. & Forrest, G.M. (1981). *Defining public examination standards*. Schools Council Research Studies. London: Macmillan Education.

French, S., Slater, J. B., Vassiloglou, M. & Willmott, A. S. (1988). *The role of Descriptive and Normative Techniques in Examination Assessment*. In: H. D. Black and B. Dockrell (Eds.), *Monograph of Evaluation and Assessment Series No. 3*. Edinburgh: Scottish Academic Press.

Gill, T., Bramley, T. & Black, B. (2007). *An investigation of standard maintaining in GCSE English using a rank-ordering method*. Paper presented at the British Educational Research Association Conference, London, September 2007.

Greatorex, J. (2003). *What happened to limen referencing? An exploration of how the Awarding of public examinations has been and might be conceptualised*. Paper presented at the British Educational Research Association Conference, Edinburgh, September 2003.

QCA (2008). *GCSE, GCE, VCE, GNVQ and AEA Code of Practice2007/8*. London: Qualifications and Curriculum Authority.

Thurstone, L.L. (1927a). Psychophysical analysis. *American Journal of Psychology*, **38**, 368–389. Chapter 2 In: L. L. Thurstone (1959), *The measurement of values*. Chicago: University of Chicago Press.

Thurstone, L.L. (1927b). A law of comparative judgment. *Psychological Review*, **3**, 273–286.

Investigation into whether z-scores are more reliable at estimating missing marks than the current method

Peter Bird Operational Research Team, OCR

Context

The awarding bodies in the UK use similar, but slightly different, methodologies for assessing missing marks (i.e. marks for candidates who are absent with good reason). In an attempt to standardise the process across awarding bodies, a z-score method of estimating missing marks (as used by other awarding bodies) was investigated to see if it was better than the current proportional estimation method being used by OCR. The proportional method requires the available marks for a candidate to be used in calculating the missing mark, and therefore this method is straightforward to apply operationally. Any new method would also be constrained by what can be achieved operationally at what is already a very busy time of year for processing. The aim of this article is to compare the two methods for a sample of specifications, to highlight any issues and differences in the accuracy of estimating marks. Further, more in depth work, could then be undertaken if required.

Introduction

Two GCSE specifications and three A-level specifications were used to investigate whether a z-score estimation method was better than the current estimation method in use. The subjects were chosen because of their very different characteristics. This study was designed to be an exploration of the likely issues and problems from each method before a more in-depth analysis was carried out.

The 'current proportional method', which has been in use for many years, assumes that a candidate will perform equally well on the unit/components that they are missing as they did on the units/components for which they have marks. The 'z-score method' assumes that the relative position of a candidate's mark in relation to all other candidates taking the same unit/component stays the same for both unit/components. In short: i) the existing method assumes the same proportional score in relation to the maximum mark on the missing component(s) as on the components taken; ii) the z-score method assumes the missing mark lies the same number of standard deviations from the mean on the missing component as on components taken.

Each method was compared by treating in turn all candidates as having missing marks. Estimates were then calculated. This was repeated for each unit/component within each specification.

At OCR, missing marks have been estimated for many years on the assumption that a candidate performs equally well on the unit/components for which we have marks for them, as they do on the missing unit/component. The reliability of this method relies on assuming there is a good correlation between the unit(s) being predicted from and to, and that the distribution characteristics of each unit are similar. This method does not take into account whether the marks already achieved come from a distribution with the same distributional characteristics as the one which is being estimated, that is, the obtained mark may have come from a skewed distribution, such as a coursework unit, or from a tiered paper, and the estimate may be required for a unit which has a bell-shaped distribution.

The new proposed method of using z-scores is a method which takes into account how well the candidate for which we are estimating a missing mark has performed on other components in relation to all other candidates taking the same unit/component. It effectively gives a higher z-score to a candidate who has achieved a mark in the top end of the mark distribution, and similarly a lower z-score to a candidate who achieved a mark at the bottom end of the mark distribution. For a normal distribution we would expect 68% of candidates to lie within the mean +/- one standard deviation, and 95% of candidates to lie within the mean +/- two standard deviations. A mark is transformed to a z-score by subtracting the mean and dividing by the standard deviation from the distribution it comes from.

Example of applying both methods

Specification with three components. Candidate has component 3 missing.

Component	*Mark Achieved*	*Max*	*Mean**	*Std Dev**	*Calculated z-score*
1	12	40	20	10	=(12-20)/10= -0.8
2	17	50	25	12.5	=(17-25)/12.5= -0.64
3	Missing	30	15	7.5	

* assume bell shaped distributions

Current method to predict missing mark on component 3

$$= \frac{\text{Marks gained on components 1 \& 2 } = 12+17}{\text{(Max Mark on component 1 \& 2 } = 40+50} \times \text{(Max mark on missing component 3 } = 30)$$

= (29/90) × 30 = 9.66 = rounded to 10 marks.

Z score method to predict missing mark on component 3

= [(combined z-score of component 1 & 2) × std dev of component 3] + (mean of component 3)

= [(-0.71) × 7.5]+15=9.675 = rounded to 10 marks.

Where combined z-score component 1 & 2

= [z-score component 1 × (max component 1) / (max component 1+2)) +
[z-score component 2 × (max component 2) / (max component 1+2))]

= [(-0.8 × (40/(40+50))) + (-0.64 × (50/(40+50)))] =
(-0.35) + (-0.35) = -0.71

By using bell shaped distributions for all components with the mean set at half the maximum marks and the standard deviation set at half the mean mark, the estimates for both methods came out very similar.

Effect of z-score process

To see the effect of the process, random data have been generated to create an example of a typical written paper mark distribution with mean 50, standard deviation 15. These have then been converted to z-scores (Figure 1 below).

A ceiling at a mark of 60 was introduced to create a skewed distribution as might be seen in coursework mark distributions. This produced a mean of 43.4 and standard deviation of 11 (below).

The effect of using coursework to predict a mark on the written paper in the example above is that even if a candidate achieves the maximum mark on coursework, they are effectively capped for their estimated mark on the written paper to about 70 out of 100 (because the maximum z-score they can achieve is around +1.5).

The effect of using written papers to predict coursework in the example above is that anyone achieving over approximately 70 marks on the written paper will be estimated as achieving the maximum mark on the coursework.

Combining units/components

Using a combination of different types of mark distribution is more likely to produce less reliable mark estimates than estimating using similar types of distribution. In order to combine z-scores from different units/components to create one z-score, the individual z-scores are weighted according to the relative weightings of each unit/component to each other.

For example,

If a candidate's marks produced z-scores of +1 and +1.5 on units with weightings of 20% and 30% respectively, the combined z-score is $[(+1 \times 20)/(20+30)]+[(+1.5 \times 30)/(20+30)]=(+0.4)+(+0.9)=+1.3$.

Issues with cohorts used for estimating

Coursework marks may be used from a distribution which contains both foundation and higher tier candidates so the mean and standard deviations would not be truly representative of a particular tier cohort. In a unitised scheme, you cannot guarantee the cohort from which an estimate is obtained is the same as the original cohort for the missing unit, particularly with early takers or re-sitters being included. For this analysis it was assumed any mark estimates would be based on the distribution of the missing unit within the same session as the aggregation of unit results was requested.

The more the cohort used for prediction varies, the more you would anticipate that the reliability of the estimation will decrease. In the examples shown so far, these did not involve UMS marks. However, when UMS marks are used for estimating other UMS marks we have to bear in mind the marks have already been subjected to some 'stretching and squeezing' across the mark ranges. Comparisons of the differences in z-scores were looked at between those derived from weighted raw marks

