

*Lynch School of Education*

*inTASC Publications*

---

*Boston College*

*Year 2002*

---

The Influence of Computer-Print on  
Rater Scores

Michael Russell  
Boston College,



inTASC

[www.intasc.org](http://www.intasc.org)

## The Influence of Computer-Print on Rater Scores

Michael Russell  
Technology and Assessment Study Collaborative  
CSTEOP, Boston College  
332 Campion Hall  
Chestnut Hill, MA 02467



## The Influence of Computer-Print on Rater Scores

Michael Russell  
Technology and Assessment Study Collaborative  
CSTEEP, Boston College  
Released April 2002

### Abstract

This study replicates and extends the work of Powers et al. (1994) by examining the influence computer-print and handwriting have on raters' scores. This replication study employs an experimental design that presents the same set of responses to raters in four different formats. A second experiment is conducted to explore the extent to which the presentation effect can be reduced by supplemental training that focuses specifically on the causes of this presentation effect and includes practice scoring of responses presented in different formats. As Powers et al. found, the first experiment indicates that responses to composition test items presented in handwritten form receive significantly higher scores than the same responses presented in computer-print form. This effect is due to the visibility of errors and higher expectations for computer-printed responses coupled with increased identity with the writer generated by handwriting. Through supplemental training, the presentation effect was eliminated.

### Introduction

This study partially replicates and extends the work of Powers et al. (1994) and Russell & Plati (2000). As the Educational Testing Service prepared to offer the Praxis test on computer and on paper, Powers et al. conducted a small experiment to examine the influence computer-printed text had on raters' scores for a composition item. Although Powers et al. anticipated that responses presented in hand-written form would receive lower scores, the opposite occurred.

Hypothesizing that the perceived length of passages was one factor that contributed to raters' lower scores for computer-printed passages, Powers et. al. (1994) conducted a follow-up study in which all computer-printed responses were presented double-spaced. While this approach did reduce the size of the effect, computer-printed passages still received lower scores.

More recently, Russell & Plati (2000) conducted a similar study in which responses produced by students in grade four, eight and ten were presented in three forms: handwritten, single-spaced 12 point computer text, and double-spaced 14 point computer text. Like Powers et. al. (1994), Russell and Plati report that responses presented in hand-written form received significantly higher scores. However, unlike Powers et. al., Russell and Plati also found that responses presented in double-spaced form (thus making the passages appear longer), received lower scores than responses presented in single-space form.

Through interviews with the raters, Russell and Plati (2000) identified three possible reasons why raters tended to award lower scores to computer-printed responses: 1) Typos, uncapitalized letters and punctuation errors are easier to overlook when essays are handwritten than when they are typed; 2) Because readers associate typed text with final versions, they are more critical of mechanical errors and interpret these errors as a lack of careful proofreading, whereas they tend to interpret the same errors in handwriting as something the author would correct if he or she had more time; and 3) Some readers felt more connected to the writer as person as a result of viewing his or her handwriting and thus were more likely to give the writer the benefit of the doubt.

The study reported here extends the work of Powers et. al. (1994) and Russell and Plati (2000) by conducting a series of experiments that: A) explore possible causes of the presentation effect and, B) attempt to reduce or eliminate the presentation effect by formatting text with a scripted font and through training procedures that familiarize raters with the presentation effect.

## Background

Research on testing via computer goes back several decades and suggests that for multiple-choice tests, administration via computer yields about the same results, at least on average, as administering tests via paper-and-pencil (Bunderson, Inouye, & Olsen, 1989, Mead & Drasgow, 1993). However, more recent research shows that for young people who have gone to school with computers, open-ended (that is, not multiple choice) questions administered via paper-and-pencil yield severe underestimates of students' skills as compared with the same questions administered via computer (Russell, 1999; Russell & Haney, 1997). In both studies, the effect sizes for students accustomed to working computer ranged from .57 to 1.25. Effect sizes of this magnitude imply that the score for the average student in the experimental group tested on computer exceeds that of 72 to 89 percent of the students in the control group tested via paper and pencil.

A more recent study conducted during the spring of 2000 examined the mode of administration effect in grades four, eight and ten, and for special education students. Focusing on the extended composition item used as part of the Massachusetts Comprehensive Assessment System (MCAS), Russell and Plati (2000) report substantial effects in all grade levels. Moreover, for eighth grade students receiving special education services for language arts, the effect size was about 1.5 times larger than for non-special education students. When combined with the effect found for the MCAS short-answer items, the mode of administration effect could result in underestimating student performance by four to eight points on an eighty point scale.

