DO MARKING RELIABILITY STUDIES HAVE VALIDITY?

Suzanne Chamberlain

ABSTRACT

Marking reliability studies are essential to ensure that examiners' marking of high stakes public assessments is appropriate, consistent and fair to candidates. The importance of such studies is confirmed by an extensive marking reliability literature which spans the early twentieth century to the present day, and covers many different forms of examination papers.

There are many challenges in designing a marking reliability study, and perhaps the most significant of these is deciding whether to undertake research in a live or non-live examination context. While a live context has the greatest external validity, it is not always feasible or fair to candidates or examiners. Alternatively, using a non-live context allows for greater control of the research environment but brings into question the validity of the outcomes. For example, if participants are aware that the examination scripts have no bearing on candidates' real examination results, will they mark the research scripts with the same diligence as live scripts?

This paper overviews the outcomes of a questionnaire distributed to 89 examiners who took part in a controlled, experimental marking reliability study. The questionnaire was designed to compare the degree of conscientiousness applied during the experimental study and under normal live marking conditions. In doing so, the questionnaire aimed to gauge the representativeness and naturalness of participants' behaviours and thus contribute to assessing the ecological validity of the primary (marking reliability) study.

It is found that a small proportion of participants acknowledged that they responded differently during the study. It is suggested that these differences could represent a threat to the validity of the study. It is concluded, however, that this is an inevitable drawback of controlled, non-live research, and that this is counter-balanced by the important insights into the reliability of examiners' marking that such studies give us.

INTRODUCTION

The suite of public examinations of England, Northern Ireland and Wales includes the General Certificate of Secondary Education (GCSE), the General Certificate of Education (GCE), and a range of applied and vocational qualifications. These public examinations are high stakes and their outcomes serve several important social functions. Chiefly, the examination candidates use the outcomes as 'entry tickets' (Denscombe, 2000) for other educational, training or employment opportunities. They are also used as indicators of teacher and especially school performance at national and even international levels, and to monitor examination standards over time (Goldstein, 2001).

Given these important social functions, marking reliability studies have long played a crucial role in educational research and the work of awarding bodies particularly (see for example, Meadows and Billington, 2005). Such studies typically aim to explore the consistency with which examiners apply the mark scheme by measuring examiners' divergence from the 'true' script scores (usually the script scores awarded by the Principal Examiner who is responsible for setting the paper and devising the accompanying mark scheme). In some cases marking reliability studies aim to manipulate certain variables to explore their impact, while others attempt to assess the quality of operational marking (Meadows and Billington, 2005). The key concern of such studies is to measure marking reliability, but in doing so some important questions emerge relating to study design and how this impacts on the validity of the findings.

Designing a marking reliability study

One of the key decisions to be made in designing a marking reliability study is whether the research should be conducted in a live or non-live examination setting. Essentially this represents the difference between a naturalistic setting and an artificially created research environment that to a greater or lesser degree attempts to simulate what occurs in the live environment. This decision has important implications for the degree of control over the research setting, the participants, the materials and tasks, and the extent to which potential confounds can be controlled for. In turn, these determine the extent to which the conclusions drawn can be deemed to be generalisable, and externally valid.

The findings drawn from a naturalistic setting are usually highly representative and generalisable to other occasions (Schmuckler, 2001). However, the constraints of real-world processes are difficult to overcome and may introduce several confounds into a study. For example, in a marking reliability study it is highly desirable that all examiners mark the same scripts. This allows for the calculation of inter- and intra-rater reliability coefficients; the comparison of like with like; order effects to be explored or controlled for; and ensures that individual examiners are not advantaged or disadvantaged by the quality and range of candidate responses within their different script samples. Achieving this in a live setting is far more complex than within a non-live setting, and entails one of two options:

1) allocating every examiner in the study an additional sample of, say, 40 common scripts. The scripts would have to be marked at a time when examiners are already under considerable pressure to mark large batches of scripts within a tight timeframe. It would add to the administrative burden of awarding body personnel at the busiest and most crucial time of the academic year. It is also highly questionable whether it is fair to candidates to have their examiners mark copies of scripts for research purposes only; scripts that could distract them from their true goal of ensuring that the examination performance of each candidate is appropriately rewarded. (Although this has to be balanced against the goal of enhancing processes that could benefit everyone involved in the examination cycle);
2) requesting that the Principal Examiner over-marks a proportion of the script samples of every examiner included in the study. This option would have minimal impact on the live marking performance of the examiners, but, given the numbers of examiners and scripts that would be required to achieve sufficient statistical power, this option is highly burdensome for the Principal Examiner. Even with relatively small samples of, say, 20 examiners each marking 20 different scripts, this represents an additional marking load of 400 scripts for the Principal Examiner. The use of multiple samples also limits the types of inter-rater analyses that can be conducted.