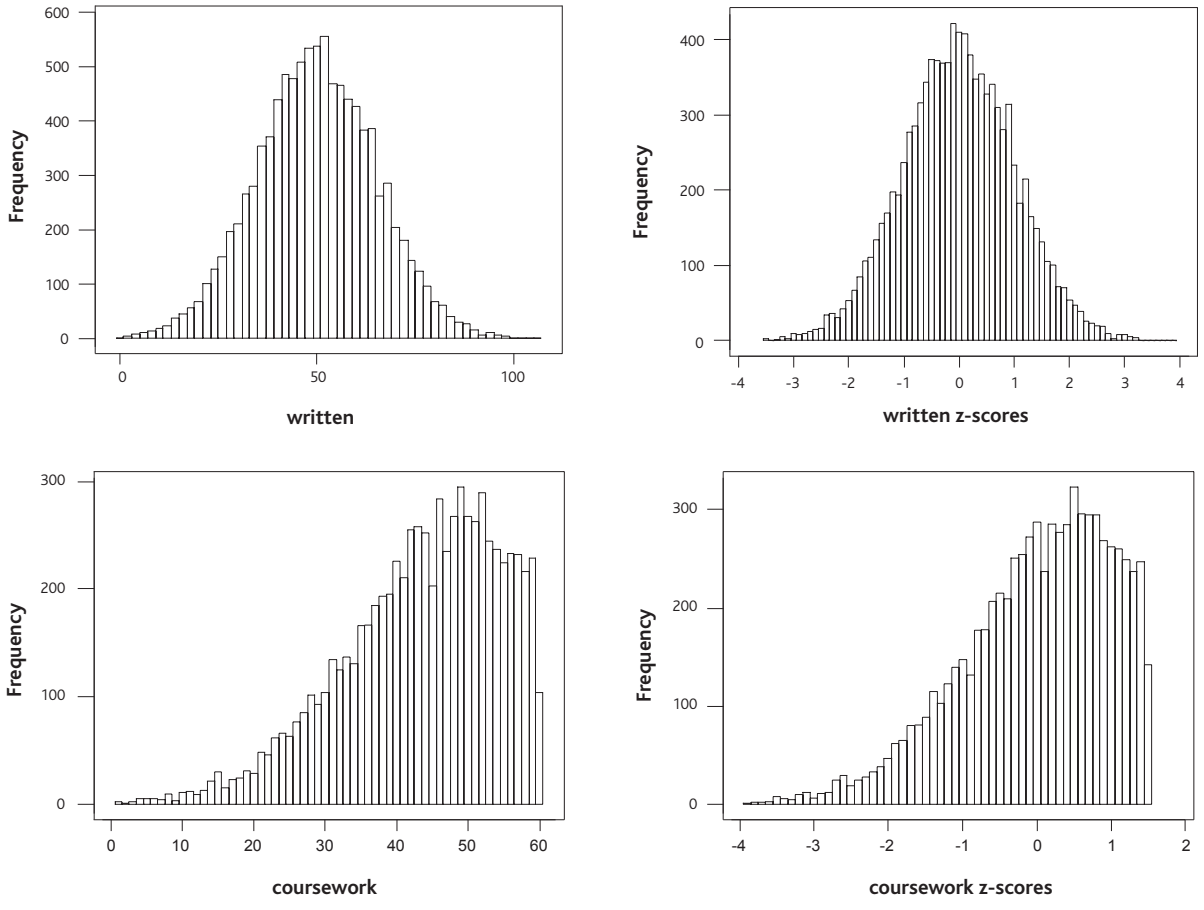


Figure 1: Effect of z-score process

Table 1: Examples of GCSE and A-level differences in cohort/prediction method

Specification type	*Predict z-scores from*	*Map z-scores onto*	*Cohort*
GCSE Linear (non tiered)	Weighted marks from Component(s)	Weighted Mark for missing component	Exactly the same cohort used to predict from and to.
GCSE Linear (tiered)	Written Component(s) from relevant tier and/or coursework	Weighted Mark for missing Component	Exactly the same cohort used to predict from and to. Z-scores could be distorted if coursework tier breakdowns not available.
GCSE Unitised (tiered)	Weighted unit marks for session these were sat in	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to. Z-scores could be distorted if coursework tier breakdowns not available.
GCSE Unitised (untiered)	Weighted unit marks for session these were sat in	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to
GCSE Unitised (Linear)	Weighted unit marks for session these were sat in	Weighted/UMS Mark for unit in aggregation session	Exactly the same cohort used to predict from and to. Z-scores could be distorted if coursework tier breakdowns not available.
A/AS-level (missing AS unit)	Weighted AS unit marks for session these were sat in.	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to
A-level (missing A2 unit)	Weighted A2 unit marks for session these were sat in	UMS Mark for unit in aggregation session	Most likely not the same cohort used to predict from and to.
Missing component within (GCSE or A-Level)	Weighted marks from Component(s)	Weighted Mark for missing component	Exactly the same cohort used to predict from and to unit

and those derived from UMS marks. These showed that for the majority of candidates there are no differences, although the z-scores varied by (+/-) 0.1 to 0.2 for approximately 10–25% of candidates.

For any readers who are unfamiliar with the concept of the Uniform Mark Scale (UMS), an excellent explanation is found in Gray and Shaw (2009).

To improve reliability of estimating for A-level you might want to look only at the best marks from all units of the candidates who are aggregating. Table 1 above outlines differences in the cohort used for different specification types and where reliability issues may exist.

Comparison of different estimation methods

In order to compare different estimation methods, missing marks were created where valid marks already existed for entire units/components, this then allowed comparisons of the estimation accuracy of each method. GCSE linear (tiered), GCSE unitised (untiered) and A/AS-level specifications were used in analysis to look at any differences between tiered/uncapped and UMS conversion specifications. To do this the following assumptions were made:

- Only candidates with complete profiles of marks were included.
- Where options exist within units, the mark used to calculate the z-score is the final weighted mark.
- The z-score is calculated from the unit in the session from which it counted towards aggregation.
- Estimation of unit UMS mark is based on using z-scores from the unit UMS distribution of missing mark in June 2007 (i.e. aggregating session).
- Where optional units exist, the estimation will be based on the marks each candidate has achieved on the units taken.
- Missing AS units are only estimated on AS units.
- Missing A2 units are only estimated on A2 units.
- Very small entry units are excluded.
- Z-score calculations were calculated using data which are shown on our exams processing system.

Estimating marks for candidates aggregating GCSE Geography 1987 in June 2007

GCSE Geography 1987 was used to evaluate the effectiveness of each estimation method as it contains a good mix of distribution types, a large number of candidates and two tiers. Candidates take either Foundation or Higher option and components as below:

Foundation: Component 1 (Foundation) + Component 3 (Foundation) + Component 5 (coursework)

Higher: Component 2 (Higher) + Component 4 (Higher) + Component 5 (coursework)

Summary statistics for each component are shown in Table 2. The foundation option papers are both skewed as candidates tend to get higher than half marks whereas the higher tier candidates' marks are well dispersed on the written paper but skewed on the coursework. The correlation between written papers is higher than between written paper and coursework which makes this a very 'typical' specification. The correlations between component 1 and 3 is +0.71;

Table 2: Summary statistics for the weighted marks for GCSE Geography 1987

Component	*01*	*02*	*03*	*04*	*05 (F=Found, H=Higher)*
MEAN	48.32	49.49	33.09	32	24.5(F)/38.8(H)
STD	12.24	11.65	8.43	7.59	8.9(F)/7.7(H)
N	17271	20591	17271	20591	17271(F)/20591(H)
MAX	90	90	60	60	50
SKEW	-0.50	0.0	-0.59	+0.1	-0.12(F)/-0.69(H)

between 2 and 4 is +0.59; and all remaining correlations are between +0.4 to +0.46.

For each candidate, an estimation of each of their marks was calculated in turn using the other available marks, that is, effectively treating each candidate as having a missing mark. Component 01 was then estimated from marks on components 03 and 05; component 02 was then estimated from marks on components 04 and 05, etc. This was carried out for both the current estimation method and the z-scores estimation method.

COMPONENT 01 (Foundation Written Paper)

The graphs in Graph 1 below show two box plots. The first plots the differences between the estimated written marks on component 1 (using the current estimation method based on component 3 [written paper] and component 5 [coursework]) and the actual marks the candidates achieved. The second shows the same estimation, but using a z-score methodology instead. The vertical axis shows the differences (estimated-actual) and the horizontal axis shows the actual mark achieved. A positive difference shows where the estimation process was over-estimating the mark and a negative difference where it was under-estimating the mark.

The edges of each box for each mark point show where the 25 and 75 percentiles of candidates' marks lie between and the horizontal line within the box is the 50 percentile point. The lines extend to contain 90% of candidates' marks. A box plot of the differences between estimating marks using the z-score method and the actual marks is shown in Graph 2. Both Graph 1 and Graph 2 are very similar, thus both methods produce very similar outcomes although the widths of the 25 and 75 percentiles are marginally smaller using the z-score estimation method.

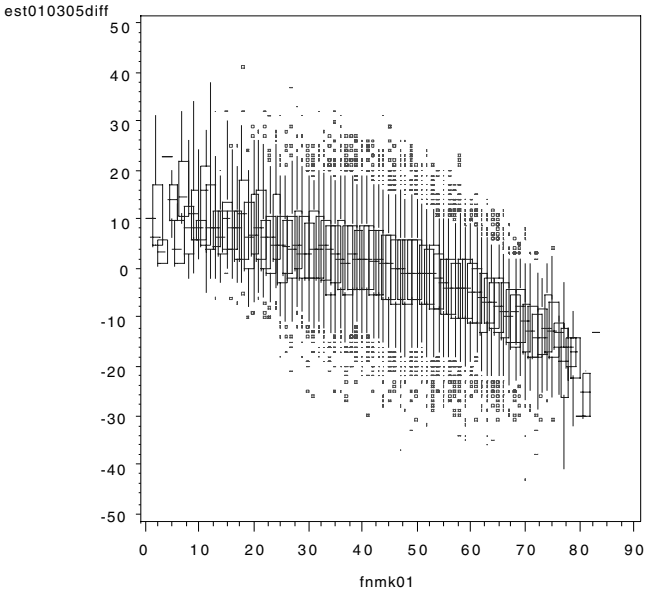
Both box plots show that a candidate would most likely achieve a higher estimated mark than they would have achieved if their actual mark was below the mean mark and a lower estimated mark than they would have achieved if their actual mark was above the mean mark. The differences vary more in magnitude towards the upper end of the mark range.

Using Linear Regression (as a possible method), it is possible to effectively scale/transform the marks in such a manner that the variation on any mark is minimised once all mark estimations have been calculated. A box plot of the differences between estimating marks using the z-score method and then applying a linear regression scaling and the actual marks is shown in Graph 3.

