Centre for Education Research and Practice

Who is the specialist?

The effect of specialisms on the marking reliability of an English literature examination

Claire Whitehouse

Summary

The effect of examiner specialisms on marking reliability is an under-researched topic. This report explores the topic using multilevel modelling of mark remark data from quality monitoring samples. It aims to provide an answer to the question: do examiners mark more accurately when they mark responses from their own specialisms?

Data from 53 examiners across more than 3,000 responses to optional questions from an A-level English literature examination were modelled. The optionality in the question paper was based on prescribed texts. Each examiner was matched to a home centre. The texts offered by a home centre (based on candidates' choices in the examination) became the examiner's specialisms. The report draws one main conclusion. Examiners who have experience within their working environment of a range of specialisms, mark more reliably than those examiners who work with a limited number of specialisms. This finding was statistically significant. Suggestions for future work in this area are provided at the end of the report.

Keywords: marking reliability, examiner specialisms, optional essays, English literature

Background

Recently, the accuracy of the marking of high stakes examinations, such as GCSEs and A-levels, has been the focus of attention for a number of stakeholders including government (na, 2012a), head teachers (na, 2012b), and the regulator (Opposs & He, 2011), as well as awarding organisations. Inaccurate marking compromises the validity of qualifications. Awarding organisations, operating in an environment of continuous improvement and transparency, must secure validity. It is, therefore, incumbent on them to consider all means available to increase the reliability of marking. To do this, it is necessary to identify and understand the variables that affect marking reliability.

To this end Suto and Nádas (2008) proposed a qualitative model with two main influences on marking reliability: examiners' expertise and the demands of the marking task. They identified a number of aspects to examiners' expertise including: marking experience, teaching experience, level of general education, subject knowledge, personality traits and training. Suto and coworkers (2008 & 2011) used multiple mark data for selected questions from GCSEs in maths and science to investigate selection criteria. The most pertinent finding from their studies was that highest level of education, regardless of the subject area, was a better predictor of marking accuracy than highest level of education in a relevant subject, previous marking experience or teaching experience. Royal-Dawson and Baird (2009) arrived at a similar conclusion in the context of the marking of Key Stage 3 English. They recommended that, except for a few highly subject specific items, the requirement of teaching experience be loosened to allow non-teaching graduates of any subject to mark responses as their marking reliability was at least as

Copyright © 2014 AQA Education and its licensors. All rights reserved.

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

good as that of those examiners with teaching experience. In a designed study using selected questions from a GCSE English examination, Meadows and Billington (2010) attempted to control for marking experience, teaching experience and subject knowledge. They concluded that trainee English teachers, undergraduates of English and undergraduates of other subjects could mark as accurately as GCSE English examiners, provided they were offered the same level of training. Some questions requiring more complex marking strategies were exempt from this conclusion.

In summary, evidence exists to demonstrate that neither subject expertise nor teaching experience is a necessary requirement for marking accuracy. This evidence, though, is limited to questions with low tariffs that do not require essay responses. There is also little evidence that supports or refutes the need for specialisation within certain subjects.

Recently, Pointer (2013) used a Rasch analysis to investigate the effects of examiners marking responses outside their specialist subjects in a combined GCSE science qualification. He concluded that marking accuracy remained stable when examiners marked outside their specialisms. Additionally, there did not appear to be any questions that caused undue difficulties for any particular group of specialists. From these findings, Pointer recommended that candidates' responses need not be directed to examiners based on their science specialism(s).

The components in Pointer's (2013) study contained compulsory questions only. Assessments that contain optional questions may also encourage the formation of groups of specialist examiners. For example, English literature requires students, and therefore teachers and examiners, to be familiar with a number of specified texts. As it is not possible to study all of the texts, selection begins in the classroom. Through their teaching experience, examiners may feel that they are specialists in certain texts.

Whilst operationally cumbersome, the presence of optional topics within a question paper offers benefits to teachers and students. Teachers can select content that is appropriate to their own teaching, to the resources available in their school or college and to their students' interests and learning needs. In the examination, students choose the option(s) they believe allow them to perform to the best of their ability. In some subjects, such as English literature, reducing optionality would be detrimental for both consequential validity and face validity (Messick, 1989). Teachers may be discouraged from delivering the most educationally beneficial courses as a consequence of qualifications without optional routes (Bridgeman, Morgan, & Wang, 1996).