There are two other important hurdles to overcome in a live setting. Firstly, there is the question of how candidates' 'true' scores are derived when the scripts have been marked by multiple examiners. If each examiner in a live marking reliability study marks the same batch of scripts, how should the final scores be best determined for these candidates, given that the scores of multiple examiners are likely to vary? It is generally accepted that multiple ratings of candidates' performances enhance the reliability of assessment outcomes (Cronbach, 1971; Thorndike and Hagen, 1977). In the case of an examination, however, where candidates are able to appeal against their outcomes, having multiple assessments may cast doubt over the awarded mark, and the mark may be difficult to defend in the context of an appeal. Using multiple ratings to create an average score for each candidate is also undesirable as averaging reduces the spread of marks and would complicate the grading process (Brooks, 1980). Additionally, it is questionable whether it is fair to the remaining candidates whose scripts were not included in the study and were marked by one examiner only. The other alternative, using multiple copies of the same scripts, would entail photocopying the original script and distributing copies to each examiner. As photocopied scripts are easily identifiable among a batch

of original scripts this brings into question the reliability of the examiners' assessments and the external validity of the findings[1]. The study scripts may be given more or less consideration than examiners would usually apply to their live marking allocation.

For these reasons it therefore may be preferable to undertake a purposely designed study outside of the live marking setting. A designed study has the following advantages:

1) it enables greater control over the research environment, and potential confounds may be controlled for;
2) with no real-world constraints (other than costs, time and resources etc), the size of the examiner and script samples can be determined with the required statistical power in mind;
3) using non-live scripts will have no impact on candidates or their examination outcomes;
4) it can be conducted at a time to suit examining personnel;
5) experimental interventions can be introduced to the study without fear of impacting on live examination processes and outcomes;
6) the study participants are able to mark the same batch of common scripts, allowing the full range of inter- and intra-rater analyses to be conducted; and
7) the sample of scripts can be purposely selected to test the application of particular aspects of the mark scheme or scripts within a particular mark range.

A designed study can create the conditions for the gathering of high quality, statistically robust and reliable examiner marking reliability data. However, any study requires a trade-off between the naturalness of the setting and the degree of control required over the processes and the participants' actions (Kvavilashvili and Ellis, 2004; Schmuckler, 2001). Any advantages of a designed study have to be counter-balanced against the drawback of the non-live setting and the impact this may have on the representativeness of the findings and the external validity of the conclusions. Equally, the advantages of a live setting have to be balanced against a potential loss of control of the research environment and procedural constraints.

External and ecological validity

The literature repeatedly states that validity is not a property of the research instrument (or test or assessment) *per se*, but rather a property of the inferences drawn from the findings of a study (e.g. Cronbach, 1971; Messick, 1989; Schmuckler, 2001). It is not an observable, measureable artefact, but a corpus of substantive arguments that verify the adequacy and appropriateness of the inferences made (Cronbach, 1971). Demonstrating validity therefore requires that a set of arguments or evidence is collected until a convincing saturation point has been reached to confirm that the research findings are valid for the intended purpose. Good descriptions of the participants and the processes followed throughout the study are also useful for judging validity (Keeves, 1988).

The natural versus controlled dilemma with which the researcher has to grapple, is essentially a dilemma about ecological validity. Debates about ecological validity have been prominent in the psychology literature where the use of laboratory-based, experimental human research has brought into question the extent to which findings from artificial environments are generalisable outside the confines of the experiment (see for example, Kvavilashvili and Ellis, 2004). It is not simply the case that research that replicates reality has good ecological validity, while research conducted in controlled, artificial environments has poor ecological validity; there are many shades of grey in between and some aspects of the research may be more ecologically valid than others (Chow, 1987; Keeves, 1988; Kvavilashvili and Ellis, 2004; Schmuckler, 2001).

[1] This is not necessarily the case with electronic marking.