In response to these findings, Russell and Haney (2000) argue that two approaches to improving the quality of education in U.S. schools, namely standards-based testing and educational technology, currently work against each other. This conflict results from the inability of paper-and-pencil tests to provide valid measures of the writing skills of students accustomed to writing on computers. Anticipating this conflict, Alberta Learning (2000) began offering students the option of performing the province's graduation exams on paper or on computer in 1993. More recently, Russell and Plati (2000) have advocated that state testing programs that employ extended open-ended items also allow students the option of composing responses on paper or on computer. And in the near future, ETS plans to begin administering the National Assessment of Educational Progress Writing test on paper and on computer. As more testing programs offer students the option of producing essay responses on paper or on computer, the presentation effect reported by Powers et. al. (1994) and Russell and Plati (2000) raises a serious concern about the equivalence of scores. The experiments presented below explore the factors that contribute to the presentation effect and explore two approaches that testing programs might take to reduce the effect.

## Methods

As part of a larger study that focused on the mode of administration effect on the MCAS Composition items, Russell and Plati (2000) transcribed approximately 240 hand-written responses to computer text. These responses were generated by students in grades four, eight and ten and were in response to a separate extended composition item administered within each grade level. The experiments presented in this paper use a sample of the grade eight responses to explore methods of counteracting some of the factors believed to contribute to the presentation effect described above. Employing a new set of raters who had not seen these responses before and who were unaware of the presentation effect, the following experiments were conducted:

1. Altering Appearance—Several studies have shown that essays presented with neat handwriting will typically receive higher scores than will essays written with poor penmanship (Chase, 1986; Marshall & Powers, 1969; Markham, 1976; Bull & Stevens, 1979). Similarly, Powers et. al. (1994) and Russell and Plati (2000) report that altering the appearance of computer-printed text can effect the scores raters award. While the reader might expect that essays presented with neat computer-print text would in turn receive higher scores than handwritten essays, previous studies suggest that readers have higher standards for computer-printed text and thus tend to award them lower scores (Powers et al., 1994). One approach to reducing the presentation effect may be to format computer-printed text so that it appears less formal. In this experiment, the same set of student responses were presented in three ways:
  - A. Hand-written text
  - B. Double-spaced 12 point Times Font; and
  - C. Double-spaced 14 point Script Font (creating the appearance of hand-written cursive writing);

2. Spelling Errors – Powers et al (1994) and Russell & Plati (2000) both speculate that computer-printed text makes mechanical errors such as spelling and punctuation more visible and adversely affect rater scores. To examine the influence of spelling errors on rater scores, student responses were presented in three ways:
  - A. Hand-written
  - B. Double-Spaced 12 point Times Font transcribed verbatim; and
  - C. Double-spaced 12 point Time font with all spelling errors corrected.

Note that this experiment was intended to provide further insight into the role visibility of errors plays in rater scores. Because this experiment changed the actual text produced by students rather than simply altering the appearance of that text, this experiment clearly is not an appropriate method for reducing the presentation effect.

3. Training Away the Presentation Effect – Powers et. al. (1994) provide some evidence that making raters aware of the presentation effect can reduce the size of the presentation effect. For the experiment reported here, one set of raters was trained using the current MCAS training protocol. A second set of raters was trained using the current protocol and also received additional training that focused on the issues believed to influence raters' scores when reading computer-printed text. This additional training included:
  - 1) reviewing past research on the topic;
  - 2) examining a set of responses from a previous study to compare differences in the apparent lengths of the same responses presented in handwritten and computer-print forms and examining differences in the visibility of spelling, punctuation and paragraphing errors;
  - 3) scoring a sample of four responses presented in both formats and discussing differences in scores with a specific focus on the influence of appearance;
  - 4) suggesting that raters maintain a mental count of the number of mechanical errors they observe while carefully reading a response, and;
  - 5) encouraging raters to think carefully about the factors that influence their judgments before assigning a final score.

After one set of four raters received this supplemental training, both sets of raters then scored the same responses formatted in two ways:

- A. Hand-written; and
- B. Double-Spaced 12 point Times Font transcribed verbatim.

In all three experiments, all responses presented in a given format were double-scored. Following MCAS scoring procedures, scores awarded by each rater were aggregated into a single score.