In the future it will be operationally possible to send responses to specific questions to specialist examiners for on-screen item-level marking. It is currently possible to direct paper-based whole scripts to examiners based on their specialisms. Whatever the mode of marking, there are financial and resource costs associated with this facility. Before committing these resources it would be prudent to investigate the likely effects on marking reliability of allocating responses to examiners based on their specialisms.

A model of reliability and specialisms

GCE English Literature (specification B)

To gain an insight into the effects of examiners' specialisms on mark remark reliability, data from one A2 component in GCE English Literature (specification B) (LITB3) were explored using multilevel modelling (the model is given in Appendix A). The data were taken from essay responses, marked in summer 2013, that were dependent on prescribed texts. They represented an opportunistic sample of responses that were remarked for quality monitoring purposes. As this component was whole script marked on paper, monitoring samples were taken at two points during the marking period. The first sample of ten scripts was taken up to two days after online standardisation to check that training had been effective. If, after re-marking ten scripts, an examiner's marking was in doubt, the senior examiner requested a further ten scripts. Then, approximately half way through marking, the second sample was taken to ensure marking was still accurate. Each examiner selected fifty scripts to send to a senior examiner. The senior examiner re-marked fifteen of these scripts at random. If the junior examiner's marking was in doubt, a further ten scripts were re-marked.

The optional essay rubric

Candidates for LITB3 were required to answer two 40-mark essay questions, one from each section of the paper. Section A and section B offered a choice of eighteen questions and six questions, respectively. In both sections one half of the questions were based on texts from the Gothic genre and the other half were based on texts from the Pastoral genre (see Appendix B for a full list). Only questions in section A were based on specific texts. Questions in section B were more general in nature requiring candidates to consider a number of texts as they developed an argument. It would be necessary to view each script to identify how many and which texts candidates referred to in their responses to section B. Despite this, data from section B were used in this analysis as described below. According to the specification, candidates are expected to study at least three texts from one of the two genres (na, 2007).

Examiner specialisms – existing data

The particular component chosen for this analysis was part of a trial to match scripts to examiners' specialisms that was conducted by the Assessment Resourcing department in summer 2013. Examiners were surveyed as to which texts they had taught and which texts they felt confident about marking. Centres were requested to submit details about the texts their students were being taught and were expected to respond to in the examination. Examiners were then allocated scripts from centres which were a match for their specialisms. This attempt at matching examiners to their specialisms may be a confounding factor in the investigation reported here. On the other hand, a higher than random match rate between examiners and specialisms may yield effects that are measureable.

However, information elicited from the questionnaires was not used to assign specialisms to examiners in the current study because self-report data can be unreliable, particularly when there may be ulterior motives underlying the responses. Also, in this case, the data were incomplete as not all examiners and centres responded to the surveys. Therefore, an alternative approach to eliciting examiner specialism was pursued. It used recruitment and results data available in AQA's Examinations Processing System (EPS).

Home centres and examiners' specialisms – an alternative approach

Examiners' specialisms were determined by associating each examiner with a 'home centre' that had entries for the same component. AQA's EPS holds information about centres that currently employ or have employed an examiner within the last four years. Operationally, the contact centre number and previous centre numbers are used to avoid including in an examiner's allocation scripts from centres in which the examiner may have an interest.