Ecological validity is considered to be dependent upon the degree of representativeness of the study and generalisability of the findings. Representativeness refers to the naturalness or artificiality of the research setting, the research materials, the task participants are asked to complete, and the responses that the task provokes within the participants (Schmuckler, 2001). Each of these should correspond to the form in which they occur in everyday life, in being meaningful and plausible, in order to confer ecological validity upon the findings (Kvavilashvili and Ellis, 2004). Generalisability refers to the extent to which the findings explain or are consistent with comparable processes in everyday life. Generalisability is also at the heart of external validity, such that if the findings are demonstrably generalisable then they are also considered to be externally valid (Kvavilashvili and Ellis, 2004). Kvavilashvili and Ellis (2004) note that different perspectives in the ecological validity debates place greater or lesser emphasis on different aspects of the research environment; some consider the task to be the most important element by which to demonstrate ecological validity, while others emphasise the research materials or the degree of generalisability. Schmuckler (2001) suggests there are no generic criteria for determining which elements of the study are more important than others; this judgement needs to be made and defended within specific contexts.

Kvavilashvili and Ellis (2004) summarise that if the research setting utilises processes that are comparable to the real-world context, and the findings are evidently generalisable, then the conclusions drawn may be considered ecologically valid. If the study is unable to make claims to either representativeness or generalisability then the conclusions will lack ecological validity and will contribute very little to enhancing our knowledge and understanding of the phenomenon under study. In an artificial research setting it is more likely, however, that naturalness, control and, ultimately, validity have been sacrificed in some areas and maintained in others (Schmuckler, 2001). The ecological validity of a study then becomes more difficult to judge if the setting is considered to be representative, or the findings generalisable, but not both (Schmuckler, 2001).

It is argued that generalisability is the more important property of a research study, and that the goal of achieving this rests upon having good internal validity (Kvavilashvili and Ellis (2004). Although it is not considered necessary or desirable to mimic every aspect of real-world conditions (Chow, 1987), there should be harmony between the processes invoked during the study and those that would be encountered in the real-world context. The internal validity of the research should be such that the findings tell us something about human behaviour or performance in the live context, and not just about how people respond in the artificial environment. Indeed, there is limited utility in controlling the research setting and materials to such a degree that the research creates a situation that people would not experience (poor ecological validity), and thus produces findings that are not generalisable (poor external validity). Keeves (1988) suggests that this represents the difference between psychological realism (replicating the important psychological processes) and mundane realism (unnecessarily replicating as faithfully as possible all the conditions of the real-world context). An artificial environment may not therefore be a barrier to provoking realistic and generalisable human responses, as long as sufficient psychological realism is achieved.

Assessing ecological validity

The aim of the primary study (the marking reliability study) was to explore examiners' marking reliability after they had received training in the application of a particular mark scheme (see Taylor, Chamberlain and Meadows, 2008). The participants were randomly split into two groups, with one group acting as a control group, receiving training in its conventional form (face-to-face), and the second group acting as the experimental group, receiving their training via a new online examiner training system. The research environment was strictly controlled so that potential sources of error and confounds were eliminated as much as possible, and any observed inter-group differences in marking reliability could be attributable primarily to the type of training received.

In order to generalise with confidence to other occasions involving different examination papers and examiners, it is necessary to demonstrate that the conclusions have external and ecological validity. Ensuring the generalisability of the study with regard to other examination papers was incorporated into the design of the study in two ways. Firstly, the study materials (candidates' examination scripts from 2007) and processes (training and marking procedures) were realistic, and representative, as much as possible, of those used in the live marking context. Secondly, the examination paper at the centre of the study – a GCSE History paper - was chosen as an example of a fairly taxing paper from the examiner's perspective. O'Donovan (2005) notes that mark schemes in the Humanities are complex to apply as they are 'content-advisory' rather than 'content-specific'. This is in contrast to examination papers in the Sciences, for example, that tend to use questions that are more clearly right or wrong. As such, the examiners' task in this study was doubly difficult – not only were they working with extended mark ranges (thus increasing the likelihood of their awarded marks differing from the 'true' marks of the Principal Examiner), they were also required to apply their own interpretations in marking candidates' responses. The complexity of the selected paper meant that the findings could be generalised to other examination papers with the same type of mark scheme[2], and with similar or less complex structures and mark ranges.

An equally pressing ecological validity question remained however. For several reasons the research project could not be conducted in real time using live candidate scripts. Not only would it have been unethical to undertake an experiment with examiner training during the marking period, it would have also placed severe constraints upon the timing, design and scope of the study. In the light of these issues, the study was designed as a stand-alone experiment, undertaken at a time when the participants were free from examining commitments (although not necessarily other paid work commitments).

During the recruitment stage of the primary study the examiners were made fully aware that:

- they were participating in a stand-alone research exercise;
- the study was designed to explore the impact of training on examiners' quality of marking;
- the outcomes of the study would have no bearing on their current or future employment as examiners;
- all participants were marking the same set of scripts in the same order;
- all participants would receive the same payment, regardless of their performance;
- the scripts were from an earlier examination series; and
- therefore the marks that they awarded had no impact on real candidates.

