Extended essay marking on screen: Is examiner marking accuracy influenced by marking mode?

Rebecca Hopkin, Martin Johnson*, Hannah Shiell, John F. Bell and Nicholas Raikes

Cambridge Assessment, Cambridge, United Kingdom

1 Regent Street, Cambridge CB2 1GG

*Corresponding author. Email: Johnson.m2@cambridgeassessment.org.uk

Abstract

Background: In the UK and elsewhere, large-scale educational assessment agencies are increasingly shifting towards having examiners mark examination scripts on screen rather than on paper. This shift has prompted questions about whether the mode of marking might influence examiner marking accuracy, particularly in relation to extended essay responses. This issue is important since it has implications for whether the stakeholders in large-scale assessments, including the candidates being assessed, can trust the marking outcomes if extended essays are marked on screen.

Purpose: The study reported in this paper aimed to investigate empirically whether the mode in which extended essays are marked influences the accuracy of marking outcomes. This study was completed as part a wider research project which looked broadly at the influence of marking mode on examiners' marking outcomes and processes for extended essays.

Sample: A sample of 12 Advanced GCE examiners participated in the study. The examiners were all relatively experienced, holding between 6 and 31 total years' experience (mean: 16.8 years) of marking for large-scale educational assessment agencies in the UK. Five of the examiners had some previous experience of marking essays on screen.

Design and methods: One-hundred-and-eighty Advanced GCE American History examination essays were selected and split into two matched samples of 90 essays. The 180 essays were blind marked on paper by the examination's Principal Examiner (PE) to establish a study reference mark for each essay. Following training and standardisation, the sample of examiners each marked one 90-essay sample on paper and one 90-essay sample on screen. To control for essay sample and for marking order a crossover research design was used and the examiners were allocated to one of four examiner marking groups. Marking accuracy was defined as the extent of agreement between the examiner marks and the corresponding PE reference marks. Based on this definition, descriptive and general linear modelling statistical analyses were used to investigate whether the magnitude or direction of examiners' marking accuracy was influenced by marking mode.

Results: No association was found between marking mode and the magnitude of marking accuracy, but an extremely weak association was found between marking mode and the direction of marking accuracy. This latter result was identified as both practically and statistically negligible. Overall the results presented no substantial evidence to indicate that marking accuracy for extended essays was influenced by marking mode.

Conclusions: The results supported the conclusion that examiners are able to mark extended essays with equal accuracy on screen as they do on paper. The proposed practical implication of this conclusion is that extended essays can be marked on screen without compromising accuracy. The need for further investigation into the influence of marking mode on examiners' marking processes for extended essays is highlighted.

Keywords: assessment; screen marking; accuracy; extended essay

Background

Technological developments are impacting upon assessment practices in many ways. For large-scale educational assessment agencies in the UK and elsewhere, a key example of such impact is the ongoing shift towards examiners marking digitally scanned copies of examination scripts on screen rather than the original paper documents. This digital shift affords opportunities to improve the efficiency and effectiveness of marking and quality assurance procedures in ways that are not possible in traditional paper-based marking systems. At the same time, however, the shift towards marking scripts on screen has prompted important questions about whether the mode of marking might influence marking outcomes, particularly in relation to essay responses. These questions stand as a principal concern for large-scale educational assessment agencies, posing potential implications in terms of both the defensibility of marking outcomes and stakeholder trust in the assessment system. To this end it is important that assessment agencies gather and publish evidence showing that the judgements made by examiners are not influenced by the mode in which the marking is carried out.

In the context of examination essay responses, the notion of 'marking outcomes' would typically refer to the final marks issued to essays by examiners. Concerns about the influence of marking mode on essay marking outcomes might therefore tend to centre on the accuracy of essay marks: whether the marks awarded by examiners are equally close to the 'true' essay mark in both marking modes.

