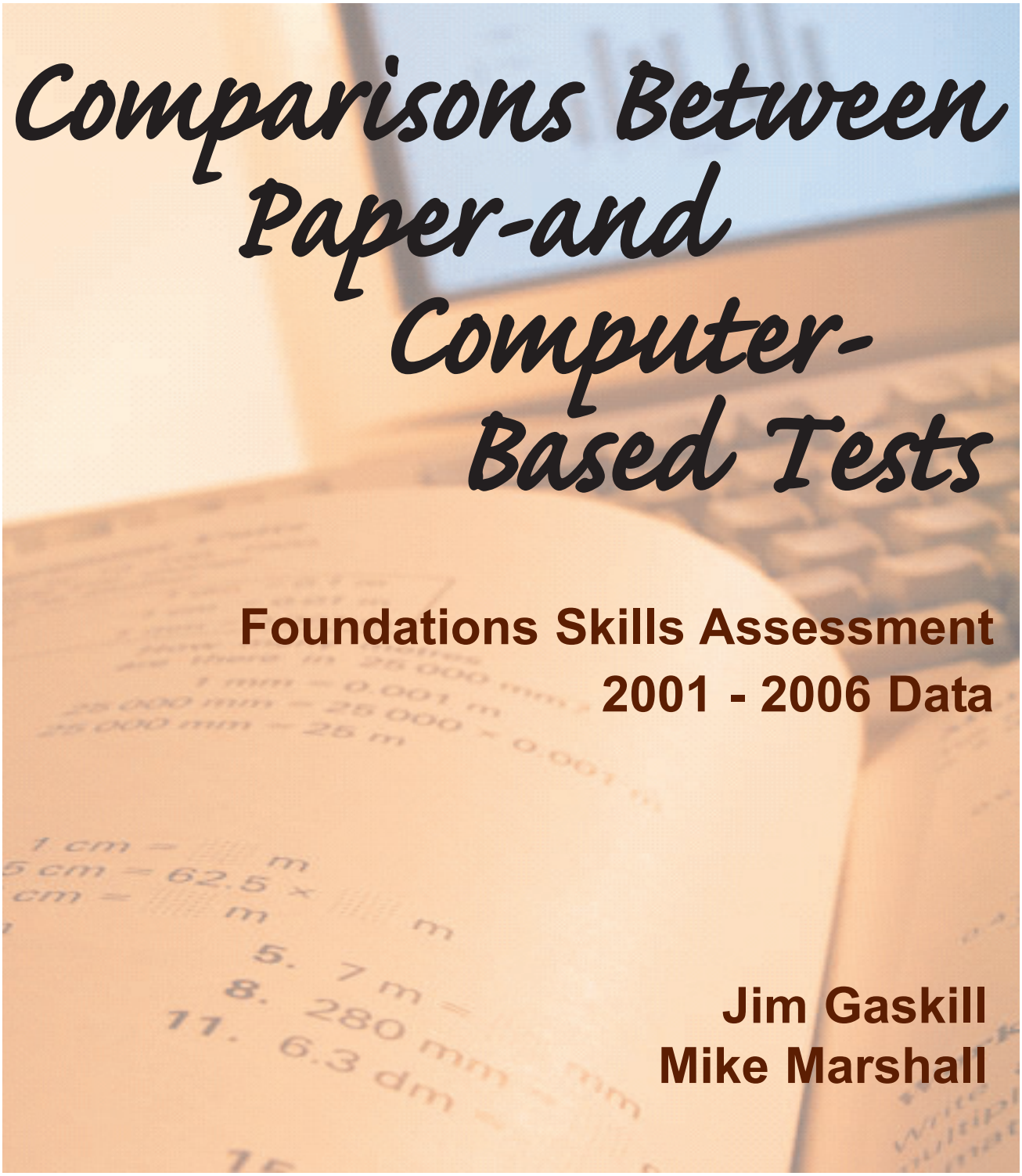
TASA INSTITUTE



Comparisons Between Paper-and Computer-Based Tests

Foundations Skills Assessment 2001 - 2006 Data

Jim Gaskill
Mike Marshall

SOCIETY FOR THE ADVANCEMENT OF Excellence in Education

Comparisons between Paper- and Computer-Based Tests

Foundation Skills Assessment – 2001 to 2006 Data

J. L. Gaskill
M. Marshall

SOCIETY FOR THE ADVANCEMENT OF Excellence in Education

ISBN 978-0-9783018

SOCIETY FOR THE ADVANCEMENT OF EXCELLENCE IN EDUCATION (SAEE)

SAEE is an independent non-profit education research agency founded in 1996. The mission of the Society is to encourage excellence in public education through the provision of research to guide policy and practice. SAEE has commissioned 35 studies to date and has a particular interest in examining innovations that may improve learning outcomes for less-advantaged students. As a registered Canadian charity, SAEE provides official tax receipts for donations to its research.

For additional copies of this report, please contact:

Society for the Advancement of Excellence in Education (SAEE)
225-1889 Springfield Road, Kelowna, B.C. V1Y 5V5
Tel. 250-717-1163
info@saee.ca
www.saee.ca

ACKNOWLEDGEMENTS

We would like to express appreciation to the Max Bell Foundation and the Society for the Advancement of Excellence in Education (SAEE) for sponsoring the Technology Assisted Student Assessment Research Institute and providing the research grant in support of this publication. Special thanks go to Brent Munro and Nancy Walt of the BC Ministry of Education who guided us through the requirements of Ministry research policy.

ABOUT THE AUTHORS

Jim Gaskill is an independent educational consultant and the Director of TASA Institute. He has taught at the secondary and university levels and has been principal of a K–12 school in the Queen Charlotte Islands. Before leaving the BC Ministry of Education, Dr. Gaskill was in charge of the development of Examinations and Assessments for the province.

Mike Marshall is Executive Director of the Applied Research and Evaluation Services (ARES) department at the University of British Columbia. Before joining ARES in 1990, he had been a secondary school teacher, department head, and principal. He has taught measurement, research, statistics and education administration courses at the university level and has played senior project management roles in large-scale provincial, national, and international assessments.

ABOUT THE TASA INSTITUTE

The Technology Assisted Student Assessment (TASA) Institute is a research initiative of the Society for the Advancement of Excellence in Education (SAEE). Established in 2004 with a research grant from Max Bell Foundation, the mission of TASA Institute is to study and advance knowledge in the development and application of assessment technology in the Canadian public education system.

The purposes of the Institute are:

1. To document trends, leading-edge prototypes, evidence regarding their effectiveness, best practice, and implications for policy in the field of technology-delivered student assessment.

2. To develop a next-generation assessment toolset and process, leveraging the considerable strengths of computer and online technologies.

3. To collaborate with Ministries of Education, school districts, testing agencies and international researchers in the piloting and evaluation of computer assisted assessment models.

4. To serve as a clearinghouse for research and provide a source of expertise to schools, districts, and ministries/departments of education on the design, implementation, and use of computer based assessment.

For more information, see: http://www.tasainstitute.com

4 Comparisons Between Paper-and Computer-Based Tests

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

8 Comparisons Between Paper-and Computer-Based Tests

EXECUTIVE SUMMARY

In 2004, one school administered the Grade 7 Foundation Skills Assessment (FSA) electronically. In 2005, the number of schools increased to eight and in 2006, 28 schools administered the FSA to at least some of their students. This report is based on 15 schools that administered the FSA electronically in at least one year. A number of schools were eliminated from this study because they administered to a very small proportion of the student enrolment. Schools that have administered the FSA electronically vary substantially in size: over 80% of all students came from only seven of the schools.

Because the electronic interface has different characteristics for each of the FSA components and because males and females may approach computer applications differently, separate analyses with gender as a factor were carried out for Numeracy Multiple-Choice, Reading Multiple-Choice, Reading constructed-response, Writing Focused Response and Writing Extended Response.

In addition, it was hypothesized that any impact of the electronic interface would be different for students of different ability levels. Students were placed in three achievement categories based on their results on the Grade 4 FSA and the analyses were carried out using Achievement Category as a factor.

Comparisons of the electronic interface and paper booklets revealed some differences that were anticipated to have an impact on student performance. Some electronic features appeared to be favourable to students and others may have disadvantaged students.

The following results of the statistical analyses cannot be generalized to all schools as the schools in the study chose to administer electronically: they were not randomly selected. In addition, in all analyses, there were significant differences among the schools.

Students did significantly better in the paper mode for Numeracy and Reading multiple-choice. The difference between paper and electronic modes was greater for males than females. While students appeared to do better in the electronic mode for Reading constructed-response, the difference was not significant possibly because of substantial and significant differences among the schools. There were no differences between paper and electronic modes for Writing focused response or Writing extended response.

Schools undertaking to administer the FSA might consider ensuring that, for Numeracy, there is adequate space for students to do calculations on paper and that it easy to copy information from the screen to the rough-work paper. Students should have some practice with the electronic versions of the tests. In the case of Reading, teachers should help students practice moving from one screen to another and moving within the screen.

Continued research is recommended because of the limited number of schools and students involved in the study. If such research is carried out, additional information should be gathered about how the schools prepared their students for the assessment, the degree to which regular instruction included computer based assignments, and the degree to which students have had access to and feel comfortable with computers.

10 Comparisons Between Paper-and Computer-Based Tests

1 Considerations in Administering by Paper or Computer

In May 2005, eight schools administered the Foundation Skills Assessment (FSA) by computer: seven schools used computer-based administration for the first time and one for the second. In May 2006, 28 schools administered the FSA by computer to at least some of their students. As the FSA has been administered annually since 2000, this presented an opportunity to see if there were any differences in achievement levels that could be associated with computer-based or paper-based administration in those same schools.

The FSA is a provincial assessment administered to approximately 45,000 students in each of grades 4 and 7 and has three components: Reading Comprehension (Reading), Writing, and Numeracy. Both Reading and Numeracy have a multiple-choice part and a constructed-response part. The Writing component has two writing prompts: a focused (short) writing task and an extended (long) writing task.

Scaled scores for the Reading and Numeracy components are derived using a 2-parameter Item Response Theory (IRT) model. Writing scores are the raw scores assigned during a centralized marking session. Each of the Numeracy and Reading constructed responses was marked on a four-category (1 to 4) holistic scale. The two Writing tasks were also marked on a four-category holistic scale with the focused response task marked once and the extended response marked by two independent markers. The extended-writing-response markers resolved any differences greater than adjacent categories. The resulting three marks, the single mark for the focused response and the two for the extended response, are summed resulting in a 10-point scale; 3–12.

Many markers return from one year to the next and marker leaders have had several years of experience marking and selecting exemplar papers. Marker Agreement Papers (MAPs) are used regularly during the marking sessions to maintain consistent marking, and team leaders rescore selected papers throughout the marking session. These procedures are used to minimize within-year and year-to-year differences due to marking.

The FSA electronic administration was done via a central server over the Internet. Throughout this report, the method of test administration will be referred to as either paper mode (paper-and-pencil) or electronic mode (electronically by computer over the Internet).

This report will provide a review of the research on comparisons of the results of paper and electronic modes, a descriptive comparison of the paper and electronic administrations, the results of the study, an interpretation of the analyses of each of the six parts of the FSA, and conclude with some recommendations for practice and future research.

12 Comparisons Between Paper-and Computer-Based Tests

Literature Review

The question as to whether or not the mode of test administration affects student responses has been addressed by a number of studies reaching back to the 1970s and even earlier. As the technology has improved immensely over the last two decades and the available interfaces have become so much simpler, the literature review will be limited to studies carried out since the mid-1980s.

A number of studies have been done at the post-secondary level (English, Reckase, and Patience, 1977), on surveys (Sun and McClanahan, 2003), on attitude tests (Hol, Vorst, and Mellenbergh, 2005), or on certification examinations such as medical association examinations (Lunz and Bergstrom, 1995). This review will, in general, focus on reports from the K–12 school system that are related to achievement tests, except where the study analyzes effects on different groupings of examinees (e.g., by gender or ability) or provides information related to the effects of particular interface features.

Several meta-analyses (Bergstrom, 1992, and Mead & Drasgow, 1993) sought to address the overall question of whether or not mode of test administration made a difference. Bergstrom (1992) conducted a meta-analysis of 20 studies from eight research reports comparing the results of paper-and-pencil tests with those of computer adaptive tests (CAT). She notes that reasons "why ability measures might not be statistically equivalent include: differences in item presentation, differences in cueing due to varying context and location of items on the CAT and differences in difficulty ordering" (p. 3). Of the 20 studies, 14 were conducted at the college level. Three of the 20 studies showed significant differences; two of those were from the K–12 system. The remaining six studies from the K–12 system were from two research reports. For convenience, the study numbers used here are those used in the meta-analysis.