Given these controls, it seemed that only the examiners' professionalism and commitment to the task motivated them to complete their marking to the best of their abilities, and to give each script the consideration it required. In order to support the ecological validity of the conclusions it was necessary to explore whether the non-live research environment had shaped examiners' behaviours to any significant degree. In particular it was necessary to investigate whether they had completed their marking with a comparable level of diligence as they would during live marking, and thus whether their behaviour during the study was representative of their normal behaviours. This not only applied to examiners exerting less care and consideration, but also, perhaps if they were not assured by the confidentiality of the study, to examiners being excessively careful in their responses to the scripts.

METHOD

A short, closed-question postal questionnaire survey was designed to assess the naturalness (or otherwise) of participants' behaviours during the primary (marking reliability) study. The questionnaire asked participants

[2] The findings would not necessarily be generalisable across different types of marks schemes. For example, introducing a new form of examiner training may have less of an effect on a points-based mark scheme than a model answer-based mark scheme.

to compare how they usually respond to marking scripts under normal live conditions to how they responded to marking scripts during the non-live research study. Evidence of good consistency in participants' behaviours could then be applied to support the ecological validity of the conclusions drawn from the study.

The key concept to be measured by the questionnaire was 'conscientiousness'. 'Conscientiousness' was operationalised as incorporating thoroughness, consistency, effort, confidence, doubt and decisiveness (e.g. Barrick and Mount, 1991; Costa and McCrae, 1992). Thirteen items were developed to assess the presence of these traits during live marking and the experimental study. The questionnaire also included two global self-ratings of 'effort'; one rating of effort during normal conditions and one of effort under the study conditions. The items and questionnaire as a whole were concerned with one aspect of ecological validity; namely the naturalness and representativeness of the responses that the task provoked within the participants (Schmuckler, 2001).

Personality inventories often depend on self-reports of traits such as 'conscientiousness' (e.g. Costa and McCrae's (1992) NEO Five Factor Inventory (NEO-FFI)), and previous studies have found this attribute to be positively correlated with marking reliability (Meadows and Billington, 2007) and occupational performance more generally (e.g. Barrick and Mount, 1991). In this case the questionnaire was not seeking to use 'conscientiousness' as a predictor of marking reliability or performance, but as a check on inter- and intra-examiner consistency over the two occasions (live and experimental) and between the two groups (control and experimental). Although there were some potential limitations to using a questionnaire (see Discussion and Conclusions), it was considered the most appropriate means of gathering evidence to inform the validity debate.

To counter-balance some of the potential limitations, all questionnaire responses were anonymised to assure examiners that their responses could not be used to make inferences about their professionalism (i.e. that they had not been conscientious during live marking). It was envisaged that their anonymity would allow examiners to be as truthful as possible about their experiences. Consequently no classification data were gathered at this stage (such as age, gender, marking experience etc), other than to identify which mode of training the examiner had received in the primary study[3]. As group allocation was random, and the groups relatively heterogeneous, no systematic group effects were anticipated. However, classifying participants by group allowed us to consider whether the nature of the task itself had encouraged certain behaviours; for example whether using an unfamiliar online system appeared to increase stress, or decrease reliance on the mark scheme, among the experimental group.

The questionnaire was distributed to all 89 participants of the primary (marking reliability) study upon completion of their marking. A response rate of 100% was secured.

PARTICIPANTS[4]

For inclusion in the primary (marking reliability) study, participants had to fulfil four criteria:

1) they had experience of marking GCSE History in at least one summer examination period;
2) they assessed themselves as having a broad subject matter expertise (the examiners were specialists in domains of History other than that included in the study);
3) participants should not have marked or be at all familiar with the examination paper being used in the study to ensure that the findings were not confounded by prior knowledge or experience; and finally
4) that they had received an examiner rating of grade C or above in the most recent examination series (an in-house examiner rating system whereby grade A indicates

[3] This was achieved by using different coloured questionnaires for each group of participants.
[4] Of the primary (marking reliability) study and secondary (validity) study.

excellent marking performance and conduct of administration duties, and grade C indicates satisfactory performance; grade D examiners are not normally re-employed unless further training is completed).

Basic classification data were gathered at the participant recruitment stage. Table 1 shows that the groups were comparable in terms of their age and gender distribution, examining experience and employment status.

Table 1. The study participants.