To understand why a mode-related influence on essay marking accuracy might exist, it is important to explore the relationship between marking outcomes and examiner comprehension. Essay marking can be conceptualised in at least two different ways. Marking which focuses on constructing a *global* appreciation of essay quality in a holistic sense might differ from *atomistic* essay marking practices which build a profile of essay quality through giving accumulated credit for discrete components of an essay performance (Wood 1991). Research which considers global essay marking practices suggests that comprehension is a central part of assessment, with the marking of essays requiring an examiner to initially comprehend a text (Huot 1990; Sanderson 2001). An explanation of 'comprehension' is offered by Johnson-Laird (1983), who proposes that a reader attains comprehension by building a mental model of the text as a whole. Wider literature proposes that this mental model building involves the reader integrating textual information with their own existing knowledge (Weir and Khalifa 2008), while also using their working memory to retain the spatial location of key concepts in the text (Fischer 1999; Kennedy 1992). These propositions together suggest that in order to comprehend an essay, and consequently reach a marking outcome, examiners are required to align their knowledge of the essay mark scheme with the textual content of the response, while also remembering the incidence of key essay features.

In the course of marking an essay, examiners draw upon a number of manual and cognitive marking processes to aid their development of comprehension (Crisp and Johnson 2007; Johnson and Nádas 2009a). It is here that marking mode might interfere most tangibly with marking accuracy, with screen and paper-based marking modes each possessing characteristics or 'affordances' (Gibson 1979) that can affect examiners' marking processes.

For this study, manual marking processes refer to reading behaviours such as navigation, manual interaction (i.e. physical contact with the text) and annotation. It is recognised widely in the literature that if these reading behaviours are hindered or modified, for example through a change of reading mode, then they will interfere directly with readers' comprehension (for example, De Cosio and Dyson 2002; Dillon 1994; O'Hara and Sellen 1997; Piolat, Roussey and Thunin 1997; Pommerich 2004).

Cognitive marking processes are more difficult to objectively define; for this study they are considered to include the processing demands which contribute to examiner cognitive workload. This notion of cognitive workload is clearly defined by Noyes, Garland and Robbins (2004: 111) as the "interaction between the demands of a task that an individual experiences and his or her ability to cope with these demands". Literature proposes that reading on screen presents a higher cognitive workload than reading on paper (Wästlund et al. 2005) and, furthermore, that this increased cognitive workload reduces readers' ability to comprehend a text (Just and Carpenter 1987; Mayes, Sims and Koonce 2001).

These messages from existing literature about reading imply that essay marking mode may influence both cognitive and manual marking processes, which in turn may influence examiners' comprehension. In the context of essay marking it would be expected that any such influence on comprehension levels would become apparent through the accuracy of examiners' essay marks. In this sense, literature about reading offers theoretical support for the concern that the shift from paper to screen marking may influence essay marking accuracy. However, direct empirical investigation of this concern is very limited, with little publicly available research exploring the influence of marking mode. Given the potentially critical implications of the outcomes of such investigation, this stands as a considerable deficit in the field of educational assessment research.

An important step forward in addressing this deficit has been made in a small number of recent screen essay marking research studies (Coniam 2009; Fowles 2008; Johnson, Nádas and Bell 2009; Shaw and Imam 2008). Each of these studies has investigated whether the mode in which essays are marked influences the accuracy of the marking of those responses. Considering essays with lengths in the region of 150 to 600 words, these studies report a negligible mode-related influence on marking accuracy and should therefore go a considerable way towards reassuring those with concerns that the shift from paper to screen marking might influence the accuracy of examiners' marking.

Although the findings of these studies contrast with the theoretical evidence from literature about reading, the Johnson, Nádas and Bell (2009) study suggests that there might still be some unresolved issues which require further research. In addition to investigating the influence of marking mode on General Certificate of Secondary Education (GCSE)[1] English Literature essay marking accuracy, the Johnson, Nádas and Bell (2009) study investigated the influence of marking mode on examiners' manual and cognitive marking processes (Johnson and Nádas 2009b). While the study results identified no mode-related influence on examiner accuracy, they revealed a systematic mode-related influence on both manual and cognitive marking processes. The study found that examiners behaved differently across screen and paper marking

[1] GCSEs are the main form of Level 1 and 2 national examinations taken at the end of compulsory schooling in the UK. Higher tier GCSEs are usually taken by more able students as they are only awarded between GCSE grades D and A*.

modes in terms of their navigation, annotation and manual interaction behaviours (Johnson and Nádas 2009a). Furthermore, examiners experienced heightened cognitive workload while marking on screen (Johnson and Nádas 2009b). These findings led the authors to question whether marking accuracy would remain unaffected by marking mode for essays longer than approximately 600 words, since the marking of such essays may significantly influence manual marking processes, and may result in even greater cognitive workload for examiners.

