British Journal of Educational Technology *Vol 33 No 2 2002* 173–188

Responding to free response examination questions: computer versus pen and paper

Robert MacCann, Benjamin Eastment and Samantha Pickering

Dr Robert MacCann, Ben Eastment and Samantha Pickering form a team comprising the Measurement and Research Services Unit at the NSW Board of Studies. Address for correspondence: Dr Robert MacCann, Head, Measurement and Research Services, NSW Board of Studies, GPO Box 5300, Sydney, NSW, Australia 2001. Email: maccann@boardofstudies.nsw.edu.au

Abstract

A two-part study, involving 14- to 15-year-old high school students, compared two modes of responding to free response examination questions: by computer and, the traditional way, by pen and paper. In Part 1, both the computer group ($n = 57$) and the pen and paper group ($n = 52$) were formed by random assignment. They answered three essay questions from a 1997 external English test which were holistically marked. In Part 2, the computer group ($n = 88$) and the pen and paper group ($n = 53$) chose their preferred method of response. They answered two essay questions from a 1999 external English test, which were analytically marked using standards-referenced criteria. For four of the five questions, there were no significant differences between the pen and paper response marks and the computer response marks when presented in their original formats. When the pen and paper responses were word-processed, markers tended to award higher marks to the handwritten scripts.

Background

The widespread availability of the personal computer is leading to changes in the ways that students compose and submit written tasks. Increasing numbers of students now are more comfortable in composing a written response directly into a word processor rather than writing the response in longhand and getting it typed when finished. In some schools, many students have been using laptops to compose and edit their written responses throughout their entire stay at school. It is argued that they are comfortable working in this way and that to compose an essay in longhand now would be somewhat unnatural. In the US, Russell and Plati (2001) have argued that requiring students to generate responses to open-ended items by pen and paper under-estimates the achievement of students accustomed to writing using a computer. In Great Britain, Hartley (1993) refers to a student who was forbidden by the teacher to use a computer when writing essays, as the examinations required longhand. The present study took place

Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford, OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA.

in New South Wales, Australia where the system of external examining is currently based on the students writing out their answers in longhand. Many other countries have a similar system.

On the other hand, many students do not have ready access to a personal computer and have only had experience in responding by longhand. To them, it would be natural to continue to respond to examination questions in this way. Further, there is the mass of students who would have had varying degrees of exposure to word processors but would not have had the extensive experience in their use of the first group mentioned above. This leads to the question of whether all groups of students can be accommodated in public examinations by responding to the questions in the mode that best suits them and whether this is fair to all groups.

The argument for claiming computer-using students would be disadvantaged if denied their use in examinations needs to be made explicit. Presumably it rests on the assumption that their longhand essay writing skills would have deteriorated somewhat through lack of practice, or that composing an essay by word-processing somehow interferes with and reduces performance in composing an essay by longhand. To test this assumption requires a design with two groups equivalent in measured achievement at a given time, before embarking on different treatments: one continuing in longhand essay writing and the other responding by computer. After a given time, perhaps corresponding to six years of secondary school, the groups would be retested in responding by longhand. If the computer group mean was significantly lower, one could claim evidence of this kind of disadvantage. There have been few studies of this nature. Dalton and Hannafin (1987) conducted a study where the time period was a year and the computer group worked two exercises each week, but found no significant difference between the two groups when tested on a longhand essay writing task.

An opposing viewpoint is that groups using computers in examinations may be advantaged in relation to others responding by longhand, through being able to use the special features of computers. Respondents by computer can set down their ideas more quickly, knowing that individual sentences and the essay as a whole can be re-organised and improved through the cutting and pasting of text and so on. If sufficiently experienced in computer usage, respondents could use these features to gain an advantage.

This study was intended to address three questions that arise from a consideration of the experimental designs in the literature:

1. Do raters vary when judging handwritten versus computer printout of the same text?
2. For equivalently matched groups, does writing on computer versus writing on paper give similar mean scores *when presented in the same format?*
3. For equivalently matched groups, does writing on computer versus writing on paper give similar mean scores *when presented in their original formats?*

The studies discussed below show the range of designs that have been employed to deal with these three questions. The first question was addressed directly in a study by

Arnold, Legas, Obler, Pachero, Russell and Umbdenstock (1990). Essay responses that were originally handwritten were word-processed and marked using the same procedures as the originals. They found that the original handwritten essays received higher scores on average than the word-processed versions. Similar studies by Sweedler-Brown (1991, 1992) also found that handwritten essays were favoured over their word-processed equivalents.

In Wolfe, Bolton, Feltovich and Welch (1994), two self-selected groups attempted a writing task, one using a computer, the other using pen and paper. The responses were then transcribed to the other form; ie, the computer responses were handwritten and the handwritten responses were word-processed. Although pen and paper produced scripts received higher scores in handwritten form than word-processed form, the opposite occurred for computer-produced scripts. The latter received higher scores in word-processed form than handwritten form.