		Experimental (online training)		Control (face-to-face training)		All	
		N	%	N	%	N	%
Age	35 or under	17	34.7	15	37.5	32	36.0
	36-45	13	26.5	7	17.5	20	22.5
	46-55	9	18.4	6	15.0	15	16.9
	56-65	9	18.4	10	25.0	19	21.3
	66 or over	1	2.0	2	5.0	3	3.4
Gender	Female	24	49.0	19	47.5	43	48.3
	Male	25	51.0	21	52.5	46	51.7
Years examining	3 or fewer	28	57.1	22	55.0	50	56.2
	4-7	14	28.6	11	27.5	25	28.1
	8-11	1	2.0	1	2.5	2	2.2
	12 or over	6	12.2	6	15.0	12	13.5
Employment status[5]	PT (inc self)	10	20.4	9	22.5	19	21.3
	FT (inc self)	33	67.3	23	57.5	56	62.9
	Retired	6	12.2	8	20.0	14	15.7
Total		49	55.1	40	44.9	89	100.0

VALIDITY STUDY RESULTS

The examiners were asked to give a global rating of the effort they expend on marking under normal conditions and the effort expended during the experimental study. A scale of 1 to 10 was used, with 1 denoting 'very little effort' and 10 denoting 'tremendous effort'. Chart 1 shows that while none of the examiners rated their effort as below 5 in either context, the distribution of effort is different for live and study marking. Under study conditions more examiners assessed their effort as being in the range of 5-7, and far fewer rated their effort as 9 or 10. A paired samples *t*-test was statistically significant suggesting that there were systematic differences within individuals' ratings of their levels of effort during live marking and the study (t_{88}=6.654, p=0.001). The effect size (Cohen's D = 0.71) suggests that the context within which the examiner is working (live or non-live) has a 'large' affect on the degree of effort they apply to their marking (Clark-Carter, 2003).

[5] In addition to employment as an examiner.

Chart 1. Examiners' self-reported ratings of effort applied to 'live' and 'study' script marking (N=89).

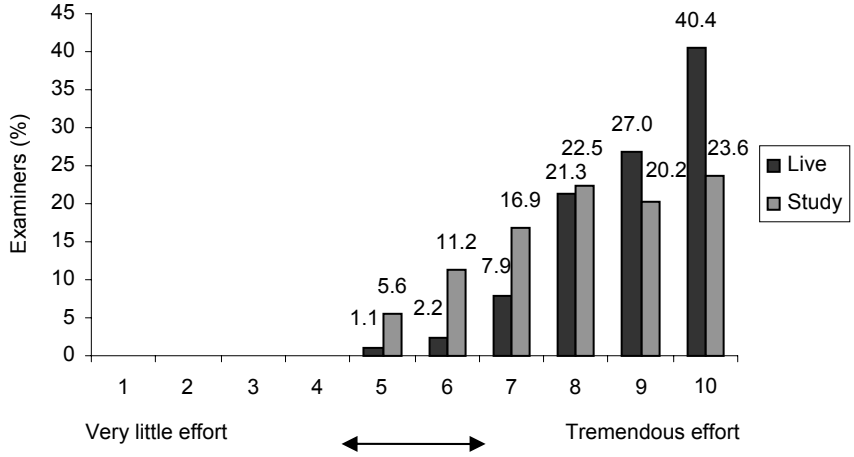

Table 2 details the differences between the levels of effort during live and study marking. Of all participants combined, just over four out of ten reported that they gave the study marking no more or less effort than they would apply to live marking. Around half of all participants applied less effort to marking the study scripts, while only five participants applied more effort (5.6%). Of those applying less effort to the study scripts, the change in the level of effort was mostly one or two points on the 10-point scale.

The table also shows that the two groups (control and experimental) responded in slightly different ways to marking the study scripts. The control group participants were more likely to maintain normal levels of effort during the study (52.5%), while the experimental group participants were more likely to report lower levels of effort than they would normally apply (57.1%). The observed differences between the two groups' effort ratings for the study are statistically significant and thus cannot be attributed to chance alone (t_{87}=2.236, p=0.028, Cohen's D = 0.48).

Table 2. Change in self-reported ratings of effort between 'live' and 'study' marking.

