communicative function due to the fact that they might represent the fluid thoughts of an examiner at a point in time during decision making, containing tacit features that support examiner thinking, and leading to them being difficult to infer meaning from. It is clear that these characteristics could limit the ability of someone to use the annotations at face value to make valid inferences about an assessed performance.

Teachers were more likely than teacher-examiners to expect annotations to provide information that could be used for formative purposes (e.g. showing explicitly where a performance could be improved). This difference in perspective is potentially important since it affects the degree to which annotations should be expected to function as tools to support transparent communication. Since examiner annotations are primarily concerned with the functions of supporting examiner thinking and communicating the reasoning behind a judgement, formative annotating is an extraneous purpose which would possibly confound the primary function of the activity and would therefore be inadvisable. In order to mitigate potentially invalid actions based on script annotations, it is advisable that teachers and candidates are informed about why it would be inappropriate for examiners to make formative annotations on scripts.

Despite the inevitably individualised characteristics of examiner annotations there is still scope for the meanings of annotations to be made more explicit to those who have access to them. This is as true for examiners who are engaged in marking a particular examination paper as it is for the teachers who can read the annotations when they access requested scripts. The inclusion of abbreviated annotation terms and shared meanings might be a useful addition to mark schemes but it is very important to recognise that this is only of superficial importance compared with the insights gained from annotations when teachers have a deep understanding of the mark scheme.

This project contributes to a growing understanding of how annotations function and suggests that the primary concern should be that annotation use be fit for purpose. Whilst validity requires that information relating to an assessment is as transparent as possible, and annotations can assist in this process, it is also important to make the limits of annotations explicit to those who receive them on returned scripts.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC.: American Educational Research Association.

Cambridge Assessment (2009). *The Cambridge Approach: Principles for designing, administering and evaluating assessment*. Cambridge: A Cambridge Assessment Publication.

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

Engeström, Y. (2001). Expansive Learning at Work: toward an activity theoretical reconceptualization. *Journal of Education and Work*, **14**, 1, 133–156.

Johnson, M. & Nádas, R. (2009). Marginalised behaviour: digital annotations, spatial encoding and the implications for reading comprehension. *Learning, Media and Technology*, **34**, 4, 323–336.

Johnson, M. & Shaw, S. (2008). Annotating to comprehend: a marginalised activity? *Research Matters: A Cambridge Assessment Publication*, **6**, 19–24.

Lave, J. & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Mislevy, R.J., Behrens, J.T., Bennett, R.E., Demark, S.F., Frezzo, D.C., Levy, R., Robinson, D.H., Rutstein, D.W., Shute, V.J., Stanley, K. & Winters, F.I. (2007). *On the roles of external knowledge representations in assessment design*. University of Maryland: National Center for Research on Evaluation, Standards, and Student Testing.

Shaw, S. & Johnson, M. (2009). *Annotating on-screen: the influence of reading environment on annotative practice and assessor comprehension building*. A paper presented at the International Association for Educational Assessment Annual Conference, Brisbane, September.

ASSURING QUALITY IN ASSESSMENT

Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology

Nicholas Raikes, Jane Fidler and Tim Gill Research Division

This article is based on a paper presented to the annual conference of the British Educational Research Association held in Manchester, UK, in September 2009.

Summary

When high stakes examinations are marked by a panel of examiners, the examiners must be standardised so that candidates are not advantaged or disadvantaged according to which examiner marks their work.

It is common practice for awarding bodies' standardisation processes to include a 'standardisation' or 'co-ordination' meeting, where all examiners meet to be briefed by the Principal Examiner and to discuss the application of the mark scheme in relation to specific examples of candidates' work. Research into the effectiveness of standardisation meetings has cast doubt on their usefulness, however, at least for experienced examiners.

In the present study we addressed the following research questions:

1. What is the effect on marking accuracy of including a face-to-face meeting as part of an examiner standardisation process?