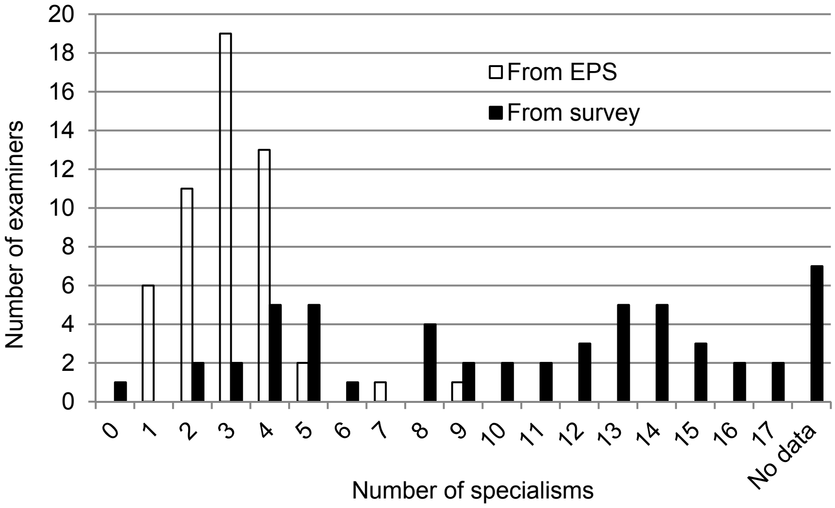
Only examiners who could be associated with a home centre were sampled for this investigation. Almost two-thirds of examiners (63.1% or 53 out of 84) could be associated with a home centre that had entries for the AQA specification in summer 2013.

Next, item-level data provided information about the two questions to which each candidate in an examiner's home centre responded. The same questions in the examiner's monitoring samples were flagged as being marked by a specialist. This flag became a binary variable affixed to responses from section A and section B of the question paper.

With section B questions, there is a necessary assumption that if any text of a given genre (Gothic or Pastoral) in section A, is an examiner's specialism, then that examiner will be a specialist with respect to a response of that genre in section B, provided candidates from their home centre responded to the same question. The risk in making this assumption is that some responses may be categorised as being specialisms of the examiner who marked them when they are not or, at least, not entirely. The opposite may also be true, that some responses are categorised as not being an examiner's specialism when they are. This risk is felt to be outweighed by the benefits of the larger dataset and offset by the fact that section B is promoted as being skills based (for example, making connections between texts).

A continuous variable that quantified the extent of an examiner's specialisms was created using only the responses to questions in section A. Using the home centre information stored in EPS, the number of specialisms ranged from one to nine with a mode of three. Most examiners (92.5%) had four or fewer specialisms; see Figure 1. About a third (37.8% or 601 out of 1,591) of the responses from section A in the sample were a specialism of the examiner who marked them. Because it used information from section A of the question paper only, this method underestimated the number of specialisms. Nonetheless the number of specialisms was valid as an indication of relative coverage of texts taught, if not as an absolute measure.

Figure 1: Distributions of the number of specialisms calculated from two sources of data



In contrast, when the calculation of number of specialisms was based on the information from the survey, the distribution was much wider, ranging from zero to seventeen. Only half (48.5%) of the responses marked by the sampled examiners were their specialisms. This suggested a poor match rate between examiners and centres in the trial conducted last summer. In part, this is explained by response rates to the surveys being less than 100% (for example, seven out the 53 examiners in the sample did not respond).

