THE INFLUENCE OF PRESENTATION UPON EXAMINATION MARKS

J Hughes
School of Computing
University of Dundee
Dundee
jhughes@computing.dundee.ac.uk
www.computing.dundee.ac.uk

S Akbar
School of Computing
University of Dundee
Dundee
s.akbar@dundee.ac.uk
www.computing.dundee.ac.uk

ABSTRACT

An investigation was made into how the appearance of answer scripts provided in a computing-related assessment can affect the grades that students are then awarded, in particular comparing typed and traditional handwritten formats. The comparison was made using the answers to an Information Technology examination taken by first year computing undergraduate students. Each student completed the Information Technology exam using a PC and basic text editor to type their answers. Original typed scripts were then transcribed into a handwritten version and a grammatically-correct and spell-checked version. All three versions were marked by independent assessors. In an adaptation of the methodology adopted by Russell and Tao (2004b), each assessor was assigned one presentation format only.

Analysis of the marks awarded indicated that the original typed answers were rated significantly more highly than handwritten versions of the same answers. This may be because the assessors were more comfortable marking an answer produced in a word processed format, since they were specialists in the computing field. However, the set of "perfected" typed answers (spelling and grammatical mistakes had been removed) surprisingly were awarded lower grades than the original typed answers that did contain spelling and grammatical mistakes. Possible explanations for this are considered, for example, grammatically-correct scripts may have indicated a high level of professionalism to the assessors and hence raised their expectations of in-depth answers. Variation between marks awarded by assessors was higher than expected.

Keywords

Assessment; e-examination

1. INTRODUCTION

Research comparing handwriting with word processing for the presentation of assessment answers can broadly be identified as having two main phases. In the initial phase of the early 1990s, a small number of groups investigated the effects of two modes of answering an examination questions – handwritten or word processed – upon students' scores. One notable work in this phase was by Powers, Fowles, Farnum and Ramsey (1994). These authors converted a sample of original handwritten essay answers into word processed versions and transcribed a sample of original word processed essay answers into handwritten versions. Analysis of the scoring revealed that handwritten answers were awarded higher average scores than word processed answers, irrespective of the original mode in which the answers were produced. This replicated findings of Arnold, Legas, Obler, Pacheco, Russell and Umbdenstock (1990) that student papers converted to word-processed versions received lower scores than did the original handwritten versions. Arnold et al suggested that the reviewers may have had higher expectations of the word processed work, less empathy with the authors of word processed work or may have been less likely to give the benefit of any doubt. Powers et al suggested other possible factors including lack of evidence in the word processed versions of evidence to revise work, greater visibility of typographical errors, or *apparently* shorter answers.

A second phase of research in this area resulted in a series of publications from 2004 onwards. Principal authors have included Russell and Tao (2004a, 2004b) who found some evidence in support of the Powers et al work of 1994, suggesting that the lower scores for computer printed work may have resulted from factors identified by increased visibility of any errors, higher expectations of readers and less empathy with students ("felt a stronger connection to the writer because of the handwriting").

Shaw (2005) conducted a small study in which a group of IELTS examiners were presented with handwritten and typed format scripts. One factor commented upon was the legibility of the scripts: one examiner observed that "Lots of typos have a more negative affect than lots of handwritten ones …" and Shaw noted that markers give 'benefit of doubt' where legibility is an issue. Concerned that an effect appears to result from the mode of presenting assessment answers, Russell and Tao (1994b) were prompted to investigate if training of markers could eliminate any such differentials in assessment. The authors concluded this was possible although they recommended more research in this area. More recent work in the United Kingdom by researchers at the University of Edinburgh (Mogey, Paterson, Burk and Purcell, 2010) compared transcribed scripts of first year students in a mock examination: handwritten scripts were transcribed into typed format and typed scripts were transcribed into handwritten format. Mogey et al found "weak evidence" that handwritten scripts generally scored slightly higher than typed scripts.

2. RESEARCH QUESTION

Most of the studies described in section 1 applied to students of non-technical subjects. Powers et al (1994) used a sample of college or university students principally from business, social sciences, humanities/arts and education. Russell and Tao's 2004 work featured school students of Language Arts. Mogey et al (2010) investigated the work by divinity students. Although almost by definition most 'digital native' students are familiar with word processing, it seems unlikely that students of non-technical subjects will be as confident and competent using a keyboard to present answers as students of computing, who typically use keyboards on a daily basis for most aspects of their practical learning. The work described in this paper is part of an investigation comparing on-paper assessment with on-screen assessment in the subject area of computing. In particular for this paper, the focus was whether handwritten scripts are judged differently to typed scripts in the subject area of computing. In the style of Mogey et al (2010), the research question was: is the mark awarded to an examination answer influenced by its mode of presentation?

3. METHODOLOGY

The methodology adopted was based upon the study conducted by Russell and Tao (2004b). Assessments of 40 undergraduate students in their first year of study were used to test the comparison of handwritten and computer-based examination answers. All were studying computing at the University of Dundee in the School of Computing. All the students had regularly been using computers throughout the academic session and therefore were all experienced with typing, although not all were touch-typists.