2. How does the effect on marking accuracy of a face-to-face meeting vary with the type of question being marked (short-answer or essay) and the level of experience of the examiners?
3. To what extent do examiners carry forward standardisation on one set of questions to a different but very similar set of questions?

We found that while standardisation improved marking accuracy for both new and experienced examiners, marking both short-answers and structured, factual essays, the benefit of including a face-to-face meeting in the standardisation process was variable, small and questionable. We also found that the effects of standardisation on one set of questions – with or without a meeting – carried forward into improved marking accuracy on other, very similar questions, implying that some transferable examiner learning had taken place and that the impact of – and need for – standardisation might decrease with examiner experience.

We concluded that it would be reasonable for examining bodies to explore whether standardisation can be achieved using more cost-effective and efficient methods than face-to-face meetings.

Background

The regulatory authorities for public examinations in England, Wales and Northern Ireland prescribe that awarding bodies must have a standardisation process that is "designed to make sure that all examiners mark candidates' work consistently and accurately [and which] establishes a common standard of marking that should be used to maintain the quality of marking during the marking period." (Qualifications and Curriculum Authority, 2009, section 4.14).

Research into the effectiveness of standardisation meetings has cast doubt on their usefulness, at least for experienced examiners. For example, Baird *et al*. (2004) found neither consensual meetings – where the examiners mutually agreed a common interpretation of the mark scheme – nor hierarchical meetings, where the Principal Examiner tried to impose his interpretation of the mark scheme on to the other examiners, improved the marking reliability of experienced GCSE History examiners. Similarly, Greatorex and Bell (2008) found that a standardisation meeting on its own had little effect on the reliability of experienced examiners of AS Biology. Greatorex *et al*. (2007) compared the pre- and post-standardisation meeting marking accuracy of experienced examiners of GCSE mathematics and physics with that of mathematics and physics graduates who lacked both teaching and examining experience and who would therefore not normally have been eligible to mark the examinations. They found that for the questions that the researchers had previously judged to entail more complex cognitive marking strategies, the standardisation meeting led to a much greater improvement of the graduates' accuracy than of the experienced examiners' accuracy. However, the improvement shown by graduates might also have occurred if other standardisation methods had been used, and might not be dependent on a standardisation *meeting* being held.

Method

Choice of examination

Two A-Level psychology units were chosen for the research, one assessed using short-answer questions, the other assessed using essay questions. We chose A-Level psychology because this subject uses both these types of question and because there is a large entry and correspondingly large pool of examiners.

Choice of examination questions

The short-answer examination we selected contained a number of discrete sections, each of which consisted of compulsory questions on a single topic. Two of the sections had identically structured questions, and by selecting these sections for the study and standardising examiners on only one of them, we could investigate the extent to which standardisation on one set of short answer questions carried over to other very similar questions answered by the same candidates. This would help us understand whether generic marking skills were developed through standardisation that lessened the impact of and need for standardisation in subsequent sessions, as examiners gained experience.

The essay examination gave candidates a choice of questions, so each question was answered by a different sub-group of candidates. We therefore investigated the carrying-forward of standardisation using essays from examinations held in consecutive years, selecting the closest matching questions for use in the study (question 4 in each case).

Some details concerning the chosen questions are given below:

Short answer questions

Questions which required candidates to write a sentence or two.

Short-Answer Collection 1 *Examiners were standardised on these*		*Short-Answer Collection 2* *Examiners were not standardised on these*	
Topic: Cognitive Psychology		Topic: Social Psychology	
Question	Mark tariff	Question	Mark tariff
1, 2a, 2b & 3	2 each	13, 14a, 14b, 15	2 each
4	4	16 4	

Essay questions

Questions which required candidates to write a page or two of factual information.

Essay Collection 1 *Examiners were standardised on these*		*Essay Collection 2* *Examiners were not standardised on these*	
Examination 1		Examination 2	
Question	Mark tariff	Question	Mark tariff
4a, 4b	12 each	4a, 4b 12 each	

Participants

Twenty-four psychology examiners were recruited for the study, none of whom had operationally-marked the examinations. Twelve of the examiners had experience of marking other psychology A-Level examinations; the other twelve examiners were new to examining, having been recruited for operational work but not yet deployed.