Studies 5 and 6 (Olson, Maynes, Slawson, and Ho, 1986) used grade 3 and 6 mathematics items, respectively, from the California Assessment Program. There was no significant difference reported for either study. Studies 11 through 14 (Baghi, Gabris, and Ferrara, 1991) used mathematics and reading items from the Maryland Functional Test. In Study 11, the mathematics CAT was administered first and in Study 12 the order was reversed. There was no significant difference between results in either study. In Study 13, the reading CAT was administered first and in Study 14 the order was reversed. In both cases there was a significant difference favouring the paper-and-pencil mode. In these two studies, scrollable text was required.

Another meta-analysis (Mead & Drasgow, 1993) used 115 sets of results from 24 research reports. They found little effect of medium of administration for power tests but cautioned

> Our conclusion – that a computerized version of a timed power test can be constructed to measure the same trait as a corresponding paper-and-pencil form – should not be taken for granted for any computerized test. We would expect the finding to generalize only to timed power tests that were computerized by the careful processes use by the researchers whose studies we have examined (p. 456).

Mead & Drasgow (1993) also note that "with additional studies the content of the tests could have been examined as a moderator. For example, it seems likely that reading extended passages on a computer display would be more difficult than reading from paper" (p. 456). They could not examine this variable due to the limited number of studies available but note that "Kiely et al. (1986), for example, investigated three different ways of presenting paragraph comprehension

items on a monitor …. All three modes, taken as a whole, were harder in the computerized version (d = -0.31)" (p. 457). In the case of speeded tests, where even physical differences between marking a bubble and selecting the response with a keystroke could have an effect, they found substantial differences (p. 453).

Gender and racial-ethnic groups were studied by Gallagher, Bridgeman, & Cahalan (2002). Operational testing data were drawn from GRE General Test, SAT® I: Reasoning (SAT), Graduate Management Admissions Test (GMAT®), and Praxis® Professional Assessments for Beginning Teachers. Comparing computer-based testing (CBT) with paper-based testing, they found that

> Although differences were quite small, some consistent patterns of changes in impact were found for racial-ethnic and gender groups. African-American examinees and, to a lesser degree, Hispanic examinees appear to benefit slightly from the CBT format. Where significant differences in impact were found for these groups, all indicated reduced impact as a result of the change to the CBT format. On all CBTs analyzed by race-ethnicity, African-American and Hispanic examinees performed better than or as well as they performed on the paper-based tests (p. 144).

They also note that "it is not clear whether the general lack of gender and racial-ethnic differences noted in the study resulted from minimal differences on all of these characteristics of the testing experience, or whether negative impact on one characteristic was counter balanced by positive impact on another" (p. 144). They end with a caution,

> … even though mean impact appears to be reduced for some groups of examinees, it is also unclear how the rank ordering of individuals is affected. It is conceivable that the mode of administration may interact with individual differences in test takers in such a way that rank ordering of students changes (p. 145).

Significant differences between administration modes on a practice Graduate Record Examination were found by Goldberg and Pedulla (2002). They also found a significant interaction between computer familiarity and test mode on the Quantitative subtest. They found that "time constraints negatively impacted test performance in both computerized modes" (p. 1065).

In a two-part study involving ACT tests of English, Reading, and Science Reasoning, Pommerich (2002, 2004) found that the initial results for English and Science Reasoning favoured the computer-based test. Some reasons given for this were that the paper-and-pencil tests underlined the particular phrase being referenced by the question while the computer highlighted the phrase and the relevant section of the text was centered in the screen with the options aligned with the highlighted phrase. It was also noted that completion rates were higher for the computer-based English and Science tests but lower for the Reading, which involved scrolling. A second study, following a redesign of the interface to more closely reflect the format of the paper-and-pencil tests, found that the differences were reduced. However, the English and Science Reasoning results were still better on the computer-based tests and the Reading results were lower.

Reflecting on the results following the redesign of the interface, Pommerich (2002) stated

> … it is important to note that the effect [of redesign] was not always the intended effect. This suggests that examinees are sensitive and respond to how information is presented on computer, but not always in ways that are predictable (p. 36).

and that

> In some cases, the results appeared influenced by better pre-test training on how to use the functions necessary to take the test on computer, improved navigation speed and navigation capabilities and making information about the test session more readily available (p. 36).

Pommerich (2002) noted that "computerized administration is more of an issue for complex tests that contain information that cannot all be displayed on-screen at once for an item" (p. 3), such as text-based passages and items requiring multiple figures or tables.

Schwartz, Rich, and Podrabsky (2003) compared computer-based and paper-based tests item by item using a Differential Item Functioning (DIF) procedure comparing results to the standardized group. The items were from the grade 4–9 InView aptitude test. All items flagged for DIF required scrolling (p. 6).

A study based on the 2003 Kansas large-scale assessment program found that there was no significant difference between paper-and-pencil and computer based testing (CBT) modes. Of the 48 schools that administered the CBT, 12 volunteered to administer the paper-based assessment as well. All but three of the schools administered the paper-based assessment after the CBT. It is possible that students were less motivated in the second (paper) administration and that finding no difference may be a result of lower paper-based performance than might have been expected. In addition to examining overall differences in results, gender, SES (free, reduced, or no school lunch), and academic placement (regular, gifted, disabled) were also examined, with no significant differences being found. Finally, they performed item-by-item DIF analysis and only nine items were flagged. Although they found a tendency for the flagged items to be large and require scrolling, they note that many items not flagged also required scrolling.

While differences in impact between modes of testing have been found for reading, Mazzeo, Druesne, Raffield, Checketts, & Muhlstein (1992) found that, after redesigning the computer interface based on an initial study, the results of the second study suggest that differences for the English Composition Examination were eliminated but not differences for the Mathematics Examination. "The results of both studies underscore the need to determine empirically, (rather than just to assume) the equivalence of computer and paper versions of an examination" (Abstract).

Clariana & Wallace (2002, p. 597) found that there was an interaction between mode and high- and low-attaining students: the computer test favoured high-attaining students over the paper-based test. They state

> Based on our review and these results, we anticipate that computer familiarity is the most fundamental key factor in the test mode effect, especially for unfamiliar content and/or for low attaining examinees (especially an issue for students with reduced computer access, such as women and minorities). In general, higher-attaining students will adapt most quickly to any new assessment approach (p. 600).

While most studies have been based on tests utilizing multiple-choice items, Russell (1999) and Russell and Haney (1997) studied differences in constructed-response items following disappointing results on a statewide test of writing in a school that had focused on the writing process. Russell and Haney (1997) hypothesized that there was a mismatch between the writing

processes carried out by the school and those of the writing test. In particular, the Advanced Learning Laboratory School (ALLS) used computers widely in their writing process while the statewide assessment required responses using paper and pencil. They found that multiple-choice test results on NAEP items did not differ much by mode of administration. However, for students accustomed to writing on computer, the results of test responses written on computer were substantially higher than those written by hand. Their results suggested that even with short constructed responses, the mode of administration may affect a student's performance particularly if the mode is a familiar one.

In a follow-up study, Russell (1999) included another school to increase the variance in computer experience among the students and included measures of keyboarding skills and of past experience using computers. Using items from the Massachusetts Comprehensive Assessment System (MCAS) and the National Assessment of Educational Progress (NAEP), Russell (1999) found that there was a positive computer group effect on the science test but that there were no overall group effects on the language arts tests. When keyboarding skill was used as a control variable, it was found that there was a substantial negative effect for students with keyboarding skills 0.5 standard deviations below the mean and, conversely, that there was a moderate positive effect for students 0.5 standard deviations above the mean. In order to minimize the effects of computer presentation, students taking a test on computer were given a hard copy of the test booklet. In comparing the results of this study to those found previously (Russell and Haney, 1997), Russell noted that the larger effect size found in the earlier study might be explained by those results coming from a formal test situation while this study was described to students as a practice for the spring state tests. He also noted that, since the state still required paper-and-pencil responses for written work, teachers at the ALLS spent considerable time in preparing students on paper-and-pencil writing.

Many of the studies noted above have considered interface complexity–for example, the impact of scrolling on items that require more space than can be shown in the computer screen frame, or scrolling through extended reading passages. Kobrin & Young (2003) studied the cognitive equivalence of reading between computer and paper modes of administration by using verbal protocols. "If it is found that participants answering the computerized test items engage in cognitive processes that are irrelevant to the construct of reading comprehension, such as processes that reflect working memory or spatial ability components, the construct validity of the computer test scores may come into question" (p. 116). Some of the differences are difficulty in reading text from a computer screen, and the inability to mark or underline text or to see the entire passage and all of the questions at once. They hypothesized that differences in process should be observable in the verbal protocols. The study was carried out using relatively easy and familiar material with 48 juniors and seniors of an eastern university. The subjects were assigned to four groups of 12: protocol–computer then paper, protocol–paper then computer, non-protocol computer then paper, and non-protocol–paper then computer. The protocol was therefore administered to 24 students. Although no significant achievement effects were noted (no ANOVA tables were in the article), the mean scores for protocol students and non-protocol students (computer and paper combined) had mean results on the first test of 5.04 and 5.12 respectively, but on the second test the mean scores were 5.37 and 4.92 respectively: the protocol subjects improved and the non-protocol students declined (Table 2, p. 125).

Summary of Literature

Where the impact of mode of test administration has been significant, the differences have been generally in favour of paper-based testing and most often for reading (Bergstrom, 1992; Mead &

Drasgow, 1993). In general, these differences have been attributed to reading lengthy passages where scrolling is required. This is supported by Pommerich (2002), who found that, when items on an English test were placed near the location of the information required, differences in mode favoured the computer and, in the same study, differences for reading favoured paper. Poggio, Glasnapp, and Yang (2005) and Schwartz, Rich, and Podrabsky (2003) found that all items that showed differential item functioning required scrolling.

Differences between modes of administration in mathematics testing do not appear often. However, Mazzeo, Druesne, Raffield, Checketts, & Muhlstein (1992) were able to eliminate differences in English composition by redesigning the computer interface, but the redesign did not remove the differences found for Mathematics.

Designing the computer interface to closely match that of the paper version is important. This conclusion is supported by Pommerich (2002, 2004), who also noted that the effect of the design is not always the intended one.

One study (Russell and Haney, 1997) raised the issue of a match between instruction and testing. Students in a school where computer-based essay writing was emphasized did poorly on a paper test of essay writing. When the test was repeated on computer, the students improved. In a later study including the same school (Russell, 1999); the expected differences did not appear. It was noted that the school where the initial differences were found now emphasized writing on paper to meet the state requirements.

Noting the general lack of racial or gender differences Gallagher, Bridgeman, & Cahalan (2002) wondered if a negative characteristic was balanced by a positive characteristic. They continue in this vein, commenting that, even if they found no group differences, the rank ordering of individuals might change. This is an important consideration. If group scores are important, then eliminating mean differences may be sufficient. On the other hand, if individual decisions are being made, then it is important to know what the impact is on individual students or at least on students with particular attribute profiles. They recommended that future studies should control for computer access and familiarity.

That other factors are involved is supported by Goldberg and Pedulla (2002), who found that there was a significant interaction between computer familiarity and test mode and that the relationship was non-linear. Time constraints also negatively impacted performance in computer modes. Russell (1999) found that differences in keyboarding skills impacted responses to essay writing on the computer. Clariana & Wallace (2002) also noted differential effects due to ability.

Clariana & Wallace (2002) stated that there was a need for research to demonstrate the match between paper and computer modes and that equating studies should be carried out.

This current study provides an opportunity to compare mode effects on multiple-choice Numeracy and Reading items, multiple-choice Reading, constructed-response Reading items, short Reading constructed responses, a several-paragraph focused writing task, and a multi-paragraph extended writing response. These effects will be studied overall, by gender, and by ability grouping.

2 Paper and Electronic Modes

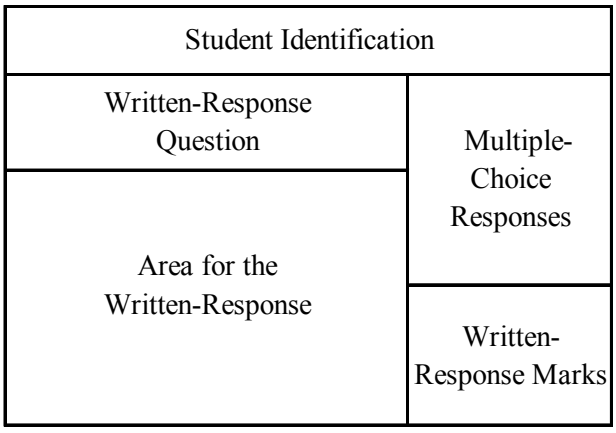
COMPARISON OF THE FSA PAPER MATERIALS AND THE ELECTRONIC INTERFACE

In paper mode the questions for all FSA components were contained in one booklet. Students responded to the Numeracy and Reading questions on two double-sided, personalized response sheets and to the writing prompts in a personalized response booklet. Wherever possible, reading questions were presented on the page facing the passage. If the reading passage extended onto a second page the questions followed on the third.

