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.

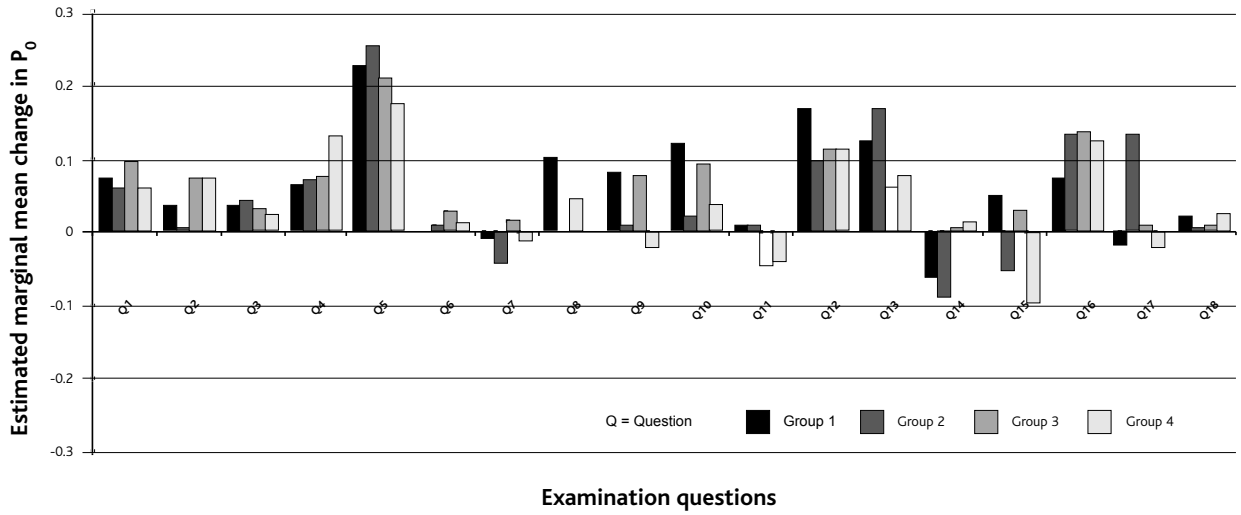

Figure 1: Graph showing changes in accuracy after training for individual AS-level biology questions

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)

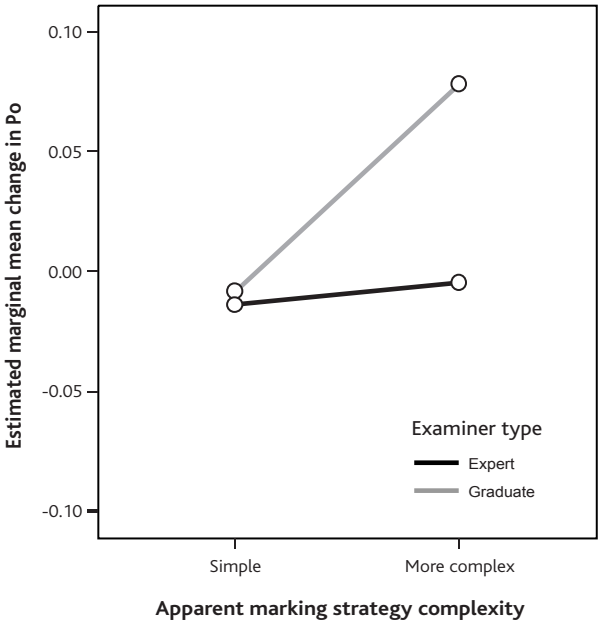

Figure 2: Graph showing estimated marginal mean changes in P_o values for expert and graduate maths examiners for questions with different apparent marking strategy complexities.

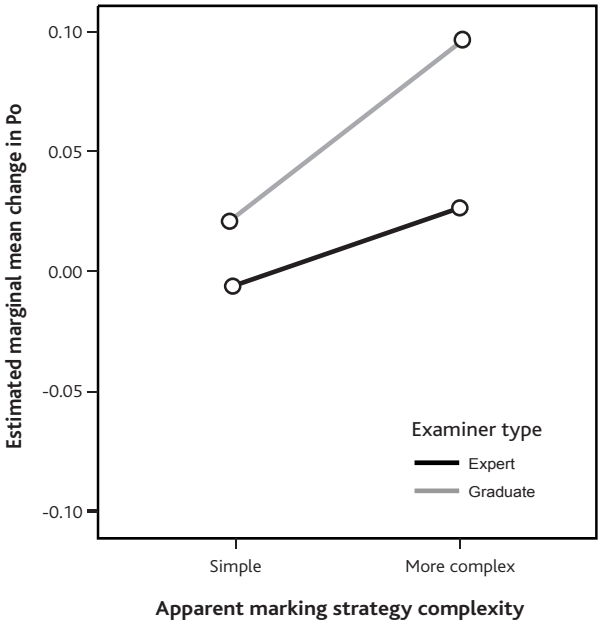

Figure 3: Graph showing estimated marginal mean changes in P_o values for expert and graduate physics examiners for questions with different apparent marking strategy complexities.

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.