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

Cambridge Assessment

A paper presented at the British Educational Research Association Annual Conference, September 2009, Manchester

Contact:

Nicholas Raikes
Research Division
Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Email: Raikes.N@CambridgeAssessment.org.uk

Abstract

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 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.

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.

Introduction

Background

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. 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, 2008, section 4.14).

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. 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.

Research questions

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?

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.

The essay examination gave candidates a choice of questions, so each question was answered by a different sub-group of candidates. We therefore used essays from examinations held in consecutive years, selecting the closest matching question 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 Topic: Cognitive Psychology		**Short-Answer Collection 2** Examiners were **not** standardised on these 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

Essay Collection 1 Examiners were standardised on these Examination 1		**Essay Collection 2** Examiners were **not** standardised on these 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 live-marked the examinations. Twelve of the examiners had experience of marking other psychology A-Level examinations; the other twelve examiners were brand new to examining, having been recruited for live 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 live 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 live 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 live examinations once all live 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 live marks, so that the Pre-and Post-batches were as similar as possible.

Pre-Standardisation batches:

Batch *Short-1i* 50 answers to each question in Short-Answer Collection 1	**Batch *Essay-1i*** 25 answers to each question in Essay Collection 1	Examiners were to be standardised on these questions
Batch *Short-2i* 50 answers to each question in Short-Answer Collection 2	**Batch *Essay-2i*** 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):

Batch *Short-Si* 5 answers to each question in Short-Answer Collection 1	**Batch *Essay-Si*** 5 answers to each question in Essay Collection 1
Batch *Short-Sii* 5 answers to each question in Short-Answer Collection 1	**Batch *Essay-Sii*** 5 answers to each question in Essay Collection 1
Batch *Short-Siii* 10 answers to each question in Short-Answer Collection 1	**Batch *Essay-Siii*** 10 answers to each question in Essay Collection 1

Post-Standardisation batches:

Batch *Short-1ii* 50 answers to each question in Short-Answer Collection 1	**Batch *Essay-1ii*** 25 answers to each question in Essay Collection 1	Examiners were standardised on these questions
Batch *Short-2ii* 50 answers to each question in Short-Answer Collection 2	**Batch *Essay-2ii*** 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-1i first, then Essay-1i, then Short-1ii, then Essay-1ii

(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-Si, then opened the envelope containing the Team Leader's marks and explanations for Short-Si. They were instructed to compare the Team Leader's marks with their own and read the explanations.	
(7)	All Examiners marked batch Short-Sii.	
(8)		A2 & B2 examiners opened the envelope containing the Team Leader's marks and explanations for batch Short-Sii. They were instructed to compare the marks with their own and read the explanations.
(9)	All examiners marked batch Essay-Si, 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-Sii	
(11)		A2 & B2 examiners opened the envelope containing the Team Leader's marks and explanations for batch Essay-Sii. 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-Sii and Essay-Sii 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-Siii and Essay-Siii. 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 Siii marking and answer questions	

Stage 3: Post-Standardisation

(15) Examiners marked the Post-Standardisation scripts in the following order: Short-1ii first, then Essay-1ii, then Short-2ii and finally Essay-2ii.

(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 live 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

All marks were keyed into SPSS for 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, i.e. all were positive numbers. These absolute differences gave the size of the difference, and when averaged do 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 red lines correspond to the results from the examiners who attended the meeting ("Face to face" Standardisation type), the blue 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. Full details of the results of the ANCOVA analysis are given in Appendix A.

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.

