Investigating the Effect of Computer-Administered Versus Traditional Paper and Pencil Assessments on Student Writing Achievement

Robert Laurie¹, Beatrice L. Bridglall², and Patrick Arseneault³

Abstract

The effect of using a computer or paper and pencil on student writing scores on a provincial standardized writing assessment was studied. A sample of 302 francophone students wrote a short essay using a computer equipped with Microsoft Word with all of its correction functions enabled. One week later, the same students wrote a second short essay using paper and pencil with access to dictionaries. Mean scores were compared for essays on each medium as well as scores on six specific criteria. There was no significant difference between the overall mean scores on the paper and pencil essays and those written using a computer. Significant differences favoring the paper and pencil essays were seen on the ideas, punctuation, and syntax criteria. A significant difference in favor of the computer written essays was seen on the orthography criterion. Possible practical implications and suggestions for future research are discussed.

Keywords

large-scale writing assessments, paper and pencil writing assessments, computer-assisted writing assessments

The role of assessment in teaching and student learning has particular salience in light of several overarching goals. These include the importance of affording students instruments that accurately measure their evolving knowledge, skills, and abilities, and provide the basis for validation and accreditation (Darling-Hammond & Bransford, 2005). The former is considered low-stakes given goals of increasing academic/cognitive abilities, whereas the latter is considered high-stakes given the potential impact of these decisions on students’ education trajectories. Efforts at designing assessments that serve these goals include the use of computers in gauging what students know and can do. Barton and Coley (1994) reported that computers are increasingly used to assess how students craft responses to certain questions, particularly in writing. Their report, now more than two decades old, underscores the fact that these concerns are not new. The implications of the practice of utilizing technology to assess students are considerable and can be organized around the use of the technology itself, concerns for equity, efficiency, and the practical considerations of actually shifting from the use of paper-based assessments to those that are technology-based (Russell, 2002a; Sandene, Horkay, Bennett, Allen, Braswell, Kaplan & Oranje, 2005). Inadequate attention to these concerns can result in significant challenges for assessment designers, educators, and students.

With regard to the technology itself, the measuring of skills not utilized in traditional paper and pencil assessments (such as those necessary for using technology in learning and problem solving) or vice versa can be an issue (Sandene et al., 2005). Similarly, the accuracy of assessment results may be in question when we consider that some students do not have extensive access to or skills with computers. In addition, the low cost of delivering and scoring large-scale assessments via computers may be outweighed by the dilemmas of transitioning from a paper-based system into one that involves relevant facilities, equipment, and synergies between administrators and schools (Sandene et al., 2005). Another issue of particular concern to educators and assessment designers is whether scores derived from traditional paper and pencil assessments and the use of computers are equivalent. Specifically, it is important to ensure the

¹University of New Brunswick, Fredericton, New Brunswick, Canada
²NYU-Shanghai Initiative, China
³Ministère de l’Éducation et du Développement de la petite enfance, Fredericton, New Brunswick, Canada

Corresponding Author:
Robert Laurie, Canadian Research Institute for Social Policy, University of New Brunswick, Keirstead Hall, Suite 300, Fredericton, New Brunswick, Canada E3B 5A3.
Email: robert.laurie@unb.ca

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://www.creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (http://www.uk.sagepub.com/aboutus/openaccess.htm).
assessment of intended constructs instead of computer skills unrelated to the constructs. Older studies (Bunderson, Inouye & Olsen, 1989) suggest that mean scores between paper and pencil assessments and computer-administered assessments were often less equivalent than assumed. Specifically, scores on computer-administered assessments were lower than those on traditional paper and pencil assessments. Even so, these researchers found that the score differences were of little import given how small they were. A later study by Mead and Drasgow (1993) suggests that computerized assessments were slightly more difficult than those administered via paper and pencil, leading to the conclusion that there were no differences between the two media when assessments were carefully constructed (Mead & Drasgow, 1993). A significant medium effect, however, was found for speeded tests administered on computers.