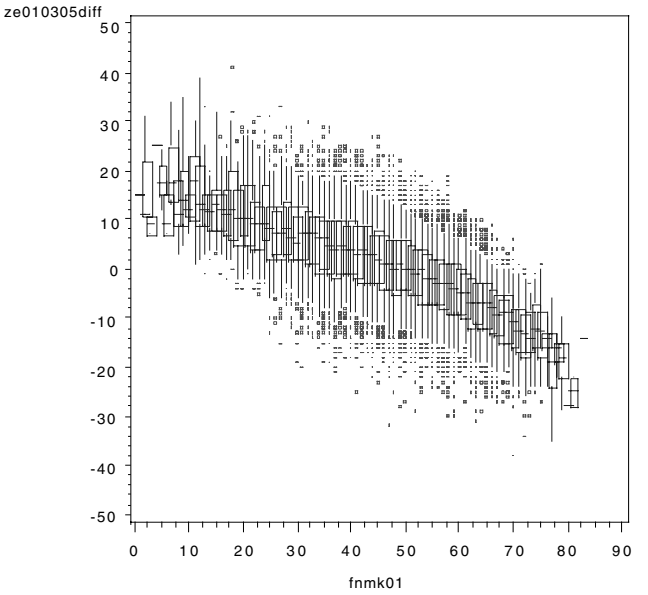
A simple linear regression line was calculated from the differences in Graph 2 treating *ze010305diff* as the dependent variable and *fnmk01* as the independent variable. All differences were then adjusted by subtracting the outcome of this line of best fit for each actual mark.

Using this method, we would be around 90% confident that most mark estimates are within a certain mark range. In this example, the majority of estimates would lie within approximately 10 marks of their actual mark. If this were to be applied to the current estimation method, it would produce similar results.

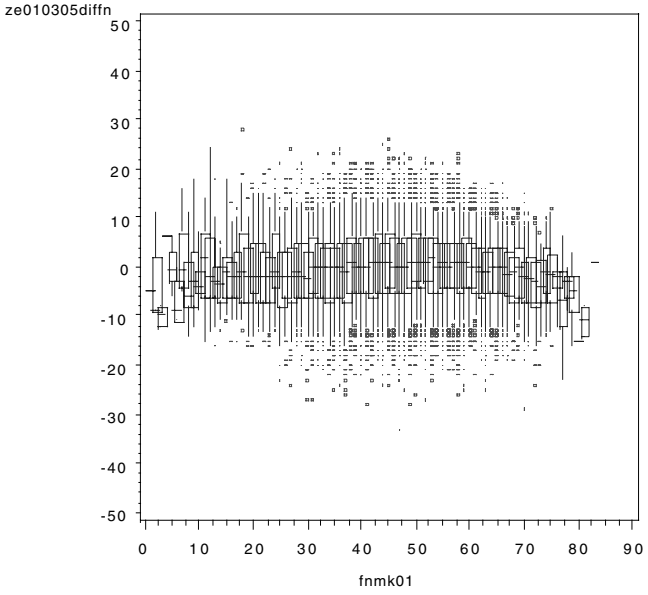
A summary of the differences of the three methods estimation (est), z-scores estimation (ze) and z-scores estimation followed by linear regression scaling (zen) is shown in Table 3. For this 'closed linear cohort' the average z-score estimated difference is 0 marks as by definition transforming to z-scores would do this. Comparisons of the 10, 25, 50, 75 and 90 percentile differences show that as each method is applied, the size of the errors between estimated and actual marks generally decreases.



Graph 1: Plot of the differences between (estimated–actual) against actual mark for component 01 (using current estimation rules)



Graph 2: Plot of the differences between (estimated–actual) against actual mark for component 01 (using Z-scores)



Graph 3: Plot of the differences between (estimated–actual) against actual mark for component 01. (Using linear regression to scale marks after z-score estimation has taken place)

Table 3: Summary of differences between estimated and actual marks for each estimation method for Geography 1987/01

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	01	-1.2	9.9	-14	-8	-1	5	11
ze	01	0.0	9.1	-12	-6	0	6	12
zen	01	0.0	7.5	-10	-5	0	5	10

COMPONENT 02 (Higher Written Paper)

This process was repeated for estimating the marks on component 02 using z-scores from the marks on component 4 (written) and 5 (coursework). A box plot of the differences between estimating marks using the current method and the actual marks is shown in Graph 4 and differences between estimating marks using the z-score method and the actual marks is shown in Graph 5.

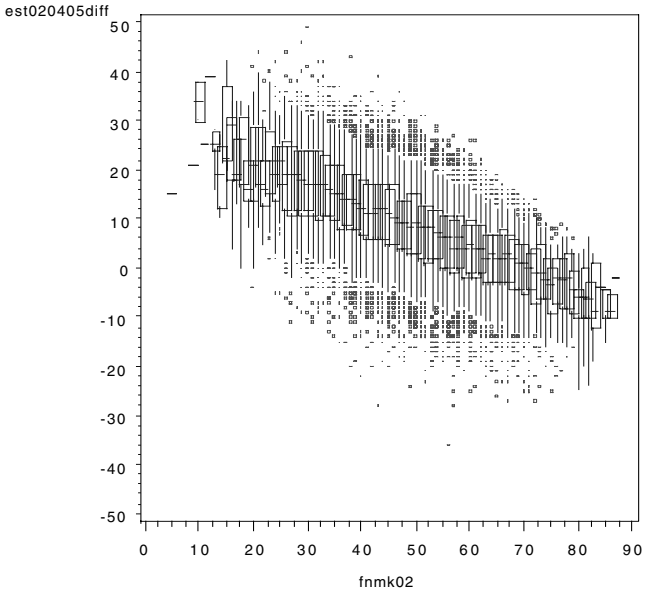
The current estimation process estimates more marks above their actual marks, whereas the z-score method ensures approximately the same number of mark estimates are above and below their actual marks. In contrast to component 1, you do not see the 'dipping' of the plot towards the end of the mark range so the size of the differences are more proportional across the entire mark range.

This example shows how using marks from a coursework distribution (which have a high mean in relation to the maximum mark) as part of the prediction for a written paper (where the mean is closer to half the maximum mark) will over-estimate the marks under the current methodology.

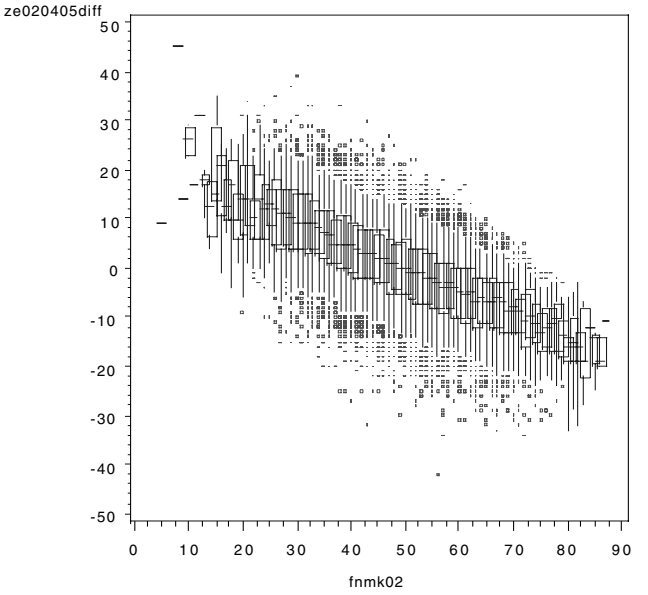
Linear Regression was used to scale/transform the marks in such a manner that expected variation on any mark was minimised once all mark estimations have been calculated. A box plot of the differences between estimating marks using the z-score method (and then applying a linear regression scaling) and the actual marks is shown in Graph 6.

Using this method, we would be reasonably confident that the mark estimate is within approximately 10 marks of their actual mark. If this were to be applied to the current estimation method, it would produce similar results.

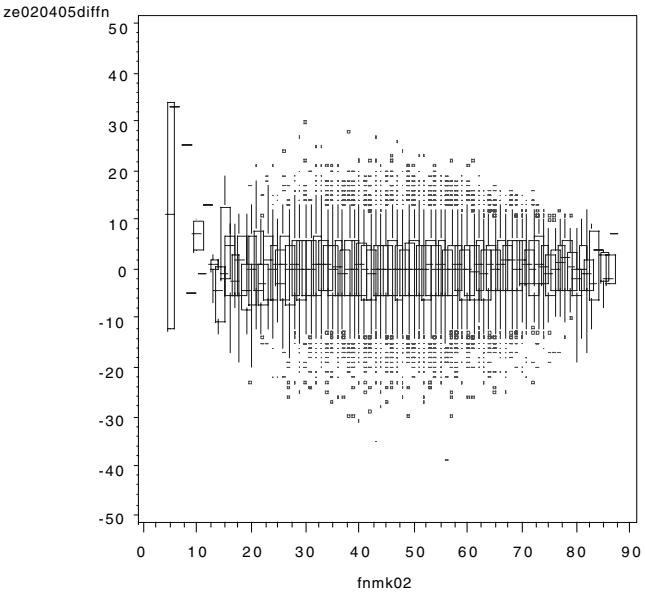
A summary of the differences of the three methods estimation (est), z-scores estimation (ze) and z-scores estimation followed by linear regression scaling (zen) is shown in Table 4. Comparisons of the 10, 25, 50, 75 and 90 percentile differences show that as each method is applied, the size of the errors between estimated and actual marks generally decreases.



Graph 4: Plot of the differences between (estimated–actual) against actual mark for component 02. (Using current estimation rules)



Graph 5: Plot of the differences between (estimated–actual) against actual mark for component 02. (Using Z-scores)



Graph 6: Plot of the differences between (estimated–actual) against actual mark for component 02. (Using linear regression to scale marks after z-score estimation has taken place)

Table 4: Summary of differences between estimated and actual marks for each estimation method for Geography 1987/02

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	02	8.4	9.7	-4	2	8	15	21
ze	02	0.0	9.4	-12	-6	0	6	12
zen	02	0.0	7.6	-10	-5	0	5	10

COMPONENTS 03, 04 and 05

For summary data of the differences seen for the remaining components please contact the author.