In summary, previous research has established that marking mode has no influence on examiners' marking accuracy when marking essays shorter than approximately 600 words. However, some of this research also proposes that the influence of mode on examiners' manual and cognitive marking processes might be heightened for essays beyond this length. This highlights a need to investigate the influence of mode on marking accuracy for extended essays, broadly defined in this study as essays longer than approximately 600 words.

Purpose

The main purpose of the present study was to investigate empirically the influence of marking mode on examiner marking accuracy for extended essays. Addressing an area with limited existing research literature, this study used Advanced Level General Certificate of Education (GCE)[2] extended essays to investigate the research question: 'Is marking accuracy for extended essays influenced by marking mode?' This study was completed as part of a wider research project which looked broadly at the influence of marking mode on examiners' marking outcomes and processes for extended essays.

Design and Methods

This study replicated the research design of the Johnson, Nádas and Bell (2009) study using Advanced GCE extended essays taken from a unit in American History. During this examination the candidates were given 90 minutes to write two extended essays, with each essay response being awarded a maximum of 60 marks which fit into one of seven different Band categories (Band 1 highest quality, Band 7 lowest quality).
As a precursor to the study, 913 Advanced GCE American History paper scripts from the June 2009 examination session were scanned before being marked operationally on paper. 831 of these scripts contained responses to the particular essay question that was to be the focus of this investigation. The operational marks awarded for these essays in the June 2009 session were then used to select a total of 180 essays, divided into two matched samples of 90. Analysis of the 180 sample essays showed that the average essay length within the sample was approximately 900 words (5.3 sides of A4 text). This analysis confirmed that the sample essays in this study were, on average, substantially longer than the sample essays in the earlier screen marking studies (for

[2] The Advanced Level General Certificate of Education (GCE) is a Level 3 national examination. It is usually studied over a two year period and is widely recognised in England, Wales and Northern Ireland as being the standard entry qualification for assessing the suitability of applicants for academic courses in UK Universities.

which the maximum average length within a sample was approximately 600 words or 3.4 sides of A4 text).

The 180 sample essays were blind marked on paper by the examination's Principal Examiner (PE) to establish a study reference mark for each essay. These PE reference marks validated the sample construction; essays from all mark scheme Band levels were represented in both samples and the mean marks for each sample were closely matched (Table 1). Independent-samples t-test analysis showed that the small differences between the mean marks of the two samples were not statistically significant (t (178) = -1.30, p = .20).

Table 1. Mean and standard deviation of PE reference marks by essay sample.

	N	PE reference marks	
		Mean	Standard deviation
Sample 1	90	33.98	9.67
Sample 2	90	35.98	11.01

The study involved a purposive sample of 12 Advanced GCE American History examiners from within a large UK-based educational assessment agency. The examiners were all relatively experienced, with each having between 6 and 31 years' marking experience (mean: 16.8 years), and between 2 and 10 years' experience of marking the Advanced GCE American History paper (mean: 5.9 years). Five of the examiners had some previous experience of marking essays on screen.

Each of the 12 examiners was issued one 90-essay sample to mark on paper and one 90-essay sample to mark on screen. To control for any potentially confounding effects of essay sample or marking order, the examiners were randomly allocated to one of four examiner marking groups (Table 2). This crossover design would allow subsequent analyses to specifically isolate the influence of marking mode.

Table 2. Examiner marking groups and essay allocation design.

Examiner marking group	First marking	Second marking
1	Sample 1 - Paper	Sample 2 – Screen
2	Sample 2 - Paper	Sample 1 – Screen
3	Sample 1 - Screen	Sample 2 – Paper
4	Sample 2 - Screen	Sample 1 – Paper

The marking software used for the study was a specially designed system based on an operational version which had already been used in examination marking with short response items. Whilst the software system was specifically built to enable the multiple distribution of essays demanded by the study research design, the software user interface mirrored that of the operational version. The software allowed

the examiners to download and navigate essay scripts as scanned PDF files. In the software environment the examiners had access to an assortment of marking tools; these included a variety of pre-specified annotation 'stamps', a final comments facility, and a zoom function. As well as the script currently being marked, examiners were able to access essays that they had marked previously.