The examiners were randomly assigned to experimental groups of six as follows:

	New Examiners	Experienced Examiners
Attends standardisation meeting	Group A1	Group B1
No meeting	Group A2	Group B2

In addition to these twenty-four examiners, two Team Leaders from the operational examinations were recruited, one from the short-answer examination, the other from the essay examination. These Team Leaders had each been responsible for supervising a team of examiners in the operational marking and were chosen based on the recommendations of the Principal Examiners and Professional Officer.

The role of the Team Leaders in the study was to standardise the other examiners and to provide reference marks for each answer against which the examiners' marks could be compared.

Overview of the sequence of events for Examiners

1. *Examiners marked pre-standardisation batches of scripts.*
   The marks from these scripts were used to calculate the examiners' pre-standardisation marking accuracies on each collection of questions (in relation to the Team Leaders' reference marks).
2. *Examiners were standardised, with or without a meeting according to their experimental group.*
3. *Examiners marked post-standardisation batches of scripts.*
   The marks from these were used to calculate the examiners' post-standardisation marking accuracies on each collection of questions (again in relation to the Team Leaders' reference marks).

Materials

Scripts

A random sample of scripts, stratified by grade, was drawn from the operational examinations once all marking and grading were complete.

The scripts were scanned and the marks and examiner annotations electronically deleted from the resulting images. The images relating to the questions chosen for use in the study were then printed out to give 'clean' hard copies. All participants marked the same answers, so twenty-six copies were printed.

The clean answers were divided into a number of batches, as shown below. The answers used in standardisation were selected by the Team Leaders. The pre- and post-standardisation batches were selected by the researchers and were matched by operational marks, so that the pre-and post- batches were as similar as possible.

Pre-standardisation batches:

Batch **Short-1-Pre** 50 answers to each question in Short-Answer Collection 1	Batch **Essay-1-Pre** 25 answers to each question in Essay Collection 1	Examiners were to be standardised on these questions
Batch **Short-2-Pre** 50 answers to each question in Short-Answer Collection 2	Batch **Essay-2-Pre** 25 answers to each question in Essay Collection 2	Examiners were **not** to be standardised on these questions

Batches for use in standardisation:
(Question collections 1 only. The standardisation procedure, described below, required three standardisation batches of each answer type)

Batch **Short-Stand-i** 5 answers to each question in Short-Answer Collection 1	Batch **Essay-Stand-i** 5 answers to each question in Essay Collection 1
Batch **Short-Stand-ii** 5 answers to each question in Short-Answer Collection 1	Batch **Essay-Stand-ii** 5 answers to each question in Essay Collection 1
Batch **Short-Stand-iii** 10 answers to each question in Short-Answer Collection 1	Batch **Essay-Stand-iii** 10 answers to each question in Essay Collection 1

Post-standardisation batches:

Batch **Short-1-Post** 50 answers to each question in Short-Answer Collection 1	Batch **Essay-1-Post** 25 answers to each question in Essay Collection 1	Examiners were standardised on these questions
Batch **Short-2-Post** 50 answers to each question in Short-Answer Collection 2	Batch **Essay-2-Post** 25 answers to each question in Essay Collection 2	Examiners were **not** standardised on these questions

Materials written by the Team Leaders

The Team Leaders were commissioned to write:

- an *Introduction to Marking* for new examiners;
- a *Mark scheme Rationale* explaining to examiners how the mark schemes for the chosen questions should be applied;
- written explanations for the marks they awarded to the first and second standardisation batches of short answers and essays. Copies of these would be placed in sealed envelopes for the examiners to open and read when directed, as described below under 'Experimental Procedure'.

Additional materials supplied to participants

- Copies of the question papers
- Copies of the relevant parts of the mark schemes
- Instructions

Experimental Procedure

Stage 1: Pre-standardisation

(1) The pre-standardisation batches were posted to the examiners, together with copies of the questions and mark schemes.