Powers, Fowles, Farnum and Ramsey (1994), conducted two studies in which the same students answered two extended response questions, one by computer and the other by pen and paper. All computer responses were later handwritten and all pen and paper responses were later word-processed. In the first study, a strong effect was found in favour of handwritten responses over word-processed responses. When originally handwritten responses were word-processed, the mean was lowered substantially. When the originally word-processed responses were later handwritten, the mean increased slightly. This interaction was statistically significant. To remove the strong effect of markers giving higher marks to handwritten essays, a new set of markers were given a different training program and the experiment repeated with the same essay responses. This ameliorated the effect but did not remove it.

The second question, concerning a judgement of the quality of the writing in different modes under the same format, has also been addressed in the literature. In the second study by Powers *et al.* (1994), the computer responses were favoured over the pen and paper responses (both averaged over the two formats). For each of the handwritten and word-processed formats, the computer response average was higher than the pen and paper average. Snyder (1993) studied two groups that were equivalent on various assessment measures, one responding by pen and paper, the other by computer. In a similar design by Russell (1999), the two groups were randomly assigned. In both studies, the pen and paper responses were word-processed to give the same format as the computer responses, and compared to the latter. In Snyder, the computer response marks were significantly higher. In Russell, there was little difference in two language tests but in an open-ended Science test, the computer responses were favoured. In a later study, Russell and Plati (2001) found (at both Grade 8 and Grade 10 level) that computer responses were favoured over pen and paper responses when both were presented in word-processed format. The latter three studies seemingly assumed that the appropriate comparison was between computer produced essays and pen and paper essays *that were word-processed.* If the findings relating to the first question are justified, it would have been more beneficial for the pen and paper group to have their original (not word-processed) essays compared to the computer responses.

The third question was addressed in a large scale study by Bridgeman and Cooper (1998). Here 3470 students took both equated versions of the Graduate Management Admissions Test (GMAT), answering one by computer and the other by pen and paper. Both versions required two 30 minute essays. The pen and paper responses (in their original form) gained higher marks than the computer responses. This result did not vary across gender, ethnicity or ESL. A different result was obtained in the Powers, Fowles, Farnum and Ramsey (1994) second study. This found no significant difference between handwritten pen and paper responses and word-processed computer responses. A third result was obtained by the Wolfe, Bolton, Feltovich and Bangert (1996) study in which students attempted two pre-equated narrative questions, answering one by computer and the other by pen and paper, in their original formats. Conflicting results were obtained, with an interaction between the question and the mode of response—pen and paper produced essays scored higher on one question and computer responses scored higher on the other. One would not have expected this result if the questions were effectively pre-equated.

The above review suggests some consistency in the literature for the first two questions, but not much consistency for the third question. For Question 1, a consistent effect is that when originally handwritten responses are word-processed and re-marked, the marks are generally lowered. However, when responses by computer are later handwritten and re-marked, the new marks can be either lower (Wolfe *et al.*, 1994) or higher (Powers *et al.*, 1994). For Question 2, most studies seem to favour responding by computer, when the results are presented in word-processed format, although some comparisons gave no significant difference. For Question 3, when the two methods were compared in their original formats, one study favoured pen and paper (Bridgeman and Cooper, 1998), one gave no difference (Powers *et al.*, 1994) and one gave mixed results within the one study, depending on the essay question (Wolfe *et al.*, 1996). A summary of the salient features of these studies is presented in Table 1.

One of the strongest positive correlates of high essay scores is the length of the essay in words (Page, 1966; Page, 1968; Page and Petersen, 1995). Reid and Findlay (1986, 12) argue that "The longer essays correlate significantly with quality writing because they demonstrate development within paragraphs, structural completeness and scribal fluency...". Reviews by Cochrane-Smith (1991) and Hawisher (1988) indicate that responding by computer tends to produce longer answers for essays in general. The studies by Powers *et al.* (1994), Russell (1999), and Russell and Plati (2001) also indicate that this holds for examination settings. This suggests that responding by computer has the potential to lead to higher marks through producing longer essay responses.

Method

To examine the above issues under conditions in the NSW education system, a two-part study was conducted at a large independent school in Sydney, Part 1 using 1997 test data and Part 2 using 2000 test data.