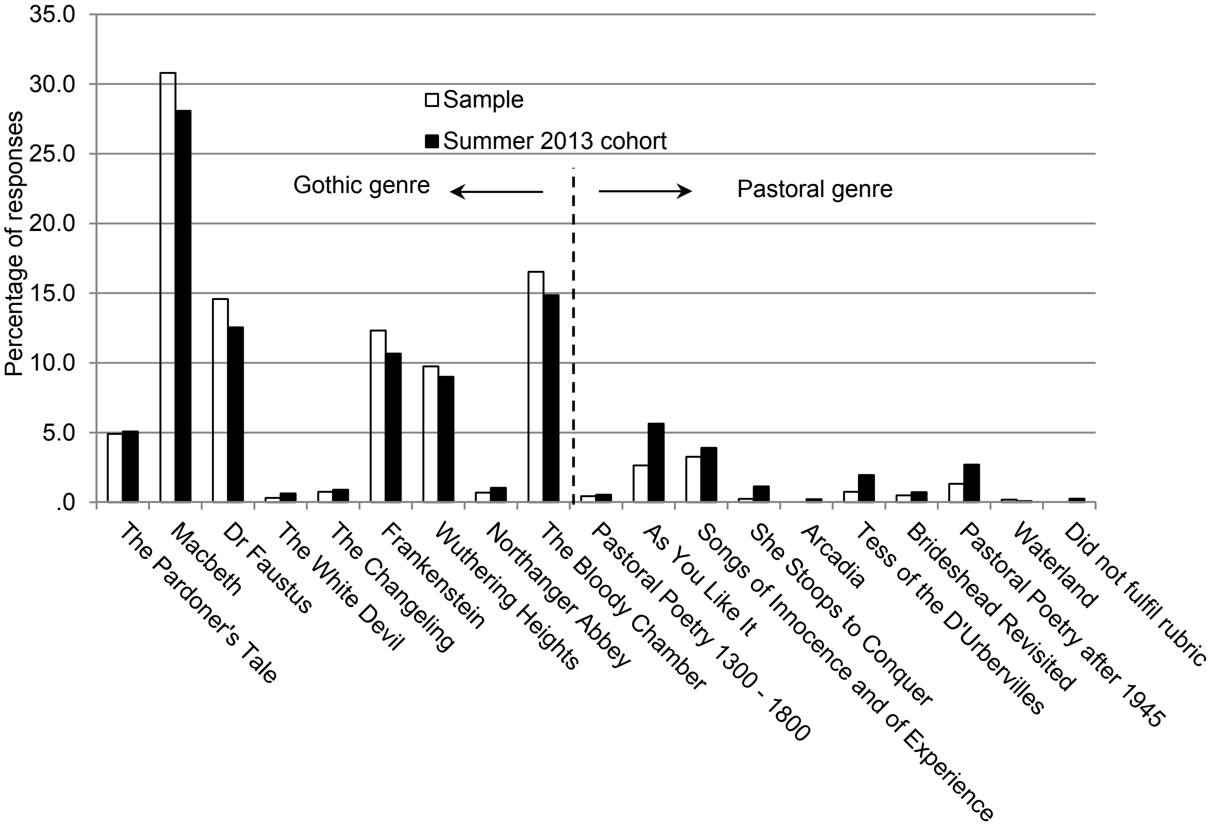
Figure 2 shows the percentages of responses to each of the texts made by the entire summer 2013 cohort (14,526 candidates) and those in the monitoring samples (1,591 responses to section A) marked by the 53 examiners in this study. The five most popular texts are slightly over-represented in the sample. Consequently, the least popular texts are under-represented, particularly those in the Pastoral genre, with *Arcadia* being absent entirely.

As many data as possible were retained in the sample so that the likelihood of observing any effects was maximised. Consequently, data from the first phase monitoring sample for all examiners were used. This included three examiners whose employment was terminated after the first monitoring sample as their marking was considered to be irretrievably deviant from the standard.

Data and model

Mark remark data from the first and second phase monitoring samples of the 53 examiners were modelled using the MLwiN software (Rasbash et al., 2000). For simplicity one model is presented, but variations on this model were investigated and found to lead to the same discussion points. Following a hierarchical definition of the 'true' mark and marking reliability, the response variable was formulated as the absolute mark difference between the senior examiner's mark and the junior examiner's mark for a response.

Figure 2: Comparison of texts responded to by entire cohort and in sample



A total of 3,175 responses from both sections of the question paper were nested within 53 examiners to give a two-level linear model. The average number of responses per examiner was 59.91 with a standard deviation of 15.92. Main effects were fitted and interactions considered. However, the interactions did not provide additional information, so the model presented in Appendix A includes only the main effects.

The model included two independent variables to explain the possible effects of examiner specialism. The first variable was binary and indicated whether the response was on a text which was an examiner's specialism or not. Just over half (52.7% or 1,673 of 3,175) of the responses were an examiner's specialism. The second was a continuous variable measuring the number of specialisms for each examiner as described in the previous section.

A further five independent variables were included in the model as follows.

- A binary variable indicating whether the absolute mark differences were from responses in the first phase monitoring sample or from the second phase monitoring sample. Small changes in marking accuracy over time have been demonstrated in the past (Pinot de Moira, Massey, Baird, & Morrissy, 2002; Pinot de Moira, 2013) as have training effects (Suto et al., 2011).