For all of the items, the scoring criteria developed for MCAS were used (Massachusetts Department of Education, 2000a). The MCAS scoring guidelines for

the composition items focused on two areas of writing, namely Topic/Idea Development and Standard English Conventions. The scale for Topic Development ranged from 1 to 6 and the scale for English Conventions ranged from 1 to 4, with one representing the lowest level of performance for both scales. Table 1 presents the category descriptions for each point on the two scales.

**Table 1: Category Descriptions for MCAS Composition Rubrics**

| <b>Score</b> | <b>Topic Development</b>  | <b>English Standards</b>   |
|--------------|---|--|
| 1            | Little topic/idea development, organization, and/or details<br><br>Little or no awareness of audience and/or task     | Errors seriously interfere with communication AND<br><br>Little control of sentence structure, grammar and usage, and mechanics  |
| 2            | Limited or weak topic/idea development, organization, and/or details<br><br>Limited awareness of audience and/or task | Errors interfere somewhat with communication and/or<br><br>Too many errors relative to the length of the essay or complexity of sentence structure, grammar and usage, and mechanics |
| 3            | Rudimentary topic/idea development and/or organization<br><br>Basic supporting details<br>Simplistic language         | Errors do not interfere with communication and/or<br><br>Few errors relative to length of essay or complexity of sentence structure, grammar and usage, and mechanics                |
| 4            | Moderate topic/idea development and organization<br><br>Adequate, relevant details<br>Some variety in language        | Control of sentence structure, grammar and usage, and mechanics (length and complexity of essay provide opportunity for students to show control of standard English conventions)    |
| 5            | Full topic/idea development<br><br>Logical organization<br>Strong details<br>Appropriate use of language              |  |
| 6            | Rich topic/idea development<br><br>Careful and/or subtle organization<br>Effective/rich use of language               |  |

In addition to the general descriptions, MCAS also provides anchor papers and benchmark papers for each category. These anchor and benchmark papers provide concrete examples of each performance level. The anchor and benchmark papers were first introduced to raters during the common scoring training session and were available to raters throughout the scoring process.

In total, 12 raters were employed for this study. Eleven of the twelve raters were teachers in K–12 schools. The final rater was an advanced graduate student in education.

To be clear, all 12 raters received the same 3 hours of score training that was based on training software provided by NCS Pearson (2000) for the Massachusetts Department of Education. Eight of the twelve raters then scored the same 60 responses which were presented in four forms: Handwritten, Single-space 12 point Times New Roman font, Single-Spaced 14 point Lucida Handwriting font, and Single-spaced 12 point Times New Roman font with all spelling corrected. A spiral design was employed so that all 8 raters scored responses in all four formats but scored each response only once.

The remaining four raters received an additional hour of training that focused specifically on the presentation effect found in past studies. These four raters then scored responses presented in handwritten and single-spaced Times New Roman font. Again, a spiral design was employed so that all four raters scored responses in both formats but scored each response only once.

Finally, when transcribing responses from their original handwritten form to computer text, responses were first transcribed verbatim into the computer. The transcriber then printed out the computer version and compared it word by word with the original, making corrections as needed. A second person then compared these corrected transcriptions with the originals and made additional changes as needed. Following this process, a sample of 10 responses was checked a third time. Out of 3,524 words of text, only three errors were found and in two cases a word that had been misspelled in the original was spelled correctly in the transcribed text. Thus, while slight differences may exist between the original handwritten and the transcribed versions, these differences are likely to have a very minor effect on rater's scores.

## Results

Inter-rater reliability for responses presented in each format were generally adequate, but not strong. Table 2 displays the correlation coefficients for responses scored in each format. Table 3 displays the percent agreement and disagreement for each format. It is interesting to note that the lowest reliability occurred with the handwritten responses scored by raters who were provided with additional training that focused on the presentation effect. Also note that the Massachusetts Department of Education (1999; 2000b) reports inter-rater reliability as percent agreement within one point and typically reports agreement to be above 90%. Although the correlation coefficients reported in Table 2 are all less than .8 and many are below .7, the percent agreement within one point was 100% for both Topic Development and English Conventions.