Before starting their marking all of the examiners attended a two-day meeting at the study office base. The first day of the meeting was used to provide tailored examiner training in how to use the on screen marking software, including time to individually mark 20 practice essays. The second day dealt with examiner standardisation in both paper and screen marking modes. The process of standardisation included interactive discussions about the mark scheme, as well as PE exemplifications of the expected marking standard.

Usual paper marking practice for the sample of examiners conforms to a devolved marking model, where they receive scripts sent from the coordinating assessment agency and mark them to a deadline before returning them to the agency. In order to replicate the normal marking experience as much as possible the examiners were encouraged to complete their marking away from the study office base. For paper marking this was possible for all of the examiners. However, for screen marking the examiners were required to verify that their computer systems complied with the minimum requirements for the online screen marking software. Eleven examiners' computer systems conformed to these requirements, leaving one examiner to complete their marking in the study office base.

Results

The marking completion rates were high. All 12 examiners marked their full allocation of 90 paper essays but eight of the examiners were unable to complete a small amount of their screen essay allocation for technical reasons[3]. This left 13 unmarked essays across all examiners and an overall marking completion rate across modes and examiners of 99.4 per cent.

Table 3 presents the mean and standard deviation of examiner essay marks by marking mode for the 12 examiners, both individually and overall. For the 12 examiners overall the mean examiner mark was less than half a mark higher on screen than on paper, presenting no evidence of a substantive mode-related influence on examiner marking.

[3] There were some unanticipated problems with the multiple script delivery feature of the software. This software feature was specially designed for use in this study, enabling the demands of the research design to be met.

Table 3. Mean and standard deviation of examiner marks by marking mode and examiner.

Examiner	Paper examiner marks				Screen examiner marks			
	Essay sample	N	Mean	Standard deviation	Essay sample	N	Mean	Standard deviation
1	1	90	33.33	10.12	2	89	34.52	9.31
2	1	90	37.67	8.56	2	89	39.76	8.43
3	1	90	32.68	7.48	2	89	32.30	7.06
4	2	90	35.69	10.03	1	87	36.46	7.92
5	2	90	33.87	8.00	1	90	33.76	7.93
6	2	90	35.90	9.69	1	88	37.16	9.02
7	2	90	32.81	7.97	1	90	34.01	8.29
8	2	90	35.10	10.46	1	90	35.11	9.74
9	2	90	30.39	6.66	1	88	32.10	6.04
10	1	90	37.13	9.81	2	88	35.28	9.39
11	1	90	38.34	9.78	2	90	38.30	9.37
12	1	90	37.04	9.97	2	89	36.67	8.35
All	All	1080	35.00	9.36	All	1067	35.45	8.72

These initial analyses of the raw examiner marks offer a clear starting point for the investigation of mode-related influences on marking accuracy; however, they hold a notable limitation. Without comparison to any marking standard the raw examiner marks can provide little indication of the accuracy of marking. For example, leniently marked essays may be compensated for by severely marked essays in the calculation of a mean. As a result, the correspondence between mean examiner marks across modes could not be claimed as a correspondence in accuracy.

There are many definitions of marking accuracy (Bramley 2007). In this study it was assumed that the PE reference marks represented the 'true' marks for the sample essays. Marking accuracy was therefore defined as the extent of agreement between examiner marks and the corresponding PE reference marks. In light of this definition, further analyses considered two distinct measures of marking accuracy:

(1) Absolute mark difference:

The absolute difference between an examiner mark and the corresponding PE reference mark. This measure assigns all differences a positive value, regardless of their direction. Absolute mark differences therefore provide a clear indicator of the *magnitude* of marking accuracy: smaller absolute mark differences represent greater marking accuracy.

(2) Actual mark difference:

The actual difference between an examiner mark and the corresponding PE reference mark. This measure assigns a negative value to marks below the

reference mark and a positive value to marks above the reference mark. Actual mark differences therefore provide a useful indicator of the *direction* of marking accuracy: negative actual mark differences represent severe marking and positive actual mark differences represent lenient marking.

Table 4 presents descriptive analyses of the overall absolute and actual mark differences for all the examiners, by marking mode. As well as presenting the mean absolute and actual mark differences for each marking mode, the table presents the standard deviation (SD) and median of these mark differences. The standard deviation statistics show the distribution of the mark differences, giving an idea of how much variation there was from the mean result. The median statistics show the mid point of the complete mark difference distribution.