Response Sheet Format

Numeracy and reading questions were arranged in two sections, each containing a series of multiple-choice items and one constructed-response item. Each side of the double-sided response form was associated with one of the sections of questions. Only the response sheet was sent to the marking centre. The marks for the constructed responses were recorded on a space provided and the whole sheet was then scanned for both multiple-choice and constructed-response data together with the student identification number. Figure 1 shows a diagram of the layout of a reading and numeracy response sheet.

Figure 1 Diagrammatic Representation of a Reading and Numeracy FSA Student Response Sheet



The electronic interface for the FSA has been designed to present the test questions in a format that matches the paper version as closely as possible. Both Numeracy and Reading questions have an additional feature in the electronic version: a check box beside the question stem permits

students to flag the item for later reconsideration.

Reading Component

When the electronic Reading test was started, the student saw the statement, "Click on the blue title heading to read *Passage Title*" and several questions. Clicking on *Passage Title* which was in blue, split the window into two panes with the passage in the right pane and questions in the left.

Figure 2 shows a diagrammatic representation of a reading screen. A navigational bar at the top of the screen had four hyperlink boxes on the left side: Box 1–*Instructions* displayed the instruction page; Box 2–*For Reference* displayed relevant data or required formula; Box 3–*Exam Questions* caused a return to the current exam question from one of the other pages; and Box 4–*Review/Index* opened up a review page where all item stems were presented in index fashion. On that page, a yellow question mark beside the question number indicated that a question was yet unanswered; a red flag indicated a question that the student checked for reconsideration; and the question stem was a hyperlink back to that particular question. On the Reading screen additional control boxes were placed on the right: Box 5–*Close This Window*>> closed the right pane showing only the questions; Box 6–*Open Half Screen* reduced the size of the passage pane to one half of the screen; and Box 7–<< *Open Full Screen* permitted access to more of the passage at once. These last three dynamic buttons were useful because the line dividing the screen into the two parts can be *clicked* and *dragged* to change the relative width of the two panes. The <<*Back* and *Next*>> buttons at the bottom permitted students to move back to the previous item or forward to the next one. At the bottom left of the question pane there was a histogram with text above indicating that the student was on question number X out of a total of Y.

When a complete passage could not be presented in a pane, a scroll bar appeared on the right side of the right pane so that the student could move up and down through the text. To some extent, when the panes were resized, text flowed to fit the resized panes. If the size of the question pane was reduced sufficiently, text in the multiple-choice options would not resize and was hidden. When the passage pane was reduced and there were no other constraints, the text flowed smoothly to fit the pane size. In some circumstances, such as with double-columned text and pictures or diagrams which could not be resized, reducing the width of the passage pane would hide some text. Whenever text was hidden, a horizontal scroll bar was introduced into the pane. Figure 2 shows the placement of the two horizontal scroll bars, only one of which was displayed at any one time.

The Reading Component had two items that required a constructed response. In those cases, only the one item was displayed (see Figure 3) and a response box opened. There was no vertical scroll bar in the question pane. The response box could not be resized and so the lower left horizontal scroll bar appeared when the left pane was reduced too much. If the student entered more text than could be displayed in the response window, a vertical scroll bar appeared to the right of the response window. Text entry was straight forward with two depressions of the <Enter> key to indicate paragraph separation. *Click*, *Drag* and *Release* highlighted text which could then be moved with the mouse to a new location or the text could be *Cut* or *Copied* and *Pasted* elsewhere.

Figure 2 Diagrammatic Representation of a Computer Screen for the Electronic Version of the Reading Component Showing the Relative Locations of a Passage, Passage Scroll-bar, Items, and Other Navigational Features.

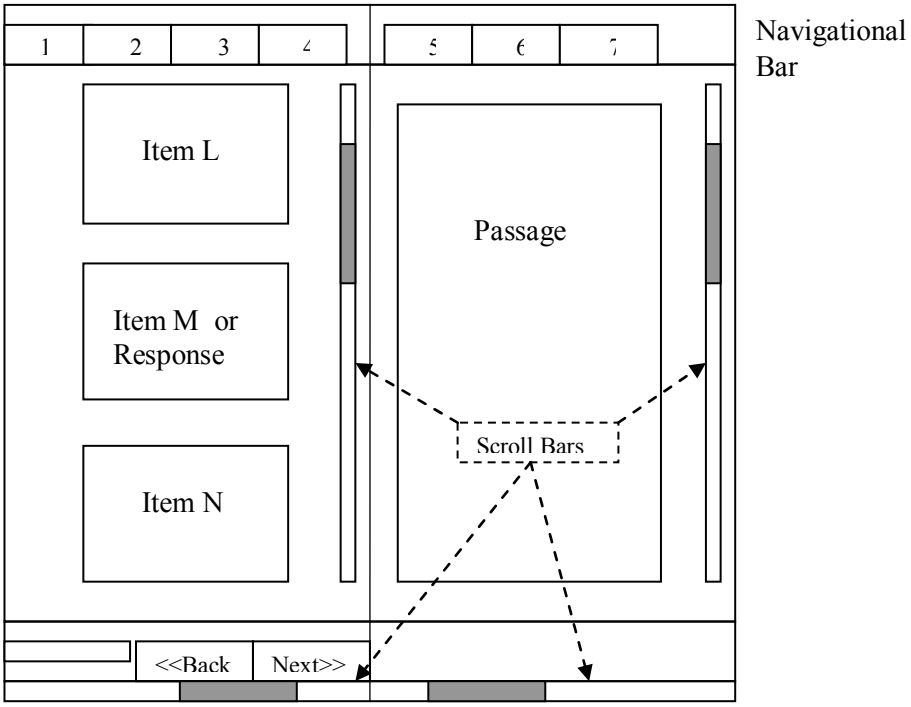
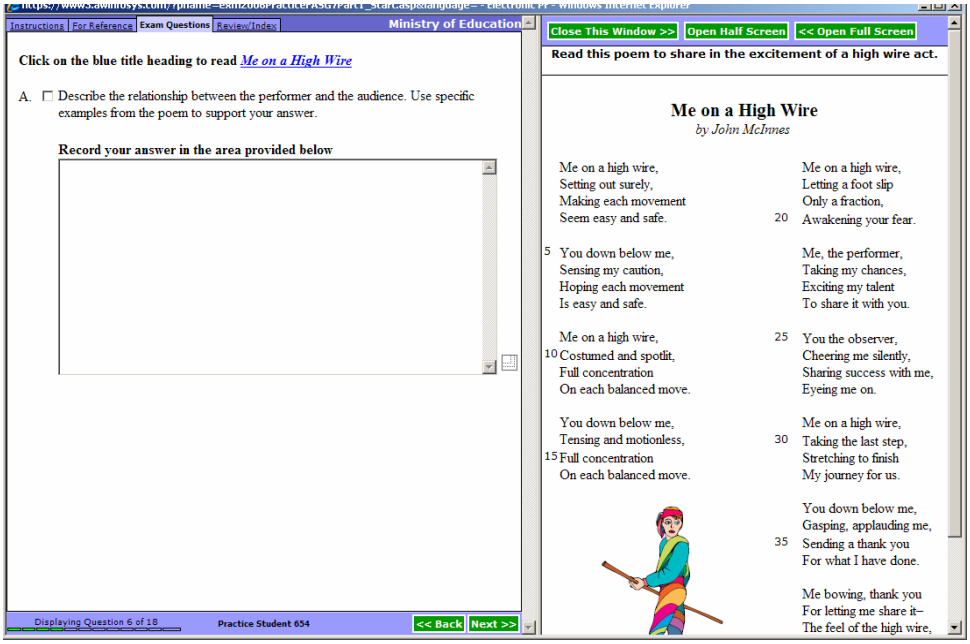


Figure 3 A Sample of a Reading Screen Requiring a Constructed-Response.



The Reading Component also had a "Connections" section that required the student to respond to questions on two passages. One of the questions required a constructed response. In this component, when the passage was opened the <<*Back* and *Next*>> buttons were hidden and the lower left scroll bar had to be used to get to these buttons. Once the student got to the page where the constructed-response was entered, the student could not move from one passage to the other without going back to a previous screen where the passage hyperlinks are displayed.

Numeracy Component

In comparison to Reading, the screen for electronic Numeracy questions was relatively straightforward. The items were virtually identical in format to the paper questions except that they were presented one at a time. There were no horizontal scroll bars or multiple panes. The right hand vertical scroll bar appeared only when the item together with the options could not fit vertical dimension of the screen. As noted earlier, students had to respond to questions requiring a constructed response directly on the paper response form; no electronic interface was required.

Writing Component

In preparation for the Writing Component students were lead through a pre-writing activity which included reading through five simple criteria to be considered by the student, who then engaged in a structured planning session, and wrote the actual response. Straightforward electronic pages are provided for each of these activities as they are in the paper mode. The actual writing page displays the prompt and a single large response box.

NAVIGATION IN PAPER AND ELECTRONIC MODES

Navigation is a term used in electronic applications to refer to the movement from page-to-page and within the pages of the application to select desired material. It is not used as often to refer to moving from page-to-page and identifying material for future reconsideration in paper applications. Yet the layout of paper examinations is developed with a view to minimizing the need to copy information and to move back and forth within the examination. It is useful then, in the light of some of the findings in the literature, to compare the features of the two modes that may limit or facilitate students in demonstrating their best achievement.

Reading Multiple-Choice (RMC)

In the paper mode, multiple-choice Reading items were presented as close to the passage as possible. If the passage was short enough, then passage and items were on facing pages. Lengthier passages span two facing pages with associated items often overleaf, requiring the student to flip pages back and forth, and limiting the possibility of placing the question near the relevant text. Questions can be marked for reconsideration but if a response is changed on the response form then nearly complete erasure of the old response is required. Further, in order to relocate items to be reconsidered, the student must go through all the pages: missing a question marked for review is quite possible. Questions not answered show as a blank on the response sheet.

In the electronic mode, multiple-choice questions are always next to the passage because of the capability to scroll the passage. The Review/Index page clearly shows items marked for review

that were missed and provides direct access to any item by hyperlink. Differences in changed responses are absolute. Students will not place the item response opposite the wrong question on a response form. In this mode, questions and passages are presented in a plane vertical to the desk, with all material a constant distance from the student.

From the above comparison, it would appear that the electronic interface for RMC items provided easier access to the material, reduced the potential for mistaken responses and blurred corrections. It would be reasonable to hypothesize that Reading achievement in the electronic mode would be higher than in the paper mode.

Reading Constructed-Response (RCR)

In the paper mode, the student responded directly on the separate response sheet (see Figure 1). The sheet could be placed as the student desired in relationship to the passage. This facilitated copying relevant text references correctly.

In the electronic mode, students could scroll through the passage where required, placing passage text close to the response box. Reading responses were relatively short, seldom requiring scrolling within the response box.

As pointed out by Russell (1999) and Russell and Haney (1997) students instructed in one mode and tested in another may be disadvantaged. The schools of this study were self-selected, likely because staff and students were relatively familiar with computers. In these schools, many of the students' written assignments would likely be done on a computer. If this was the case for several years, it could well be that students could have been disadvantaged when required to respond in handwritten form.

In addition, even though markers were trained to ignore readability of handwriting, the clarity of the electronic text may have provided an advantage; perhaps more so for boys than girls. It would be reasonable to expect that RCR would show increased achievement over paper and that the difference would be greater for males.

Numeracy Multiple-Choice (NMC)

There were no discernable differences in the navigation in paper and electronic modes. However, nearly all numeracy questions require some paper-and-pencil work in order to find the answer. In the paper mode, questions could be worked out on the same page and often on the same line as the question. In the electronic mode, any required information had to be transferred to the working paper, the solution found, and the result selected on the screen. Opportunities for making mistakes were more frequent in electronic mode.

It is hypothesized that achievement in the electronic mode would be less than in the paper mode. Further, that the difference would be greater for students of lower ability.

Numeracy Written-Response (NCR)

There was no electronic mode for NCR. However, NCR was analyzed in relation to performance in paper mode years and electronic mode years. It is hypothesized that there would be no difference in NCR for electronic and paper modes.

Writing Focused Response (WFR)

The focused response or short writing section is called the Connections section because students were required to make connections between two passages. In the paper mode, students could move their response booklet in relation to the two passages from which they were expected to take specific references.

In the electronic mode, students could not switch from passage to passage directly from the response screen; they had to use the <<*Back* button to get to the previous screen and then activate the hyperlink for the desired passage. This may not have been obvious to the students.

Although it may be expected that WFR achievement might be improved over paper mode for the same reasons that RCR would, the difficulty in navigation would mitigate against improved achievement in the electronic mode. It is hypothesized that there would be no difference in achievement by mode for WFR for females but that males, who may be less intimidated by complex navigation, would show improvement in the electronic mode.

Writing Extended Response (WER)

In the paper mode, the planning page and the criteria could be used flexibly by the students and compared side by side with the essay.

In the electronic mode, the planning and criteria pages could only be accessed by the << *Back* button or through the Review/Index page. In this part of the assessment, it is likely that the response would extend beyond the response box requiring additional facility with navigation.