(2) Examiners were instructed to mark the pre-standardisation batches in the following order: Short-1-Pre first, then Essay-1-Pre, then Short-2-Pre, then Essay-2-Pre.

(3) Examiners returned their marked pre-standardisation batches.

(4) The remaining materials were posted to examiners.

Stage 2: Standardisation
The standardisation procedure was the same for all examiners, except for the inclusion of a standardisation meeting for examiners in experimental groups A1 and B1.

	Groups A1 & B1	*Groups A2 &B2*
(5)	All examiners were instructed to read *Introduction to Marking* and the questions, mark schemes and mark scheme rationale.	
(6)	All examiners marked batch Short-Stand-i, then opened the envelope containing the Team Leader's marks and explanations for Short-Stand-i. They were instructed to compare the Team Leader's marks with their own and read the explanations.	
(7)	All examiners marked batch Short-Stand-ii.	
(8)		A2 & B2 examiners opened the envelope containing the Team Leader's marks and explanations for batch Short-Stand-ii. They were instructed to compare the marks with their own and read the explanations.

(9)	All examiners marked batch Essay-Stand-i, opened the envelope containing the Team Leader's marks and explanations, compared the marks with their own and read the explanations.	
(10)	All examiners marked batch Essay-Stand-ii.	
(11)		A2 & B2 examiners opened the envelope containing the Team Leader's marks and explanations for batch Essay-Stand-ii. They were instructed to compare the marks with their own and read the explanations.
(12)	A1 & B1 examiners attended a standardisation meeting, at which their marking of Short-Stand-ii and Essay-Stand-ii was discussed and the correct marks provided and explained. At the end of the meeting the examiners were also supplied with copies of the written explanations and marks previously given to the non-meeting groups, so that all had the same materials.	
(13)	All examiners marked batches Short-Stand-iii and Essay-Stand-iii. They were instructed to enter their marks into spreadsheets and email them to the appropriate Team Leader.	
(14)	Team Leaders phoned each examiner individually to discuss their Stand-iii marking and answer questions.	

Stage 3: Post-standardisation

(15)	Examiners marked the post-standardisation scripts in the following order: Short-1-Post first, then Essay-1-Post, then Short-2-Post and finally Essay-2-Post.
(16)	Examiners returned all their marked scripts.

Additionally, the Team Leaders marked the pre- and post-standardisation batches to provide reference marks for use in the analysis. Each Team Leader marked only short answers or essays according to their specialism.

The standardisation meeting

Examiners in groups A1 and B1 attended a standardisation meeting in Cambridge, led by the two Team Leaders. After a preliminary welcome, a brief presentation was given by one of the Team Leaders recapping the material contained in the *Introduction to Marking* document. Consecutive sessions were then held for the short-answer and essay questions, each led by the appropriate Team Leader and conducted as similarly as possible to the operational standardisation meeting. During these sessions examiners went through the second standardisation batches and the Team Leader led a discussion of the examiners' initial marks and provided and explained the 'correct' marks. Examiners had ample opportunity to ask questions.

Analysis

The 'absolute difference' between each examiner's mark for an answer and the reference mark was calculated – this was simply the value obtained by subtracting examiner-mark from reference-mark and discarding the sign, that is, all were positive numbers. These absolute differences gave the size of the difference, and when averaged did not cancel out as actual differences might.

The mean absolute difference was calculated for each examiner on each question in the pre- and post-Standardisation collections. Means were also calculated at the level of experimental group, and batch.

Analysis of covariance (ANCOVA) was performed to test whether post-standardisation differences between the experimental groups were statistically significant, having controlled for pre-standardisation differences.

Results and discussion

The charts in this section show the pre- and post-standardisation mean absolute-difference between examiner-mark and reference-mark for each experimental group. The solid lines correspond to the results from the examiners who attended the meeting ('Face-to-face' standardisation type), the dotted lines to those from the examiners who did not attend the meeting ('Remote' standardisation type). Statistical significance information from the ANCOVA analyses are given underneath the charts, where ✓ indicates $p < 0.05$, i.e. where examiner experience, or standardisation type, or different combinations of these two factors ('interaction') resulted in statistically significant differences in post-standardisation absolute-differences.