- A binary variable indicating whether the examiner provided either both monitoring samples or only the first sample so as to control for extremes of examiner performance. Recall that three examiners were stopped from marking after the first phase because of poor performance.
- A categorical variable for optional question number because examiners may experience reduced performance due to a lack of engagement with a text in the same way that students can.
- The question facility as a continuous variable because previous work has suggested that difficult items may be more difficult to mark (Sweiry, 2012).
- The continuous variable total script mark because Pinot de Moira (2013) suggested that, rather than the more difficult questions giving rise to low marking reliability, it may be the higher quality responses that lower reliability. Total script mark was used as an indicator of quality because the correlation between the marks for responses to the two sections was high. This variable was centred to allow findings to be presented relative to the mid-point of the mark scale.

Findings

Explained variation

After fitting the main effects, most of the variation in the random part of the model was at the level of the responses; see Table 1. In fact, the variation at response level dominated the variation at examiner level with an intraclass correlation of 11.10%. As previously observed by Pinot de Moira (2013), in multilevel modeling of features of a mark scheme, marking reliability may be limited by the idiosyncratic, but acceptable, nature of candidates' responses. Baird *et al.* (2013) took a slightly different perspective on the much greater variance residing at the lowest level of the model, attributing it to the interaction between examiners and responses: "the errors were largely due to idiosyncratic marker responses to individual parts of candidates' performances."

Despite the limitations of the model presented here, and the variations on it that were considered, some of the independent variables affected the dependent variable in a consistent manner. Not all independent variables showed a statistically significant effect. Therefore, the interpretations that follow should be treated with some caution.

Table 1: Estimates for the two-level model

	β	se(β)	p
Cons	5.962	0.20	**0.00**
Question choice	-0.013	0.01	0.24
Is a specialism	0.024	0.13	0.85
Extent of specialisms	-0.273	0.10	**0.01**
Centred total script mark	0.014	0.00	**0.00**
Second phase sample	-0.405	0.10	**0.00**
First phase sample only examiners	1.804	0.68	**0.01**
Facility index	-0.057	0.11	0.62
Examiner	0.92	0.20	**0.00**
Response	7.33	0.19	**0.00**

Convergence RIGLS; Explained variation R^2 = 3.25% (Snijders & Bosker, 1999)

Examiners' specialisms

The number of specialisms an examiner had (or that his or her centre offered) appeared to exert a statistically significant influence on marking reliability. For an increase of one in the number of specialisms, the absolute mark difference decreased by 0.273 marks. An examiner from a centre offering nine specialisms could be expected to have a mean absolute mark difference some 2.184 marks lower than the mean mark difference of an examiner from a centre apparently offering one specialism. Thus, the fewer the number of specialisms an examiner had experience of, the more difficulties they experienced with their marking. Or, conversely, the broader the teaching delivered by the examiner's centre, the fewer difficulties he or she experienced with their marking.

The number of texts that candidates from an examiner's home centre responded to in the examination was used as a proxy for the number of specialisms that examiner had. As noted above, this number is likely to be an underestimate.

In contrast to the number of specialisms an examiner has, whether or not a response was an examiner's specialism appeared to have little influence on marking reliability and was without statistical significance. This was so with or without Extent of specialisms in the model. Thus it appears that exposure to a broad range of texts is more likely to improve marking reliability than does marking a particular text.

It is suggested here that the higher the number of texts offered in an examiner's home centre, then the greater the focus on the skills of interpretation, not just for the students but for the examiner too. Experience of many texts may enable an examiner to appreciate the skills in common necessary to analyse, interpret and evaluate a range of texts. At the opposite end of the spectrum, an examiner working in a home centre that offers few texts may place their attention on favoured interpretations of a text rather than skills or may not yet have appreciated that the necessary skills can be applied consistently across all texts.

There may also be a parallel here with the findings of Suto and Nádas (2008). They argued that the level of general academic ability found in the graduate population, from which examiners are drawn, was both necessary and sufficient to be able to use a mark scheme as it was intended. In other words, examiners in possession of an education that involved generic skills of analysis, evaluation and interpretation that could be transferred between subjects experienced fewer difficulties in marking.