Table 4. Absolute and actual mark differences between examiner and PE marks by marking mode.

	Marking mode	
	Paper	Screen
N	1080	1067
Absolute mark difference		
Mean	5.82	5.74
Standard deviation	4.86	4.45
Median	4.5	5
Actual mark difference		
Mean	0.02	0.47
Standard deviation	7.59	7.25
Median	0	1

Descriptive analyses of absolute mark differences revealed that in both marking modes half of all examiner marks were awarded within five marks of the corresponding PE reference mark. Furthermore, a disparity of just 0.08 marks between mean absolute mark differences was identified across modes. Given the 60-mark range available for the essays, absolute mark differences across the 12 examiners therefore suggested close equivalence in the overall magnitude of marking accuracy on paper and on screen.

Descriptive analyses of actual mark differences added greater depth to this picture. On paper the overall median absolute mark difference was 0 and mean absolute mark difference 0.02, indicating a balance of leniency and severity in marking. In contrast, on screen the overall median absolute mark difference was 1 and mean absolute mark difference 0.47, indicating a very slight tendency towards more lenient marking on screen. These relationships are presented in Figure 1, which compares the distribution of actual mark differences by marking mode. The slight tendency towards more lenient marking on screen is clearly apparent: the longest bar on paper is the bar crossing 0 (representing differences in the range -2 to 2 marks), but the longest bar on screen is the bar directly above this (representing differences in the

range 2 to 6 marks). However, the length of the bars is very similar in the screen and paper marking modes, reflecting the slightness of the mode-related difference. Overall, despite no substantive mode-related differences in the magnitude of marking accuracy, actual mark differences across the 12 examiners suggested a tendency for slightly more lenient marking on screen than on paper.