In summary, it seems that for 'closed cohort' linear specifications, z-scores would ensure the mean difference between the estimated and actual mark is zero. Any deviations away from this would be balanced positively and negatively. With the current estimation method we cannot guarantee this unless we check using data from all candidates, and make any necessary mark transformations to make this so.

Estimating written papers component 01 and 02 based on written papers 03 and 04 only

It was interesting to try to estimate a written paper mark using only the mark from the other written paper taken so the effect of not using coursework marks for estimation could be seen. Table 5 shows the differences from estimating component 01 from component 03 only, and component 02 from component 04. Only some data are shown here.

This produced slightly better estimates as we might expect. In particular, the mean difference dropped from +8.4 to -1.2 for component 2. In terms of the range of differences seen, component 1 had less large differences at the top end of range and component 02 produced a more even number of positive and negative differences, similar to those seen with the z-scores method.

Table 5: Summary of differences between estimated and actual marks for each estimation method for Geography 1987/01/02 (estimated from written paper only)

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	01	1.6	9.4	-10	-5	2	8	14
ze	01	0.0	9.3	-12	-6	0	6	12
zen	01	0.0	8.6	-11	-6	0	6	11

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	02	-1.2	10.4	-14	-8	-1	6	12
ze	02	0.0	10.5	-13	-7	0	7	14
zen	02	0.0	9.4	-12	-7	0	6	12

A box plot of the differences between estimating marks using the current method and the actual marks for component 01 using component 03 and then using components 03 and 05 are shown in Graphs 7 and 8 respectively below.

Estimation of marks on A-level Physics 7883 for those candidates aggregating in June 2007

Overview

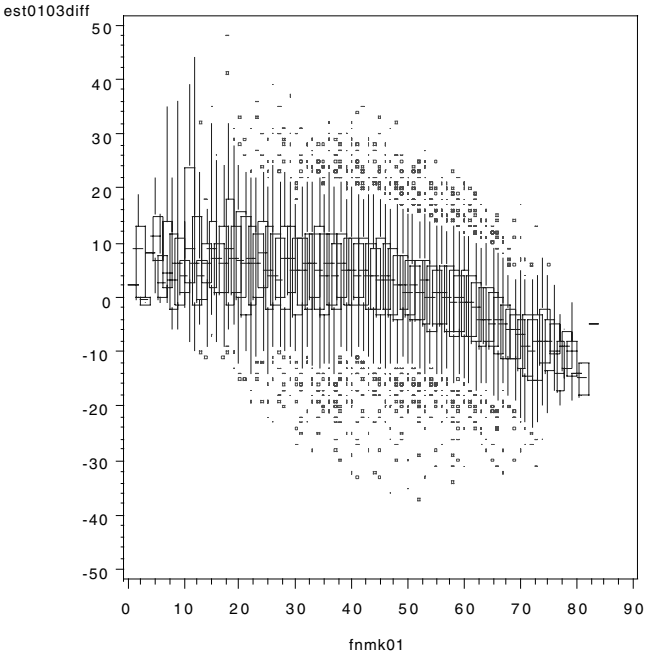
A-level Physics was used to evaluate the effectiveness of each estimation method as it contains reasonable bell shaped distributions, a reasonable number of candidates, and a range of unit types including compulsory/optional and written/coursework or practical. Only A2 units were used for estimation to minimise re-sit effects. 40% of candidates chose to take unit 2824 in both January 2007 and June 2007, whereas less than 5% and 1% did for units 2825 and 2826 respectively. The correlations between units' marks were also all fairly consistent at approximately +0.7 to +0.8. The specification is made up of three AS units 2821–2823 and three A2 units 2824–2826.

Estimation of mark on unit 2824, 'Forces, Fields and Energy'

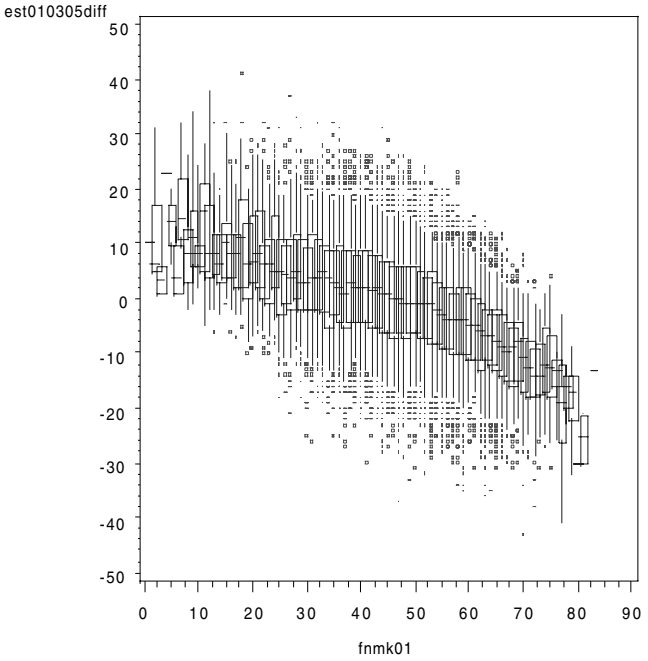
Box plots of the differences between estimating marks using the current method and the actual marks for unit 2824 are shown in Graphs 9 and 10 respectively. This unit is estimated using unit 2825 'Options in Physics' which gives candidates five choices in paper topic and unit 2826 'Unifying concepts' which includes either coursework or practical. In this example, the z-scores method underestimates the actual mark and is prone to slightly more error in estimates across the mark range.

Adjusting for the slope of the differences (zen) makes very little difference to the reliability of the estimated marks calculated using the current estimated method (est); this is shown in Table 6.

Please note that the mean of the differences using the z-scores method is not zero. This seems be an effect of not using a 'closed cohort', that is, z-scores might be pulled from more than one session, where other candidates not aggregating exist and previous attempted marks exist, and these are mapped onto the final aggregation session unit distribution which again may include candidates not aggregating.



Graph 7: Plot of the differences between (estimated–actual) against actual mark for component 01 (Using current estimation rules using component 03 only to estimate from)



Graph 8: Plot of the differences between (estimated–actual) against actual mark for component 01 (Using current estimation rules using components 03 and 05 to estimate from)

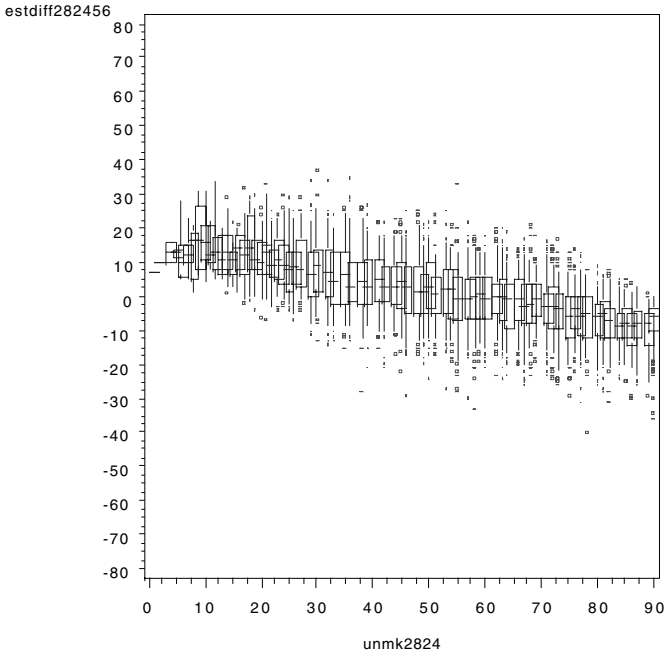
Table 6: Summary of differences between estimated and actual marks for each estimation method for unit 2824

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2824	0.28	10.13	-12	-7	0	7	13
ze	2824	-4.95	10.48	-18	-12	-5	2	9
zen	2824	-0.01	9.71	-12	-6	0	7	12

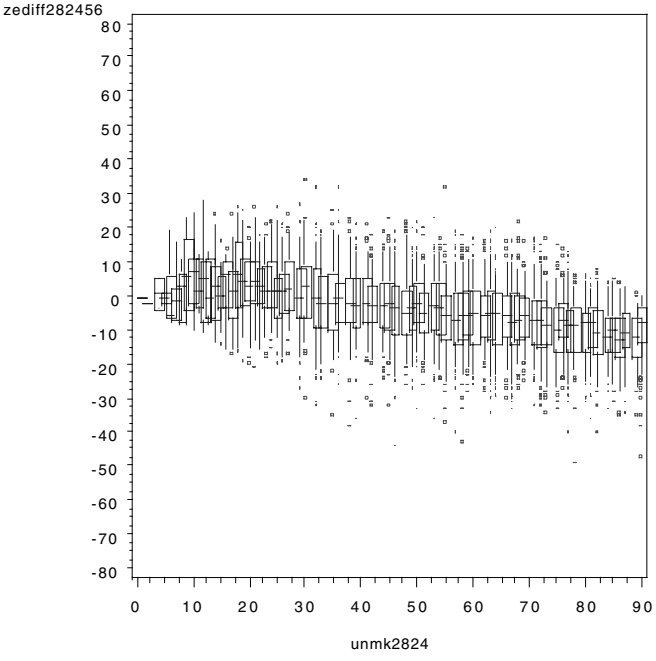
Estimation of mark on unit 2825, Options in Physics

Box plots of the differences between estimating marks using the current estimation method and the actual marks for unit 2825 are shown in Graphs 11 and 12 respectively. Unit 2825 is estimated using unit 2824 and 2826.