In their study focussing on whether college-level history students in the USA made the best choices when responding to optional questions, Bridgeman, Morgan and Wang (1996) made two recommendations that are relevant to the current discussion. They advocated the development of mark schemes that are applicable to all optional questions in a question paper such that the "quantity and quality of required evidence for a particular score should be consistent." They also strongly favoured examiners marking across topics even if this meant marking a topic in which they did not consider themselves a specialist.

> *"There must be a single group of raters that establishes a single standard. Even though raters may feel more comfortable specializing in a particular topic, consideration should be given to having all raters read all topics. This would minimize the chances that the readers of a particular topic will inadvertently start using criteria that are more or less lenient than the criteria used by raters of other topics."* Bridgeman, Morgan and Wang (1996)

LITB3 has generic mark schemes for both sections of the question paper with small amounts of indicative content. This meets Bridgeman *et al.'s* (1996) call for consistent evidence across optional texts to be rewarded the same mark.

How other variables influenced marking reliability

The key finding, that the number of examiners' specialisms has a statistically significant effect on marking reliability, was made whilst controlling for other factors, which are described in the 'Data and Model' section. Observations on the effects of these variables are supported by the existing literature as discussed in this section.

Quality of response versus difficulty of question

Both the centred final mark and the facility index appeared to influence the marking reliability, albeit in opposite directions. Centred final mark did so with statistical significance and facility index without. As the centred final mark increased the absolute mark difference also increased. This variable was included in the model to give an overall indication of the quality of responses. Though its influence was small in the current model, the trend of better responses being problematic to mark is similar to Pinot de Moira's (2013) finding when she modelled the effects of different features of a mark scheme on marking reliability.

The absolute mark difference decreased as the facility index increased. Or, as item difficulty for candidates increased, the marking reliability decreased. Sweiry (2012) found the same, though with a lower level of assessment and much lower question tariffs.

Post-standardisation training effect

Marking reliability was higher when the response marked was from an examiner's second phase monitoring sample. The extent of this increased reliability appeared to be of the order of 0.4 of a mark. Using data taken from the first and second phase monitoring samples is akin to taking snapshots at the beginning and towards the middle of the marking period. Ideally, there would be no change in marking accuracy between the two monitoring samples. This, however, assumes that examiners all achieve the marking standard immediately after standardisation, which is not always the case (Suto et al., 2011). Nonetheless an increase in accuracy is better than a decrease and may be attributed to the additional guidance offered as part of the first phase monitoring by senior examiners post-standardisation.

Related to this is that being an examiner with highly deviant marking decreased marking reliability statistically significantly. Those examiners who ceased marking after the first phase sample induced an increase in the absolute mark difference of about 1.8 marks.

Conclusions

Limitations of the data and the model

This exploration of the effects of examiner specialisms on marking reliability had a number of limitations. Not least of these was that it considers only one component and therefore lacks generalisability. The effects observed for LITB3 may not be present in other components or, indeed, other components may exhibit an effect of specialism that LITB3 does not. Extending this study to more components would increase the number of examiners in the higher level(s) of the model. This would increase the power of the model (McCoach, 2010).

Including more components would also mean considering a range of rubrics and mark schemes related to optional questions. Some of the rubrics may provide less conservative information about the number of specialisms examiners' schools and colleges offer their students than did LITB3's. The information, though, would still be relatively difficult to access and would be retrospective. It may, also, not be reliable enough for operational use in future examination series. Different styles of mark scheme (for example, lacking a generic marking grid) may yield results that are dissimilar to those reported here.

The method of associating examiners with a home centre described in this paper relies on accurate, up-to-date information being stored in EPS. The validity of the assumption that an examiner was involved in the teaching of all the texts on which the candidates from the home centre answered questions needs to be checked. In home centres with smaller cohorts this is likely to be the reality. When a larger cohort is involved the examiner may have had direct involvement in the teaching of only some of the texts assigned as specialisms. However, he or she would probably have taken part in planning the curriculum for the subject and developing a consistent approach to its teaching.