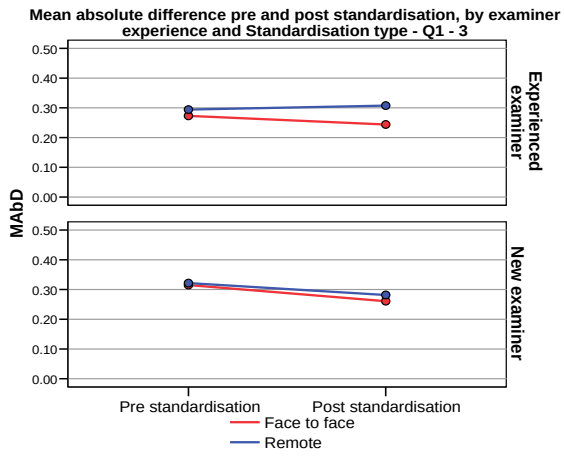
Short-answer questions

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 1: 2-mark questions

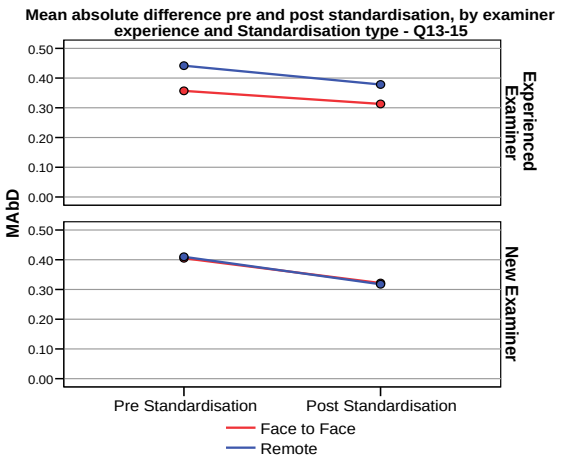
Examiners were standardised on these



Statistical significance

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

Examiners were **not** standardised on these



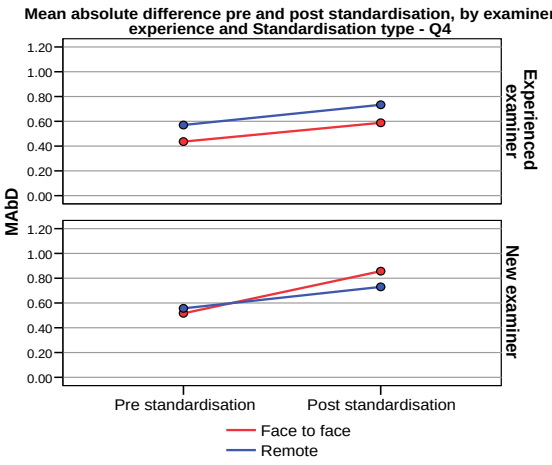
Statistical significance

Examiner experience	×	p=.096
Standardisation type	×	p=.084
Interaction	✓	p=.044

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 2: 4-mark question

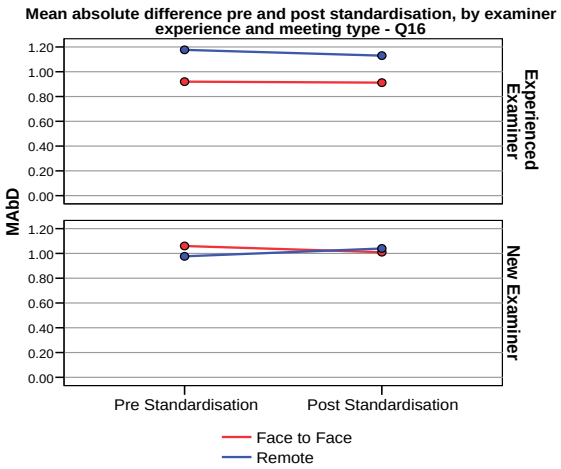
Examiners were standardised on these



Statistical significance

Examiner experience	✓	p=.002
Standardisation type	×	p=.947
Interaction	✓	p=.002

Examiners were **not** standardised on these



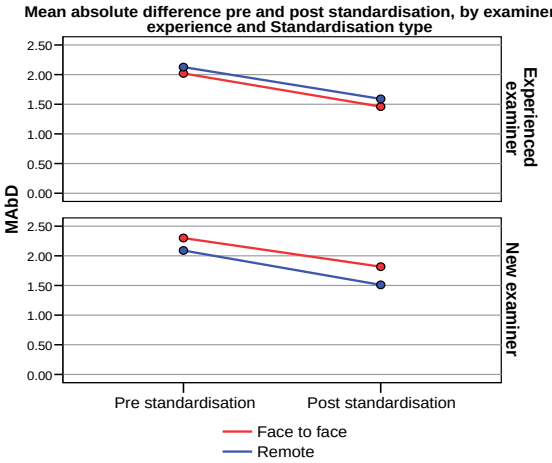
Statistical significance

Examiner experience	×	p=.934
Standardisation type	✓	p=.040
Interaction	×	p=.135

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, this greater improvement for the remotely Standardised examiners was statistically significant regardless of Examiner Experience.