**Table 2: Inter-rater Reliability Correlation Coefficients**

|                          | Hand | Single-Space | Script Text | Spell Check | Informed Hand | Informed Computer |
|--------------------------|------|--------------|-------------|-------------|---------------|-------------------|
| <b>Topic Development</b> | .77  | .67          | .73         | .77         | .64           | .79               |
| <b>Conventions</b>       | .60  | .66          | .62         | .60         | .55           | .53               |
| <b>Total Score</b>       | .74  | .67          | .74         | .75         | .68           | .74               |

**Table 3: Inter-Rater Reliability Percent (%) Agreement and Disagreement**

| Score Differential       | Hand | Single-Space | Script Text | Spell Check | Informed Hand | Informed Computer |
|--------------------------|------|--------------|-------------|-------------|---------------|-------------------|
| <b>Topic Development</b> |      |              |             |             |               |                   |
| -2                       | 0    | 2            | 2           | 2           | 0             | 0                 |
| -1                       | 40   | 37           | 15          | 40          | 18            | 20                |
| 0                        | 52   | 42           | 60          | 52          | 38            | 47                |
| 1                        | 8    | 20           | 22          | 7           | 43            | 33                |
| 2                        | 0    | 0            | 2           | 0           | 0             | 0                 |
| <b>Conventions</b>       |      |              |             |             |               |                   |
| -2                       | 0    | 0            | 0           | 0           | 0             | 0                 |
| -1                       | 20   | 22           | 12          | 12          | 17            | 13                |
| 0                        | 62   | 65           | 68          | 73          | 58            | 65                |
| 1                        | 18   | 13           | 20          | 15          | 25            | 22                |
| 2                        | 0    | 0            | 0           | 0           | 0             | 0                 |
| <b>Total Score</b>       |      |              |             |             |               |                   |
| -3                       | 0    | 0            | 0           | 0           | 0             | 0                 |
| -2                       | 17   | 18           | 8           | 10          | 7             | 0                 |
| -1                       | 22   | 25           | 8           | 28          | 15            | 33                |
| 0                        | 47   | 33           | 55          | 50          | 27            | 30                |
| 1                        | 8    | 13           | 18          | 8           | 42            | 18                |
| 2                        | 7    | 10           | 8           | 3           | 10            | 18                |
| 3                        | 0    | 0            | 2           | 0           | 0             | 0                 |

Table 4 presents the summary statistics for responses scored in each format. As was found in the two previous studies, when scored by raters who were unaware of the presentation effect, single-spaced computer text received lower scores for both topic development and standard English conventions than did the exact same responses presented in their original handwritten form. However, when scored by raters who received training on the presentation effect, this difference was reduced greatly. In table 3, also note that responses presented as scripted text received the highest scores.

**Table 4: Summary Statistics for Responses Scored in Each Format**

|                            | <b>Hand</b> | <b>Single-Space</b> | <b>Script Text</b> | <b>Spell Check</b> | <b>Informed Hand</b> | <b>Informed Computer</b> |
|----------------------------|-------------|---------------------|--------------------|--------------------|----------------------|--------------------------|
| <b>Topic Development</b>   |             |                     |                    |                    |                      |                          |
| Mean                       | 8.85        | 8.03                | 9.10               | 8.37               | 8.78                 | 8.73                     |
| St.Dev.                    | 1.70        | 1.74                | 1.81               | 1.75               | 1.58                 | 1.93                     |
| Min                        | 6           | 5                   | 5                  | 5                  | 5                    | 5                        |
| Max                        | 12          | 11                  | 12                 | 12                 | 11                   | 12                       |
| <b>English Conventions</b> |             |                     |                    |                    |                      |                          |
| Mean                       | 6.45        | 5.98                | 6.65               | 6.47               | 6.28                 | 6.48                     |
| St.Dev.                    | 1.24        | 1.28                | 1.15               | 1.03               | 1.04                 | 1.07                     |
| Min                        | 4           | 3                   | 4                  | 4                  | 4                    | 4                        |
| Max                        | 8           | 8                   | 8                  | 8                  | 8                    | 8                        |
| <b>Total Score</b>         |             |                     |                    |                    |                      |                          |
| Mean                       | 15.30       | 14.02               | 15.75              | 14.83              | 15.07                | 15.22                    |
| St.Dev.                    | 2.71        | 2.72                | 2.65               | 2.34               | 2.43                 | 2.69                     |
| Min                        | 10          | 8                   | 9                  | 10                 | 9                    | 9                        |
| Max                        | 20          | 19                  | 20                 | 20                 | 19                   | 20                       |

### Impact of Altering Appearance

To examine the effect altering the appearance of the response had on rater scores, a repeated measures analysis of variance was performed with the Handwritten, Single-Spaced and Scripted font responses. As Table 5 displays, the results showed a significant effect of the format in which responses were scored for the total score and both sub-categories.