Table 1: Summary of some selected studies

Study	*Sample sizes*	*Students*	*Results*
Arnold *et al.* (1990)	300 scripts in both WP and HW formats	Mixed ages, college (mostly part time)	PP(HW) > PP(WP)
Sweedler-Brown (1991)	61 scripts in both WP and HW formats	1st year college	PP(HW) > PP(WP)
Sweedler-Brown (1992)	27 scripts in WP, HW (neat), and HW (messy) formats	1st year college	PP(HW-neat) > PP(WP)
Wolfe *et al.* (1994)	2 groups: 80 in PP, 77 in PC. Scripts were transcribed to the other format	Grade 10 school	PP(HW) > PP(WP) PC(HW) < PC(WP)
Powers *et al.* (1994)	32 cases do 2 essays, one by PP, the other by PC. Each was transcribed to the other format.	Mixed ages, college	Ex 1: PP(HW) > PP(WP) nsd: PC v PP (averaged) Ex 2: HW > WP (averaged) nsd: PP(HW) v PC(WP) PC > PP (averaged)
Snyder (1993)	2 groups: 25 cases use PP, 26 use PC. HW converted to WP.	Year 8 (13-y-o)	PC(WP) > PP(WP)
Russell (1999)	2 groups for each course Lang Arts 1: 57 in PP, 60 in PC Lang Arts 2: 45 in PP, 55 in PC Science: 51 in PP, 51 in PC	Grade 8 school	LA1: nsd PC(WP) v PP(WP) LA2: nsd PC(WP) v PP(WP) Sci: PC(WP) > PP(WP)
Russell *et al.* (2001)	Grade 8: 85 in PP, 59 in PC	Grade 8 school	G8: PC(WP) > PP(WP)
	Grade 10: 74 in PP, 71 in PC	Grade 10 school	G10: PC(WP) > PP(WP)
Wolfe *et al.* (1996)	437 do Task 1 (PC), Task 2 (PP) 337 do Task 1 (PP), Task 2 (PC)	Grade 10 school	T1: PC(WP) > PP(HW) T2: PC(WP) < PP(HW) ie, interaction: task & method
Bridgeman *et al.* (1998)	3470 cases do both PP & PC forms of GMAT	College students taking GMAT	PP(HW) > PC(WP)

PP(HW): pen and paper method, handwritten format
PP(WP): pen and paper method, word-processed format
PC(WP): computer method, word-processed format
PC(HW): computer method, handwritten format
nsd: no significant difference

Part 1

In Part 1, Year 9 students (ages 14–15) attempted a 1997 Year 10 external English test under public examination conditions at the same time as the Year 10 students. Two groups were randomly formed, having near identical means and standard deviations on the school assessment mark, one responding by computer ($n = 57$) and the other by pen and paper ($n = 52$). The computer group used the school's networked PCs in three classrooms.

The computer responses were printed and stapled into answer booklets and were later handwritten by students in other classes into further copies of the answer booklets. The pen and paper group handwrote their responses directly into the answer booklets. Their responses were later word-processed by experienced clerical staff and the printed responses stapled into the answer booklets. The students and clerical staff involved in the transcription were instructed to reproduce all the idiosyncrasies of the original text. These scripts were marked by six markers as part of the normal marking operation. The marking used an holistic marking procedure, with five categories, A, B, C, D and E. Markers would first assign a script to one of these categories and then indicate whether the script was high, middling or low within the category. General descriptors were given for each category and prior to the commencement of marking, sample scripts were extracted that served as exemplars of typical work within each category. Procedures were established to ensure that no marker received the same script in both forms.

Each student responded to three extended response questions which will be referred to here as Sections A, B and C. Section A: Reading (10 marks), allowed 15 minutes to respond to stimulus material. Section B: Literature/Mass media (30 marks), based on a long piece of stimulus material, allowed 30 minutes for writing. Section C: Writing (30 marks), allowed 30 minutes to write a story based on any one of nine pieces of stimulus material.

Part 2

In Part 2, Year 9 students attempted two essays from the 1999 external test, a year after the examination was originally held. They will be referred to here as Sections D and E. In contrast to Part 1, the groups were not formed by random assignment. Two groups were identified, differing slightly on a school assessment measure: those who wished to respond by computer and those who wished to respond by pen and paper. Using frequency distributions of the groups on the school assessment, students were systematically deleted from the larger group (the computer group) to give similar frequency distributions on the school assessment.

The computer group ($n = 88$) brought their own laptops and were supplied with floppy disks with template documents on which to prepare their responses. Most used a recent version of Microsoft Word to respond, but some had older or alternative software. Students responding by pen and paper ($n = 53$) were issued with plenty of writing paper to ensure equity with the computer group.

The computer responses were later handwritten by students in other classes at the school and the pen and paper responses were later word-processed by hired clerical staff. In transcribing the scripts, care was taken to preserve peculiarities in the spelling, punctuation and layout.

Section D (10 marks, allow 15 minutes) required at least 150 words in writing a formal letter to a newspaper editor arguing for or against the viewpoint expressed in stimulus material. Section E (20 marks, allow 25 minutes) required the writing of a formal speech designed to persuade a committee. Whereas the Part 1 scripts were marked holistically, the Part 2 scripts were marked strictly according to given criteria. For Section D, these four criteria were sentence structure, grammar, language and vocabulary and effective communication. For Section E, the seven criteria were form, spelling, punctuation, paragraphing, language and vocabulary, sustained development of ideas, and effective communication. The marking was conducted by six experienced markers, all of whom had marked at least one of the two questions during the actual test marking in 1999. A summary of the requirements for the five questions is given below in Table 2.

Results

Comparison of presentation formats and response modes

The data analysis design required an unbalanced repeated measures ANOVA which was performed using the General Linear Model (GLM). The full ANOVA tables are given in the Appendix—the probability levels will be repeated in the text where appropriate.