	Difference between ratings	Control group		Experimental group		All participants	
		N	%	N	%	N	%
Greater effort during study	2	0	0.0	1	2.0	1	1.1
	1	2	5.0	2	4.1	4	4.5
No change	0	21	52.5	18	36.7	39	43.8
Less effort during study	1	10	25.0	12	24.5	22	24.7
	2	4	10.0	10	20.4	14	15.7
	3	3	7.5	5	10.2	8	9.0
	4	0	0.0	1	2.0	1	1.1
Total		40	100	49	100	89	100

The questionnaire contained 13 items that were designed to measure different aspects of conscientiousness. Exploratory factor analysis was used to explore whether the pattern of participants' responses suggested the existence of any meaningful sub-themes[6]. The analysis offered four factors with eigenvalues greater than 1. Factors 3 and 4, however, consisted of two and one item respectively, and so were excluded from the

[6] Although factor analysis assumes the use of interval data, it is permissible to use ordinal data, as used here, as long as the categories used are consistent with the underlying metric scale (Kim and Mueller, 1978).

analysis. The three excluded items referred to the speed with which participants marked scripts, and the extent to which they revisited marked scripts and provided annotations.

Table 3 shows how the remaining ten items were split between two factors. The first of these, termed 'self-assuredness' consisted of six items that together explained 29.2% of the variance (Cronbach's alpha = 0.43). The second factor, 'comprehensiveness', consisted of four items and explained 16.8% of the variance (Cronbach's alpha = 0.52). Given the relatively limited reliability coefficients, the relatively small proportion of variance explained by both factors (46%), and the small number of items in each factor, the factors are applied to organise the discussion only and not to compute indices of 'conscientiousness'.

Table 3. Factor analysis of 'conscientiousness' items (N=89).

When marking scripts, how would you rate yourself in terms of the following in comparison to your usual standard of marking?			
Factor 1: Self-assuredness	Factor loadings	**Factor 2: Comprehensiveness**	Factor loadings
Indecisive	0.735	Thorough	0.689
Confident	-0.729	Consistent	0.617
Doubtful of right mark	0.700	Concerned to award right mark	0.527
Stressed	0.673	Focused	0.521
Referral to mark scheme	0.668		
Time on each script	0.658		

Table 4 outlines the participants' responses to the self-assuredness items. Interestingly, some of the responses appear to conflict with the participants' global ratings of effort, with the experimental group suggesting greater levels of conscientiousness on some measures than that implied by their effort ratings. For example, 41% of experimental group participants, compared to 25% of the control group, reported that they spent more time marking each script in the study than they would under live conditions. Similarly, 51% referred to the mark scheme more often than they would usually (compared to 40% among the control group participants). However, the experimental group also expressed greater levels of doubt that they were awarding the 'right' marks (46.9% compared with 37.5% among the control group), which may correspond with the increased levels of referral to the mark scheme and the time spent on each script.

On each of the items relating to 'self-assuredness' approximately half of all participants suggested that their responses to the study did not differ to how they would respond under live conditions. Interestingly, only four out of ten participants were less stressed by the study marking than they would be by live marking. Almost four out of ten also reported lower levels of confidence during the study. These latter findings are surprising given that the examiners' performances were of no consequence outside the confines of the study.

Only the item referring to the time spent on each script produced a statistically significant difference between the two participant groups ($\chi^2(2, N = 89) = 8.173$, $p = 0.017$). The distribution of responses suggested that the control group spent less time, and the experimental group spent more time marking each script than would be expected had all other things been equal. As noted above, this may be related to higher levels of doubt among the experimental group participants.

Table 4. Participants' responses to items relating to 'self-assuredness'.

	Control group (%)			Experimental group (%)			All participants (%)			All participants	
	More than usual	Less than usual	No more or less than usual	More than usual	Less than usual	No more or less than usual	More than usual	Less than usual	No more or less than usual	Chi-square test statistic (P)	Effect size (Cramer's Phi)
Indecisive	35.0	22.5	42.5	36.7	12.2	51.0	36.0	16.9	47.2	1.731 (0.421)	0.14
Stressed	20.0	45.0	35.0	20.4	36.7	40.8	20.2	40.4	38.2	0.558 (0.756)	0.08
Confident	15.0	40.0	45.0	6.1	32.7	59.2	10.1	36.0	52.8	2.871 (0.238)	0.18
Doubtful of right mark	37.5	15.0	47.5	46.9	10.2	42.9	42.7	12.4	44.9	0.975 (0.614)	0.11
Referral to mark scheme	40.0	2.5	57.5	51.0	6.1	42.9	46.1	4.5	49.4	2.179 (0.336)	0.16
Time on each script	25.0	0.0	75.0	40.8	10.2	49.0	33.7	5.6	60.7	8.173 (0.017)	0.30