Nevertheless, given conflicting results between the two mediums and the evolving nature of technology, it is crucial that we continue to add to our understanding regarding whether computer-administered tasks are equivalent to those that are paper-based. It is equally important to understand factors that influence comprehension, reasoning, and problem solving pursuant to writing tasks, for example, that contribute to each medium’s use. In this vein, subsequent studies of score equivalence between computer-administered and paper and pencil assessments have continued to inform the assessment domain (Hargreaves, Shorrock-Taylor, Swinnerton, Tait, & Threlfall, 2004). With respect to comprehending computer-administered tasks, an earlier study by Belmore (1985) noted that participants did not gain a strong understanding of the information presented on video display terminals. However, this dynamic occurred only when participants first viewed the material on the computer, suggesting that the ease of utilizing the computer was facilitated by viewing the task on paper first. An alternative hypothesis may hinge on the fact that participants were not afforded opportunities to practice on the computer. This may lead to Belmore’s (1993) suggestion (and substantiated by other researchers, including Cushman, 1986; Muter, Latremouille, Treurniet, & Beam, 1982; Muter & Maurutto, 1991; Oborne & Holton, 1988) that participants’ understanding of the tasks may have been comparable in both mediums if they had opportunities to practice.

With regard to reasoning, the population sampled by Askwall (1985) tended to search for more information and in a longer span of time when paper assessments were utilized. Alternatively, participants in Weldon, Mills, Koved, and Schneiderman’s (1985) study were able to solve problems in less time with paper-based assessments. Similarly, skilled writers tended to craft their work in 50% less time on paper than on computers (Gould, 1981). Students also needed less time to respond to questions on paper than online (Hansen, Doring, & Whitlock, 1978). These discrepancies between the efficacy of computer and paper assessments were attributed to the state of technology, particularly the inadequate interface design prevalent during the 1980s and early 1990s (Ziefle, 1998). In an effort to test this, Gray, Barber, and Shasha (1991) substituted non-linear text with dynamic text, and practice effects notwithstanding, discovered that participants’ information searching capacity improved.

The early studies of the efficacy of paper- and computer-based tasks noted above suggest a preference for utilizing paper-based assessments, particularly when we consider other salient factors that affect results, such as the visual quality of the two mediums (Ziefle, 1998), how tasks are understood, and the accuracy and speed with which they can be executed. In the last decade, however, the user/computer interface has become more user friendly, sophisticated, and prevalent. This has prompted current research comparing the equivalence of computer- and paper-based scores on complete tasks rather than discrete components, including reading speed (Noyes & Garland, 2008). Even so, differences in the execution of tasks vis-à-vis the two mediums were found. Mayes, Sims, and Koonce (2001), for instance, found that reading on the computer tends to be slower than reading on paper. Despite this, participants’ learning or comprehension, as demonstrated by their scores from both online and paper-based tasks, was similar in both mediums (Bodmann & Robinson, 2004; Garland & Noyes, 2004; Mason, Patry, & Bernstein, 2001; Mayes et al., 2001; Noyes & Garland, 2003; van de Velde & von Grünau, 2003). In contrast, Wästlund, Reinikka, Norlander, and Archer’s (2005) study did resonate with earlier studies in that participants’ comprehension of tasks on paper appeared to be stronger when compared with the computer results.

Similarly, van de Velde and von Grünau (2003) did not perceive variations in eye movement patterns in the two mediums. In a study comparing student performance in language arts, science, and mathematics, Russell (1999) reported a positive effect for student performance in science using computers, no effect in language arts, and a negative effect in mathematics. Russell (2001) later reported that the administration effect becomes meaningful and in favor of computer-administered tests when students achieve keyboard typing speeds of 20 to 24 words per min.

Given the discrepancy between these studies, it is important to continue efforts at understanding the effects of these mediums on teaching and learning, particularly in light of increased demands on educators and students, and the ease some users, particularly students, now have manipulating rapidly evolving technologies. In this vein, this article details a study within the larger context of provincial assessments developed by the francophone sector of the New Brunswick Department of Education and Early Childhood Development in Canada. Specifically, we investigated whether student scores derived from a common assessment in writing but administered via computer were equivalent to those derived from a traditional paper and pencil assessment.
**Background to the Study**