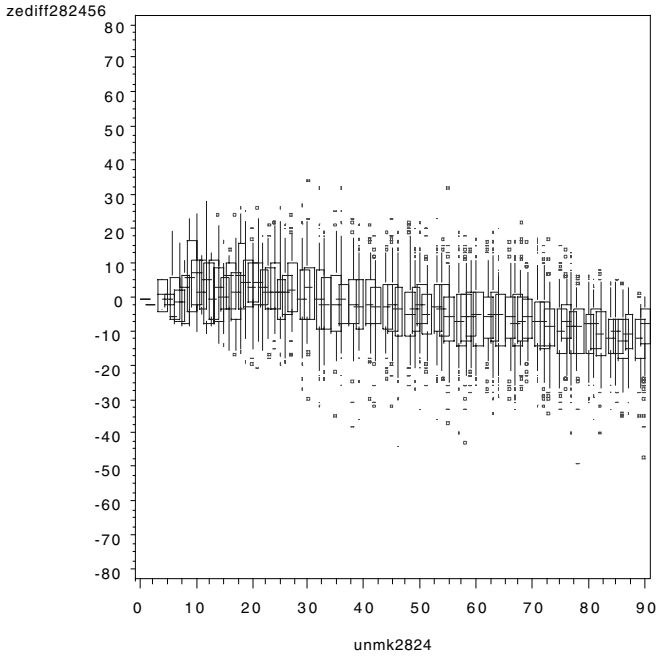
For this unit, the z-scores method is vastly better at estimating the marks than the current estimation method as on average it is only over estimating by 2 rather than 4 marks as shown in Table 7.



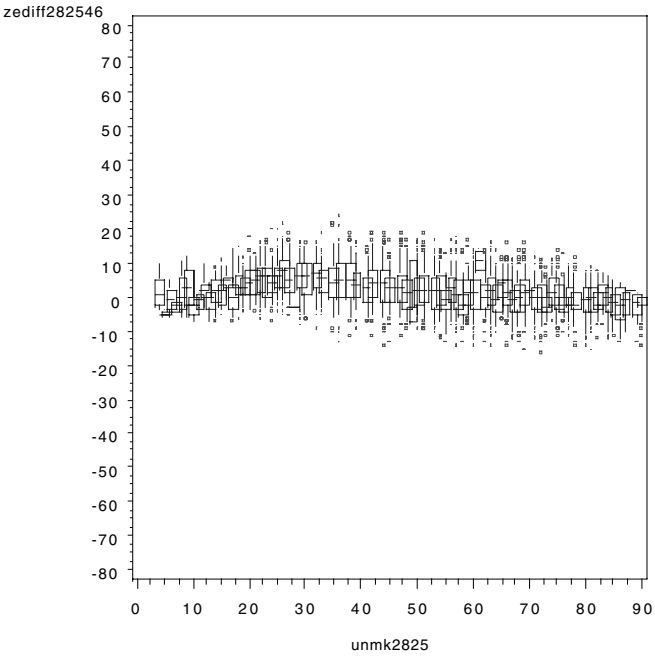
Graph 9: Plot of the differences between (estimated–actual) against actual mark for unit 2824. (Using current estimation rules)



Graph 10: Plot of the differences between (estimated–actual) against actual mark for unit 2824. (Using Z-scores)



Graph 11: Plot of the differences between (estimated–actual) against actual mark for unit 2825 (Using current estimation rules)



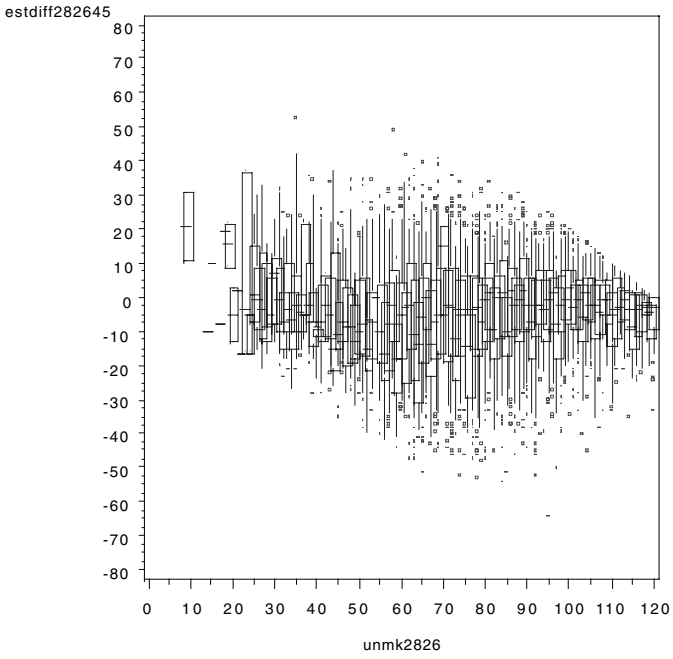
Graph 12: Plot of the differences between (estimated–actual) against actual mark for unit 2825 (Using Z-scores)

Table 7: Summary of differences between estimated and actual marks for each estimation method for unit 2825

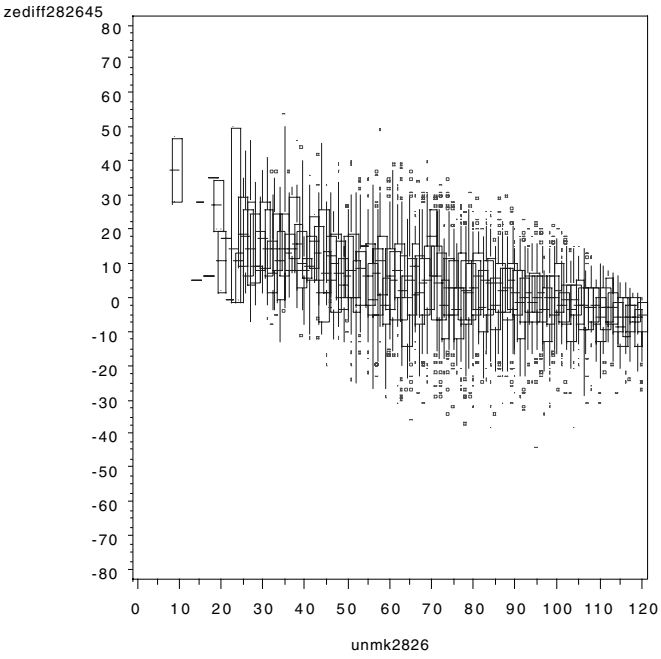
				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2825	3.19	10.56	-10	-4	3	10	17
ze	2825	1.92	5.74	-5	-2	1	6	10
zen	2825	-0.02	5.40	-7	-4	0	4	7

Estimation of mark on unit 2826 'Unifying concepts in Physics'

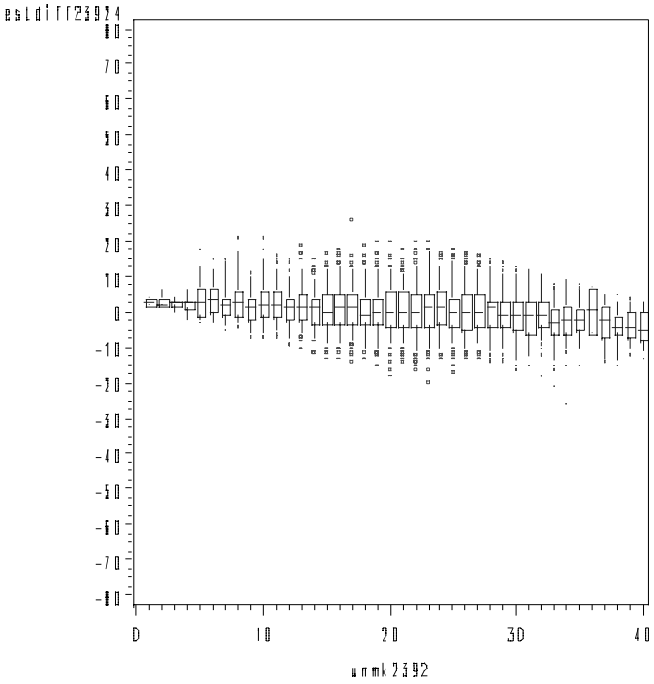
Box plots of the differences between estimating marks using the current method and the actual marks for unit 2825 are shown in Graphs 13 and 14 respectively. Unit 2826 is estimated using units 2824 and 2825. For unit 2826, the estimation of the marks using both methods varies more considerably than the previous units across the entire mark range, as we might expect, as this contains some centre assessed work (Table 8).