The first thing to note from the charts is that in almost all cases standardisation had a beneficial effect in bringing examiners' marks closer to the reference marks, regardless of whether examiners attended the meeting. The ANCOVA analysis helps determine whether meeting attendance had an *additional* effect on marking accuracy, over and above that derived from undertaking the remote standardisation tasks, and whether this varied with examiner experience.

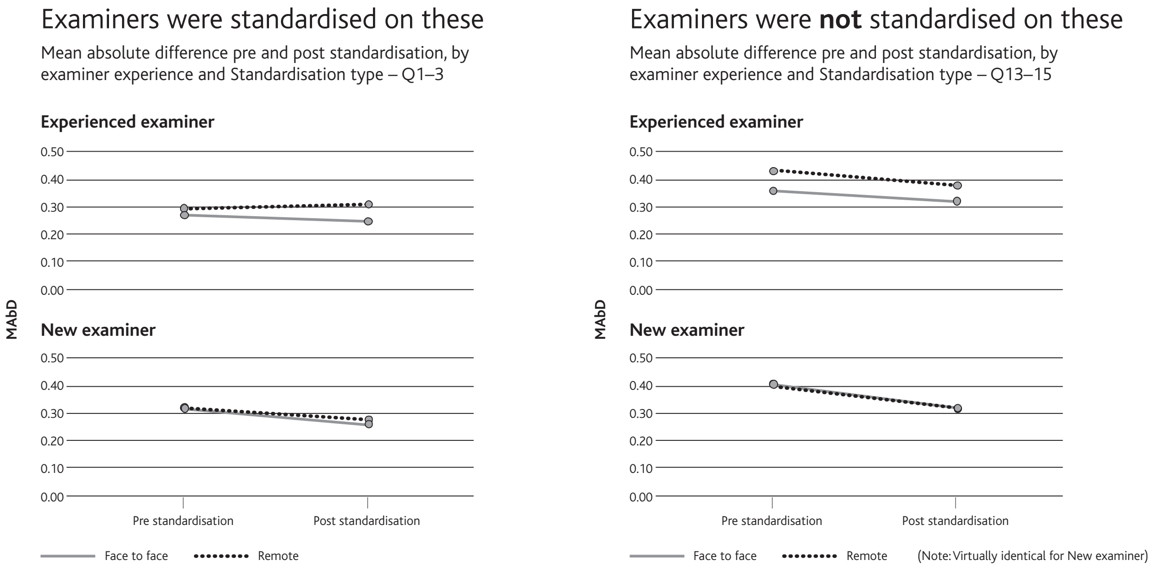
Short-answer questions

Figure 1 shows the pre- and post-standardisation mean absolute-differences for each experimental group on the 2-mark questions. The charts on the left show the results on the standardised questions, those on the right give the results on the un-standardised questions. In both cases the experienced examiners' results are presented in the top charts.

There was a slight but statistically significant benefit (in terms of reducing mean absolute differences) in attending the standardisation meeting for the standardised questions only. For the un-standardised questions, attending the meeting did not provide a general significant benefit, but there was a significant but very small interaction between standardisation type and examiner experience: from the diagrams it is apparent that there is no difference between the lines for the new examiners, but those for the experienced examiners are a little less than parallel.

Figure 2 shows the results for the 4-mark question. Clearly standardisation had unintended consequences for question 4: marking accuracy worsened! This is the only question for which this is the case. Examiner experience had a significant effect, with the experienced examiners' accuracy worsening slightly less; attending the meeting had a particularly negative effect on the new examiners. On question 16, the 4-mark question on which examiners were not standardised, meeting attendance resulted in a very slight, but statistically significant, improvement.