In efforts to address equity and fairness concerns in the context of technology use in assessing what students know, the francophone sector of the New Brunswick Department of Education and Early Childhood Development in Canada is responsible for ensuring the comparability of computer-administered and paper and pencil assessments. As part of its assessment program, the Department assesses student writing skills at the end of Grade 8. The assessment is low-stakes for students unless school districts decide to include the results in their final grades. Results from the assessment are widely available at the school, district, and provincial levels, which render them high-stakes for educators and administrators at all levels. Ensuring comparability of the same assessment administered using two different mediums is crucial to the credibility of the provincial assessment program. As such, the aim of this study is to test the effect of computer-administered versus traditional paper and pencil assessments on the scores of the Grade-8 essay assessment as administered in May of the 2010-2011 school year.

**Method**

**Participants**

Participating Grade-8 students were sampled at the classroom level. Each of the five francophone school districts was asked to contribute two or three classes from at least two schools. Participation was voluntary at the school and classroom levels. School principals consulted with teachers and students prior to identifying participating classes. All students in the selected classes were part of the sample, which consisted of 302 students (174 females, 128 males) out of the 2,047 students in the 2010-2011 Grade-8 cohort. These students came from classes in 12 francophone schools in New Brunswick and represented a variety of demographic backgrounds. New Brunswick schools follow a fully inclusive approach toward education in which all students regardless of their demographic background, special needs, and so on are taught together in heterogeneous classes with respect to ability. Notwithstanding this inclusionary approach, the New Brunswick francophone student population is very homogeneous with respect to language and ethnicities/nationalities. The province is the only officially bilingual Canadian province and as such, has a dual education system based on language: Francophone students attend French schools, whereas anglophone students attend English schools. Moreover, New Brunswick immigration rates are very low (Statistics Canada, 2010), and most immigrants attend English schools.

New Brunswick francophone students may be exempted from participating in provincial assessments under exceptional circumstances. The exemption policy in use for all provincial assessments was respected in this study. A total of 26 students from the 12 participating classes were exempted from the provincial assessment. Thus, the sample size of 302 students represents the total number of students who actually participated in the study and not the total number of students in the 12 participating classes. Because this study was focused on comparing student performance on a writing task using either a computer or the more traditional paper and pencil assessment medium, we did not disaggregate data relative to student background or special needs.

Eight of the participating classes were already using the Desire2Learn (D2L) platform during regular classroom activities. The four classes that had limited exposure to D2L were provided with additional support, including “online mentors” from their school district and the Department. A parallel practice version of the writing assessment was made available online several weeks before the actual assessment. All 12 classes used this practice assessment whose results contributed to students’ cumulative grade point averages at the teacher’s discretion.

The 302 participating students completed both a computer-administered writing task and a paper and pencil writing task. First, they completed a computer-administered writing task whose psychometric properties paralleled that of the paper and pencil writing task on the provincial exam. One week later, all Grade-8 francophone students, including the 302 students in this study, were administered the paper and pencil writing task. To reduce stress, increase student engagement, and compensate for possible medium bias, students were informed that the higher of the two scores would be considered for the official student, class, school, district, and provincial reports.

**Design of the Writing Assessments**

The writing assessment required students to write a 200-word essay based on one of two proposed writing prompts, which they selected at the start of the assessment. Students were allocated 2½ hr to produce their essay. To ensure the 302 participating students did not have an unfair advantage on the paper and pencil version by having seen the writing prompts 1 week prior to provincial assessment, the choices of prompts differed for the 2 versions of the assessment. A validation committee composed of teachers, district literacy specialists, and the provincial language arts consultants for curriculum and assessments ensured the prompts were equivalent. This ensured that any change in the results between both writing tasks was not due to the difficulty of the prompts.

For the paper and pencil version of the assessment, students were given a booklet that included directives and lined pages on which to write their essay. The directives in the booklet guided the students according to the following sequence:

1. Choosing one of the proposed writing prompts;
2. Reading the criteria used to score the essay;
3. Using a guide (checklist) to help them organize their ideas for the essay;
4. Writing the first version of their essay on the “Draft” pages of the booklet; and
5. Writing the final version of their essay on the “Final version” pages of the booklet.