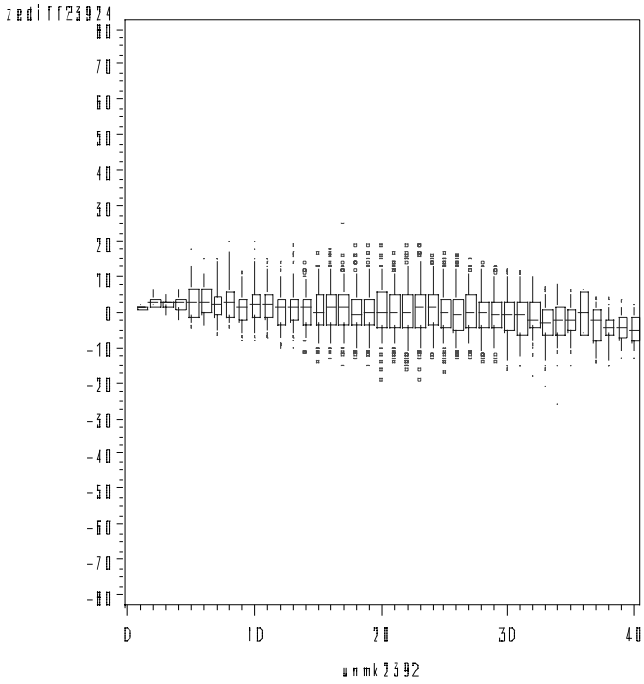
Graph 13: Plot of the differences between (estimated–actual) against actual mark for unit 2826 (Using current estimation rules)



Graph 14: Plot of the differences between (estimated–actual) against actual mark for unit 2826 (Using Z-scores)



Graph 15: Plot of the differences between (estimated–actual) against actual mark for unit 2392 (Using current estimation rules)



Graph 16: Plot of the differences between (estimated–actual) against actual mark for unit 2392 (Using Z-scores)

The z-scores method tends to be more reliable at estimating at the top end of mark range but over-estimates at the bottom end. For this unit, the z-scores method looks more reliable.

Table 8: Summary of differences between estimated and actual marks for each estimation method for unit 2826

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2826	-4.07	14.70	-24	-14	-3	6	14
ze	2826	2.72	12.35	-13	-5	2	11	18
zen	2826	-0.01	11.47	-15	-7	0	8	14

From the Physics units, it seems that estimating coursework from written papers is always going to be more prone to error as some candidates prefer written papers to coursework. It also seems that estimating based on an optional paper (unit 2825) produces slightly less accurate estimates than those based on a compulsory paper (unit 2824).

A compulsory paper is a measure of the candidates' abilities compared to each other whereas the relative positioning of candidates for a paper which has been chosen may allow more variation, particularly in the smaller entry option papers. The re-sitting of unit 2824 most likely allows any candidates who were not in their correct relative position the first time they sat the unit to improve their z-score for this unit.

A-level French 7861 and AS Business Studies 3811

Units in A-level French 7861 and AS Business Studies 3811 were also estimated. Please contact the author for summary statistics. In French, the estimates were mixed as each unit is testing very different traits, speaking, listening, reading and writing. For some units the current estimation method was better, in others the z-scores method was better, but it seems estimating from distributions with high mean marks in relation to the max mark to distributions with mean marks closer to the half the max marks tends to over-estimate, and vice-versa. In Business Studies the estimates for both methods were very similar for all AS units.

GCSE Religious Studies 1030 (Short course)

Overview

GCSE Religious Studies (short course) contains ten papers/units 2391–2400. Each candidate must take two of these units. There are no tiers, no capping, and each paper produces similar looking distributions with correlations between marks of +0.7 and +0.8 for those candidates aggregating. Estimation of units will therefore be dependent on paper choice.

Unit 2392 – 'Christian Perspectives'

Box plots of the differences between estimating marks using the current method and the actual marks for unit 2392 are shown in Graphs 15 and 16 respectively, in these graphs the marks are estimated using unit 2394. The graphs show that the estimation of the marks using the current estimation method is very close to those estimated by the z-score method. The summary statistics for estimating all marks on this unit using the corresponding unit which was taken are shown in Table 9. In this unit, both the current and z-score methods on average underestimated the marks by around 2 marks. Very similar summaries of differences are found on all units in this specification.

Table 9: Summary of differences between estimated and actual marks for each estimation method for unit 2392 (estimated from other available unit)

				Percentile				
method	*comp*	*mean*	*std*	*10%*	*25%*	*50%*	*75%*	*90%*
est	2392	-2.27	5.67	-9	-6	-2	1	5
ze	2392	-2.25	5.64	-9	-6	-2	2	5
zen	2392	0.00	5.50	-7	-4	0	4	7

Conclusion

There is no perfect system when it comes to estimating marks, as candidates perform differently on different units/components. The current estimation process and the z-scores method both rely on the correlation between units/components being as close to one another as possible, but in practice this is never met. The z-scores method does take into account the relative positioning of candidates in respect to other candidates but it is also affected by different shaped distributions and estimates can be artificially capped. It does, however, try to address the over-inflating of written paper marks where a skewed coursework distribution is used to estimate these.

On linear specifications, z-scores would ensure the mean difference between the estimated and actual mark is zero and thus the direction of any errors in estimating marks would be balanced both positively and negatively across the mark range. This cannot be guaranteed with the current estimation method. However, for unitised schemes (which are continuing to increase in number) it is less clear, as in some cases the estimates were very similar; in some cases better and in some cases worse. This is very much dependant on the types of units, correlations between units marks and distribution types.

Unitised schemes by their nature allow candidates to take units throughout the course of study; allow more unit choice; and include a larger number of types of units. Part of the benefit of using z-scores is that it is able to put a measure on the relative position of how well one candidate does in respect to another taking the same paper. However, this benefit becomes less apparent when the candidates taking any one unit are not the same as those taking another unit.

Both methods suffer from different amounts of over-estimating candidates' marks at the lower end of the mark range and under-estimating candidates' marks at the top end of mark range. The z-score method would not always work in all cases, as it would require a minimum number of candidates entered on a particular unit/component to produce sensible z-scores.

A method to improve on the estimations by effectively applying statistically determined scaling adjustments on the marks to counter the effect of under/over-estimating of marks was suggested. To create these scaling adjustments regression analysis was used. Regression analysis can in its own right estimate marks as it takes into account the correlation between the unit marks. The downside of using this method is that it would require the majority of marks to be available before any estimation of missing marks could take place. Its biggest downfall would most likely be the set-up and processing time required on our exams processing system. Further work using regression analysis to estimate marks is planned.

Overall, it seems both the current method and the proposed z-score method produce similar outcomes for unitised schemes. Most of the new GCSE specifications will be unitised, not linear. Therefore, the benefits of changing the current estimation method do not appear to be that great, and brings into question the amount of effort required to bring in a new method which will make no significant improvement on the current method.

References:

Gray, E. & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication*, **7**, 32–37.

EQUITY ISSUESS

'Happy birthday to you'; but not if it's summertime

Tim Oates Assessment Research & Development, **Dr Elizabeth Sykes** Independent Consultant in Cognitive Assessment, **Dr Joanne Emery, John F. Bell and Dr Carmen Vidal Rodeiro** Research Division

For years, evidence of a birthdate effect has stared out of qualifications data for the United Kingdom; summer-born children appear to be strongly disadvantaged. Whilst those responsible for working on these data have, through mounting concern, periodically tried to bring public attention to this very serious issue, it has been neglected by agencies central to education and training policy. Following a flurry of press interest during 2007 and 2008, it has – justifiably – become a key part of the recommendations which may flow from the Rose Enquiry of the primary curriculum.

Researchers at Cambridge Assessment have had a long interest in the birthdate effect because it is so readily observable in the assessment data that they have worked with (Bell and Daniels, 1990; Massey, Elliott and Ross, 1996; Bell, Massey and Dexter, 1997; Alton and Massey, 1998). More recently, Cambridge Assessment decided to review the issue with the intention to advance the understanding of the extent and causes of the birthdate effect in the English education system (Sykes, Bell and Vidal Rodeiro, 2009). A number of hypotheses have been advanced for its cause – clarity in understanding this fully is a vital part of determining possible remedies. Although the review focuses on understanding the birthdate effect in England, it uses international comparisons as one means of throwing light on key factors.

This article outlines the findings of the review. There is robust evidence from around the world that, on average, the youngest children in their year group at school perform at a lower level than their older classmates (the 'birthdate effect'). This is a general effect found across large groups of pupils. In the UK, where the school year starts on September 1st, the disadvantage is greatest for children born during the summer months (June, July, August). Individual summer-born pupils may be progressing well, but the strength of the effect for the group as a whole is an issue of very significant concern. Since the effect of being the youngest in the year group holds in other countries where the school year begins at other times in the calendar year, medical/seasonality hypotheses regarding pre-natal

exposure to viral infections during the winter months for summer-born children can be ruled out as a major explanation of this effect.

As would be expected, given that one year is a smaller proportion of the total life of a sixteen year old than for a four year old, the birthdate effect is most pronounced during infant and primary school but the magnitude of the effect gradually and continually decreases through Key Stage (KS) 3, 4, and A-level. This pattern is particularly evident in research by the Institute of Fiscal Studies (Crawford, Dearden, and Meghir, 2007). The disadvantage for August-born children over September-born children in attainment dropped from an average of 25% at KS 1 to 12% at KS 2, to 9% at KS 3, to 6% at KS 4 and to 1% at A-level. Despite this decrease, the effect remains significant at GCSE, A-level and in respect of entry into higher education. Likewise, analysis of the results from all of the GCSE examinations taken by over half a million candidates born in England, Wales and Northern Ireland within the same academic year showed a consistent depression in grades achieved for students born from September through to August. In addition, the same pattern of depression was detected in the number of subjects undertaken. Despite decrease in magnitude, the birthdate effect persists until the end of higher education (Alton and Massey, 1998).