The two methods of response (computer vs pen and paper) and two formats for presentation (handwritten vs typed) gave four mean scores for each question. To facilitate comparisons among these means, they were converted to standard score form (z scores). These z scores were then graphed with the same scale on the vertical axis so that the results for each question could be visually compared (see Figures 1 and 2 below).

In Figure 1, it can be seen that Sections A and B show a similar pattern. In each case, the line connecting the two means from the handwriting format is above the line connecting the two means from the typed format. These lines do not deviate significantly from parallel, indicating no interaction between method of response and presentation format. Although the handwritten means are consistently above the typed means, these differences do quite not reach statistical significance in these cases ($p = 0.14$ for A; $p = 0.10$ for B).

For A, the computer response mean (averaged over both methods of presentation), although higher, is not significantly higher than the pen and paper response mean. For B, however, the average computer response mean is significantly higher than the pen and paper response mean ($p < 0.05$).

For Section C the pattern is quite different. The means for the typed responses (averaged over both methods of presentation), are significantly above those for the handwritten responses ($p = 0.002$). As in the previous cases, the lines do not deviate significantly

Table 2: Summary of questions

Question	*Mark value*	*Suggested writing time*	*Task*
A	10 marks	15 mins	Stimulus material: contains text, photo and map of Snowy River, NSW. Students asked to write on how the words, pictures & other features in the material encourage people to appreciate their environment and to visit such places.
B	30 marks	30 mins	Stimulus material: story on visit to grandmother in nursing home. Quest. asks what do we learn about the characters & how does the writer bring them to life. Suggestions to consider include the characters' feelings, their relationships, the writer's language.
C	30 marks	30 mins	Stimulus material: 9 separate items, 5 comprising a few sentences of text setting a scene, 4 comprising drawings of scenes. Quest: write a story on one of these items. Do not write a poem or draw a cartoon.
D	10 marks	15 mins	Stimulus material: an article entitled "A rage for curiosity" which criticises young people for being bored and lacking curiosity. Quest: write a formal letter to a newspaper editor, arguing for or against the views expressed.
E	20 marks	25 mins	Quest: Think about a particular challenge that appeals to you or a personal goal you would like to achieve. Imagine that your school is offering support under a new Personal Challenge Scheme. Write a formal speech to be delivered to a committee of parents and teachers, convincing them that you should receive support under this scheme. Your speech could include: • a description of your challenge or goal • reasons why this challenge or goal is important to you • an explanation of how you plan to pursue your challenge or goal • your arguments why the committee should support you.

from parallel, indicating no interaction between method of response and presentation format. In addition, for Section C, responding by computer gives significantly higher means than responding by pen and paper ($p = 0.000$).

From Figure 2, it can be seen that Section D has a similar pattern to Sections A and B in the 1997 data. For Section D, however, the difference between the handwritten means (averaged over the two response methods) and the typed means is statistically significant ($p < 0.05$). In addition, there is no significant difference between the two response methods (computer vs. pen and paper).