With the exception of not requiring a draft, the computer-administered version of the assessment provided specific instructions regarding the use of D2L in a similar format:

1. Opening and using MS Word to write the essay;
2. Using an adapted version of the checklist;
3. Saving the essay file on the computer upon completion of the task; and
4. Uploading the essay file through D2L.

For the paper and pencil assessment, students had access to dictionaries, grammar manuals, or any other references normally accessible during school assessments. For the computer-administered assessment, students had access to the software’s spell check function and other word processing functions available with MS Word. The usual rules and restrictions governing provincial assessments were enforced in the course of administering the assessments via computer and paper and pencil. Because the use of the D2L platform requires full access to the Internet (through the Department’s portal), students were instructed not to access any program or website that allows emails, messaging, or other means to communicate or exchange information during the assessment. These restrictions were enforced by the examination supervisor and the monitoring of Internet activities for all D2L accounts in use by students.

Instruments, Marking, and Scores

There exists variability in the use of the terms scores and marks (Almond, 2009; Luyten & Dolkar, 2010; MacCann & Stanley, 2010). In the context of this article, marks are taken to be the direct numerical value stemming from the judgment of markers. We refer to scores as pertaining either to individual assessment criteria or to the result on the overall assessment, the latter reported on a percentage scale.

Six criteria based on the language arts curriculum are used to score the essays. The first three comprise the essay’s content or function, whereas the last three comprise its form:

1. **Ideas** or quality of the narrative: includes the description of the space, period, characters, and events;
2. **Structure of the narrative**: includes the sequence of the narrative (introduction, development, and conclusion), the use of paragraphs, and a logical use of time adverbs, connecting words, and appropriate phrasing;
3. **Vocabulary**: includes inappropriate wording, repetitions, or lexical errors;
4. **Punctuation**: counting the number of punctuation errors;
5. **Syntax**: includes missing words, inappropriate syntax, verbs, and misuse of pronouns, adverbs, and connecting words; and
6. **Orthography** or spelling: includes normal spelling, gender, number, verbs, and so on.

The first three criteria (content or function) are marked using a holistic approach based on four performance levels: “Superior,” “Expected,” “Acceptable,” or “Insufficient.” Each level is converted into numerical scores of 4, 3, 2, and 1, respectively. The last three criteria (form) are marked analytically reflecting the number of errors for each criterion as tallied by the markers.

Computer-administered and paper and pencil assessments were marked simultaneously using the same rubrics for both assessments, in the same marking center. Markers included Grade-8 teachers from various schools across the province, recently retired teachers, and bachelor of education students from the Université de Moncton. Literacy learning specialists from each of the five school districts were assigned as head markers. The support staff for the marking session included the provincial learning assessment specialist responsible for the Grade-8 language arts provincial exam, the marking site manager, and clerical staff. Because of the 302 additional writing pieces from the computer-administered assessment, additional personnel included two provincial technology learning specialists and two other staff members with expertise in assessment and evaluation and in language arts from the assessment and evaluation directorate.

The marking process followed for the 2011 Grade-8 language arts assessment writing essay was the same as that used by the assessment and evaluation directorate in previous years. All markers received the same detailed systematic training on the common scoring rubric for both assessments prior to marking student essays. The addition of the computer-administered version required a marking process adapted to the specificities of this assessment and designed to be equivalent to the one used for the paper and pencil version. Markers for the computer-administered version were given additional instructions on how to use D2L as a marking tool and how to save their marks.

There was a different head marker for each of the two versions. For both versions, markers were assigned to mark specific criteria. A first group of markers was tasked with marking the first three criteria (ideas, structure, and vocabulary), whereas a second was tasked with marking the last three criteria (punctuation, syntax, and orthography). For the paper and pencil version, bundles of about 30 randomly selected booklets were distributed to all markers. A tracking sheet containing the unique booklet number for each booklet in the bundle was attached to the bundle. Once all the booklets in a bundle were marked by the first group of markers, the bundle was then redistributed to the other group of markers so that
the other three criteria could be marked. Marked booklets were identified accordingly on the tracking sheet to ensure that all booklets were scored. This process allowed a quick and effective way to ensure that all booklets were marked.