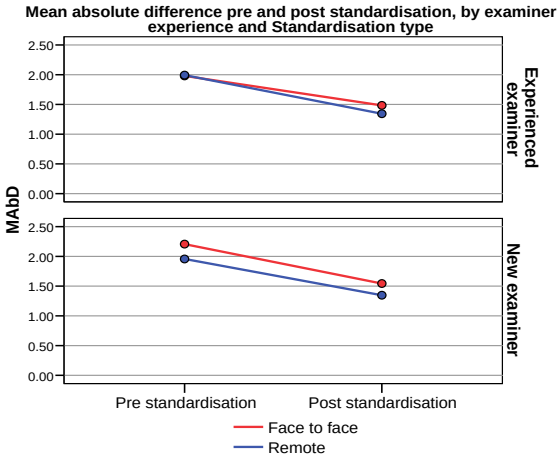
Figure 3: Essay questions

Examiners were standardised on these



Statistical significance		
Examiner experience	×	p=.094
Standardisation type	×	p=.282
Interaction	✓	p=.008

Examiners were **not** standardised on these



Statistical significance		
Examiner experience	×	p=.745
Standardisation type	✓	p=.045
Interaction	×	p=.795

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;
- 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, both the new and experienced 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 Team Leaders were not experienced at leading Standardisation, a task carried out in live examining 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 live marking mitigates against this risk;
- Both the meeting and the remote Standardisation tasks differed from normal live practice. Cambridge Assessment does not currently Standardise essay marking remotely, and where remote methods are used for short-answer Standardisation this is in the context of online marking, where examiners can be monitored and supported more effectively than when marking on paper. Live 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 a live examination.
- 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 live marking examiners will be given additional Standardisation if necessary and will be removed from the marking panel if their accuracy remains unsatisfactory. Additionally, examiners' live marking is sampled on several occasions after initial Standardisation, to check that accuracy levels are maintained. For these reasons live marking is likely to be more accurate than was found in this study.

References

Baird, Jo-Anne, Greatorex, Jackie and Bell, John F. (2004) *What makes marking reliable? Experiments with UK examinations*. Assessment in Education Vol. 11, No. 3, pp. 331-348.

Greatorex, Jackie and Bell, John F. (2008) *What makes AS marking reliable? An experiment with some stages from the standardisation process*. Research Papers in Education.

Greatorex, Jackie, Nádas, Rita, Suto, Irenka and Bell, John 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 (April 2008) *GCSE, GCE , GNVQ and AEA Code of Practice*. QCA, 83 Piccadilly, London W1J 8QA.

Appendix A: ANCOVA Results

Standardised 2-mark questions

Covariance summary table

Dependent Variable: abs_diff_post

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	3.217(a)	4	.804	3.456	.008
Intercept	243.947	1	243.947	1048.135	.000
abs_diff_pre	.695	1	.695	2.985	.084
Ex_Group	.032	1	.032	.139	.710
Mtg_type	2.000	1	2.000	8.594	**.003**
Ex_Group * Mtg_type	.512	1	.512	2.199	.138
Error	1069.458	4595	.233		
Total	1420.000	4600			
Corrected Total	1072.675	4599			

a R Squared = .003 (Adjusted R Squared = .002)

Marginal Means

2. Mtg_type

Dependent Variable: abs_diff_post

			95% Confidence Interval	
Mtg_type	Mean	Std. Error	Lower Bound	Upper Bound
Face to face	.253(a)	.010	.232	.273
Remote	.294(a)	.010	.275	.314

a Covariates appearing in the model are evaluated at the following values: abs_diff_pre = .3022.

3. Ex_Group * Mtg_type

Dependent Variable: abs_diff_post

				95% Confidence Interval	
Ex_Group	Mtg_type	Mean	Std. Error	Lower Bound	Upper Bound
Experienced examiner	Face to face	.245(a)	.015	.215	.275
	Remote	.308(a)	.014	.280	.335
New examiner	Face to face	.261(a)	.014	.233	.288
	Remote	.281(a)	.014	.254	.309

a Covariates appearing in the model are evaluated at the following values: abs_diff_pre = .3022.

Un-Standardised 2-mark questions

Covariance summary table

Dependent Variable: abs_diff_post

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	9.794(a)	4	2.449	8.367	.000
Intercept	292.416	1	292.416	999.257	.000
abs_diffq13_pre	6.487	1	6.487	22.167	.000
Ex_Group	.809	1	.809	2.765	.096
Mtg_type	.876	1	.876	2.994	.084
Ex_Group * Mtg_type	1.183	1	1.183	4.044	**.044**
Error	1344.650	4595	.293		
Total	1866.000	4600			
Corrected Total	1354.444	4599			

a R Squared = .007 (Adjusted R Squared = .006)

Marginal Means

3. Ex_Group * Mtg_type

Dependent Variable: abs_diff_post

Ex_Group	Mtg_type	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Experienced Examiner	Face to Face	.316(a)	.017	.283	.350
	Remote	.376(a)	.016	.345	.407
New Examiner	Face to Face	.322(a)	.016	.291	.352
	Remote	.317(a)	.016	.287	.348

a Covariates appearing in the model are evaluated at the following values: abs_diff_pre = .4054.