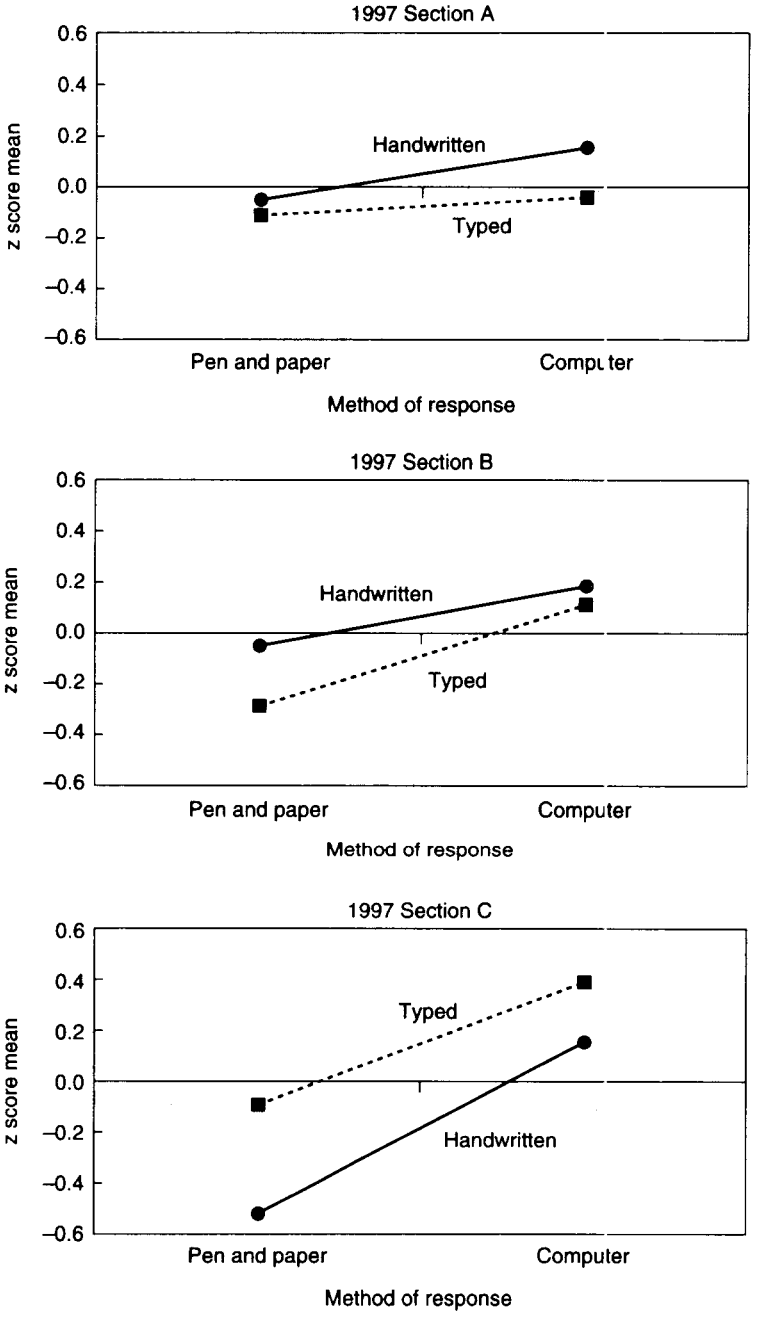

Figure 1: Part 1 comparisons

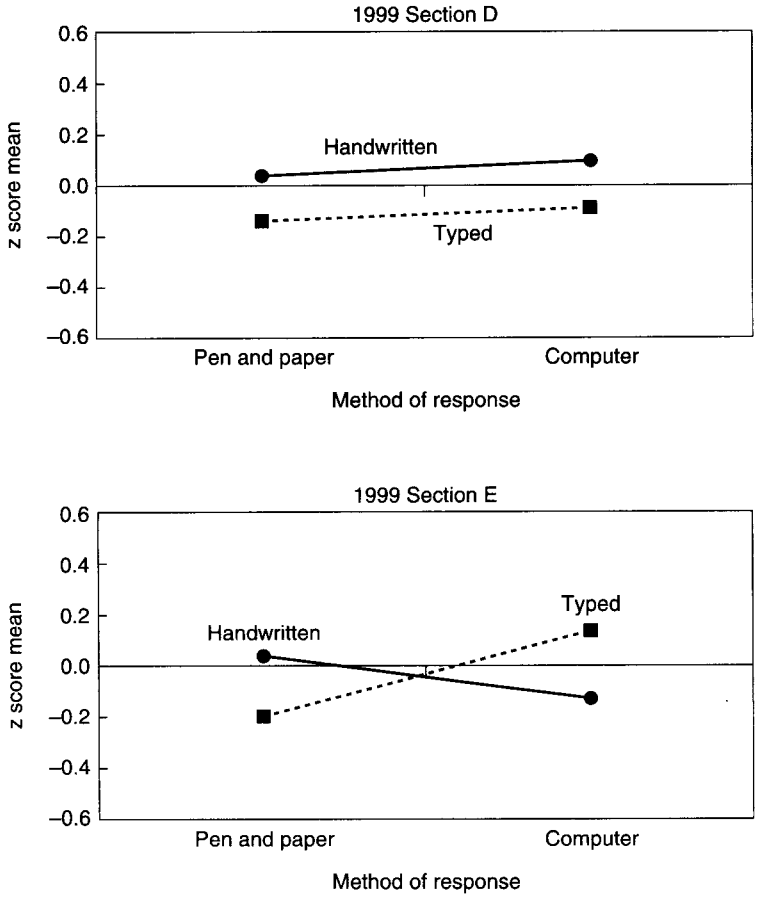

Figure 2: Part 2 comparisons

For Section E, the pattern is different to that of the other questions. There is a statistically significant ($p < 0.05$) interaction between the method of response and the presentation format. For responding by pen and paper, the mean of the handwritten responses is higher than the mean of the typed responses. The opposite occurs for responses by computer. This pattern also occurred in Wolfe *et al.* (1994).

Comparison between methods in their original forms

The statistical tests described above have compared a computer response mean with a pen and paper response mean with these means averaged across the handwritten and typed formats. But in a practical marking setting, the comparison of interest is responses by pen and paper (in handwritten format) versus responses by computer (in typed format). These means were statistically compared and in most cases there were no significant differences between the two methods (A, $p = 0.975$; B, $p = 0.363$;

D. $p = 0.469$; and E. $p = 0.569$). Only for Section C was the comparison highly significant in favour of responding by the computer ($p = 0.000$).

Length of response

This issue was tested with the data from Part 2. Three indices of length were obtained—a word count, the number of pages when typed and the number of pages when handwritten. When the pen and paper responses were later word-processed to Times Roman 12 pitch single spacing, the number of physical pages for the responses was substantially reduced. For Section D, the average reduction was 73.1% and for Section E the reduction was 73.9% for word processing based on single spacing. Conversely, when the computer responses were later handwritten, the former as a percentage of the latter was 76.6% for D and 76.1% for E. Thus, as a rough figure, one could say that a single-spaced typed response would be about three quarters of a written response. Given that the length based on physical appearance could have influenced the markers, it may have been better to use one and a half spacing in word-processing the written responses.