To examine whether the scores differed significantly between the handwritten and single spaced text or between the handwritten and scripted text, Tukey's method of adjusting for multiple comparisons was employed (Glass & Hopkins, 1984). As table 6 indicates, responses presented as computer text received significantly lower

scores than did the same responses presented as scripted computer text. Moreover, responses presented as scripted computer text did not differ significantly from the handwritten responses but did receive significantly higher scores than did the regular computer text. Thus, it appears that altering the appearance of computer printed text by using a script font, thus making the response appear more similar to a handwritten response, may eliminate the presentation effect.

**Table 5: Results of Repeated Measures ANOVA for Altering Appearance**

|                              | <b>F(2, 118)</b> | <b>Sig.</b> |
|------------------------------|------------------|-------------|
| <b>Altering Appearance</b>   |                  |             |
| Topic Development            | 14.07            | <.001       |
| Standard English Conventions | 8.58             | <.001       |
| Total Score                  | 16.45            | <.001       |

**Table 6: Results of Altering Appearance Contrasts**

|                            | <b>Mean Difference</b> | <b>Std. Error</b> | <b>Significance</b> |
|----------------------------|------------------------|-------------------|---------------------|
| <b>Topic Development</b>   |                        |                   |                     |
| Hand vs Computer           | .82                    | .32               | .028                |
| Hand vs Script             | -.25                   | .32               | .71                 |
| Script vs Computer         | 1.07                   | .32               | .002                |
| <b>English Conventions</b> |                        |                   |                     |
| Hand vs Computer           | .47                    | .22               | .09                 |
| Hand vs Script             | -.20                   | .22               | .64                 |
| Script vs Computer         | .67                    | .22               | .008                |
| <b>Total Score</b>         |                        |                   |                     |
| Hand vs Computer           | 1.28                   | .49               | .024                |
| Hand vs Script             | -.45                   | .49               | .63                 |
| Script vs Computer         | -1.73                  | .49               | .001                |

### Impact of Corrected Spelling

To examine the effect spelling errors had on rater's scores, a repeated measures analysis of variance was performed with the Handwritten, Single-Spaced and Single-Spaced with Spelling Corrected responses. Table 7 indicates that there were significant differences in the scores awarded to responses presented in handwritten form, verbatim computer text, or as computer text with spelling corrected.

To examine whether the scores differed significantly between the handwritten and spell-checked text or between the verbatim computer text and spell-checked text, the Tukey method of adjusting for multiple comparisons was again employed. As table 8 indicates, responses presented as verbatim computer text received lower scores than did the same responses presented as spell-checked computer text, but the difference was not statistically significant. Conversely, responses presented as handwritten text received higher scores than did the same responses presented as spell-checked computer text, but again the difference was not statistically significant. Thus, it appears that correcting spelling may have a small effect on rater's scores for responses presented in computer text, but this difference is not statistically significant and accounts for only a portion of the presentation effect, at best.

**Table 7: Results of Repeated Measures ANOVA for Correcting Spelling**

|                              | <b>F(2, 118)</b> | <b>Sig</b> |
|------------------------------|------------------|------------|
| <b>Correcting Spelling</b>   |                  |            |
| Topic Development            | 10.08            | <.001      |
| Standard English Conventions | 5.86             | .004       |
| Total Score                  | 10.88            | <.001      |

**Table 8: Results of Spell-Checking Contrasts**

|                            | <b>Mean Difference</b> | <b>Standard Error</b> | <b>Significance</b> |
|----------------------------|------------------------|-----------------------|---------------------|
| <b>Topic Development</b>   |                        |                       |                     |
| Hand vs Spell              | .48                    | .32                   | .28                 |
| Computer vs Spell          | -.33                   | .32                   | .54                 |
| <b>English Conventions</b> |                        |                       |                     |
| Hand vs Spell              | -.02                   | .22                   | .99                 |
| Computer vs Spell          | -.48                   | .22                   | .07                 |
| <b>Total Score</b>         |                        |                       |                     |
| Hand vs Spell              | .47                    | .47                   | .59                 |
| Computer vs Spell          | -.82                   | .47                   | .20                 |

### Impact of Supplemental Training

The final experiment examined whether the presentation effect could be reduced or eliminated through training. Table 9 displays the results of t-tests that compare scores awarded by raters who received additional training. Table 9 also displays the summary statistics for the scores awarded to the same responses by raters who did not receive additional training. To assist in interpreting the mean score difference, Table 9 also displays Glass's delta effect size (Glass & Hopkins, 1984).