Table 5 shows the responses to the four items of factor 2 relating to 'comprehensiveness'. These items suggest greater consistency between the experimental and control group than noted above for the 'self-assuredness' items. Around seven or eight out of ten participants suggested that their focus, thoroughness and consistency were no different from that applied during live marking. Interestingly, although the participants were fully aware that the scripts were from an earlier examination series, and that their marking therefore had no impact on real candidates, just over half of all participants were as concerned about awarding the 'right' mark as they would be during live marking. Three out of ten participants were less concerned than usual, and one in ten reported that they were more concerned that they identified the 'right' mark. The level of consistency in the participants' responses to the artificial marking environment is unexpected and surprising, and perhaps reflects general levels of professionalism and conscientiousness among examiners (see e.g. Meadows and Billington, 2007).

None of the chi-square analyses were statistically significant, suggesting that the distribution of responses between the control and experimental groups were no different than what might have been expected, all other things being equal. Indeed, the test statistic for the 'concerned' item is remarkably low (0.409) and indicates only a marginal difference between the observed and expected frequencies of the two groups.

Table 5. Participants' responses to items relating to 'comprehensiveness'.

	Control group (%)			Experimental group (%)			All participants (%)			All participants	
	More than usual	Less than usual	No more or less than usual	More than usual	Less than usual	No more or less than usual	More than usual	Less than usual	No more or less than usual	Chi-square test statistic (P)	Effect size (Cramer's Phi)
Focused	5.0	10.0	85.0	12.2	20.4	67.3	9.0	15.7	75.3	3.714 (0.156)	0.20
Thorough	15.0	7.5	77.5	8.2	16.3	75.5	11.2	12.4	76.4	2.316 (0.314)	0.16
Consistent	10.0	12.5	77.5	4.1	16.3	79.6	6.7	14.6	78.7	1.377 (0.502)	0.12
Concerned to award right mark	12.5	35.0	52.5	10.2	30.6	59.2	11.2	32.6	56.2	0.409 (0.815)	0.07

DISCUSSION AND CONCLUSIONS

In experimental research it is important to consider the extent to which the processes and materials used in the study reflect what occurs in the real-world setting, and the findings can thus be considered as generalisable. If the degree of control over the research environment is such that it becomes wholly distinct from the real-world setting the findings will have poor ecological validity (naturalness and representativeness) and poor external validity (generalisability). If, on the other hand, the essential psychological processes are replicated, and the human responses that the research materials provoke are comparable to those that would be observed in the real-world setting, the conclusions drawn may be considered ecologically and externally valid.

It was noted that the processes and materials used in this marking reliability study were as representative as possible of those used in the live context. Evidence of the internal and external validity of the findings is documented elsewhere (Taylor *et al*, 2008), and appears relatively secure. The question remained however, whether the non-live setting would impact on the participants' responses to the task and thus undermine any claims to ecological validity. It was easy to envisage how this might be the case; examiners are used to marking in highly pressurised environments, both in terms of the time available and the internal and external significance attached to the assessments that they make of candidate scripts. In contrast, this study held few of those pressures. It was therefore necessary to check that as a whole and as two distinct groups, the participants' approaches to the task were not markedly different from those under live conditions.

The findings of the questionnaire suggested that there were two elements to the participants' 'conscientiousness': those of self-assuredness, the degree to which respondents felt confident about the task, and comprehensiveness, which referred to the thoroughness with which the participants tackled their marking. The former element appears to consist mainly of affective items which concern the way participants felt about the task (Ajzen, 1988). In contrast, the latter element 'comprehensiveness' appears to consist of conative items detailing the participants' behavioural tendencies and commitment to the task (Ajzen, 1988). It is the latter element that is arguably the more significant of the two, having a greater and more direct impact on the ecological validity of the findings. Further research in this area may benefit from developing and using conative items alone.

The chi-square analyses suggested that all but one of the items produced no statistically significant differences between the observed and expected frequencies of the participants' responses. The time that

each participant spent marking each script produced the only significant difference and the largest test statistic, although the effect size suggested only a small effect in relation to this item. The distribution of responses implied that a larger number of experimental participants devoted more time to marking than the control group participants. It was suggested that the increase in time associated with the experimental group may be explained by a greater sense of doubt, and increased referral to the mark scheme. Neither of these latter findings was statistically significant, but it may be the case that the experimental intervention – the online training which they were experiencing for the first time – increased their sense of doubt, over and above that observed among the control group who were very familiar with the form of training they received.