In the extended response, any impact of handwriting over computer text would be increased. Further, the added navigational flexibility would seem to favour males. It is hypothesized that males would show improved achievement in the electronic mode while females would show reduced achievement.

3 Study Design

AVAILABLE DATA

The design of this study was constrained to data already gathered by the British Columbia Ministry of Education (MOE). Individual schools determined whether or not their students would write the FSA electronically or not. Some schools elected to administer electronically to only a selected few students. In all, 28 schools administered the Grade 7 FSA electronically in at least one of the years 2004, 2005, and 2006 to at least some of their students: several did so twice and one school administered electronically three times. Of the 28 schools, four administered the FSA electronically to one student, four schools were distance education schools and 5 others administered electronically to fewer than 70%. In two very small schools (fewer than 10 students per grade), data was missing for one or more components. The Numeracy analyses were carried out on 13 schools with a total of 769 students who wrote electronically, the Reading and Writing analyses were carried out on 15 schools with a total of 779 and 725 students respectively who wrote electronically. Table 2 shows the numbers of students in each component, mode, and gender.

The electronic administration of the FSA became an option in 2004. Prior to 2004, all schools administered in the paper mode. Table 1 shows the years in which the FSA was administered electronically for each school in the study.

Table 1 The Years in Which FSA was Administered Electronically in Each School of the Study.

	School	1	2	3	4	5	6	7	8	9	10	11	12	13	14**	15**
	2004	*				E										
Year	2005		E		E	E	E			E						
	2006	E	E	E		E		E	E		E	E	E	E	E	E

* Blank cells indicate that the FSA was administered in paper mode for those schools and years.

** Not included in Numeracy because gender was not available.

While information about computer access and familiarity was not available, the FSA tests do provide a rich set of data. The FSA has three components: Numeracy, Reading, and Writing. Each of these three components has two parts: Numeracy Multiple-Choice (NMC), Numeracy Written-Response (NCR), Reading Multiple-Choice (RMC), Reading Written-Response (RCR),

Writing Focused Response (WFR)–several paragraphs, and Writing Extended Response (WER)–a more lengthy response. In addition, for grade 7 students taking the FSA in 2003, 2004, 2005, and 2006, FSA achievement in grade 4 was also available, thus permitting an indicator of ability level.

Because any individual student wrote only in paper or electronic mode, in this study the unit of comparison was the school. That is, the students writing in the electronic mode in a given school were compared to the students who wrote in the paper mode in that same school. Only schools that administered the FSA in both electronic and paper modes were included in this study.

Most schools experience considerable year-to-year variation in average student achievement: a cohort effect. Therefore, it was important to include as many school cohorts in the analysis as possible. Although the FSA has been administered from 2000, data for Writing were only available from 2001 onward so each school included in this study had six cohorts.

Student responses were collected electronically for all parts except NCR where all students were required to respond on paper as there was no way to electronically capture diagrams, equations, and interspersed comments. Analysis of the NCR is useful though, as it provides an indication of the level of achievement of the schools included in the study with that of the rest of the province and also provides a check of the level of achievement between the paper years and the electronic years for the thirteen included schools. All analyses were carried out with SPSS (Statistical Package for the Social Sciences) version 14.

The analysis was broken into two parts. Part 1 dealt with the question of whether or not administering a test electronically affects the achievement of the students and Part 2 considered whether any there were any differential effects related to ability level. Both parts of the study included gender and school as factors. The reason for breaking the study into two parts will become clear when the study design is described.

In Part 1, Grade 7 FSA results from 2001 through to 2006 were included. There were three factors in this part of the analysis: mode, gender, and school. Because the schools are self-selected the ANOVA design set school as a fixed effect. The factors of mode and gender were also set as fixed effects**.** Table 2 shows the number of students by gender and mode for each of the FSA components.

Table 2 Numbers of Students in Part 1 of the Study by Mode, Component, and Gender

	Paper			Electronic		
	Num.	Read.	Writ.	Num.	Read.	Writ.
Males	1457	1464	1482	401	402	371
Females	1379	1424	1437	368	377	354
Total	2836	2888	2919	769	779	725

Table 3 Numbers of Students in Part 2 of the Study by Mode, Component, Gender and Grade 4 Achievement Level

		Paper			Electronic		
	Ach Lev	High	Average	Low	High	Average	Low
Numeracy	Males	187	255	273	125	120	91
	Females	207	248	282	102	95	104
	Total	394	503	555	227	215	195
			Paper Total	1452		Electronic Total	637
Reading	Males	121	135	183	121	105	120
	Females	174	171	145	125	92	93
	Total	295	306	328	246	197	213
			Paper Total	929		Electronic Total	656
Writing	Males	62	352	39	40	261	18
	Females	138	342	15	88	200	11
	Total	200	694	54	128	461	29
			Paper Total	948		Electronic Total	618

The analysis of Part 2 took advantage of the fact that the majority of students who wrote the Grade 7 FSA in 2004, 2005, or 2006 wrote the Grade 4 FSA by paper in 2001, 2002, and 2003 respectively. A comparison of Table 2 with Table 3 shows that the number of students in Part 1 is substantially higher than in Part 2. The major difference is that the grade 7 cohorts for the 2001, 2002, and 2003 assessments (all in paper mode) could not be included because there were no matching grade 4 FSA results: the number of students included in the paper group of Part 2 is approximately half that in the Part 1 analyses. The impact is that the estimates of achievement in paper mode for the schools included in Part 1 of the study are more stable. This difference is the reason that the study was broken into two parts. The additional attrition of students in the electronic mode in the Part 2 analyses can be attributed to immigration, emigration, and changes in local decisions about including and excluding special needs and second language students. For students in these categories, one or the other of the grade 4 or grade 7 results was missing.

Rescaling of Components and Parts of Components

Although IRT parameters were available from the MOE for each of the items in each year, those parameters reflected equating from year-to-year. Equating from year-to-year is intended to permit tracking of changes in performance in the system. Because data for the electronic mode tended to come from later years and paper from earlier years, any systemic changes would be reflected in the data. Therefore, the entire (i.e. population) data set for each year and component part and total component was rescored. The scaled score distribution for each component part and overall

component was set to a mean of 500 and a standard deviation of 100. This was done in order that a common scale could be used for each of the components and that effect sizes would be similar.

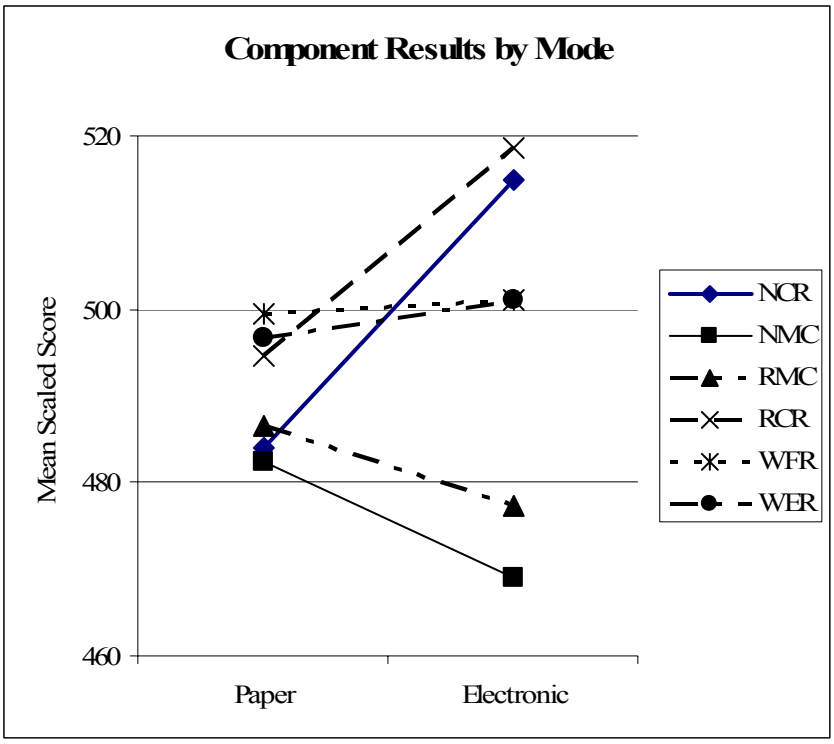
Table 4 shows the means and standard deviations for each of the FSA component parts for paper and electronic modes. Considering only the paper mode, the mean of the component parts is less than the provincial mean of 500 for NCR, NMC, and RCR and virtually the same as the provincial mean for both writing parts. This indicates that the study schools are somewhat lower-achieving schools than is typical in the province. The standard deviations of the component parts are very similar to those of the provincial standard deviation of 100. In general, it can be said that although the students in the schools in the study may be slightly lower in achievement in Numeracy and Reading they exhibit similar variability.

Table 4 Means and Standard Deviations for Each FSA Part for Paper and Electronic Modes.

		NCR	NMC	RMC	RCR	WFR	WER
Paper	M	484	482.4	486.6	494.6	499.4	496.7
	SD	98.1	96.3	101.3	100.7	102.3	103.7
	N	2836		2888		2919	
Electronic	M	514.9	469	477.2	518.6	501	501.1
	SD	96.9	97.4	107.3	112.9	97.1	90.7
	N	769		779		725	
Total	M	490.6	469	484.6	499.7	499.7	497.6
	SD	98.6	97.4	102.7	103.9	101.3	101.3
	N	3605		3667		3644	

Figure 4 shows a graph of the mean scaled scores for paper and electronic modes on each of the component parts of the assessment.

Figure 4 Mean Scaled Scores for Paper and Electronic Mode by Component Part.



ANALYSES: PART 1

For each of the six component parts, a three-factor analysis of variance (ANOVA) was carried out. The three factors were mode, gender, and school. School was included as a factor because schools self-selected whether to administer the FSA electronically or not, students are embedded within a school as part of a cohort, size of individual schools varied substantially (the smallest schools had a cohort enrolment less than 10 and the largest had cohort enrolments greater than 200), and computer access, and teacher and student readiness, may vary from school. In addition, each school would then act as its own control.

In the following discussion of the results of the analyses, differences between modes will be P–E and between genders will be M–F, so positive differences will mean that paper results and male results are respectively higher. Because this analysis is exploratory, no attempt has been made to set levels of confidence for multiple comparisons; effects that appear with a confidence level greater than 95% confidence will be noted. In addition, there has been no attempt to adjust for the fact that the results for each part are based on multiple measures for a given student: the results may be correlated

Numeracy

Although Numeracy is typically administered at the end of the assessment, it will be treated first because in the administration of the FSA, Numeracy constructed-responses must be done on

paper even when the administration of the other components is electronic. It was anticipated that this would provide a benchmark, although somewhat limited, of the overall difference in ability of students in these schools in relation to the rest of the province and of the extent of any difference in the electronic cohorts compared to the paper cohorts. Table 5 shows that levels of male and female performance on the NCR, while close, significantly favoured females (M–F difference is -6.3; p = 0.000). The table also shows that the overall P–E difference between students in the two modes was -30.9 (p = 0.006): results for students in the electronic mode years are substantially higher even though they responded as usual in paper mode for this part of the assessment. There was also a significant mode-by-school interaction (p = 0.000). Of interest is that the difference between modes for females (-29.6) is less than that for males (-32.1) but the interaction was not significant. The full ANOVA for these results can be found in Table 12 of the Appendix.