Figure 1: 2-mark questions


Examiners were standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q1–3

Examiners were **not** standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q13–15

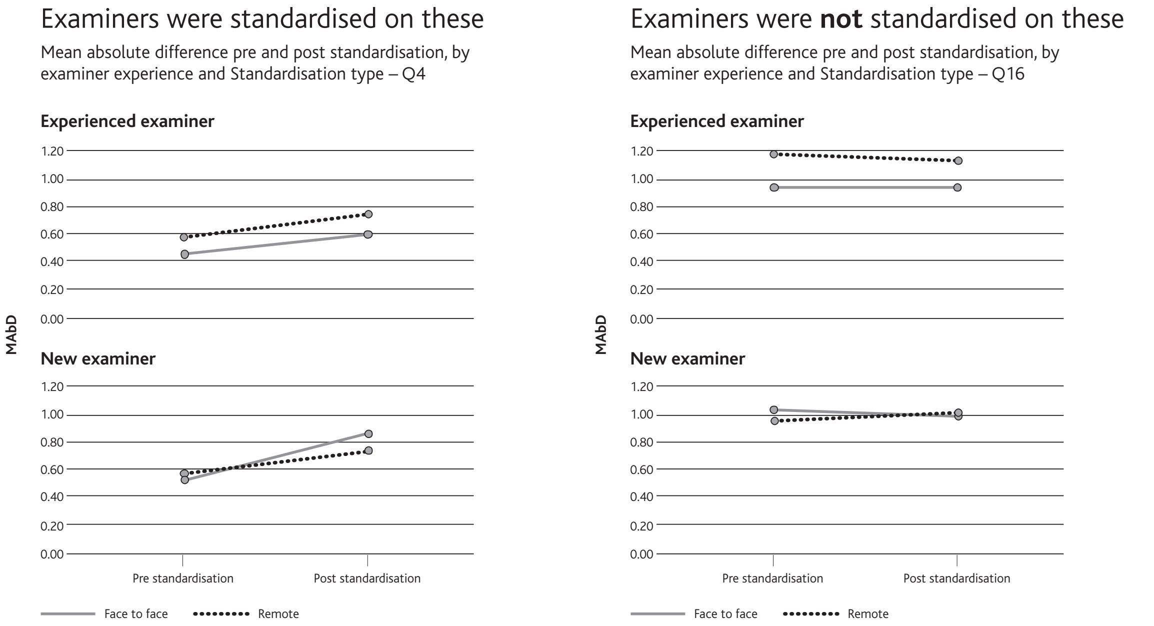
Statistical significance

Examiner experience	✕	p=.710
Standardisation type	✓	p=.003
Interaction	✕	p=.138

Statistical significance

Examiner experience	✘	p=.096
Standardisation type	✘	p=.084
Interaction	✔	p=.044

Figure 2: 4-mark question


Examiners were standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q4

Examiners were **not** standardised on these

Mean absolute difference pre and post standardisation, by examiner experience and Standardisation type – Q16