Though flawed, using the information stored in EPS uses fewer resources than surveying examiners and centres and matching up the information from the two sources. The survey data are reliant on a high rate of return and self-reporting. Additionally, matching the texts examiners said they had taught in the survey with what the questions candidates from their home centre responded to suggested the match rate between examiners and scripts in the trial was less than 50%.

In previous studies, as well as the current one, preparation of data from monitoring samples has been a bottleneck. The data were keyed manually, which took about seven and a half person days. In future, based on code generated by CERP's placement student (Maunders, 2014), it will be possible to automate the extraction of these data from the Examiner Extranet. For a similar volume of data the preparation time would decrease by approximately 95%.

Recommendations

The following three recommendations are made with the component LITB3 in mind. However, they could, tentatively, be extended to other components that contain optional essay questions.

Firstly, targeting texts by examiner specialism is unnecessary as it is unlikely to result in meaningful improvements to marking reliability. Secondly, if feasible, the employment of examiners who have experienced a range of specialisms in their working environment should be encouraged. Such examiners appear to experience fewer difficulties with their marking. This recommendation may be difficult to implement and there will always be a broad distribution of examiner backgrounds. Therefore, the final recommendation is to encourage a focus, at standardisation and in the mark scheme, on rewarding the skills that candidates need to demonstrate regardless of their choice of optional essay.

Future work

The topic of examiner specialisms is only one strand of research into how examiners' expertise affects marking reliability. Further research could extend our knowledge of the effects of examiner specialisms by exploring the different rubrics and mark schemes of other components. Or, it could shift the focus to factors within examiners' working environments that may influence their marking reliability.

There is a school of thought that suggests that effective schools (however effective may be defined) offer a broad range of teaching to their students. This has been articulated as "curriculum extension" (Rudd, Aiston, Davies, Rickinson, & Dartnall, 2002) and "Ensuring high curriculum coverage or opportunities to learn" (Calman, 2010). In the current study, it is noticeable that examiners from sixth form colleges have, on average 5.2 specialisms, almost twice as many as those from academies and secondary comprehensives who have an average of 2.9 specialisms. It may be that centres' strategies for the depth and breadth of their curricula affect examiner performance; thus making this area a rich vein for future research.

References

Baird, J.-A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability. A comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling.* Ofqual/13/5261. University of Oxford: Oxford University Centre for Educational Assessment.

Bridgeman, B., Morgan, R., & Wang, M. (1996). *The reliability of document-based essay questions on advanced placement history examinations.* RR-96-5. Princeton, New Jersey: Educational Testing Service.

Calman, R. C. (2010). *Exploring the Underlying Traits of High-Performing Schools*. Ontario, Canada: Education Quality and Accountability Office.

Maunders, N. (2014). *Extracting first phase and second phase sample data from the examiner extranet.* (No. CERP_MO_NM_30042014). Manchester, UK: AQA Centre for Education Research and Practice.

McCoach, D. B. (2010). Hierarchical Linear Modeling. In *G.R. Hancock & R.O. Mueller (Eds), The Reviewer's Guide to Quantitative Methods in the Social Sciences.* New York and London: Routledge.

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English.* RPA_10_MM_RP_028. Manchester, UK: AQA Centre for Education Research and Practice.

Messick, S. (1989). Validity. In *R.L. Linn (Ed.) Educational Measurement.* 3rd ed., pp. 13–103. New York: Macmillan.

na. (2007). GCE AS and A level Specification English Literature B. AS exams 2009 onwards. A2 exams 2010 onwards. (version 1.4). AQA: Manchester, UK. Retrieved from http://filestore.aqa.org.uk/subjects/AQA-2745-W-SP-10.PDF on 7 May 2014.

na. (2012a). *Education Committee - First Report. The Administration of examinations for 15-19 year olds in England.* London: The Stationary Office.

na. (2012b). *England's "examinations industry": deterioration and decay. A report from HMC on endemic problems with marking, awarding, re-marks and appeals at GCSE and A level, 2007-12.* Market Harborough: Headmasters' and Headmistresses' Conference.

na. (2013). GCE (A-level) June 2013 English Literature B LITB3 (Specification 2745) Unit 3: Texts and Genres Final mark Scheme. AQA: Manchester, UK.