Data from 13 LEAs providing GCSE results (undertaken in 1990 to 1994) revealed that birthdate effects were still very evident when all subjects were considered. Summer-borns were the lowest attainers in 10 LEAs and Autumn-born children were the highest attainers in 9 of the Authorities. If gender was included in comparisons then summer-born boys had the greatest disadvantage and autumn-born girls had the greatest advantage. Significantly, it was noted that the difference between these 2 groups was about 1 grade at GCSE in each of 9 subjects taken (Sharp, 1995).

Similarly, the IFS researchers (Crawford, Dearden and Meghir, 2007) found that approximately 6% fewer August-born children reached the expected level of attainment in the three core subjects relative to September-born children (August-born girls 55%; August-born boys 44%; September-born girls 61%; September-born boys 50%). Moon (2003) concludes: 'If all the pupils in this cohort who were born in the spring or summer terms were to perform at the level of the autumn-born pupils, it would mean that 213 pupils out of a total of 308 improving their GCSE results by an average of 1.5 grades'. The magnitude of the effect has important implications for pupils' successes and for schools' overall results.

If the birthdate effect is serious in mainstream education, then it can be argued that it is most serious for those who are struggling in the education system. A disproportionately high percentage of relatively young children in the school year also are referred for special educational needs and many of these appear to be misdiagnosed (Sharp, 1995). The birthdate effect may operate in teachers' identification of children in need of special education. Teachers may not be making sufficient allowances for the level of attainment against specific curriculum outcomes of the younger members of their classes.

Beyond GCSE, education becomes more selective with choices being made about further participation. Unfortunately, the birthdate effect seems to have serious consequences. The percentage of GCSE students going on to take at least one A-level drops from 35% in September-born students to 30.0% for August-born students (Alton and Massey, 1998). Likewise, September-born students are 20% more likely to go to university than their August-born peers. The Higher Education Funding Council has concluded that '…if all English children had the same chance of going to university as those born in September then there would typically be around 12,000 extra young entrants per cohort, increasing young participation by 2 percentage points…' (HEFCE, 2005).

Given the existence of this effect, it is necessary to identify the underlying cause. There are competing theories regarding birthdate effects. One is the 'length of schooling' hypothesis – when school admissions are staggered over the year then the youngest have the least schooling. Another is the 'relative age' hypothesis – even with the same length of schooling, the youngest in a year group will be, on average, less mature – cognitively, socially and emotionally – than their older classmates, leading to unequal competition in all three domains that could impact negatively on the younger group. Although it is sometimes difficult to disentangle these two hypotheses, evidence tends to support the latter. Using a common start date does not solve the problem of this type of disadvantage (Daniels, Shorrocks-Taylor and Redfern, 2000).

Teacher expectancy effects may contribute to birthdate effects – teachers may not take children's relative levels of maturity into account when making assessments of their ability and may therefore label younger children as less able than their older peers.

Evidence from developmental psychology suggests that children between the ages of 4 and 5 may not be ready, developmentally, for formal education. Birthdate effects appear to be greatly reduced in countries where formal education begins at a later age. There needs to be a careful consideration of what is best for all children in the early years of schooling, based on solid evidence from psychological research.

The review described here is far more than a simple rehearsal of the findings of a series of relevant studies. It allows an understanding of the accumulation of evidence in respect of the birthdate effect and certain explanations of why it occurs to be discounted. Crucially, the review considers the whole of the education system and this reveals two critical issues. First, that the birthdate effect persists throughout education and training. Secondly, that a strong selection effect may be in operation at all stages – that is, summer-borns are not progressing onto certain routes and into certain levels of education. This effect is not obvious from individual studies limited to specific phases of education. It explains why the summer-borns who get through to the highest level of education are doing well: it is vital to recognise that disproportionately fewer summer-borns actually get to this level *at all*.

Although the existing research is illuminating in respect of the extent of the birthdate effect and of its causes, there is still a need to identify remedies. We believe that work on remedies is not yet sufficiently advanced; substantial, urgent work is required on the means of devising adequate approaches. Although this review was focussed primarily on UK research, it also noted the effect is present in other countries. However, as Bedard and Dhuey(2006) noted, the effect varies from country to country and there is scope for more international work to identify potential solutions to this problem.

From this review, and from the work of comprehensive reviews of the quality of primary and early years education, it is likely that adequate remedy will lie not only in development of a strategy regarding *when* formal schooling should start, but also – at least – in respect of: specific balance in respect of curriculum elements devoted to cognitive, emotional and social development; the training requirements of teaching and support staff; curriculum frameworks; inspection foci; pupil grouping strategy; management of differentiation; and the articulation between early years units and compulsory schooling.

References

Alton, A. & Massey, A. (1998). Date of birth and achievement in GCSE and GCE A level. *Educational Research*, **40**, 1, 105–9.

Bedard, K. & Dhuey, E. (2006). The Persistence of Early Childhood Maturity: International Evidence of Long-Run Age Effects. *The Quarterly Journal of Economics*, 2006, **121**, 4, 1437–1472.

Bell J.F. & Daniels S. (1990). Are Summer-born children disadvantaged? The birthdate effect in education. *Oxford Review of Education*, **16**, 1, 67–80.

Bell J.F., Massey A. & Dexter T. (1997). Birthdate and ratings of sporting achievement: analysis of physical education GCSE results. *European Journal of Physical Education*, **2**, 160–166.

Crawford, C., Dearden, L. & Meghir, C. (2007). *When you are born matters: The impact of date of birth on child cognitive outcomes in England*. The Institute of Fiscal Studies: London.

Daniels, S., Shorrocks-Taylor, D. & Redfern, E. (2000). Can starting Summer-born children earlier at infant school improve their National Curriculum results? *Oxford Review of Education*, **26**, 2, 207–20.

Higher Education Funding Council for England (HEFCE) (2005). *Young Participation in Higher Education*. Report Ref. 2005/03. HEFCE, Bristol.

Massey, A., Elliott, G. & Ross, E. (1996). Season of birth, sex and success in GCSE English, mathematics and science: some long-lasting effects from the early years? *Research Papers in Education*, **11**, 2 129–50.

Moon, S. (2003). Birth date and pupil attainment. *Education Today*, **53**, 4, 28–33.

Sharp C. (1995). What's age got to do with it? A study of patterns of school entry and the impact of season of birth on school attainments. *Educational Research*, **37**, 251–265.

Sykes E., Bell J.F. & Vidal Rodeiro, C.L. (2009). *Birthdate Effects: A Review of the Literature from 1990–on*. Research Report. Cambridge Assessment: Cambridge.

Cambridge Assessment Parliamentary Research Seminar Series – *Better training: Better teachers?*

Sylvia Green Research Division

This seminar series is organised by Cambridge Assessment and hosted by Barry Sheerman MP, Chair of the Children, Schools and Families Select Committee and held in the House of Commons. The aim of the series is to bring together members of the research, academic and education communities as well as policy makers and influencers. This provides the opportunity for those working in educational research to present new ideas and evidence to key decision influencers as well as providing a forum for discussion on important topical issues in the field of education. Previous seminars have covered topics such as, *Aspects of Literacy, New Approaches to National Assessment* and *What makes a good teacher?*

The latest seminar took place in February and focused on the issue of effective teacher training. Over 140 teaching professionals attended, including researchers, practitioners and those involved in the delivery of teacher training in both Initial Teacher Training (ITT) and Continuing Professional Development (CPD) contexts. The seminar was entitled, *Better Training: Better Teachers?* This was a topical area since the select committee is undertaking an inquiry into ITT and CPD for teachers and teacher education is a key policy area. There were two guest speakers. The first to present was Professor John Furlong, Director of the Oxford University Department of Education. The second speaker was Dr David Pedder, Lecturer in Educational Leadership and School Improvement at the University of Cambridge.

PROFESSOR JOHN FURLONG

Professor Furlong addressed a series of questions and began by asking, *What is the role of initial teacher education in improving the quality of teaching and learning in our schools?* and *Is the Teacher Supply Model fit for purpose today?* The difficulties he outlined were: the impact of the economic downturn on supply; hidden and suppressed shortages; implications of the changing gender and age structure of the profession; the flight from private schools; local pressures on school funding; impact of the collapse of the housing market on job mobility.

He also questioned whether we have the right routes into teaching and whether they really bring in different populations. He asked what the right balance of different populations entering the profession should be, whether the quality was the same for each of the routes and why those routes have to be so separate. The data he presented on the quality of the intake into the profession indicated that 58% of those in primary and 54% in secondary had 'good' degrees. Interestingly, the data also showed that in 2007 the average UCAS tariff for undergraduate teacher training was 198 (equivalent of C, D, D), while for Mathematics it was 395 and for European Languages 434. He asked if it was time to abandon the BEd or dramatically increase its intake quality.

The question of quality of provision was discussed and the issue of Teach First was raised. The question was whether there was any evidence to say that Teach First was an effective strategy in raising the quality of entrants to the profession. Another area of the 'quality' discussion related to whether the current approach to quality control is fit for purpose. A great deal of teaching is described as satisfactory and we need to have control approaches (standards, regulatory and inspection frameworks, self assessment documents) that will enhance quality beyond 'satisfactory'. Data were presented on the link between teacher education quality and educational research and the trend was for institutions scoring highly for research to be more highly rated for the quality of their teacher training provision. This led to a discussion about who our teacher educators are and how we recruit and develop them. A survey conducted by Dr Viv Ellis from the University of Oxford, asking *Who are our HEI Teacher Educators?* found that in three months last year a survey of advertised jobs showed that there were 65 posts advertised of which 50% were permanent and 25% were hourly paid temporary workers with pro-rata salaries of £28,000–£35,000.