Standardised 4-mark question

Covariance summary table

Dependent Variable: abs_diffq4_post

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12.651(a)	4	3.163	6.012	.000
Intercept	337.368	1	337.368	641.253	.000
abs_diff_q4_pre	2.804	1	2.804	5.330	.021
Ex_Group	4.838	1	4.838	9.196	**.002**
Mtg_type	.002	1	.002	.004	.947
Ex_Group * Mtg_type	5.011	1	5.011	9.525	**.002**
Error	602.393	1145	.526		
Total	1233.000	1150			
Corrected Total	615.044	1149			

a R Squared = .021 (Adjusted R Squared = .017)

Marginal means

1. Ex_group

Dependent Variable: abs_diffq4_post

Ex_Group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Experienced examiner	.662(a)	.031	.601	.723
New examiner	.792(a)	.030	.734	.850

a Covariates appearing in the model are evaluated at the following values: abs_diff_q4_pre = .5235.

3. Ex_Group * Mtg_type

Dependent Variable: abs_diffq4_post

Ex_Group	Mtg_type	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Experienced examiner	Face to face	.594(a)	.046	.504	.685
	Remote	.730(a)	.042	.648	.812
New examiner	Face to face	.857(a)	.042	.775	.939
	Remote	.728(a)	.042	.645	.810

a Covariates appearing in the model are evaluated at the following values: abs_diff_q4_pre = .5235.

Un-Standardised 4-mark question

Covariance summary table

Dependent Variable: abs_diffq16_post

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7.737(a)	4	1.934	1.937	.102
Intercept	556.056	1	556.056	556.847	.000
abs_diffq16_pre	1.111	1	1.111	1.113	.292
Ex_Group	.007	1	.007	.007	.934
Mtg_type	4.200	1	4.200	4.206	**.040**
Ex_Group * Mtg_type	2.238	1	2.238	2.242	.135
Error	1143.373	1145	.999		
Total	2366.000	1150			
Corrected Total	1151.110	1149			

a R Squared = .007 (Adjusted R Squared = .003)

Marginal Means

2. Mtg_type

Dependent Variable: abs_diffq16_post

			95% Confidence Interval	
Mtg_type	Mean	Std. Error	Lower Bound	Upper Bound
Face to Face	.962(a)	.043	.878	1.046
Remote	1.084(a)	.041	1.004	1.164

a Covariates appearing in the model are evaluated at the following values: abs_diffq16_pre = 1.0383.

Standardised essay questions

Covariance summary table

Dependent Variable: abs_diffq_post

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	22.116(a)	4	5.529	2.920	.020
Intercept	1138.481	1	1138.481	601.306	.000
abs_diffq_pre	.673	1	.673	.355	.551
Ex_Group	5.330	1	5.330	2.815	.094
Mtg_type	2.193	1	2.193	1.158	.282
Ex_Group * Mtg_type	13.323	1	13.323	7.037	**.008**
Error	2167.884	1145	1.893		
Total	5134.000	1150			
Corrected Total	2190.000	1149			

a R Squared = .010 (Adjusted R Squared = .007)

Marginal Means

3. Ex_Group * Mtg_type

Dependent Variable: abs_diffq_post

Ex_Group	Mtg_type	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Experienced examiner	Face to face	1.462(a)	.087	1.291	1.632
	Remote	1.590(a)	.079	1.434	1.746
New examiner	Face to face	1.814(a)	.080	1.658	1.970
	Remote	1.511(a)	.079	1.355	1.667

a Covariates appearing in the model are evaluated at the following values: abs_diffq_pre = 2.1391.

Un-Standardised essay questions

Covariance summary table

Dependent Variable: abs_diffq_post

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	19.101(a)	4	4.775	2.562	.037
Intercept	858.467	1	858.467	460.555	.000
abs_diffq_pre	10.192	1	10.192	5.468	.020
Ex_Group	.197	1	.197	.106	.745
Mtg_type	7.522	1	7.522	4.036	**.045**
Ex_Group * Mtg_type	.125	1	.125	.067	.795
Error	2134.264	1145	1.864		
Total	4495.000	1150			
Corrected Total	2153.364	1149			

a R Squared = .009 (Adjusted R Squared = .005)

Marginal Means

2. Mtg_type

Dependent Variable: abs_diffq_post

Mtg_type	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Face to face	1.511(a)	.058	1.396	1.625
Remote	1.348(a)	.056	1.239	1.458

a Covariates appearing in the model are evaluated at the following values: abs_diffq_pre = 2.0365.