Opposs, D., & He, Q. (2011). *The Reliability Programme. Final report.* Ofqual/11/4828. Coventry: Ofqual.

Pinot de Moira, A. (2013). *Features of a Levels-Based Mark Scheme and their Effect on Marking Reliability.* CERP_TR_APM_03042013. Manchester, UK: AQA Centre for Education Research and Practice.

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

Pointer, W. H. (2013). *A Rasch analysis of the quality of marking of GCSE Science.* CERP_TR_WHP_05112013. Manchester, UK: AQA Centre for Education Research and Practice.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., … Lewis, T. (2000). *A user'sguide to MLwiN. Version 2.1.* University of London: Multilevel Models Project, Institute of Education.

Royal-Dawson, L., & Baird, J.-A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice*, *28*(2), 2–8.

Rudd, P., Aiston, S., Davies, D., Rickinson, M., & Dartnall, L. (2002). *High Performing Specialist Schools: What Makes the Difference?* Slough: NFER.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage Publications.

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

Suto, I., Nádas, R., & Bell. (2011). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*, *26*(1), 21–51.

Sweiry, E. (2012). Conceptualising and minimising marking demand in selected and constructed response test questions. Presented at the Association for Educational Assessment (AEA) Europe Annual Conference, Berlin, Germany.

Appendix A Details of model

$$\text{absmkdiff}_{\textit{responseid, examiner}} \sim \text{N}(XB,\ \Omega)$$

$$\begin{aligned}\text{absmkdiff}_{\textit{responseid, examiner}} = {} & \beta_{0\textit{responseid, examiner}}\text{cons} + \text{-}0.013(0.011)\text{question}_{\textit{responseid, examiner}} + \\ & 0.024(0.131)\text{Is a specialism}_{\textit{responseid, examiner}} + \text{-}0.273(0.099)\text{extentspeccon}_{\textit{examiner}} + \\ & 0.014(0.003)\text{totalmarkcentred}_{\textit{responseid, examiner}} + \\ & \text{-}0.405(0.099)\text{Second phase sample}_{\textit{responseid, examiner}} + \\ & 1.804(0.675)\text{First phase sample only}_{\textit{examiner}} + \text{-}0.057(0.114)\text{facility}_{\textit{responseid, examiner}}\end{aligned}$$

$$\beta_{0\textit{responseid, examiner}} = 5.962(5.774) + u_{0\textit{examiner}} + e_{0\textit{responseid, examiner}}$$

$$\left[u_{0\textit{examiner}}\right] \sim \text{N}(0,\ \Omega_u)\ :\ \Omega_u = \left[0.916(0.204)\right]$$

$$\left[e_{0\textit{responseid, examiner}}\right] \sim \text{N}(0,\ \Omega_e)\ :\ \Omega_e = \left[7.333(0.186)\right]$$

*-2*loglikelihood(IGLS Deviance)* = 15439.720(3175 of 3175 cases in use)

Appendix B

Table B1: Texts used in GCE English Literature (specification B) examination (LITB3) in summer 2013

Genre	Title	Author
Gothic	*The Pardoner's Tale**	Geoffrey Chaucer
	*Macbeth**	William Shakespeare
	*Dr Faustus**	Christopher Marlowe
	*The White Devil**	John Webster
	The Changeling	Thomas Middleton & William Rowley
	Frankenstein	Mary Shelley
	Wuthering Heights	Emily Brontë
	Northanger Abbey	Jane Austen
	The Bloody Chamber	Angela Carter
Pastoral	*Pastoral Poetry 1300 – 1800**	Various
	*As You Like It**	William Shakespeare
	*Songs of Innocence and of Experience**	William Blake
	*She Stoops to Conquer**	Oliver Goldsmith
	Arcadia	Tom Stoppard
	Tess of the D'Urbervilles	Thomas Hardy
	Brideshead Revisited	Evelyn Waugh
	Pastoral Poetry after 1945	Various
	Waterland	Graham Swift

Texts marked with an asterisk (*) are from the period 1300 – 1800.