Table 5 Means and Differences Between Means of NCR Scores by Gender and Mode

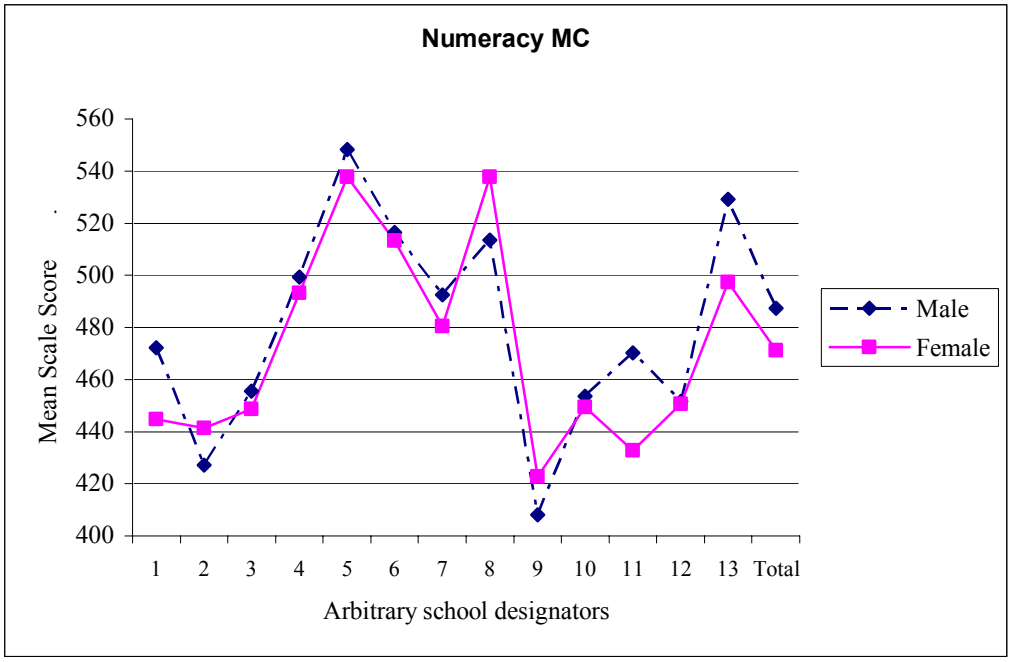
	P	E	Gender	P - E	Prob.
M	480.6	512.7	487.5	-32.1	-
F	487.6	517.2	493.8	-29.6	-
Mode	484	514.9		-30.9	0.000
M - F	-7	-4.5	-6.3		
Prob.	-	-	0.006		GM=490.6

Table 6 Means and Differences Between Means of NMC Scores by Gender and Mode

	P	E	Gender	P-E	Prob.
Male	491.3	473.3	487.4	17	
Female	473	464.3	471.2	8.7	
Mode	482.4	469		13.4	0.005
M-F	18.6	16	16.2		
Prob.			0.788		GM=479.6

Because the results of the constructed-response part of this component have already been introduced, the results of the Numeracy multiple-choice will be presented next. The means and differences between means of NMC are shown in Table 6. There is a mode difference (13.4, p = .005) favouring paper. That difference is greater for males (17.0) than females (8.7) but as Table 13 in the Appendix shows there is a non-significant mode-by-gender interaction (p = 0.206). The difference between genders (16.2) favouring males, although larger than the mode difference, is not significant (p = 0.788). While the non-significance may appear contradictory, there is a significant schools factor (p = 0.000) which removes much of the gender variance (see the small gender mean squares in Table 13 in the Appendix) and the graphical representation in Figure 5 which shows a much larger variation among schools than within schools. The significant gender effect can therefore be attributed to some larger school having considerable gender differences. There is a significant mode-by-school interaction (p = 0.000) and a non-significant mode-by-school-by-gender interaction (p = 0.851).

Figure 5 Comparisons of Male and Female Mean Scale Scores of Numeracy Multiple-Choice by School. Schools are in Arbitrary Order.



It is interesting to note that although the gender differences are quite small for both NCR and NMC, the constructed-response difference favours females and the multiple-choice difference favours males. Additional comments about this will be made in the Discussion section.

Reading Multiple-Choice

There is a significant difference between modes of 9.4 (p = 0.009) and a difference favouring females but not significant (-12.7, p = 0.118). There is a main effect of schools (p = 0.000). There is no significant mode-by-gender interaction (p = 0.797), gender-by-school interaction (p = 0.654), mode-by-school interaction (p =0.848) or mode-by-school-by-gender interaction (p = 0.629). The full ANOVA table can be seen in Table 15 in the Appendix.

Table 7 Means and Differences Between Means of RMC Scores by Gender and Mode

	Paper	Electronic	Gender	P-E	Prob.
Male	480.5	470.8	478.4	9.7	-
Female	492.9	484.1	491.1	8.8	-
Mode	486.6	477.2		9.4	0.009
M-F	-12.5	-13.3	-12.7		
Prob.	-	-	0.118		GM = 484.6

Reading Constructed-Response

The means and differences between means of the Reading constructed-response scores are shown in Table 8. While there is a substantial overall difference in mode in favour of electronic it is not significant (-24.0, p = 0.563). The difference in mode, favouring electronic, is larger for females (-33.7) than for males (-15.4) but the gender-by-mode interaction is not significant (p = 0.079). There is a significant gender effect (-34.7, p = 0.000) favouring females. The gender difference is larger for electronic (-49.2) than for paper (-30.9). The lack of a significant gender-by-mode interaction may be explained because of the significant mode-by-school interaction (p = 0.000) and the significant gender-by-school interaction (p = 0.032). Table 17 in the Appendix shows the full ANOVA table.

Table 8 Means and Differences Between Means of RCR Scores by Gender and Mode

	Paper	Electronic	Gender	P-E	Prob.
Male	479.3	494.8	482.7	-15.4	-
Female	510.3	544	517.3	-33.7	-
Mode	494.6	518.6		-24	0.453
M-F	-30.9	-49.2	-34.7		
Prob.					GM =
	-	-	0.000		499.7

Writing Focused Response

Table 9 shows the means and differences between the means of WFR. The small differences in means across the table may well reflect the fact that a very large proportion of students were in a single score category: 72% received a score of 2. Even so, there is a nearly significant 0.091 mode effect and the gender-by-mode interaction is nearly significant (0.095) as well. There is a significant three-way interaction gender-by-mode-by-school (0.024). Table 19 in the Appendix shows the complete results for the ANOVA.

Table 9 Means and Differences Between Means of WFR Scores by Gender and Mode

	Paper	Electronic	Gender	P-E	Prob.
Male	498.5	501.1	499.1	-2.6	
Female	500.2	500.8	500.4	-0.6	
Mode	499.4	501		-1.6	0.091
M-F	-1.7	-0.3	-1.3		
Prob.					GM =
			0.784		499.7

Writing Extended Response

The results of the Writing extended-response part are shown in Table 10. There is no difference between electronic and paper modes for females (-0.5) and a small difference favouring electronic for males (-8.3). These results are similar for the WFR results. The difference favouring the males

in the electronic mode (8.5) is interesting in that this difference does not appear in the WFR results but the interaction of gender-by-mode is not significant. However, there is a significant three-way interaction of gender-by-mode-by-school (0.018). The full results of the ANOVA are in Table 20 in the Appendix.

Table 10 Means and Differences Between Means of WER Scores by Gender and Mode

	Paper	Electronic	Gender	P-E	Prob.
Male	496.9	505.2	498.6	-8.3	-
Female	496.4	496.9	496.5	-0.5	-
Mode	496.7	501.1		-4.4	0.093
M-F	0.5	8.5	2.1		
Prob.	-	-	0.166		GM = 497.6

ANALYSES: PART 2

The goal for the analyses in Part 2 was to determine whether or not differences between achievement in paper and electronic modes would change by achievement level. In all the analyses for Part 2, the Grade 4 FSA achievement on the overall component was used as the basis for achievement. Several factors made this a reasonable decision. Each component of the Grade 4 FSA has a structure similar to its Grade 7 counterpart. Tables of specifications are similar as are the relative numbers of multiple-choice and constructed-response questions.

It was decided to create categories of student performance rather than use the Grade 4 measure as a covariate. In this way, non-linear relationships, if they existed, could be demonstrated. Each of the Grade 4 FSA components was split into three achievement levels: upper, middle, and lower thirds; categories 1, 2, and 3 respectively. Each of the Grade 7 students was assigned an achievement category level based on grade 4 achievement. Table 11 shows the correlations between the scores in each component part of the Grade 7 study and the corresponding total component scores from Grade 4. Although NCR was only administered in the paper mode and is not considered in this section, the correlation is included for completeness.

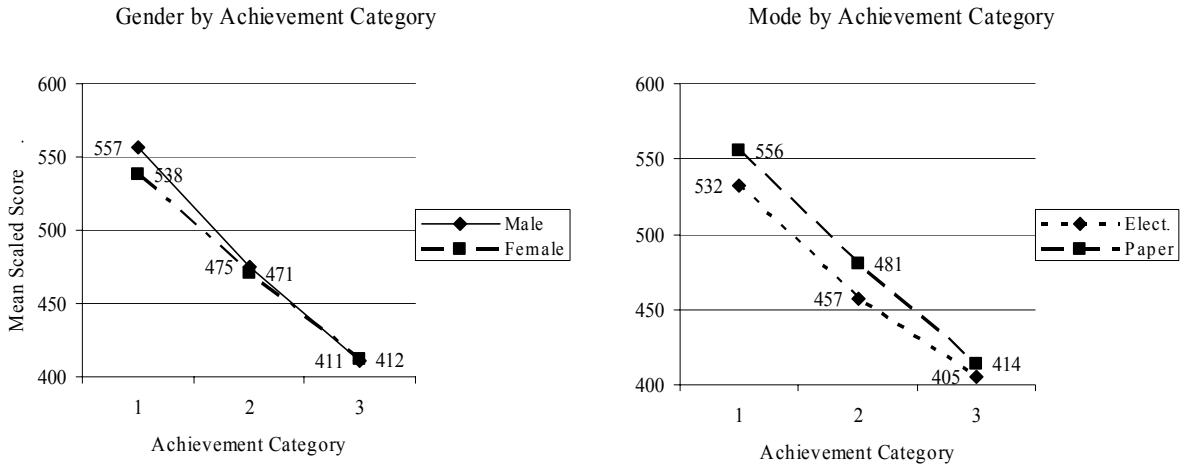
Table 11 Table of Correlations Between the Scores of the Component Parts of the Grade 7 FSA and the Grade 4 Total Scores of the Respective Components

	Gr. 7 Component Part					
	NMC	NCR	RMC	RCR	WFR	WER
Correlation	0.646	0.431	0.676	0.426	-0.002	-0.002
N	157219	157219	118110	118110	117866	117866

The lack of correlation between the Grade 4 Writing total score with each of the writing focused response and extended response scores can be attributed to nearly 75% of the students getting the same score in Grade 4 on a possible 10-point scale. All 75% were assigned to Category 2. As can be seen from Table 3, only 5.3% of the students received scores that placed them in Category 3. Because of the lack of correlation between Grade 4 and the Grade 7 results in Writing, neither focused response nor extended response parts of Writing will be considered in this section.

When considering the categorical information, there is no need to discuss the main effect: that there are significant differences in achievement among the categories is expected from the way the categories were defined. What is of interest is whether there is a marked change in the difference between the achievement level of categories 1 and 2 and the change between categories 2 and 3.

Figure 6 Numeracy Multiple-Choice Two-Way Interactions—Gender-by-Category and Mode- by-Category Showing Differences Within Each Category.

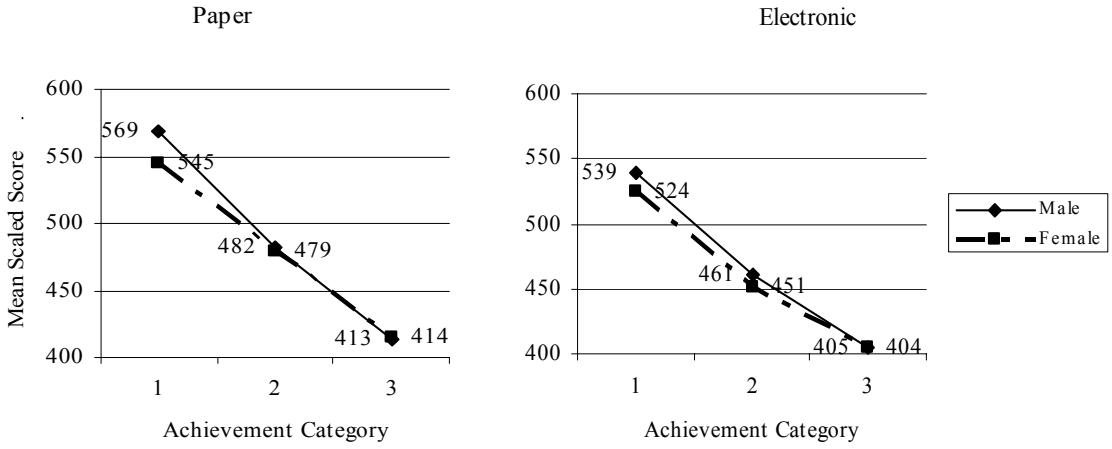


M - F	18.6	4.3	-0.3	P - E	23.7	23.7	9.0

While there was no overall gender difference, as discussed earlier in the Part 1 analysis, the graph of the gender-by-category interaction shown in Figure 6 indicates that the difference between males and females in Category 1, the high achievement level, is greater than in the other two (the interaction is significant, $p = 0.004$). A second interaction, mode-by-category, is also significant ($p = 0.007$). In this case, the difference between paper and electronic modes is smaller for Category 3 students. The full ANOVA can be found in Table 14 in the Appendix.

Figure 7 shows the differential results of males and females in each category and for each mode. The differences noted for mode in the two-way interaction above, are reflected in the lower overall location of the graph of the electronic mode when compared with the paper mode. The male-female difference in the paper mode, while large for Category 1 students, virtually disappears for Categories 2 and 3. However, in the electronic mode the difference is smaller for Category 1 students, is somewhat maintained for Category 2, and then disappears in Category 3. While of interest, the gender-by-mode-by-category interaction is not significant ($p = 0.149$).

Figure 7 Numeracy Multiple Choice Three-Way Interaction – Mode-by-Category-by-Gender Showing Differences Within Each Category.