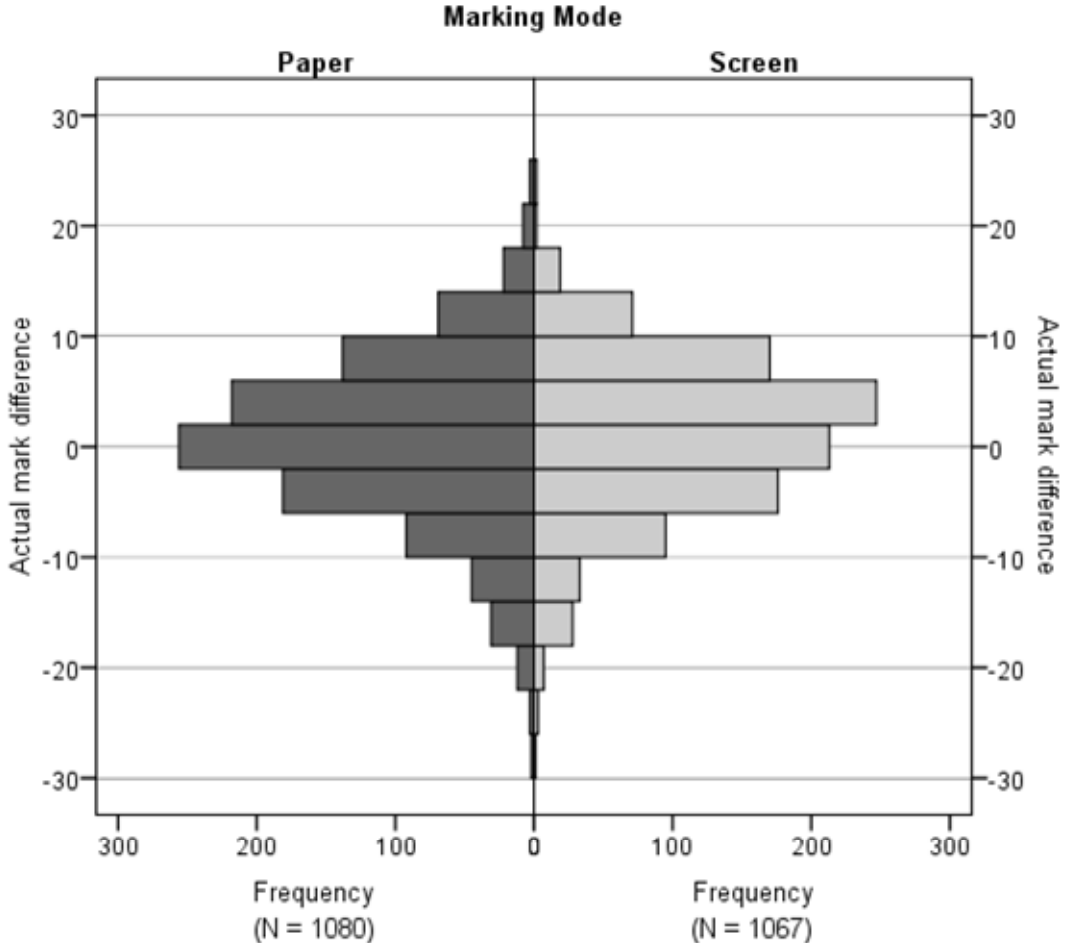

Figure 1. Distribution of actual differences between examiner and PE marks by marking mode.

Note: Each bar is four marks wide.

To enhance the certainty of these tentative descriptive outcomes, subsequent analyses were undertaken to test the statistical significance of any relationships between marking mode and marking accuracy. The specific statistical method adopted for these analyses was general linear modelling. A key advantage of this method was its ability to model the structure of the marking data, allowing exploration of the statistical association between marking mode and marking accuracy while controlling for features of the research design.

Two general linear models were fitted, the first fitting absolute mark differences (Model 1.1) and the second actual mark differences (Model 1.2) between the examiner marks and the PE reference marks. The specific design features controlled for in each of these models were individual examiner, essay sample and

individual essay within the essay sample. The final specifications of the general linear models are presented in Table 5.

Table 5. General linear model specifications for absolute and actual mark differences between examiner and PE marks.

	Dependent variable	Key independent variable	Control variables
Model 1.1	Absolute mark difference	Marking mode	Examiner Essay sample Individual essay (*nested in essay sample*)
Model 1.2	Actual mark difference		
Dependent variable = key independent variable + control variables + error			

Table 6 displays the outcomes of the general linear models fitted in this analysis. Results of Model 1.1 identified no statistically significant association between absolute mark differences and marking mode. This key outcome reiterated the findings of the descriptive analyses, confirming that there were no statistically significant mode-related differences in the overall magnitude of marking accuracy.

Table 6. Results for general linear models of absolute and actual mark differences between examiner and PE marks.

ANCOVA table (N = 2147)							
Variable	DF	Model 1.1: Absolute mark difference			Model 1.2: Actual mark difference		
		Type III SS	*F*	*p*	Type III SS	*F*	*p*
Marking mode	1	4.23	0.26	0.61	106.10	4.14	0.04
Examiner	11	789.19	4.34	< 0.001	10481.91	37.20	< 0.001
Essay sample	1	61.07	3.70	0.05	3002.49	117.20	< 0.001
Individual essay *(nested in essay sample)*	1	13453.51	4.57	< 0.001	54497.48	11.95	< 0.001
Error	1955	32308.83			50083.57		

Note: ANCOVA, analysis of covariance; DF, degrees of freedom; SS, sum of squares.

Results of Model 1.2 identified that there was a statistically significant association between marking mode and actual mark differences, at the five per cent level. Controlling for features of the research design, the examiners were on average 0.44 marks (B = -0.44, 95% CI = ± 0.43) more lenient in their marking on screen than on paper. However, the effect size for this result, another statistical indication of the estimated magnitude of the relationship, was almost negligible (partial eta squared =

0.002), highlighting that the strength of statistical association between marking mode and actual mark differences was extremely weak. This association was explored further by calculating adjusted means for actual mark differences by marking mode (Table 7).

Adjusted means are values predicted using the outcomes of a general linear model. In this analysis adjusted means for actual mark differences represent the predicted mean actual mark differences by marking mode once all other variables in the general linear model were controlled for. On paper the adjusted mean actual mark difference was 0.02, highlighting that marking was equally lenient and severe on paper. On screen the adjusted mean actual mark difference was 0.47, highlighting a very slight tendency towards more lenient marking on screen. These values confirmed that the scale of association between marking mode and actual mark differences was very small.

Table 7. Adjusted means for actual mark differences between examiner and PE marks by marking mode.