A slightly different process was used for the computer-administered essays. Essay files were stored on the D2L platform in a similar bundle system where only one marker could access, open, score, and save the results of a single bundle. Online markers identified which bundle they marked for tracking purposes. The number of online markers was about one-fifth of the total number of markers for the Grade-8 writing assessment.

Marking reliability was ensured by the head markers through rigorous training using common exemplars and answering individual queries during the marking session. In addition, the head markers conducted random reliability checks throughout the marking session by having all markers mark the same student essay. Feedback was provided immediately to all markers to increase their reliability. Final scores for the paper and pencil version obtained using an optical score reader and those of the computer-assisted version as extracted from D2L were brought together in a common data set.

**Results**

### Overall Writing Scores

The overall scores for the writing assessment were calculated on a percentage point scale based on each criterion. Table 1 presents the means, standard deviations, and the standard error of the mean for each version of the writing assessment. The correlation between the scores for both versions is positive, strong, and significant ($r = .81$, $p < .01$), which is not surprising because of the paired-samples design used in the study; one would expect student performance in a writing task to be stable over a 1-week period.

A paired-samples $t$ test was conducted to compare the means of the overall scores for both versions. The $t$ test results showed a non-significant difference between the computer-assisted and the paper and pencil versions in favor of the paper version, $t(301) = -1.605$, $p > .05$, suggesting that the assessment format does not affect overall student scores.

|                | N  | M   | SD  | SE  |
|----------------|----|-----|-----|-----|
| $M$ scores—computer | 302| 60.17 | 17.69 | 1.02 |
| $M$ scores—paper/pencil | 302| 61.19 | 17.93 | 1.03 |

### Writing Criteria Scores

A Wilcoxon rank test was conducted to evaluate the effect of the testing medium on the results for each of the six essay criteria. For the ideas criterion, the results showed a significant difference in favor of the paper and pencil format ($z = -2.05$, $p < .05$). The mean of the ranks in the paper and pencil format was 79.5, whereas the mean of the ranks for the computer-assisted format was 76.0. For the punctuation criterion, the results showed a significant difference in favor of the paper and pencil format ($z = -4.85$, $p < .01$). The mean of the ranks in the paper and pencil format was 125.5, whereas the mean of the ranks for the computer-assisted format was 121.0. For the syntax criterion, the results showed a significant difference in favor of the paper and pencil format ($z = -4.26$, $p < .01$). The mean of the ranks in the paper and pencil format was 129.4, whereas the mean of the ranks for the computer-assisted format was 121.3. For the orthography criterion, the results showed a significant difference in favor of the computer-assisted format ($z = -3.61$, $p < .01$). The mean of the ranks in the paper and pencil format was 118.4, whereas the mean of the ranks for the computer-assisted format was 142.6.

**Discussion, Limitations, and Future Research**

The use of technology has become central to students’ lives including when it comes to reading, writing, calculating, and thinking (Collins & Halverson, 2010). In a study of Grade-8 students and their interactions with technology, Clarke and Besnoy (2010) reported that students enjoyed reading with Personal Digital Assistants (PDAs) and felt that they had greater control over the format of the text environment and reading process, and also that using technology connects more closely with and is more relevant to their daily lives. Notwithstanding the fact that the Clarke and Besnoy (2010) study pertained to reading and ours to writing, the use of technology is certainly relevant to Grade-8 students, and using it can be an authentic approach to assessing their writing skills.

In this study, we report that the use of computers to assess the writing skills of Grade-8 students does not significantly affect their overall scores on a provincial assessment. Overall results from writing tasks implemented in two mediums—using a computer and traditional paper and pencil—showed no significant difference. This strongly suggests that allowing students to use computers equipped with the usual text editing functions does not jeopardize comparisons with previous or concurrent paper and pencil assessments. However, when the results from each of the six individual scoring criteria (which made up the overall score) were also compared with respect to the testing medium, four of the six criteria differed significantly including all three “form” criteria (punctuation, syntax, and orthography). Of the four criteria that resulted in significantly different scores, orthography was not only the only one that led to favorable results when using the computer but also the one impacted the most by the testing medium as evidenced by the highest values in the
mean ranks of the Wilcoxon test. Not surprisingly, orthography scores where higher when using the computer, which may reflect the effect of having the correction functions embedded in the software students used to write their essay. Such computer assistance was not available in the paper and pencil assessment, although dictionaries were made available. Interestingly, students obtained significantly better results on the punctuation and the syntax criteria on the paper and pencil test. The reasons for this are unclear, but it may be that students demonstrated a blind confidence in the computer software, perhaps erroneously thinking that it would correct mistakes other than orthography. Such an unjustified confidence may lead to carelessness, inattention to, or disregard for punctuation and syntax. Conversely, it is possible that students did not know which option to choose when presented with multiple suggestions for a given correction, although this is unlikely because if true, it may not have led to a significant difference in favor of the paper and pencil format where similar errors would not have been pointed out to them.