The above is concerned only with the physical appearance and not the length of response (word count). The latter is given below in Figure 3 which shows boxplots of the number of words for each response method for each question.

Each boxplot shows a rectangle giving the difference between the 25th and 75th percentiles, with the median score represented by the horizontal line in the middle of the rectangle. Outliers are shown by circles with the associated case number. The maximum and minimum values that are not outliers are shown as other horizontal lines (the whiskers).

For Section D, the computer responses are generally longer with a median of 273 words, compared to 237 words for pen and paper. They are also more variable in length, as indicated by the wider box and the greater distance between the whiskers. In particular, there are two outriders in the response by computer that are well above any other student (cases 111 and 123). Whereas the median candidate typed 273 words, case number 111 typed 613 words.

For Section E, a similar pattern emerges, but with the difference in the medians not so pronounced (298 words for computer, 283 words for pen and paper). The effective maximum (excluding outriders) for the computer responses, however, was substantially above the that for the pen and paper responses. The difference in mean word count was significant ($p < 0.05$) for Section D but did not reach statistical significance for Section E.

Reliability of marking

As the markings in both Part 1 and Part 2 replicated the single marking of the external tests, no direct estimate of the marker reliability was available. However, an estimate of marker reliability of the marking population from which the markers were drawn can be obtained. The marker reliability for another external examination question, the holistically marked Writing (which was double marked) was 0.63. This value, while not

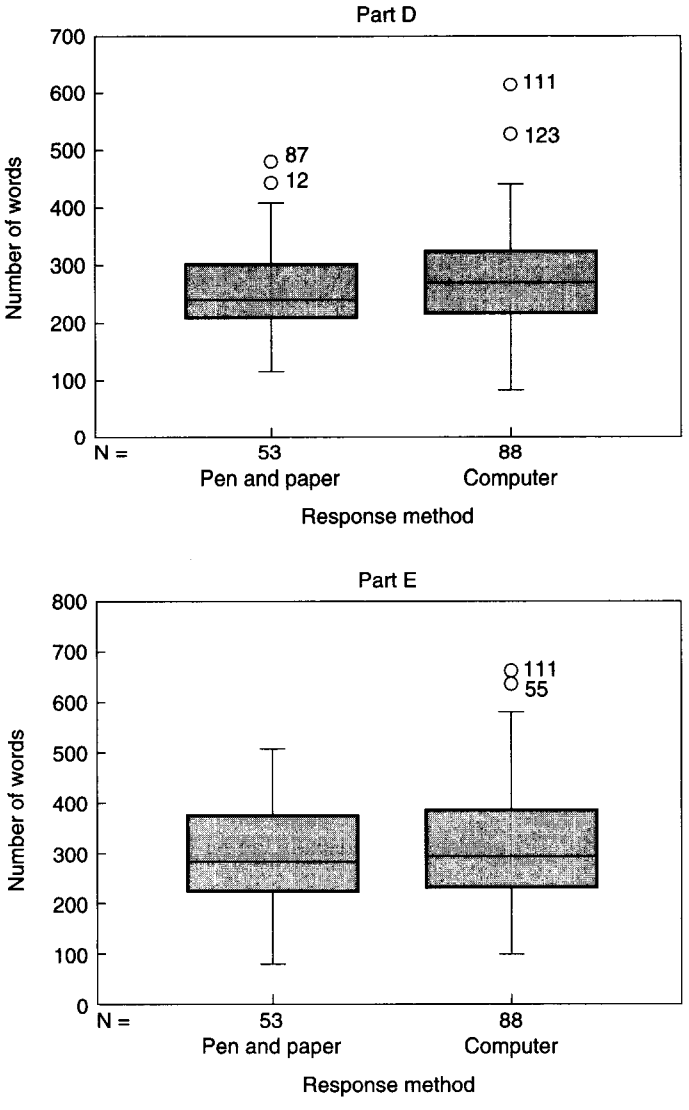

Figure 3: Boxplots of the word count for each method within each question

high in absolute terms, is consistent with other studies of essay marking reliability (eg, Cohen and Deale, 1977; Page, 1994). The small group of markers involved in these experiments were far more experienced than the average marker, suggesting that the value of 0.63 would be a conservative estimate.

Discussion

The first two questions considered in this paper assume that no interaction exists between response method and format. For example, in asking whether responding by

pen and paper or computer gives higher quality writing, one assumes that whatever result is obtained would hold no matter what format is used. Thus the result in Section E gives no clear interpretation. This interaction between method of response and presentation format also occurred in two other studies reviewed, Wolfe *et al.* (1994) and Powers *et al.* (1994). In the present study the original format always received a higher average mark than the transcribed format. This could be explained by postulating that the transcription process degraded the responses in some subtle way, despite the care taken to reproduce all the peculiarities of the spelling, grammar and layout. However, this fails to explain why Section D showed no interaction while Section E did. The differences between the questions can be seen from Table 2. In D, the students were required to read a long piece of stimulus material, whereas E required no stimulus material. In D, the students had the stimulus material as a constant point of reference which could be used to structure their responses, whereas E required the students to create a suitable challenge from their imagination. Another ostensible difference between these two questions was the length of response, with D having a shorter expected writing time of 15 minutes compared to 25 minutes for E. However it is difficult to link any of the differences between the questions to a plausible explanation of the interaction.