The examination (of an I.T. module) took place at the end of semester 2 in academic session 2008/9. Questions were designed to assess knowledge and understanding of databases, human computer interaction, programming in general and experimental methods. The questions were delivered and answers obtained using an assessment management system (Questionmark Perception) that permits a secure scheduled delivery of exams, locked down to prevent internet access. Use of this system for high-stakes assessment had been piloted the previous semester with a different IT class. The examination was set and quality assurance checks performed in the normal way.

Students were provided with the short-answer style examination questions on-screen and were able to type answers using a rudimentary text editor built into the system. No spell checker is provided with the system, therefore spelling mistakes were as likely to occur as in a handwritten exam. The questions were presented one at a time with a feature which enabled the students easily to answer questions in whichever order they preferred (Figure 1).

In addition, students were provided with a standard answer booklet to use if they wished to sketch out answers, prepare rough notes or create diagrams. Answers typed were stored on the hard drive. Library and Learning Centre staff confirmed that all candidates' answers were stored, and later provided academic staff with the answer files. Anonymised answer scripts were printed and, for the purposes of the actual examination, distributed to academic staff for marking.

For the purposes of the research, the typed answers then were transcribed by a set of paid volunteers into handwritten format. In order to reflect the general quality and style of student handwriting, each transcriber was given a sample of each student's handwriting in order that they could mimic the handwriting style. The transcribers also were instructed to reproduce the exam exactly as originally typed and not to correct spelling or grammatical mistakes. However any obvious typographical errors were to be corrected. For example a mistake such as "teh" would have been corrected to "the" as it was judged to be a mistake due to the student's typing and therefore would not have occurred in a handwritten assessment. In addition to this, a further transcription was produced in which spelling mistakes and typographical errors were eliminated whilst keeping the typed format (Figure 2).

Figure 1. Examples of on-screen assessment question and answer system

ORIGINAL	PROCESS	TRANSCRIPTION
TYPED		HANDWRITTEN
TYPED		SPELL-CHECKED TYPED

Figure 2. Transcription processes

Computing PhD students familiar with I.T. were recruited as assessors. An agreed marking scheme was used to support independent marking and a moderator was used to resolve any discrepancies. Altogether there were 240 scripts: 40 original typed for each of two assessors; 40 transcribed (handwritten) for each of a further two assessors; an additional set of 40 were produced for each of a final two assessors to analyse the effect of correction of spelling mistakes. Scripts were allocated for marking to ensure that each assessor saw only one presentation format, i.e. to keep the markers as blind as possible to the expected outcomes of the study (Figure 3). Any question originally leading to the creation of a diagram was excluded from the subsequent analysis, thus resulting in a possible maximum score of 89 to be awarded. Scores awarded by each rater in a given pair then were aggregated into a single score by adding the two single scores.

Exam	Presentation format		
Level 1 I.T. (40 scripts)	Handwritten	Typed	Spell-checked typed
	Assessor A	Assessor C	Assessor E
	Assessor B	Assessor D	Assessor F

Figure 3. Order of presentation to assessors

4. RESULTS

Handwritten scripts generally scored lower than either typed or spell-checked typed scripts (Figure 4). The results showed a significant effect of the format for the total score: $F(2,117) = 6.081$, $p < .01$. Post-hoc comparisons (Tukey HSD) confirm that the scores differ significantly for (i) handwritten and typed scripts and (ii) typed and spell-checked typed (Figure 5).

	N	Mean	SD
Handwritten	40	69.1	22.57
Typed	40	86.6	23.32
Spell-checked typed	40	73.6	24.01

Figure 4. Descriptive statistics for marks scored in each format

	Mean Difference	Std. Error	Sig.
Presentation format: F(2,117) = 6.081, *p* < 0.01			
Handwritten vs typed	-17.50[*]	5.21	.003
Handwritten vs spell-checked typed	-4.50	5.21	.664
Typed vs spell-checked typed	13.00[*]	5.21	.037

*. The mean difference is significant at the 0.05 level.

Figure 5. Comparisons between marks scored in each format

This evidence suggests that higher marks were awarded to any form of typed presentation for exam scripts than to handwritten formats, but that spell-checking and removing errors resulted in less benefit than uncorrected typing. Mark variability might be due to the differences in the way that the different markers interpreted or applied the marking scheme. This possibility was also considered by Mogey et al (2010) and therefore, replicating that approach, the distribution of marks awarded by each marker was investigated (Figure 6). A broadly similar range, skew and standard deviation was shown across the markers.
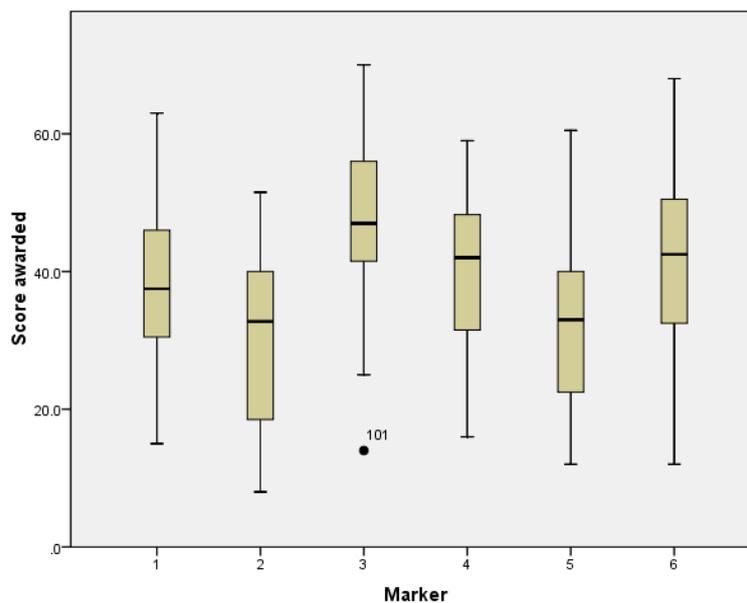