As Table 9 shows, for raters that received supplemental training, there were only slight differences between the scores awarded to the responses presented in handwritten and computer text form. And none of these differences were statistically significant. Moreover, the scores awarded by the "trained" raters more closely resembled the scores awarded to the handwritten responses by raters who only received traditional training.

**Table 9: Results of Traditional Versus Supplemental Training Experiment**

|                               | <b>Mean</b> | <b>St. Dev.</b> | <b>Difference</b> | <b>Effect Size</b> | <b>t-static</b> | <b>Sig.</b> |
|-------------------------------|-------------|-----------------|-------------------|--------------------|-----------------|-------------|
| <b>Additional Training</b>    |             |                 |                   |                    |                 |             |
| Hand Topic Dev.               | 8.78        | 1.58            |                   |                    |                 |             |
| Comp Topic Dev.               | 8.73        | 1.93            | -.05              | .03                | 0.16            | .88         |
| Hand Conventions              | 6.28        | 1.04            |                   |                    |                 |             |
| Comp Conventions              | 6.48        | 1.07            | .20               | .19                | 1.04            | .30         |
| Hand Total                    | 15.07       | 2.43            |                   |                    |                 |             |
| Comp Total                    | 15.22       | 2.69            | .15               | .06                | .32             | .75         |
| <b>No Additional Training</b> |             |                 |                   |                    |                 |             |
| Hand Topic Dev.               | 8.85        | 1.70            |                   |                    |                 |             |
| Comp Topic Dev.               | 8.03        | 1.74            | .82               | .48                |                 |             |
| Hand Conventions              | 6.45        | 1.24            |                   |                    |                 |             |
| Comp Conventions              | 5.98        | 1.28            | .47               | .38                |                 |             |
| Hand Total                    | 15.3        | 2.71            |                   |                    |                 |             |
| Comp Total                    | 14.0        | 2.72            | 1.3               | .47                |                 |             |

## Discussion

The experiments presented above were intended to explore possible causes of the presentation effect and to explore approaches that might reduce or eliminate the size of the presentation effect. As occurred in the two previous studies (Powers, et al, 1994; Russell & Plati, 2000), a statistically and practically significant presentation effect was found when responses were presented in their original handwritten format and in transcribed computer print. This presentation effect resulted in higher scores awarded to handwritten responses. On average, this difference resulted in computer printed responses receiving scores 1.3 points lower than scores received by the same response presented in handwritten form.

The presentation effect, however, seemed to disappear when computer printed responses were formatted with a scripted font that resembled cursive handwriting. On the surface, then, it appears that one approach to eliminating the presentation effect is to simply format computer printed responses with a font that resembles handwriting. By doing so, not only may the response more closely resemble a response produced by hand, but the larger font size also makes the response appear longer - two factors which previous studies suggest may contribute to the presentation effect. However, interviews with raters conducted after they completed scoring provide an alternate explanation. Four of the eight raters who read responses presented in scripted font complained that the passages were “very difficult to read” and “made my [their] eyes

tired.” At the bottom of the score sheet, one rater even wrote, “Sorry-my 52-year old eyes can’t read 15 papers of script type. The best I can do is scan for predictable features.” Two other raters also indicated that they had difficulty reading the passages carefully and tended to award scores based on their general impression of the writing. Thus, while presenting responses in scripted font may eliminate the presentation effect, this improvement comes with an important cost: Raters may read responses less carefully and award scores based on a quick rather than careful read of the response.

In the two previous studies, the authors suggested that the visibility of mechanical errors combined with higher expectations for computer-printed text contribute to lower scores awarded to responses presented in computer print. The second experiment presented here provides some evidence that spelling may have an effect on raters’ scores. The magnitude of this effect was smaller for the Topic Development scores than the English Conventions scores. But, in both cases, the size of the effect was not statistically significant, although it did result in .8 point increase on average in students’ total score. Clearly, the visibility of spelling errors alone accounts for only a fraction of the effect, at best. Spelling, however, represents only one type of mechanical error. To further explore the influence of the visibility of mechanical errors on raters’ scores, additional studies should be conducted in which a fuller range of mechanical errors such as punctuation, capitalization, and spelling are corrected.