Regarding Question 1, of the five questions analysed, three questions gave similar results in terms of the most favoured format (A, B and D). It would seem that for these questions, it is better to have the responses presented in the handwritten format rather than typed. This is consistent with most other studies cited earlier. The exception was Section C where the typed responses averaged significantly higher than the written responses for both methods of responding, against the trend of results in the literature. However this question was unusual in that it was extremely open-ended with holistic marking. The students were simply asked to write a story based on any one of nine quite different pieces of stimulus material (see Table 2). In these circumstances, the candidates would spend less time on techniques such as planning and integrating material into a reasoned response and more time on the creative flow of ideas. This type of question is now no longer in use under the current system of standards-referenced examining, in which the questions are now more structured, stating explicitly what candidates are expected to do and the marking reflects this. The responses for C were noted at the time to be much longer than the responses for A and B (although a word count was not taken), and in these circumstances, the perceived problem of the typed scripts being shorter in length may not have applied. Given that the responses were generally lengthy, it may be that for the typed responses, the advantages of neatness and readibility came into play and that the typing was thus favoured over long and harder-to-read written scripts.

For Question 2, of the four sections not incurring an interaction, two gave a significant difference between the methods when averaged over the presentation formats. These were Sections B and C. From Table 2, it can be seen that these were the two longest essays with 30 minutes writing time. It is possible that using a computer is most advantageous when used with longer essays. Several studies have shown that responses by computer under examination conditions tend to be longer than pen and paper

responses, and this also occurred in the present study. Given that an essay's score tends to be correlated with its length, one would expect the computer responses to be favoured. Perhaps these factors are most clearly exhibited in the longer samples of writing produced under longer time limits.

Question 3 considers the important practical question of comparing pen and paper responses with computer responses in their original formats. In a practical marking setting, it is unlikely that the responses by pen and paper would be word-processed and even more unlikely that the responses by computer would be written up by hand. It was found that for four of the five essays, there were no significant differences between the two methods of response. These results are influenced by two opposing factors. The computer responses tend to be longer, which works in their favour, but the pen and paper responses are handwritten, which offsets this advantage. The upshot is that when comparing a response by pen and paper in its handwritten form to a response by computer in typed form there usually is little difference between the mean scores. A good example of this occurs in Section B, where the method result favours computer responses but the format result favours the handwritten mode, providing a cancelling effect.

The above conclusion held for Sections A, B, D and E. This result, of course, represents an average effect. For individual students, one method may be highly favoured over the other. For example, from the boxplots in Figure 3, it can be seen that case number 111 is a outrider in the number of words produced on both D and E. This student is obviously a very able typist.

While these studies have provided useful information, further work is suggested. The handwritten scripts that were later typed were presented in single spaced format. This resulted in their physical page length being less than three quarters of the physical page length of a handwritten script. The question arises as to whether they would be marked more favourably if they were double spaced. In attempting to reduce the mean difference between handwritten and typed scripts, Powers *et al.* (1994) changed the spacing from single to double and, in a subsequent re-marking, did reduce (but not eliminate) this difference. However, this is confounded by the fact that the new group of markers were given a different training, so the relative contributions of each of these factors to this result are unclear. Any further work in this area could consider equalising the physical length of the pages by adjusting the margins, typeface and spacing.

The results of this study may be summarised as follows. Firstly, in marking essays there is a tendency for the handwritten format to be favoured over the word-processed format. Some reasons advanced for this effect include the following: markers have higher expectations for typed essays; markers can identify text errors more easily in typed essays; markers have more difficulty identifying the author's "voice" in typed essays. Another reason for this result could be that the typed version of the response simply appears to be physically shorter. Secondly, when the presentation formats are equalised, there is some tendency for the computer response mode to be favoured. This effect

appeared for the two longest of the five essays in this study. Finally, when the pen and paper and computer response methods were compared in their original formats, for four of the five essays there were no significant differences in their mean scores, agreeing with the Powers *et al.* (1994) study.

Acknowledgements

The work of Garry Richards, Kerry Edmeades and Jane Paleje in the planning and administration of Part 1 is gratefully acknowledged.

References

Arnold V, Legas J, Obler S, Pacheco M, Russell C and Umbdenstock L (1990) *Do students get higher scores on their word-processed papers? A study of bias in scoring handwritten versus word-processed papers* Rio Hondo College, Whittier CA. ERIC ED345818.