	Marking mode	
	Paper	Screen
N	1080	1067
Actual mark difference		
Adjusted mean	0.02	0.47
Standard error	0.23	0.22
95% Confidence interval	± 0.45	± 0.44

Overall, the outcomes of the general linear models fully reinforced those of the descriptive analyses. The magnitude of examiner marking accuracy was identified as equivalent across modes; the examiners deviated from the PE reference mark in equal magnitude whether marking on paper or on screen. However, the direction of marking accuracy displayed a very weak but significant association with marking mode, with examiners displaying slightly more leniency in their marking on screen than on paper.

Discussion

This study had a number of limitations that need consideration while discussing the research results. As a marking simulation exercise, the study differed from operational practices and contexts in the following key ways:

- The outcomes of the marking exercise had no consequence for candidates.
- The marking exercise afforded a comparatively generous time allowance.
- The total marking allocation of 180 essays was comparatively light.
- The previous marking experience of the sample of examiners was relatively high.
- The examiners were standardised twice, once in each marking mode.

- Just one example of marking software was used.

Together these issues might influence the generalisability of the results in ways that are difficult to quantify.

While acknowledging these limitations, this study set out to address a notable deficit in the research literature about marking on screen by investigating the influence of marking mode on examiner marking accuracy for extended essays. Specifically, the study sought to answer the key research question 'Is marking accuracy for extended essays influenced by marking mode?'

In the context of this study, the 'true' mark for an essay was assumed to be the reference mark awarded for it by the PE marking on paper. The accuracy of other examiner marks for the essay was defined in relation to this PE reference mark. Two measures of marking accuracy were used: *absolute* and *actual* mark differences between the examiner and PE marks. These differences represent the magnitude and direction of marking accuracy respectively; smaller values represent greater accuracy.

Statistical analyses found no evidence that the magnitude of examiner marking accuracy was influenced by marking mode: marking accuracy was as high for marking on screen as for marking on paper. Statistical analyses of the direction of marking accuracy, however, identified a statistically significant mode-related influence at the five per cent level. On average, marking on screen tended slightly towards the more lenient direction, whereas on average there was a balance of leniency and severity on paper. This apparent tendency towards leniency on screen meant that in relation to the reference marks, essays were awarded an average of 0.44 marks more out of 60 on paper than on screen.

Despite the statistical significance of the association identified between marking mode and marking leniency, interpretation of this outcome should be approached with caution. From a statistical perspective, the effect size for this result was almost negligible, highlighting an extremely weak association. Furthermore, from a practical perspective, the importance of a difference of less than half a mark out of 60 is certainly debatable, especially given the finding that the magnitude of marking accuracy was not influenced by marking mode. In light of these perspectives, the results presented no substantial evidence to indicate that marking accuracy for extended essays was influenced by marking mode.

The relatively high level of marking experience of the sample of examiners might be considered to be a potential limitation of the results, arguably constraining generalisation of the findings to other examiners with less marking experience. On the other hand, this specific sample characteristic might be considered to provide an advantage in that the results provide reassurance to stakeholders who fear that experienced examiners will not be able to make the transition to marking screen without adversely influencing the accuracy of their marking.

The results of this study both support and expand those of the existing screen essay marking research (Coniam 2009; Fowles 2008; Johnson Nádas and Bell 2009; Shaw and Imam 2008). These studies provided empirical evidence to suggest that marking mode has a negligible influence on marking accuracy for essays in the region of 150 to 600 words, a finding expanded by this study to include essays in the region of approximately 900 words.

Returning to the literature outlined in the background to this paper, the results from the present and previously cited studies comparing paper and screen marking of

essays are at odds with expectations from theory about reading in general. Messages from this literature suggest that essay marking mode may influence examiner marking processes, which in turn may influence examiner comprehension and marking accuracy. One potential explanation for the absence of any substantial mode-related influence on extended essay marking accuracy in this study might be that any mode-related influences on examiner marking processes were too small to affect their marking outcomes. Further research into the nature and extent of any mode-related influences on marking processes would therefore be of theoretical interest.

Conclusions

Within the limitations of the research design, this study presents evidence to suggest that marking accuracy for extended essays is not influenced by marking mode, supporting the conclusion that examiners are able to mark extended essays with equal accuracy on screen as they do on paper.