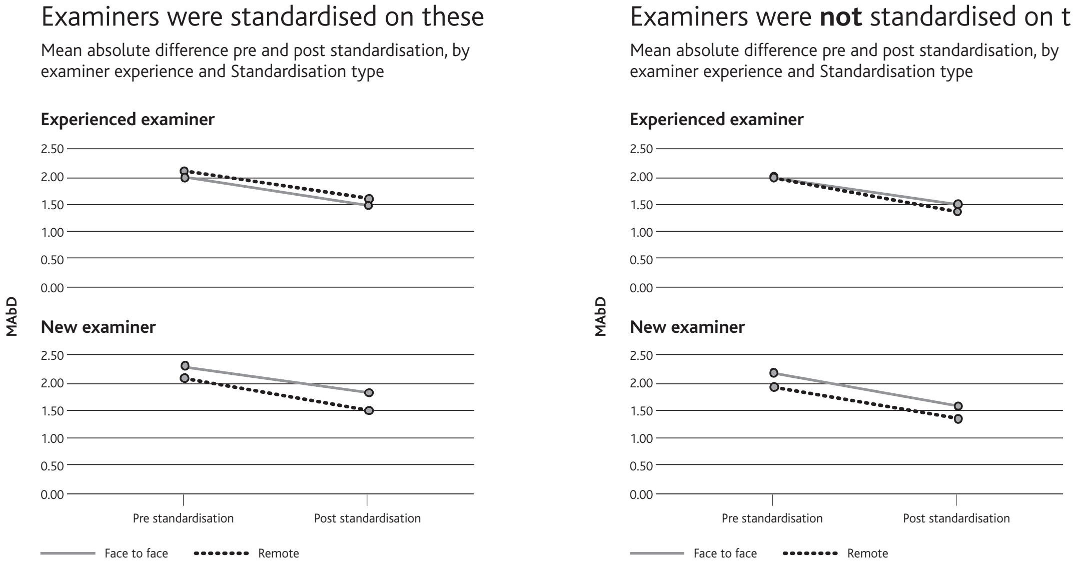
Statistical significance

Examiner experience	✔	p=.002
Standardisation type	✘	p=.947
Interaction	✔	p=.002

Statistical significance

Examiner experience	✘	p=.934
Standardisation type	✔	p=.040
Interaction	✘	p=.135

Figure 3: Essay questions



Statistical significance

Examiner experience	✘	p=.094
Standardisation type	✘	p=.282
Interaction	✔	p=.008

Statistical significance

Examiner experience	✘	p=.745
Standardisation type	✔	p=.045
Interaction	✘	p=.795

Figure 3 gives the results for the essay questions. Standardisation was clearly beneficial on both the standardised and non-standardised questions. Neither standardisation type nor examiner experience had a significant effect on the accuracy improvement on the standardised questions, but there was a significant interaction between these factors, with the remotely standardised new examiners improving more. On the un-standardised questions there was a statistically significant greater improvement for the remotely standardised examiners, with the chart suggesting that this greater improvement was shown mainly by the experienced examiners, though the interaction between experience and standardisation type was not significant.

Conclusions

On the basis of our results, we concluded that:

- Apart from the anomalous 4-mark question, standardisation improved the examiners' marking accuracy when compared with the reference marks, regardless of whether this standardisation was conducted purely remotely or with the addition of a face-to-face meeting.
- The standardisation improvement carried over into other, very similar questions, implying the examiners learnt lessons from being standardised that they were able to apply when marking other questions. This finding suggests the impact of – and need for – standardisation might reduce with examiner experience.
- Meeting attendance did not always have a statistically significant benefit, and where there was a benefit, it was very small in real terms. On the standardised questions, the meeting yielded a significant benefit on the 2-mark questions, but not on the essays, where the remotely standardised new examiners improved more than those attending the meeting. On the un-standardised essay questions, remotely-standardised examiners improved more than the meeting attendees.

From the perspective of improving marking accuracy in relation to Team Leader reference marks, the benefits of holding a face-to-face standardisation meeting therefore appear variable, small and questionable, for both new and experienced examiners, and for both essay and short-answer questions. It would be reasonable for examining bodies to explore whether standardisation can be achieved using more cost-effective and efficient methods than face-to-face meetings.

Caveats

A number of caveats must be placed on these findings.

- The essays were highly structured and factual, and marked against a prescriptive mark scheme. Findings might not be replicated with less constrained essays and marking.
- The Team Leaders were not experienced at leading standardisation, a task carried out operationally by the Principal Examiner. They were recommended to us for this task, however.
- We used only two Team Leaders, one for short-answers, the other for essays. We therefore have no way of separating any effects introduced by the Team Leaders from effects introduced by the question type. Similarly, each reference mark was produced by only one Team Leader, who may or may not have been typical – though the fact that both had been successful Team Leaders in the operational marking mitigates against this risk.
- Only twenty-four examiners took part in the study, and these examiners might not have been representative of the wider populations of experienced and new examiners.
- Both the meeting and the remote standardisation tasks differed from normal operational practice. Cambridge Assessment only uses remote standardisation methods in the context of online marking,

where examiners can be monitored and supported more effectively than when marking on paper. In the present study all marking was carried out on paper, and the standardisation tasks adapted to match as closely as possible with those used operationally with online marking. Operational standardisation meetings are conducted by Principal Examiners and focus on either the short-answer examination or the essay examination, but not both. Examiners typically mark only one examination. However, the number of questions used in the study was far fewer than would be used in an operational setting.