Bridgeman B and Cooper P (1998) *Comparability of scores on word-processed and handwritten essays on the Graduate Management Admissions Test*. ETS, Princeton NJ. ERIC ED421528.

Cochran-Smith M (1991) Word processing and writing in elementary classrooms: a critical review of related literature *Review of Educational Research* **61** 107–155.

Cohen L and Deale R (1977) *Assessment by teachers in examinations at 16+*. Schools Council Examinations Bulletin 37 Evans/Methuen Educational, London.

Dalton D and Hannafin M (1987) The effects of word processing on written composition *Journal of Educational Research* **50** 338–342.

Hawisher G (1988) Research update: writing and word processing *Computers and Composition* **5** 7–25.

Page E (1966) The imminence of grading essays by computer *Phi Delta Kappan* 48 238-243.

Page E (1968) Analyzing student essays by computer *International Review of Education* **14** 210–225.

Page E (1994) New computer grading of student prose, using modern concepts and software *Journal of Experimental Education* **62** 127–142.

Page E and Peterson N (1995). The computer moves into essay grading: updating the ancient test *Phi Delta Kappan*, March, 561–565.

Powers D, Fowles M, 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** 220–233.

Reid S and Findlay G (1986) Writer's workbench analysis of holistically scored essays *Computers and Composition* **3** (2) 6–32.

Russell M (1999) Testing on computers: A follow-up study comparing performance on computer and on paper *Education Policy Analysis Archives* **7** 20.

Russell M and Plati T (2001) Effects of computer versus paper administration of a state-mandated writing assessment. Available online at *http://www.tcrecord.org/Content.asp?ContentID=10709*.

Snyder I (1993) The impact of computers on students' writing: a comparative study of the effects of pens and word processors on writing context, process and product *Australian Journal of Education* **37** 5–25.

Sweedler-Brown C (1991) Computers and assessment: the effect of typing versus handwriting on the holistic scoring of essays *Research and Teaching in Developmental Education* **8** 5–14.

Sweedler-Brown C (1992) The effect of training on the appearance bias of holistic essay graders *Journal of Research and Development in Education* **26** 24–29.

Wolfe E, Bolton S, Feltovich B and Bangert A (1996) A study of word processing and its effects on student essay writing *Journal of Educational Computing Research*, **14** 269–284.

Wolfe E, Bolton S, Feltovich B and Welch C (1993) *A comparison of word processed and handwritten essays from a standardized writing assessment* ACT Research Report Series 93-8. ERIC ED370966.

Appendix: Repeated measures ANOVAs

Analysis of Variance for A *Source*	*DF*	*Seq SS*	*Adj SS*	*Adj MS*	*F*	*P*
METHOD	1	1.7686	1.7686	1.7686	0.65	0.422
FORMAT	1	2.0229	1.9310	1.9310	2.22	0.139
METHOD*FORMAT	1	0.4631	0.4631	0.4631	0.53	0.467
STUDENT(METHOD)	107	290.8278	290.8278	2.7180	3.13	0.000
Error	107	93.0140	93.0140	0.8693		
Total	217	388.0963				

Analysis of Variance for B *Source*	*DF*	*Seq SS*	*Adj SS*	*Adj MS*	*F*	*P*
METHOD	1	26.232	26.232	26.232	4.04	0.047
FORMAT	1	5.945	6.234	6.234	2.70	0.103
METHOD*FORMAT	1	1.775	1.775	1.775	0.77	0.383
STUDENT(METHOD)	107	694.310	694.310	6.489	2.81	0.000
Error	107	247.280	247.280	2.311		
Total	217	975.541				

Analysis of Variance for C *Source*	*DF*	*Seq SS*	*Adj SS*	*Adj MS*	*F*	*P*
METHOD	1	98.850	98.850	98.850	13.77	0.000
FORMAT	1	34.720	35.420	35.420	9.77	0.002
METHOD*FORMAT	1	2.025	2.025	2.025	0.56	0.456
STUDENT(METHOD)	107	768.251	768.251	7.180	1.98	0.000
Error	107	387.754	387.754	3.624		
Total	217	1291.601				

Analysis of Variance for D *Source*	*DF*	*Seq SS*	*Adj SS*	*Adj MS*	*F*	*P*
METHOD	1	0.561	0.561	0.561	0.17	0.677
FORMAT	1	5.419	5.065	5.065	4.56	0.035
METHOD*FORMAT	1	0.000	0.000	0.000	0.00	0.986
ID(METHOD)	139	446.274	446.274	3.211	2.89	0.000
Error	139	154.501	154.501	1.112		
Total	281	606.756				

Analysis of Variance for E *Source*	*DF*	*Seq SS*	*Adj SS*	*Adj MS*	*F*	*P*
METHOD	1	5.379	5.379	5.379	0.34	0.562
FORMAT	1	4.813	0.174	0.174	0.03	0.856
METHOD*FORMAT	1	47.322	47.322	47.322	8.99	0.003
ID(METHOD)	139	2212.259	2212.259	15.916	3.02	0.000
Error	139	731.752	731.752	5.264		
Total	281	3001.525				