In answer to the question, *What do we know about what makes effective initial teacher education?* Professor Furlong pointed to the international consensus as the importance of authentic in situ professional learning, quoting Hogan and Gopinathan (2008):

> *If student teachers are going to learn how to become effective teachers, let alone expert teachers, they need to learn from expert teachers in authentic teaching contexts, on the one hand, through close observation and a gradually expanding supporting role in the classroom, and, on the other hand, being coached and mentored, and their learning being appropriately scaffolded by expert teachers/mentors*.

He proposed the weakness of ITT is the weakness of the knowledge base of the teaching profession itself. He highlighted the need for 'a suite of "signature pedagogies" that teach people to think like, act like and be like an educator' (Shulman, 2005). Such pedagogies, he suggested, would promote deep understanding in different subjects and that we should, 'build programs of teacher education around these kinds of signature pedagogies' (Shulman, 2005).

DR DAVID PEDDER

In his presentation, Dr David Pedder focused on Continuing Professional Development (CPD) and reported on his research project, *Schools and CPD, State of the Nation*. He presented a brief overview of the design of the study but concentrated mainly on the findings and implications for policy in relation to three broad themes:

- The benefits, status and effectiveness of CPD
- The planning and organisation of CPD
- Access to CPD

The research design incorporated:

- A literature review based on 28 reports and 33 articles/conference papers.
- Surveys returned by over 1000 teachers at 151 schools with a 39% response rate overall.
- Qualitative 'snapshots' in 9 primary and 3 secondary schools.

The findings from the survey under the theme of 'Benefits, status and effectiveness of CPD' indicated the following:

- There is a lack of effective CPD in terms of levels of classroom contextualised practice, collaboration with colleagues, and research informed professional learning;
- There is a lack of effective CPD practice in terms of both the form and duration of CPD activities.
- There is little indication that current CPD is seen as having an impact on raising standards or narrowing the achievement gap. This is despite the fact that the vast majority of teachers thought that CPD would have a positive impact on pupils' learning and achievement.
- Teachers identify a wide range of benefits of CPD – however, these benefits vary significantly by school and teacher characteristics.
- School leaders report that school-based and classroom-based CPD with a clear focus on learning processes and improving pedagogy provide more value for money than CPD that takes place outside schools.

Issues raised around the second theme, 'Planning and Organisation of CPD', were:

- Relating to school contexts – strategic planning for CPD frequently does not provide for the wide range of professional development needs that exist in schools. Planning and organisation of CPD in schools tends not to be strategic and struggles to meet the competing development of individual teachers and whole-school improvement plans.
- Relating to schools as organisations – organisational choices made in schools about roles and responsibilities do not always support or help to develop CPD planning and provision.
- Relating to culture change and aspects of New Professionalism – some changes to teachers' perceptions and actions in relation to their roles and responsibilities are evident, in tune with the New Professionalism agenda. Wholesale change has not occurred.
- Relating to evaluation of CPD and follow-up – evaluation systems of CPD used in schools are insufficiently tied to considering planned outcomes, identifying specific criteria and considering value for money.

On 'Access to CPD' the survey indicated that:

- Teachers are offered a narrow range of CPD opportunities which vary significantly by experience, career stage and leadership responsibility.
- Both school-level conditions and teacher perceptions serve as barriers to CPD participation.

The findings from this research suggest that a great deal needs to be done to target CPD and to enable teachers to make the most of the opportunities on offer.

The presentations provoked lively discussion around some fundamental questions about teacher education and they will undoubtedly lead to further debate during the planned inquiry into this educational area.

For further details on the presentations see the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Events/Event_Detail?id=126302

References

Hogan, D. & Gopinathan, S. (2008). Knowledge management, sustainable innovation, and pre-service teacher education in Singapore. *Teachers and Training*, **14**, 4, 369–384.

Pedder, D., Storey, A. & Opfer, V.D. (2008). *Schools and continuing professional development (CPD) in England – State of the Nation research project*. A report commissioned by the Training and Development Agency for Schools.

Shulman, L. (2005). *The Signature Pedagogies of the Professions of Law, Medicine, Engineering, and the Clergy: Potential Lessons for the Education of Teachers*. http://hub.mspnet.org/index.cfm/11172

Research News

Cambridge Assessment Conference

The 4th Cambridge Assessment conference will take place on Monday 19th October 2009 at Robinson College, Cambridge, UK. The theme will be, *Issues of control and innovation: the role of the state in assessment systems*.

The conference marks a key point of the Cambridge Assessment Network's annual programme, bringing together leading analysts and commentators to scrutinise current developments. It will harness the insights of public policy experts, educationalists and assessment specialists to make a radical contribution to discussion and critique of issues of control of assessment. The timing of this conference is important: major changes to regulation, the shape of agencies, and to allocation and form of responsibilities are underway – mapping the consequences and implications of these changes is a vital process. A great deal is at stake in the management of large and complex systems such as educational assessment; sophisticated commentary is an essential part of both understanding the operation of new arrangements and increasing public accountability. The conference will make a major contribution to thinking in the area.

Leaders from many areas within education are invited to attend, including senior managers from schools, colleges and universities, as well as those working for national and local education bodies, professional organisations, political parties, the media, awarding bodies and employers' organisations.

The keynote speakers will be Professor Alison Wolf of King's College London and Professor Robin Alexander from the University of Cambridge. Professor Wolf will discuss, *The role of the State in educational assessment*. Professor Alexander, who is also Director of the Cambridge Primary Review, will give a presentation entitled, *Whose education system is it? Lessons on standards, quality and accountability from the Cambridge Primary Review*. There will also be eight discussion seminars, covering a range of subjects within the main conference theme. For further details see www.assessnet.org.uk/annualconference. Delegates will have opportunities to comment and ask questions during the sessions, to debate issues during their choice of seminar sessions, and to network at lunch and coffee breaks, as well as at the drinks reception at the close of the conference.

Fees

Early bird rate (for bookings received on or before 31 July 2009)	£150
Standard rate (for bookings received after 31 July 2009)	£180

For more information or a booking form, please contact:
Cambridge Assessment Network
1 Hills Road, Cambridge, CB1 2EU, UK

Email: thenetwork@cambridgeassessment.org.uk
Tel: +44 (0) 1223 553846
Fax: +44 (0) 1223 552830

www.assessnet.org.uk/annualconference

Conferences and seminars

Successful Thinking Skills for All

In January OCR hosted a conference for teachers at Rugby School in Warwickshire on successful thinking skills. Beth Black from the Research Division, Cambridge Assessment, gave the keynote address on *What do we know about Critical Thinking?*

American Educational Research Association

In April Jackie Greatorex attended the American Educational Research Association conference in San Diego and presented a paper entitled, *How do examiners make judgements about grading standards? Some insights from a qualitative analysis*. The theme of the conference was *Disciplined Inquiry: Education Research in the Circle of Knowledge*.

International Amsterdam Multilevel Conference

In May John Bell attended the International Amsterdam Multilevel Conference and presented a paper on *Evaluating multilevel models used for examination subject difficulty*.

Design & Technology Association Education and International Research Conference

Gill Elliott attended the Design & Technology Association Education and International Research Conference in Loughborough in June and presented a paper on *Issues associated with teaching practical cookery in UK schools*.

Publications

The following articles have been published since Issue 7 of *Research Matters*:

Crisp, V. (2009). Does assessing project work enhance the validity of qualifications? The case of GCSE coursework. *Educate*, **9**, 1, 16–26.

Crisp, V. & Novaković, N. (2009). Are all assessments equal? The comparability of demands of college-based assessments in a vocationally-related qualification. *Research in Post Compulsory Education*, **14**, 1, 1–18.

British Education Index

Free information services from the British Education Index

Following an intensive period of development informed by consultation with the educational research community, a new British Education Index interface was launched on the Index's website in November 2008.

The rationale of the British Education Index (BEI) remains the same – to support the professional study of education through the identification of pertinent reading matter and event-related information – and at the heart of the operation is the creation of well-structured, detailed electronic records presenting information on relevant journal articles, grey literature, conference programmes and papers, and internet resources.

However, users will now benefit from the integration of the several previously discrete information sources maintained by the BEI office at the University of Leeds.

Each record is produced and validated in Leeds by indexers who have sight of the information and who make full use of the British Education Thesaurus, enabling the most relevant information to be found quickly and easily. In addition, significant work has been carried out in harmonising the BEI authority lists to minimise duplication and ensure that search and retrieve processes are streamlined.

Simple methods may still be used to obtain the widest results but the enhanced interface now allows for precise searching to produce a more refined list of records. Furthermore, an enhanced degree of connectivity means that navigation both between and within records is efficient and quick, enabling the user to make better sense of the ever-expanding amount of information to which they now have access. While use of the full BEI database remains contingent on subscription, visitors to the new BEI site have free access to over 9000 BEI records.

These developments are consistent with the BEI's mission to be known and used as the UK's principal source of authoritative information about professional knowledge in education.

The new BEI website and search interface is at: http://www.bei.ac.uk.

For more information contact:

British Education Index
Brotherton Library
University of Leeds
Leeds LS2 9JT

Tel: 0113 3435525
email: bei@leeds.ac.uk

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

ISSN: 1755–6031