Figure 6. Boxplot of score by marker

However, when the marks were categorised (1st, 2.1, 2.2, 3rd or fail), the inter-rater reliability for each of the three pairs was found to have *fair* agreement only (Landis and Koch, 1977). This finding corresponds with that of Mogey et al (2010), who noted the importance of the "differences between the markers".

5. DISCUSSION

Analysis of the marks awarded indicated two findings. Firstly, the original typed answers were rated significantly more highly than handwritten versions of the same answers. This may be because the assessors were more comfortable marking an answer produced in a word processed format, since they were specialists in the computing field. More surprisingly, the second finding is that the set of "perfected" typed answers (spelling and grammatical mistakes had been removed) were awarded lower grades than the original typed answers that did contain spelling and grammatical mistakes. Possible explanations for this include expectations of the assessors and empathy from assessors. For the former, grammatically-correct scripts

may have indicated a high level of professionalism to the assessors and hence raised their expectations of in-depth answers – a computer-student version of the effect observed by Russell and Tao (2004a). Considering the latter finding, there may be an effect specific to computing with the juxtaposition of visibility effect (Powers et al, 1994, considered that greater visibility of spelling and punctuation errors may adversely affect rater scores) and empathy effect (Russel and Tao suggested that raters might be more lenient if they felt a stronger connection to the writer). In this case, it may be that computing markers felt a stronger connection to a writer who made spelling and punctuation errors in their typed answers than to a writer whose typing was 'perfect' but whose computing knowledge was flawed.

6. CONCLUSION

This study was designed to investigate the effect of different presentation formats upon the marks awarded to examination assessments. Results suggest that in the discipline of computing, students may benefit somewhat from typing answers rather than handwriting them, irrespective of the quality of the answer itself. Moreover, markers may favour somewhat flawed typed answers over scripts with answers with spelling or grammatical errors. Given the choice of doing high-stakes assessments on computer or by handwriting, students may well be advised to choose to type their examination answers. Examination boards may wish to consider *requiring* one presentation format, rather than giving students a choice. Finally, there is a strong indication that notable marker variability existed in this study: programme leaders may wish to review quality control processes that exist for marking.

7. LIMITATIONS

The principal limitation of this work lies in the design: selected to maximise the preservation of blind marking, it therefore was not counterbalanced. Had it been, more control could have been exercised over any differences in interpretation of the marking scheme by the different markers: each of the markers could have scored one third of each presentation format. Nonetheless, the design chosen perhaps has resulted in the identification of an unexpected issue, that of variation between the markers. Further work will include a revision to the design with further data collection and analysis. The fact that the subjects were from one discipline area only is not judged to be a limitation, since it was the presentation format for that discipline which was under investigation.

8. ACKNOWLEDGEMENT

9. REFERENCES

Arnold, V., Legas, J., Obler, S., Pacheco, M.A., Russell, C. and Umbdenstock, L. (1990) Do students get higher scores on their word-processed papers? A study of bias in scoring hand-written versus word-processed papers. Unpublished manuscript cited in Powers et al (1994).

Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics,* 33 pp159-74.

Mogey, N., Paterson J., Burk J. and Purcell, M. (2010) Typing compared with handwriting for essay examinations at university: letting the students choose. *ALT-J,* 18 (1), pp 29-47.

Powers, D.E., Fowles, M.E., Farnum, M. and Ramsey, P. (1994) Will They Think Less of My Handwritten Essay If Others Word Process Theirs? Effects on Essay Scores of Intermingling Handwritten and Word-Processed Essays. *Journal of Educational Measurement*, 31 (3), pp 220-233.]

Questionmark Perception, http://www.questionmark.co.uk, (last accessed 21.05.2010)

Russell, M. and Tao, W. (2004a) Effects of handwriting and computer-print on composition scores: a follow-up to Powers, Fowles, Farnum & Ramsey. *Practical Assessment, Research & Evaluation,* 9 (1).

Russell, M. and Tao, W. (2004b) The influence of computer-print on marker scores. *Practical Assessment, Research & Evaluation,* 9 (10).

Shaw, S.D. (2005) The impact of word processed text on marker behaviour: a review of the literature. Cambridge ESOL internal report no. 670.