The key practical implication of this finding is that extended essays can be marked on screen without compromising accuracy. This finding is of great importance to large-scale educational assessment agencies and their stakeholders, and opens the way to the expansion of screen marking to high stakes assessments involving extended essays. Caution might be urged during this transition, however, since many factors uncontrolled in this study may have a greater or lesser effect in operational contexts.

Finally, while the study reported in this paper offers a crucial insight into how marking mode might influence marking accuracy for extended essays, it needs to be acknowledged that this study is necessarily limited. By focusing mainly on marking *outcomes* this study overlooks some of the anticipated mode-related influences on examiners' manual and cognitive marking processes for extended essays. These concerns are addressed in additional studies which are part of the wider research project from which this study originates.

References

Bramley, T. 2007. Quantifying marker agreement: Terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication* 4: 22-7.

Coniam, D. 2009. A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation* 15, no. 3: 243–63.

Crisp, V. and M. Johnson. 2007. The use of annotations in examination marking: Opening a window into markers' minds. *British Educational Research Journal* 33, no.4: 943–961.

De Cosio, M. G. and M. C. Dyson. 2002. Methodology for manipulating electronic documents in relation to information retrieval. *Visible Language* 36, no.3: 282-306.

Dillon, Andrew. 1994. *Designing usable electronic text.* London: Taylor & Francis.

Fischer, M. H. 1999. Memory for word locations in reading. *Memory* 7, no.1: 79–118.

Fowles, Dee. 2008. Does marking images of essays on screen retain marker confidence and reliability? Paper presented at the International Association for

Educational Assessment Annual Conference, September 7-12, in Cambridge, England.

Gibson, James J. 1979. The ecological approach to visual perception. Boston, MA: Houghton Mifflin.

Huot, B. 1990. Reliability, validity and holistic scoring: What we know and what we need to know. *College Composition and Communication* 41: 210-3.

Johnson, M., and R. Nádas. 2009a. Marginalised behaviour: Digital annotations, spatial encoding and the implications for reading comprehension. *Learning, Media and Technology* 34, no.4: 323-36.

Johnson, M., and R. Nádas. 2009b. An investigation into marker reliability and some qualitative aspects of on-screen essay marking. *Research Matters: A Cambridge Assessment Publication* 8: 2-7.

Johnson, M., R. Nádas, and J. F. Bell. 2009. Marking essays on screen: An investigation into the reliability of marking extended subjective texts. *British Journal of Educational Technology.*

Johnson-Laird, Philip N. 1983. *Mental models: Towards a cognitive science of language, inference and consciousness.* Cambridge MA: Harvard University Press.

Just, Marcel A., and Patricia A. Carpenter. 1987. *The psychology of reading and language comprehension.* Boston, MA*:* Allyn & Bacon.

Kennedy, Alan. 1992. The Spatial Coding Hypothesis. In *Eye movements and visual cognition,* ed. K. Rayner, 379-97. New York: Springer-Verlag.

Mayes, D.K., V. K. Sims., and J.M. Koonce. 2001. Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics* 28: 367-78.

Noyes, J. M., and K. J. Garland. 2008. Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics* 5, no.9: 1352-75.

O'Hara, Kenton., and Abigail Sellen. 1997. A comparison of reading paper and on-line documents. In *Proceedings of the ACM Conference on human factors in computing systems,* 335-42. New York: ACM Press.

Piolat, A., J-Y. Roussey, and O. Thunin. 1997. Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies* 47: 565-89.

Pommerich, M. 2004. Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based texts. *The Journal of Technology, Learning and Assessment* 2, no.6: 3-45.

Sanderson, P. 2001. Language and differentiation in examining at A Level. PhD diss., University of Leeds.

Shaw, Stuart., and Helen Imam. 2008. *On-screen essay marking reliability: Towards an understanding of marker assessment behaviour.* Paper presented at the International Association for Educational Assessment Annual Conference, September 7-12, in Cambridge, England.

Wästlund, E., H. Reinikka, T. Norlander, and T. Archer. 2005. Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior* 21: 377–94.

Weir, C. J., and H. Khalifa. 2008. A cognition processing approach towards defining reading comprehension. *Research Notes* 31: 2–10.

Wood, Robert. 1991. *Assessment and testing: a survey of research.* Cambridge: University of Cambridge Press.