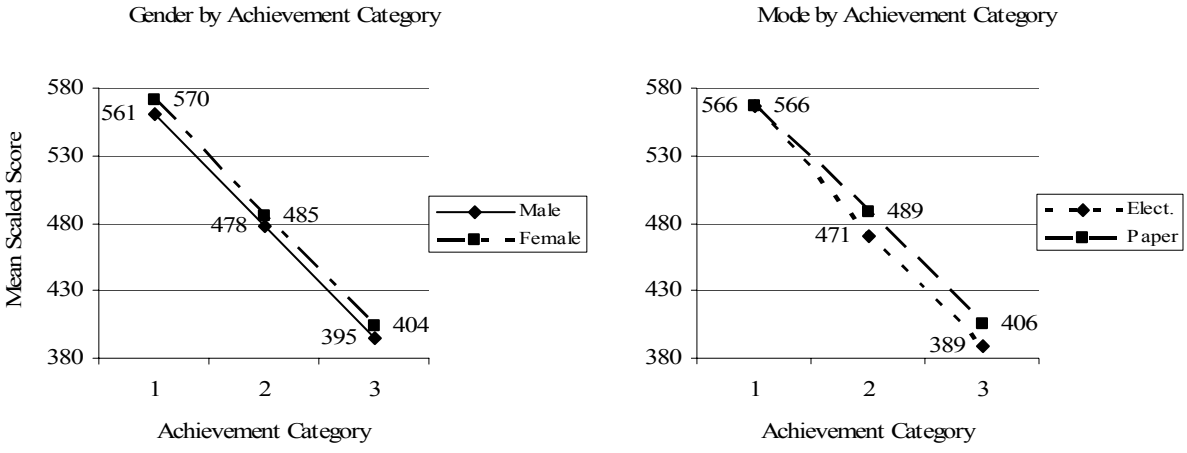
M - F	23.9	3.6	-0.9	14.1	9.6	0.8

Reading Multiple-Choice

As noted earlier, there is no significant gender difference and as shown in Figure 8 the differences are consistent from category to category. Therefore, the nearly significant ($p = 0.072$) gender-by-category interaction could result from the significant school-by-gender-by-category interaction ($p = 0.024$). There are larger differences in the mode-by-category interaction that are not significant. The significant mode-by-school interaction ($p = 0.009$) may account for this.

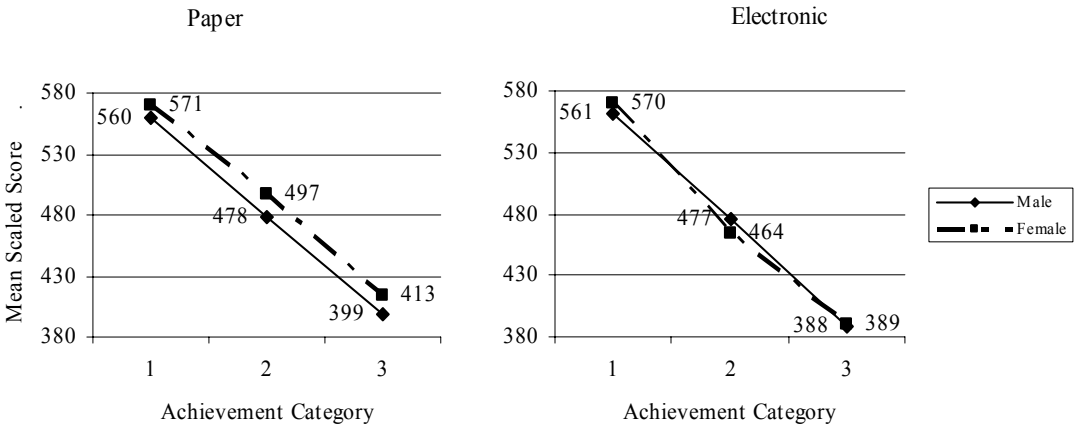
While the graphs in Figure 9 are suggestive, showing a different pattern between the two modes, the gender-by-mode-by-category is not significant ($p = 0.169$).

Figure 8 RMC Two-Way Interactions –Gender-by-Category and Mode-by-Category Showing Differences Within Each Category.



M - F	-9.9	-7.7	-9.1	P - E	0.2	17.7	16.9

Figure 9 RMC Three-Way Interactions – Gender-by-Mode-by Category Showing Differences Within Each Category.



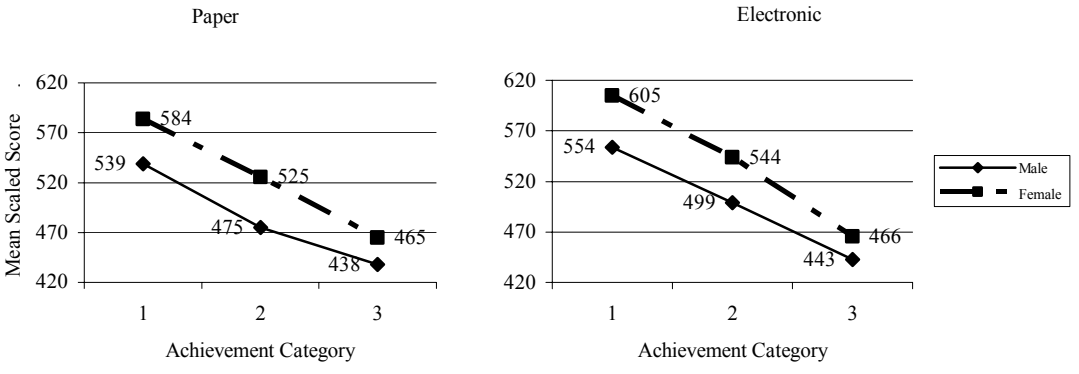
M - F	-10.7	-18.2	-14.1	-8.9	12.2	-1.2

Reading Constructed-Response

In the RCR Gender-by-Achievement-Category graph in Figure 10, there is half the difference between the genders in the lower category than in the other two. While this is of interest, the gender-by-category interaction is not significant (p =0.074). A similar type of change is shown in the Mode-by-Achievement-Category graph in the same figure but with smaller magnitudes. The mode-by-category interaction is not significant (p = 0.952).

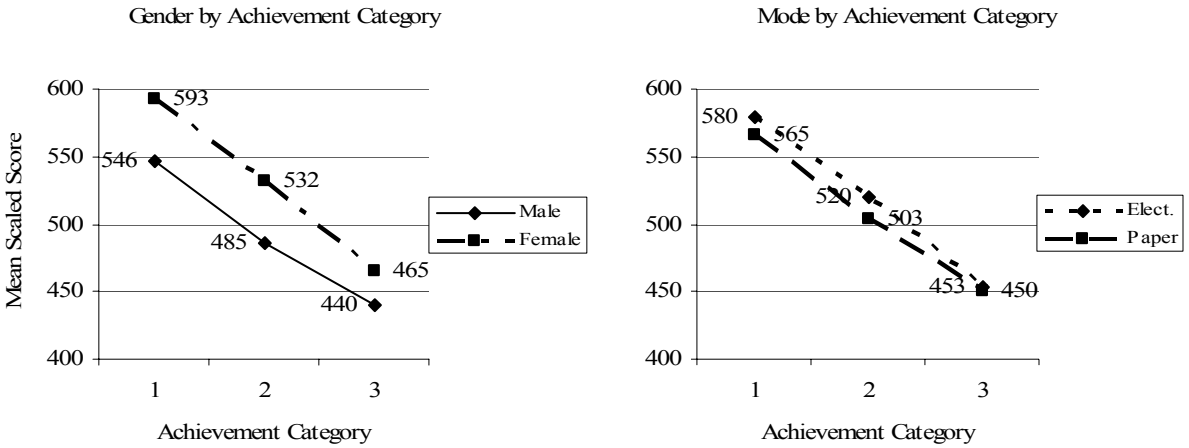
In Figure 11 the graphs reiterate the male-female differences shown in Figure 10 but with minor changes in segment slopes. The mode-by-gender-by-achievement-category interaction is not significant (p = 0.540). The full ANOVA results are in Table 18 in the Appendix.

Figure 10 RCR Two-Way Interactions – Gender-by-Category and Mode-by-Category Showing Differences Within Each Category



M - F	-44.8	-50.5	-27.1	-51.1	-45.2	-22.9

Figure 11 RCR Three-Way Interactions—Gender-by-Mode-by Category Showing Differences Within Each Category.



M-F	-46.2	-46.6	-25.4	P-E	-14.7	-16.7	-3.1

4 Discussion & Recommendations

The groups of students that took part in the paper and electronic modes were not randomly selected. However, each school in the study had results for cohorts administered in each mode. Further, the data from several years was used to limit the effects of individual cohorts. School was used as a factor for the analysis of each component part because schools self-selected themselves to take the FSA electronically and because of substantial differences in the size of schools.

Numeracy Constructed-Response

The results of the Numeracy constructed-response items were completed on paper by all students in both modes. It was anticipated that there would be little mode difference. However, the difference was significant and, in addition, there was a significant mode-by-school interaction. An inspection of the results by school showed that, even though the magnitude of the paper-electronic difference varied from school to school, in only three cases were the constructed-response means less in the electronic years compared to the paper years. Numeracy constructed-response items were among the last items administered. Although unexpected, it could be that students were either relieved that they no longer had to work on the computer or that the electronic administration maintained student engagement beyond what occurred in the paper mode years. It would have been interesting to survey students as to how they felt about interest in computers and in the electronic administration in particular.

Numeracy Multiple-Choice

The NMC results showed a significant mode difference favouring paper and that the difference is greater for males than for females the interaction is also significant. This can be seen graphically in Figure 12. The effect was anticipated because of the need to transfer information from the computer screen to paper for computation and then to find the correct answer on the screen. In paper mode, students could do their rough work on the booklet and then transfer their result to the response sheet. The response sheet could be located in close proximity to the students' rough work. It may also be the case that computer stations do not provide adequate space for students to do rough paper-and-pencil work. The differences between modes for NMC is perhaps more important if one were to accept the possibility that students may have been more motivated in the electronic mode.

It was anticipated that any difficulty in transferring data back and forth from screen to paper would impact students with lower Numeracy abilities. That is, the electronic mode would have a greater negative impact on the performance of Category 3 students than for the other two

categories. While the mode-by-achievement-category interaction was significant, the difference relatively few items that discriminated at the lower level.

Although not significant, the gender-by-mode-by-achievement-category interaction shows that while for paper the male-female difference for Category 2 students is small (3.6), in the electronic mode that difference is somewhat larger (9.6). At the next level of analysis the paper-electronic difference for males in Category 1 (29.9) is greater than for females (20.1) and the relationship is reversed for Category 2 students: paper-electronic difference for males (21.2) and females (27.2). There is an indication then that there is a greater negative impact of the electronic mode on higher performing males than moderately performing females. These differences are not likely attributed to complexities of the electronic interface, but rather to differences in attending to the details of copying information from one medium to another.

Reading Multiple-Choice

As in Numeracy multiple-choice, there is a significant difference between modes, but there is no significant difference between genders within modes. This is graphically shown in Figure 12. This finding is different from what was anticipated given that the Reading material provided easier access and that corrected responses would be clearer in the electronic mode. Navigational issues may have overridden any positive benefits of the electronic mode.