Students also scored significantly better on the ideas criterion when using the paper and pencil format. The reasons for this are far from clear. It is noteworthy that although the results for this criterion were statistically significant between testing media, the \( p \) value was near the .05 critical value, suggesting that a slight change in the results might render them not statistically significant. It is also noteworthy that the ideas criterion is the most subjective of the six scoring criteria. As such, it is vulnerable to the effects of inter-rater reliability, which this study did not control for at the criterion level. This represents one of the study’s limitations. Two factors, most likely in opposition to each other, come into play in the scoring process. It can be argued that scoring the computer-administered essays may lead to more favorable scores, because the negative effects of bad handwriting would be reduced. In contrast, it is possible that scorers prefer to score essays on paper rather than on a computer screen, which may lead to less favorable results for the computer-administered essays. The extent to which these factors exist and may cancel each other out is unknown. (For a more detailed discussion on the influence of computer print on rater scores, see Russell, 2002b.)

This study has several possible impacts on education and assessment policy and practices at the national, provincial, or state, district, and school levels despite its limitations. At the national, provincial, and state levels, this study may open the door to the implementation of large-scale assessments using computers or the evolution of existing large-scale assessments from paper and pencil to being computer-administered. Such an evolution has already happened with the Program for International Student Assessment (PISA), a worldwide study by the Organization for Economic Cooperation and Development (OECD), for example. The PISA assessments are administered every 3 years since 2000 and for the first time, in 2015, they will be computer-based.

This study may also prompt national, provincial, and state educators to review their curricula with the intention of including student learning outcomes pertaining to the use of software and their correction functions. Given the pervasive use of technology in today’s society, it behooves educators to train students how to effectively use easily accessible tools. As jurisdictions offer more and more online courses, they would also be wise to consider integrating online writing assessments to their courses where applicable. This study helps pave the way for this integration.

School districts may develop policies whereby they encourage students to bring their own computers to school if they so wish. Not all students in the same class would necessarily be using the same medium, but all could be confident that they have their preferred way of writing all the while being assessed fairly. Many jurisdictions are questioning the way they carry out writing assessments whether it be for financial reasons, about more technical aspects such as the decision to use holistic versus analytical scoring approaches (Savard, Sevigny, & Beaudoin, 2007) or the appropriateness of using only one performance task in a timed assessment (White, 1994). This article provides arguments and approaches that may encourage jurisdictions to reduce printing costs and possible travel costs for markers.

This study may also have important implications for teaching and student learning in that the criteria used for scoring the essays can be integrated within a scoring rubric. Scoring rubrics not only contribute to the reliability and validity of an assessment but also enable defensible judgment of complex competencies, which in turn enable educators to tailor instruction in efforts to promote student learning (Jonsson & Svingby, 2007). Scoring rubrics, when created and used in a systematic and rigorous manner, enable student learning vis-à-vis clearer expectations and criteria. Thus, educators can provide students with more targeted feedback as they transition from informal to more formal thinking and learning (Darling-Hammond & Bransford, 2005; De la Paz, 2009).

In addition, it is equally important for educators to be cognizant of the challenges some students may have with writing, particularly in a timed environment. In this context, Gregg, Coleman, Davis, and Chalk’s (2007) analysis suggests that students with dyslexia experienced difficulties with spelling, handwriting, complex vocabulary, and essay length. Educators and administrators can provide tailored assessments, instruction, and accommodations in continuing efforts at enabling students who experience these and other difficulties, to perform to their potential. Educators may also allow students to choose between computer and paper and pencil assessments of writing, which would give them a measure of control over the testing environment and somewhat decrease their anxiety.