Indeed, although the experimental group gave lower global ratings of effort during the study compared with live marking and compared to the control group, their responses to the 'comprehensiveness' items (and the item relating to 'time') in some respects suggested otherwise. The majority of participants across both groups reported that the focus, consistency and thoroughness with which they marked the study scripts was no greater or less than what they would apply in the live environment. This is particularly reassuring as these are highly desirable and necessary behaviours on which the reliability of the marking process partly depends (Meadows and Billington, 2005).

There are some limitations to this exploration of ecological validity. The questionnaire represented a limited operationalisation of 'conscientiousness', and explores only one aspect of ecological validity. Further work to develop conative elements of conscientiousness in the context of examining - and other measures of ecological validity - may prove useful for future studies of this kind. It is also unfortunate that it was not possible to explore the self-reported measures against the participants' actual performance in the marking reliability study. This could have provided evidence of the nature of the relationship between self-reports and performance; in this case, to determine whether those who applied higher levels of 'conscientiousness' were rewarded with an increase in marking reliability (an objective measure of the reliability of participants' self-reports). However, it was deemed important to assure anonymity in order to gain truthful insights into their approach to the task, and to avoid responses biased by concerns about perceptions of professionalism or social desirability. The mix of positive and negative outcomes implies that the anonymity of the questionnaire served its intended purpose. Finally, as completing the questionnaire was part of examiners' commitment to the study, a response rate of 100% was secured. It should be acknowledged that the requirement to complete the questionnaire may have forced some examiners to respond when they might otherwise not have done so. This may have impacted on the quality and reliability of some responses.

The primary (marking reliability) study gives us an important insight into the impact of training on the reliability of examiners' marking that would have been difficult, if not impossible, to achieve in the live context. However, a stand-alone experimental study is not without its own complications and threats to validity (including how we measure and document the threats to validity). It is evident that both types of research – live and experimental – have an important role to play in assessing quality of marking in the light of significant procedural changes.

Suzanne Chamberlain
September 2008

REFERENCES

Ajzen, I. (1988). *Attitudes, Personality and Behaviour*. Milton Keynes: Open University Press.

Barrick, M. R. and Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Brooks, V. (1980). Improving *the reliability of essay marking: a survey of the literature with particular reference to the English language composition*. (CSE Research Project Report 5) Leicester: Leicester University.

Chow, S. L. (1987). Science, Ecological Validity and Experimentation. *Journal for the Theory of Social Behaviour*, 17 (2), 181-194.

Costa, P. T. and McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI) Professional Manual*. Florida: Psychological Assessment Resources.

Clark-Carter, D. (2003). Effect size: The missing piece in the jigsaw. *The Psychologist,* 16 (12), 636-638.

Cronbach, L. J. (1971). Test validation. In Thorndike, R. L. (Ed.). *Educational Measurement* (2nd edition). Washington: American Council on Education.

Denscombe, M. (2000). Social conditions for stress: young people's experience of doing GCSEs. *British Educational Research Journal,* 26 (3), 359-374.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: scope and limitations. *British Educational Research Journal,* 27 (4), 433-442.

Keeves, J. P. (ed.) (1988). *Educational Research, Methodology, and Measurement: An International Handbook*. Oxford: Pergamon Press.

Kim, J. and Mueller, C. W. (1978). *Factor Analysis: Statistical methods and practical issues*. London: Sage Publications.

Kvavilashvili, L. and Ellis, J. (2004). Ecological validity and the real-life/laboratory controversy in memory research: A critical and historical review. *History & Philosophy of Psychology,* 6, 59-80.

Meadows, M. and Billington, L. (2005). *A Review of the Literature on Marking Reliability.* AQA Research Paper. RPA_05_MM_RP_05.

Meadows, M. and Billington, L. (2007). *The effect of marker background and training on the quality of marking in GCSE English.* AQA Research Paper. RPA_07_MM_RP_047.

Messick, S. (1989). Validity. In Linn, R. L. (Ed.). *Educational Measurement* (3rd edition). New York: American Council on Education and Macmillan.

O'Donovan (2005). There are no wrong answers: an investigation into the assessment of candidates' responses to essay-based examinations. *Oxford Review of Education*, 31 (3), 395-422.

Schmuckler, M. A. (2001). What is Ecological Validity? A Dimensional Analysis. *Infancy*, 2 (4), 419-436.

Taylor, R., Chamberlain, S. and Meadows, M. (2008). *Comparing the effects of online and face-to-face training on marking reliability*. AQA Research Paper. RPA_08_RT_053.

Thorndike, R. L. and Hagen, E. P. (1977). *Measurement and Evaluation in Psychology and Education*. London: John Wiley & Sons.