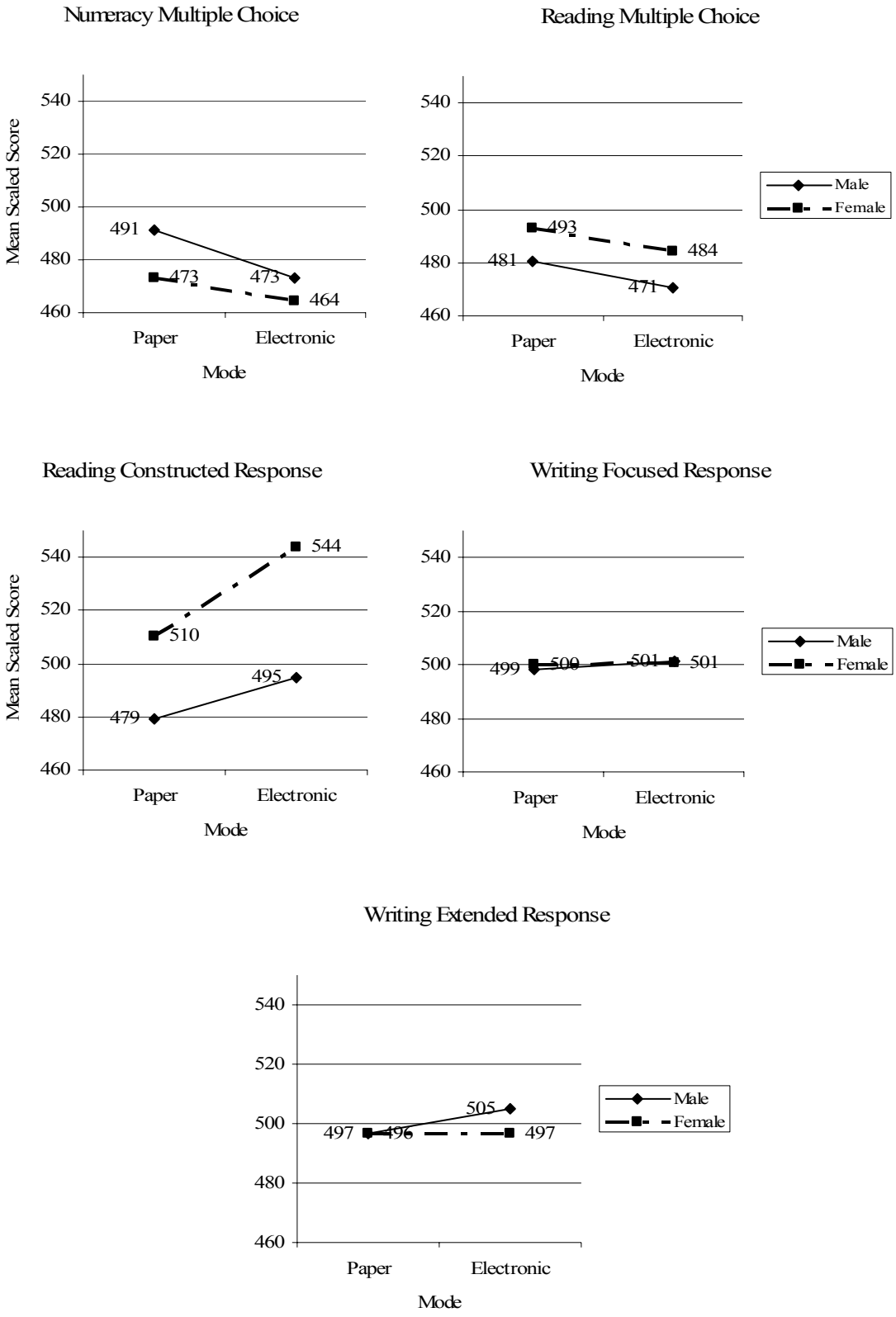
Differences in Reading results by achievement-category were not expected and, indeed, the analyses of Reading results by achievement-category show no differential impacts of mode on achievement-category.

The findings for RMC and NMC are similar, but the negative impact of the electronic mode is less for RMC than for NMC. Even though the literature suggests that splitting the screen so that text and questions can be placed side by side adds to navigational complexity, all Reading work was done on the screen. This suggests that having to work in two media, as required for Numeracy, created a greater impact than working within one medium.

Reading Constructed-Response

The significant gender difference favouring females was expected, as was the significant difference among schools. Responses to the Reading constructed-response items required students to provide support with reference to the text and it was anticipated that the electronic mode would facilitate that aspect. However, in spite of the substantial mode effect favouring electronic (-24.0) it was not significant, nor was the apparently large gender-by-mode interaction. However, the significant interactions of gender-by-school and mode-by-school imply that schools may vary considerably in preparing students for the electronic assessment, in mitigating gender differences in reading, in providing student access to school computers, in having students with access to home computers, and in giving the number and types of student assignments that may involve the use of computers.

Figure 12 Graphs of Mode-by-Gender Differences for NMC, RMC, RCR, WFR, and WER



Reading Constructed-Response

The significant gender difference favouring females was expected, as was the significant difference among schools. Responses to the Reading constructed-response items required students to provide support with reference to the text and it was anticipated that the electronic mode would facilitate that aspect. However, in spite of the substantial mode effect favouring electronic (-24.0) it was not significant, nor was the apparently large gender-by-mode interaction. However, the significant interactions of gender-by-school and mode-by-school imply that schools may vary considerably in preparing students for the electronic assessment, in mitigating gender differences in reading, in providing student access to school computers, in having students with access to home computers, and in giving the number and types of student assignments that may involve the use of computers.

Writing Focused Response

The results from the Writing focused-response items show that there is little difference between modes or genders. The size of the interaction while interesting is overshadowed by the significant gender-by-mode-by-school interaction. This interaction may reflect differences among schools in instruction, in preparation for the electronic assessment, or in student familiarity with computers.

Writing Extended Response

That trend continued with the results of the extended writing task, where there was a significant mode-by-gender-by-school interaction. There was a non-significant gender-by-mode interaction. However, while there was no difference between paper and electronic modes for females and there was an increase in electronic mode for males. This difference could reflect a greater facility with computer navigation among males, an improvement in marks because of the clarity of the text, or perhaps, a statistical artifact.

SUMMARY

The results of this analysis were tantalizing. They were consistent with the literature and with perceived differences in the paper and electronic modes of administration. But while some of the differences were significant, others were not.

While the analyses showed that there was a significant negative impact of the electronic mode on both Numeracy and Reading, the substantial improvement in the electronic schools on the Numeracy constructed-response, *which was done on paper,* in the same way as for the paper mode, is rather confusing.

It might also be the case that certain reading strategies used by readers are not currently supported in the electronic interface. For example, emergent readers may be encouraged to underline and highlight parts of the passage that they deem significant as they read. Providing a highlighting capability in the electronic mode might be of importance.

The analyses of each of the six parts of this study showed a mode-by-school interaction, a mode-by-gender-by-school interaction or both. Even if there are some overall differences between modes, there are also important school level differences. This would speak to the different

preparation provided to students for the electronic assessment, to previous access to and facility with computers, or to differences in teaching strategies and assignments involving the use of computers.

The analysis of both Writing components was hindered by little discrimination among student abilities: 74% of all 274,000 students had a focused response score of 2 on a scale of 1-4 and 60% had a score of 4 on the extended response scale of 2 to 8.

While the results of this study are of interest for the schools that took part electronically, caution must be taken in applying these findings to other schools. Within this relatively small set of schools (13 for Numeracy and 15 for Reading), over 80% of all students came from only seven schools: one school had nearly 1/3rd of the sample. The results from this study are indicative only.

Recommendations

For schools considering electronic administration of FSA, there are some strategies for student preparation that should be considered. In the case of Numeracy, students should have practice copying and checking required information from the screen. Computer stations need to provide adequate space for the student to do calculations and to copy information from screen to paper. While this may seem somewhat arbitrary and specific to the assessment, extracting correct information from a computer interface may be an important feature for students when using the Internet for any kind of numerical research. Teachers should become familiar with the navigation needs for each component and work with their students on the electronic samples supplied by the Ministry.

It may be helpful if the Ministry prepared more detailed instructions to assist teachers and students in working with the particular navigational features used in the assessment. In addition, the interface design should strike a balance between flexibility and simplicity. For example, is it necessary for the user to be able to resize the side-by-side panes for the Reading component? The Writing focused response part had a "Connections" section that required the student to respond to questions on two passages. In this component, when the passage was opened the *<<Back* and *Next>>* buttons were hidden and the lower left scroll bar had to be used to get to these buttons. Once the student got onto the page where the constructed-response was entered, the student could not move from one passage to the other without going back to a previous screen where the passage hyperlinks are displayed. Re-design of this section should be considered.

FURTHER RESEARCH

As noted earlier in the presentation of the results, the gender difference in NMC is reversed in NCR and favours females. One aspect that should be considered is whether or not there is any effect in marking due to females tending to be neater.

There are some additional analyses that could be considered but which would not be supported by the current sample size. Among the Numeracy multiple-choice items, there were several that did not require any secondary computation. That is, they did not require the transcription of data from the screen to paper to carry out the required mathematical operations. If there were no mode differences or if the difference was smaller for these items, that result would support the contention that the difference in Numeracy is a transcription problem. In turn, that would suggest that the impact would be greater for the less able students.

Among the Reading passages, the poem was the only one completely displayed on the screen and which required no scrolling. An analysis comparing the achievement on the poem and on the other passages would be of interest. If the mode difference did not appear for this passage, it would lend further support for an artifice introduced by scrolling.

As noted in the summary, the schools represented in this study are not representative of the schools in the province. There is a need to have a greater number of schools and students in any future study. As electronic assessment is used by more schools in the future, additional analyses of this nature should be conducted.

The impact of computer access and familiarity should be part of the study. In this way, the differences observed may be traced to the root cause and recommendations made for the mitigation of the impacts. Perhaps this would shed some light on the strange result of the Numeracy constructed response.

In addition to student questionnaires, an important feature of further studies would be some school-level questionnaires to determine the extent of preparation for the electronic assessment.

The effect of marking computer-based material should be researched more completely. If there is no demonstrable effect, then the focus for effect mitigation can be more appropriately directed. If there are some differential marking effects, then strategies to minimize those effects can be considered.

REFERENCES

Agency Research Consultants, (2005). 2004/2005 *Fundamental Skills Assessment: A comparison of paper and electronic results*. Report made available by British Columbia Ministry of Education, Victoria, BC, Canada.

Baghi, H., Gabryo, R., & Ferrara, S. (1991, April). *Applications of computer adaptive testing in Maryland*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Bergrstrom, B. A. (1992, April). *Ability measure equivalence of computer adaptive and pencil and paper tests: a research synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology 33*(5), 593-602.

Mazzeo, J., Druesne, B., Raffeld, P. C., Checketts, K. T., & Muhlstein, A. (1992). *Comparability of computer and paper-and-pencil scores for two CLEP General Examinations*. (Educational Testing Service Research Report 92-14). Abstract retrieved April 8, 2006, from http://www.ets.org/research/researcher/RR-92-14.html

English, R. A., Reckase, M. D., & Patience, W. M. (1977). Comparisons of paper application of tailored testing to achievement measurement. *Behaviour Research Methods and Instrumentation, 9*(2), 158-161.

Gallagher, A., Bidgeman, B., & Cahalan, C. (2002). The effect of computer-based tests on racial-ethnic and gender groups. *Journal of Educational Measurement 39*(2), 133-147.

Goldberg, A. L. & Pedulla, J. J. (2002). Performance differences according to test mode and computer familiarity on a practice graduate record exam. *Educational and Psychological Measurement 62*(6), 1053-1067.

Hol, M. A., Vorst, H. C. M., & Mellenbergh, G. J. (2005). A randomized experiment to compare conventional, computerized, and computerized adaptive administration of ordinal polytomous attitude items. *Applied Psychological Measurement*, *29*(3), 159-183.

Kobrin, J. L. & Young, J. W. (2003). The cognitive equivalence of reading comprehension test items via computerized and paper-and-pencil administration. *Applied Measurement in Education 16*(2), 115-140.

Lunz, M. E. & Bergstrom, B. A. (1995, April). Equating computerized adaptive certification examinations. Board of Registry series of studies. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

Olson, J. B., Maynes, D. D., Slawson, D., and Ho, K. (1986, April). *Comparison and equating of*

paper-administered, computer-administered and computerized adaptive tests of achievement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Pommerich, M. (2002, April). *The effect of administration mode on test performance and score precision, and some factors contributing to mode differences.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: mode effects for passage based tests. *Journal of Technology, Learning and Assessment 2*(6). Available online at http://jtla.org

Poggio, D., Glasnapp, D. R., Yang, X., and Poggio, A. J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *Journal of technology, Learning, and Assessment, 3*(6). Available from http://www.jtla.org

Russell, M. (1999). Testing on Computers: A Follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7*(20). Available online at http://epaa.asu.edu/epaa/v7n20

Russell.M., & Haney, W. (1997). Testing writing on computer: An experiment comparing student performance on tests computed via computer and via paper-and-pencil. *Education Policy Analysis Archives. 5*(3). Available online at http://epaa.asu.edu./eppaa/v5n3.html

Schwartz, R., Rich, C., & Podrabsky, T. (2003, April) *A DIF analysis of item-level mode effects for computerized and paper-and-pencil tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Sun, A. & McClanahan, R. (2003). *Is newer better: a comparison of web and paper-and-pencil survey administration modes.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

APPENDIX

Statistical Results From the Three- and Four-Factor ANOVAs

Table 12 Results of Three-Factor ANOVA on Numeracy Constructed-Response (NCR) Scores by Mode, Gender, and School

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	3324580.105(b)	51	65187.845	7.298	0	0.095	372.212	1
Intercept	207638615.4	1	207638615.4	23246.69	0	0.867	23246.69	1
Gender	68626.71	1	68626.71	7.683	0.006	0.002	7.683	0.791
NCrMode	213125.838	1	213125.838	23.861	0	0.007	23.861	0.998
School	1395097.612	12	116258.134	13.016	0	0.042	156.192	1
Gender * NCrMode	3508.145	1	3508.145	0.393	0.531	0	0.393	0.096
Gender * School	122616.319	12	10218.027	1.144	0.319	0.004	13.728	0.674
CrMode * School	490316.865	12	40859.739	4.575	0	0.015	54.895	1
School	59653.514	12	4971.126	0.557	0.878	0.002	6.679	0.331
Error	31735270.27	3553	8931.965					
Total	902676837.6	3605						
Corrected Total	35059850.37	3604						
a Computed using alpha = .05								
b R Squared = .095 (Adjusted R Squared = .082)								

Table 13 Results of Three-Factor ANOVA on Numeracy Multiple-Choice (NMC) Scores by Mode, Gender, and School