The final experiment presented here provides evidence that the presentation effect can be eliminated through training. By describing the presentation effect to raters, discussing the possible causes of the effect, providing samples of responses that appear very different when presented in handwritten and computer-printed form, by suggesting that raters maintain a mental count of the number of mechanical errors they observe while carefully reading a response, and by encouraging raters to think carefully about the factors that influence the scores they award, it appears that raters award similar scores to responses presented in both formats. For testing programs concerned about tracking trends over time, it is important to note that training raised scores for computer-printed responses to the same level as scores awarded to handwritten responses. If this finding holds for other composition items administered as part of other testing programs, this finding suggests that efforts to analyze trends may not be interrupted by allowing students to compose responses by hand or with a computer.

Both formatting computer-printed responses with scripted font and providing supplemental training eliminated the presentation effect. However, since scores should be based on a careful reading of a response and scripted text appears to make it difficult for raters to read responses carefully, providing supplemental training is a more desirable method for reducing this effect. It is interesting to note that during scoring, one rater who did not receive supplemental training on the presentation effect stated that it would have been helpful to have seen anchor papers presented in both handwritten and computer text form. This rater went on to say that she found herself applying different criteria and standards to the computer printed responses than to the handwritten responses. As this rater stated, Powers et al (1994) suggest, and this experiment demonstrates, it appears critical that raters be trained with responses presented in different modes when responses are produced and ultimately scored in different modes.

Clearly, the effect training has on reducing the presentation effect needs to be replicated with a larger sample of responses and larger groups of raters. In addition, future studies should examine this issue for a wider variety of open-response test items. Nonetheless, this study provides preliminary evidence that the presentation effect can be eliminated through training. If generalizable, this finding may clear an important obstacle to providing students with the option of composing responses to open-ended items by hand or on a computer.



## References

- Alberta Learning. (2000). Directions for Administration, Administrators Manual, Diploma Examination Program.
- Bull, R. & Stevens, J. (1979). The effects of attractiveness of writer and penmanship on essay grades. *Journal of Occupational Psychology*, 52, 53–59.
- Bunderson, C. V., Inouye, D. K. & Olsen, J. B. (1989). The four generations of computerized educational measurement. In Linn, R. L., *Educational Measurement* (3rd ed.), Washington, D.C.: American Council on Education, pp. 367–407.
- Chase, C. I. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, 23(1), 33–41.
- Glass, G. & Hopkins, K. (1984). *Statistical Methods in Education and Psychology*. Boston, MA: Allyn and Bacon.
- Markham, L. R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 13(4), 277–283.
- Marshall, J. C. & Powers, J. C. (1969). Writing neatness, composition errors, and essay grades. *Journal of Educational Measurement*, 6, 97–101.
- Massachusetts Department of Education. (1999). *Massachusetts Comprehensive Assessment System: 1998 Technical Report*. Malden, MA.
- Massachusetts Department of Education. (2000a). 1999 MCAS Sample Student Work and Scoring Guides. <http://www.doe.mass.edu/mcas/student/1999/>.
- Massachusetts Department of Education. (2000b). *1999 MCAS Technical Report*. Malden, MA.
- Mead, A. D. & Drasgow, (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114:3, 449–58.
- NCS Pearson. (2000). *Scoring MCAS Compositions: NCS Mentor™ for Massachusetts*.
- Powers, D., Fowles, M, Farnum, M, & 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), 220–233.
- Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives*, 5(3), <http://olam.ed.asu.edu/epaa/v5n3.html>.
- Russell, M. & Haney, W. (2000). Bridging the Gap Between Testing and Technology in Schools. *Education Policy Analysis Archives*, 8(19), <http://epaa.asu.edu/epaa/v8n19.html>.
- Russell, M. & Plati, T. (2000). Mode of Administration Effects on MCAS Composition Performance for Grades Four, Eight and Ten. A report submitted to the Massachusetts Department of Education by the National Board on Educational Testing and Public Policy, <http://nbtpp.bc.edu/reports.html>.

Russell, M. (1999). Testing Writing on Computers: A Follow-up Study Comparing Performance on Computer and on Paper. *Educational Policy Analysis Archives*, 7(20).