Future research can focus on shedding light on the inter-rater reliability at the criterion level, which currently represents the most important limitation of the study. Inter-rater
reliability at the criterion level would not only provide valuable information contributing to the overall validity of the assessment but would also facilitate the interpretation of the statistical analyses in cases where \( p \) values are close to critical values for significance. Future research to study the effect of the medium on the scoring process should be undertaken to better understand the results related to the ideas criterion obtained in this study. It would also be important to quantify the effects of scoring essays with poor handwriting compared with the effects of scoring essays using a computer.

As with most quantitative studies, increasing sample size is desirable even if the sample size of 302 students was sufficient to generate significant results at the \( p < .01 \) level for all three form criteria. Increasing the sample size may lead to a larger significant difference for the ideas criterion, but the practical significance of this remains to be seen in light of the inter-rater reliability results. Future research should also be undertaken to understand the reasons why students obtained significantly lower results on the punctuation and syntax criteria when using the computer.

In conclusion, overall scores on essays written on the computer or with paper and pencil were not significantly different despite showing significant differences on four of six assessed criteria.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research and/or authorship of this article.

**References**

Almond, R. J. (2009). Group assessment: Comparing group and individual undergraduate module marks. *Assessment & Evaluation in Higher Education, 34*, 141-148.

Askwall, S. (1985). Computer supported reading vs. reading text on paper: A comparison of two reading situations. *International Journal of Man-Machine Studies, 22*, 425-439.

Barton, P. E., & Coley, R. J. (1994). *Testing in America’s schools* (ETS Policy Information Center Report). NJ: Educational Testing Service, Princeton.

Belmore, S. (1985). Reading computer presented text. *Bulletin of the Psychonomic Society, 23*, 12-14.

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research, 31*, 51-60.

Bunderson, C. V., Inouye, D. K., & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational Measurement*, (3rd ed). London: Collier Macmillan.

Clarke, L. W., & Besnøy, K. D. (2010). Connecting the old to the new: What technology-crazed adolescents tell us about teaching content area literacy. *Journal of Media Literacy Education, 2*(1), 47-56.

Collins, A., & Halverson, R. (2010). The second educational revolution: Rethinking education in the age of technology. *Journal of Computer Assisted Learning, 26*, 18-27.

Cushman, W. H. (1986). Reading from microfiche, VDT, and the printed page: Subjective fatigue and performance. *Human Factors, 28*(1), 63-73.

Darling-Hammond, L., & Bransford, J. (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco, CA: Jossey-Bass.

De La & Paz, S. (2009). Rubrics: Heuristics for developing writing strategies. *Assessment for Effective Intervention, 34*, 134-146.

Garland, K. J., & Noyes, J. M. (2004). CRT monitors: Do they interfere with learning? *Behaviour & Information Technology, 23*, 43-52.

Gould, J. D. (1981). Composing letters with computer-based text editors. *Human Factors, 23*, 593-606.

Gray, S. H., Barber, C. B., & Shasha, D. (1991). Information search with dynamic text vs. paper text: An empirical comparison. *International Journal of Man-Machine Studies, 35*, 575-586.

Gregg, N., Coleman, C., Davis, M., & Chalk, J. (2007). Timed essay writing: Implications for high-stakes tests. *Journal of Learning Disabilities, 40*, 306-318.

Hansen, W. J., Doring, R., & Whitlock, L. R. (1978). Why an examination was slower on-line than on paper? *International Journal of Man-Machine Studies, 10*, 507-519.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threfall, J. (2004). Computer or paper? That is the question: Does the medium in which assessment questions are presented affect children’s performance in mathematics? *Educational Review, 46*, 29-42.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2*, 130-144.

Luyten, H., & Dolkar, D. (2010). School-based assessments in high-stakes examinations in Bhutan: A question of trust? Exploring inconsistencies between external exam scores, school-based assessments, detailed teacher ratings, and student self-ratings. *Educational Research and Evaluation: An International Journal, 16*, 421-435.