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	5409521.099(b)	51	106069.04	13.516	0	0.162	689.294	1
Intercept	183017090	1	183017090	23320.47	0	0.868	23320.467	1
Gender	567.338	1	567.338	0.072	0.788	0	0.072	0.058
NMCMode	127443.01	1	127443.01	16.239	0	0.005	16.239	0.981
School	3062924.2	12	255243.69	32.524	0	0.099	390.285	1
Gender * NMCMode	12552.551	1	12552.551	1.599	0.206	0	1.599	0.244
Gender * School	135435.97	12	11286.33	1.438	0.141	0.005	17.258	0.796
NMCMode * School	703706.66	12	58642.222	7.472	0	0.025	89.668	1
School	55675.092	12	4639.591	0.591	0.851	0.002	7.094	0.353
Error	27883649	3553	7847.917					
Total	862351037	3605						
Corrected Total	33293170	3604						
a Computed using alpha = .05								
b R Squared = .162 (Adjusted R Squared = .150)								

Table 14 Results of Four-Factor ANOVA on Numeracy Multiple-Choice (NMC) Scores by Mode, Gender, School and Achievement Category (AchCat)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	8589893.862(b)	145	59240.647	12.019	0	0.473	1742.799	1
Intercept	97536806	1	97536806	19789.19	0	0.911	19789.189	1
Gender	8.308	1	8.308	0.002	0.967	0	0.002	0.05
NMCMode	127652.26	1	127652.26	25.899	0	0.013	25.899	0.999
School	513539.34	12	42794.945	8.683	0	0.051	104.192	1
AchCat	1300532.1	2	650266.04	131.932	0	0.12	263.864	1
Gender * NMCMode	3458.38	1	3458.38	0.702	0.402	0	0.702	0.133
Gender * School	82050.732	12	6837.561	1.387	0.165	0.008	16.647	0.777
Gender * AchCat	53555.728	2	26777.864	5.433	0.004	0.006	10.866	0.847
NMCMode * School	374593.64	12	31216.137	6.333	0	0.038	76.001	1
NMCMode * AchCat	48453.825	2	24226.912	4.915	0.007	0.005	9.831	0.808
School * AchCat	190021.26	24	7917.552	1.606	0.032	0.019	38.553	0.979
Gender * NMCMode *	112861.16	11	10260.105	2.082	0.019	0.012	22.898	0.924
Gender * NMCMode *	18769.577	2	9384.789	1.904	0.149	0.002	3.808	0.397
Gender * School *	172696.05	24	7195.669	1.46	0.07	0.018	35.038	0.963
NMCMode * School *	110283.11	23	4794.918	0.973	0.498	0.011	22.375	0.801
Gender * NMCMode *	106642.17	16	6665.136	1.352	0.157	0.011	21.637	0.853
Error	9576644	1943	4928.793					
Total	485779968	2089						
Corrected Total	18166538	2088						
a Computed using alpha = .05								
b R Squared = .473 (Adjusted R Squared = .434)								

Table 15 Results of Three-Factor ANOVA on Reading Multiple-Choice Scores by Mode, Gender, and School

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	4890239.277(b)	58	84314.47	9.016	0	0.127	522.945	1
Intercept	139319178	1	139319178	14898.3	0	0.805	14898.301	1
Gender	22876.056	1	22876.056	2.446	0.118	0.001	2.446	0.346
MCMode	63542.796	1	63542.796	6.795	0.009	0.002	6.795	0.741
School	3329785.3	14	237841.81	25.434	0	0.09	356.075	1
Gender * MCMode	617.329	1	617.329	0.066	0.797	0	0.066	0.058
Gender * School	106661.88	14	7618.706	0.815	0.654	0.003	11.406	0.539
MCMode * School	81589.481	14	5827.82	0.623	0.848	0.002	8.725	0.409
Gender * MCMode * School	100810.46	13	7754.651	0.829	0.629	0.003	10.78	0.525
Error	33739659	3608	9351.347					
Total	899849109	3667						
Corrected Total	38629898	3666						
a Computed using alpha = .05								
b R Squared = .127 (Adjusted R Squared = .113)								

Table 16 Results of Four-Factor ANOVA on Reading Multiple-Choice (RMC) Scores by Mode, Gender, School and Achievement Category (AchCat)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	9043418.847(b)	147	61519.856	11.346	0	0.537	1667.796	1
Intercept	78396724.33	1	78396724.33	14458	0	0.91	14458.003	1
Gender	5257.726	1	5257.726	0.97	0.325	0.001	0.97	0.166
RMCMode	52605.106	1	52605.106	9.701	0.002	0.007	9.701	0.875
MinCode2	270579.44	14	19327.103	3.564	0	0.034	49.901	0.999
Gr4Cat	1927937.821	2	963968.91	177.776	0	0.198	355.552	1
Gender * RMCMode	3870.433	1	3870.433	0.714	0.398	0	0.714	0.135
Gender * MinCode2	58913.842	14	4208.132	0.776	0.696	0.008	10.865	0.511
Gender * Gr4Cat	28612.471	2	14306.235	2.638	0.072	0.004	5.277	0.526
RMCMode * MinCode2	145257.339	12	12104.778	2.232	0.009	0.018	26.789	0.956
RMCMode * Gr4Cat	17492.692	2	8746.346	1.613	0.2	0.002	3.226	0.342
MinCode2 * Gr4Cat	188356.626	25	7534.265	1.389	0.096	0.024	34.737	0.957
Gender * RMCMode * MinCode2	43764.539	11	3978.594	0.734	0.707	0.006	8.071	0.418
Gender * RMCMode * Gr4Cat	1833.371	2	916.685	0.169	0.844	0	0.338	0.076
Gender * MinCode2 * Gr4Cat	215273.417	24	8969.726	1.654	0.024	0.027	39.701	0.982
RMCMode * MinCode2 * Gr4Cat	105642.459	21	5030.593	0.928	0.554	0.013	19.483	0.744
Gender * RMCMode * MinCode2 * Gr4Cat	94624.026	13	7278.771	1.342	0.181	0.012	17.451	0.785
Error	7791953.966	1437	5422.376					
Total	385401936.9	1585						
Corrected Total	16835372.81	1584						

a Computed using alpha = .05

b R Squared = .537 (Adjusted R Squared = .490)

Table 17 Results of Three-Factor ANOVA on Reading Constructed-Response (RCR) Scores by Mode, Gender, and School

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	6840415.195(b)	58	117938.193	13.01	0	0.173	754.575	1
Intercept	150787102	1	150787102	16633.52	0	0.822	16633.523	1
Gender	399351.072	1	399351.072	44.053	0	0.012	44.053	1
CrMode	5102.567	1	5102.567	0.563	0.453	0	0.563	0.117
School	3447915.005	14	246279.643	27.167	0	0.095	380.344	1
Gender * RCrMode	27991.801	1	27991.801	3.088	0.079	0.001	3.088	0.42
Gender * School	229145.512	14	16367.537	1.806	0.032	0.007	25.277	0.93
CrMode * School	482609.251	14	34472.089	3.803	0	0.015	53.237	1
Gender * RCrMode * School	112180.906	13	8629.3	0.952	0.497	0.003	12.375	0.599
Error	32707434.23	3608	9065.253					
Total	955170404.8	3667						
Corrected Total	39547849.42	3666						
a Computed using alpha = .05								
b R Squared = .173 (Adjusted R Squared = .160)								

Table 18 Results of Four-Way ANOVA on Reading Constructed Response Scores by Mode, Gender,School and Achievement Category (AchCat)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	7454557.991(b)	147	50711.279	7.386	0	0.43	1085.771	1
Intercept	86078450	1	86078450	12537.49	0	0.897	12537.49	1
Gender	224501.59	1	224501.59	32.699	0	0.022	32.699	1
CrMode	18399.633	1	18399.633	2.68	0.102	0.002	2.68	0.373
School	1047894.4	14	74849.6	10.902	0	0.096	152.628	1
AchCat	782688.65	2	391344.32	57	0	0.074	114	1
Gender * CrMode	5137.575	1	5137.575	0.748	0.387	0.001	0.748	0.139
Gender * School	157303.98	14	11235.999	1.637	0.063	0.016	22.912	0.895
Gender * AchCat	35790.037	2	17895.019	2.606	0.074	0.004	5.213	0.521
CrMode * School	485790.07	12	40482.505	5.896	0	0.047	70.756	1
CrMode * AchCat	681.537	2	340.769	0.05	0.952	0	0.099	0.058
School * AchCat	162275.18	25	6491.007	0.945	0.541	0.016	23.636	0.811
Gender * CrMode * School	58691.46	11	5335.587	0.777	0.663	0.006	8.549	0.444
Gender * CrMode * AchCat	8473.671	2	4236.835	0.617	0.54	0.001	1.234	0.154
Gender * School * AchCat	184945.05	24	7706.044	1.122	0.31	0.018	26.938	0.88
CrMode * School * AchCat	147782.28	21	7037.251	1.025	0.428	0.015	21.525	0.798
Gender * CrMode * School * AchCat	74098.854	13	5699.912	0.83	0.628	0.007	10.793	0.523
Error	9865988.5	1437	6865.684					
Total	431022335	1585						
Corrected Total	17320546	1584						
a Computed using alpha = .05								
b R Squared = .430 (Adjusted R Squared = .372)								

Table 19 Results of Three-Factor ANOVA on Writing Focused Response (WFR) Scores by Mode, Gender, and School

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	758747.000(b)	58	13081.845	1.28	0.076	0.02	74.242	0.999
Intercept	134480464.7	1	134480464.7	13159	0	0.786	13158.692	1
Gender	8015.637	1	8015.637	0.784	0.376	0	0.784	0.143
WFRMode	29148.137	1	29148.137	2.852	0.091	0.001	2.852	0.393
School	193930.802	14	13852.2	1.355	0.167	0.005	18.976	0.815
Gender * WFRMode	28552.198	1	28552.198	2.794	0.095	0.001	2.794	0.386
Gender * School	98517.654	14	7036.975	0.689	0.788	0.003	9.64	0.454
WFRMode * School	228665.717	14	16333.266	1.598	0.072	0.006	22.375	0.888
Gender * WFRMode * School	255266.419	13	19635.878	1.921	0.024	0.007	24.977	0.933
Error	36638326.43	3585	10219.896					
Total	947290885	3644						
Corrected Total	37397073.43	3643						
a Computed using alpha = .05								
b R Squared = .020 (Adjusted R Squared = .004)								

Table 20 Results of Three-Factor ANOVA on Writing Extended Responses (WER) Scores by Mode, Gender, and School

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power(a)
Corrected Model	643284.702(b)	58	11091.116	1.083	0.311	0.017	62.817	0.994
Intercept	133383298.5	1	133383298.5	13025	0	0.784	13024.855	1
GENDER	19681.533	1	19681.533	1.922	0.166	0.001	1.922	0.283
EXMode	28933.834	1	28933.834	2.825	0.093	0.001	2.825	0.39
School	97323.572	14	6951.684	0.679	0.797	0.003	9.504	0.447
GENDER * EXMode	19298.52	1	19298.52	1.884	0.17	0.001	1.884	0.279
GENDER * School	119252.578	14	8518.041	0.832	0.635	0.003	11.645	0.55
EXMode * School	165261.04	14	11804.36	1.153	0.306	0.004	16.138	0.731
GENDER * EXMode * School	265269.217	13	20405.324	1.993	0.018	0.007	25.903	0.943
Error	36712817.97	3585	10240.674					
Total	939574022.7	3644						
Corrected Total	37356102.67	3643						
a Computed using alpha = .05								
b R Squared = .017 (Adjusted R Squared = .001)								

Comparisons Between Paper- and Computer-Based Tests

Foundation Skills Assessment - 2001 to 2006 Data

Jim Gaskill & Mike Marshall

Since 2004, the use of electronic administration of the Grade 7 Foundation Skills Assessment (FSA) in British Columbia has grown. Only one school was offering electronic exams in 2004 compared to 28 schools two years later. This report is based on 15 schools that administered the FSA electronically in at least one year during that period and had a fair proportion of students involved.

Because the electronic interface has different characteristics for each of the FSA components and because males and females may approach computer applications differently, separate analyses with gender as a factor were carried out for Numeracy Multiple-Choice, Reading Multiple-Choice, Reading constructed-response, Writing Focused Response and Writing Extended Response. Additionally, students were placed in three achievement categories based on their results on the Grade 4 FSA and the analyses were carried out using Achievement Category as a factor.

Student performance appeared to be impacted by certain features unique to the electronic and paper formats resulting in some advantages and some disadvantages for students.

The detailed statistical analyses provided in this report are followed by discussion and recommendations for future research on electronic FSA examinations.

This research was commissioned by the Technology Assisted Student Assessment Institute (TASA), an assessment research arm of the Society for the Advancement of Excellence in Education (SAEE). The research was supported through funding from the Max Bell Foundation.

Society for the Advancement of Excellence in Education
225-1889 Springfield Road
Kelowna BC V1Y 5V5
info@saee.ca www.saee.ca