- All participants knew that the marks did not 'count', and were only for use in the research. Whilst it is our impression that all participants were highly diligent and professional, we have no way of quantifying what effects, if any, were introduced by the low stakes nature of the exercise.

Finally, it should be noted that in operational marking settings examiners are given additional standardisation if necessary and are removed from the marking panel if their accuracy remains unsatisfactory. Additionally, examiners' operational marking is sampled on several occasions after initial standardisation, to check that accuracy levels are maintained. For these reasons operational marking is likely to be more accurate than was found in this study.

References

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

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

Greatorex, J., Nádas, R., Suto, I. & Bell, J.F. (2007). *Exploring how the cognitive strategies used to mark examination questions relate to the efficacy of examiner training*. Paper presented at the ECER conference, Ghent, Belgium in September 2007.

Qualifications and Curriculum Authority (March 2009). *GCSE, GCE and AEA Code of Practice*. London: QCA.

ASSURING QUALITY IN ASSESSMENT

A review of literature on item-level marker agreement: implications for on-screen marking monitoring research and practice

Milja Curcin Research Division

Introduction

Marking reliability contributes in important ways to the overall reliability and validity of assessment. It refers to the extent to which different examiners' marks agree with each other or with a definitive mark when they mark the same material (inter-marker agreement), and is also affected, for instance, by individual examiners' consistency throughout marking (intra-marker consistency). Validity of assessment is compromised without high marking reliability since the same mark from different examiners cannot be assumed to mean the same thing (e.g. Massey and Raikes, 2006; Cambridge Approach, 2009). However, as Wilmut *et al*. (1996) observe, "[f]or a variety of reasons, perfect reliability is not going to happen. The aim must be to get as close as possible, given irreducible constraints."

This review article focuses mainly on the literature relevant for the inter-marker agreement aspect of marking reliability in the context of on-screen marking. The increasing use of on-screen in place of paper-based marking presents new possibilities for monitoring of marking and ensuring higher agreement levels, but also raises questions with respect to the most efficient and beneficial use of marker agreement information that is routinely collected in this process, both in monitoring practice and in research.

Current Ofqual[1] regulations (Code of practice, April 2009) for on-screen marking require that the marking of individual examiners be compared to that of a senior examiner at regular intervals throughout the marking process. Although the specifics of this procedure differ across awarding bodies, this is generally implemented by means of "seeding" pre-marked "seeding scripts" (or items)[2] into live marking at regular intervals. The markers' marks are checked against the scripts'/items' "definitive marks",[3] these having been determined in advance by a single senior examiner or by a panel of senior examiners, depending on awarding body practices.

In this monitoring process, marker agreement data are collected at item level, potentially providing a rich source of information, particularly with respect to which features of items are associated with high or low marker agreement. Furthermore, since some awarding bodies use expert panels to decide on definitive marks, presumably under the assumption that groups make better decisions than individuals (cf. Levine and Moreland, 2006), it is conceivable that the group dynamics of these panels could affect the choice of the definitive marks and subsequent individual marker agreement with them. It is useful, therefore, to consider research to date on marker agreement, particularly at item level, as well as social psychology research on group dynamics, as this might inform both current marking monitoring processes and future research in this area, particularly in respect of what marker agreement levels can be

[1] Ofqual (Office of the Qualifications and Examinations Regulator) is responsible for regulating public examinations.

[2] Script: whole candidate work on one question paper. Item: candidate response on one question or question part.

[3] The definitive marks are not visible on the scripts.