MacCann, R. G., & Stanley, G. (2010). Classification consistency when scores are converted to grades: Examination marks versus moderated school assessments. *Assessment in Education: Principles, Policy & Practice, 17*, 255-272.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research, 24*, 29-39.

Mayes, D. K., Sims, V. K., & Koonce, J. M. (2001). Comprehension and workload differences for VDT and paper-based reading. *International Journal of Industrial Ergonomics, 28*, 367-378.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper and pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Muter, P., Latremouille, S. A., Treurniet, W. C., & Beam, P. (1982). Extended reading of continuous text on television screens. *Human Factors, 24*, 501-508.

Muter, P., & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited? *Behaviour & Information Technology, 10*, 257-266.
Noyes, J. M., & Garland, K. J. (2003). VDT versus paper-based text: Reply to Mayes, Sims, and Koonce. *International Journal of Industrial Ergonomics, 31*, 411-423.

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics, 51*, 1352-1375.

Oborne, D. J., & Holton, D. (1988). Reading from screen versus paper: There is no difference. *International Journal of Man-Machine Studies, 28*, 1-9.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20), Retrieved from http://epaa.asu.edu/ojs/article/view/555/678

Russell, M. (2001). *Does it matter with what I write? Comparing performance on paper, computer and portable writing devices*. Chestnut Hill, MA: CSTEEP, Boston College.

Russell, M. (2002a). *How computer-based technology can disrupt the technology of testing and assessment*. Chestnut Hill, MA: CSTEEP, Boston College.

Russell, M. (2002b). *The influence of computer-print on rater scores*. Chestnut Hill, MA: CSTEEP, Boston College.

Sandene, B., Horkay, N., Bennett, R., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online Assessment in Mathematics and Writing: Reports from the NAEP Technology-Based Assessment Project, Research and Development Series* (NCES 2005–457). Washington, DC: U.S. Government Printing Office, U.S. Department of Education, National Center for Education Statistics.

Savard, D., Sevigny, S., & Beaudoin, I. (2007). Évaluation à grande échelle: Liens entre le score holistique et les composantes de l’écriture.[Large-scale writing assessments: Relationships between the holistic score and writing components.]. *The Canadian Journal of Program Evaluation, 22*(3), 99-199.

Statistics Canada. (2010). *Visual census, immigration and citizenship, New Brunswick*. Retrieved from http://www12.statcan.gc.ca/census-recensement/2006/dp-pd/fs-fi/index.cfm?Lang=ENG&TOPIC_ID=6&PRCODE=13

van De Velde, C., & von Grünau, M. (2003). Tracking eye movements while reading: Printing press versus the cathode ray tube. [Abstract]. *Perception, 32*, 1-186.

Wästlund, E., Reinikka, H., Norlander, T., & Archer, T. (2005). Effects of VDT and paper presentation on consumption and production of information: Psychological and physiological factors. *Computers in Human Behavior, 21*, 377-394.

Weldon, L. J., Mills, C. B., Koved, L., & Schneiderman, B. (1985). The structure of information in online and paper technical manuals. In R. W. Swezey (Ed.), *Proceedings of the 29th annual meeting of the Human Factors Society* (Vol. 2, pp. 1110-1113). Santa Monica, CA: Human Factors Society.

White, E. M. (1994). *Teaching and assessing writing* (2nd ed.). San Francisco, CA: Jossey-Bass.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors, 40*, 554-568.

**Author Biographies**

Robert Laurie is adjunct professor at the University of New Brunswick, project director at the Canadian Research Institute for Social Policy and president of REL Consultants Inc. His research interests include large-scale assessments, quantitative studies and program evaluation.

Beatrice L. Bridglall is fulbright specialist in Higher Education with the Council of International Exchange of Scholars and Faculty Affiliate with the NYU-Shanghai Initiative.

Patrick Arseneault is a learning specialist and psychometrician working for the New Brunswick Department of Education and Early Childhood Development, and a lecturer at Université de Moncton, teaching quantitative research methodology. He has over 15 years of experience in learning assessment, testing methodologies and evaluation for various levels of government, including with the Canadian Forces and the territorial government of Nunavut. His research interests include Item Response Theory applications, implementation of large-scale testing protocols and test bias related to medium and gender.