University of Alberta

Comparing Performance Across Test Administration Modes In a Large-Scale Testing Environment

by

Deanna Lynn Shostak

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of requirements for degree of

Master of Education

in

Measurement, Evaluation and Cognition

Department of Educational Psychology

Spring 2014
Edmonton, Alberta

COMPARING PERFORMANCE ACROSS TEST ADMINISTRATION MODES IN A LARGE-SCALE TESTING ENVIRONMENT

Abstract

The purpose of this study was to complete a secondary analysis of the data collected for a Mathematics 30-2 field test to examine the comparability of the psychometric properties and the students' scores for a computer-based and paper-and-pencil form of the same test. A computer familiarity survey was used to explain any possible differences between the two samples. A total of 252 students responded to the paper-and-pencil field test and 378 students responded to the computer-based field test. At the item level, only one item (numerical-response item 3) had a statistical difference in difficulty with a moderate effect size. At the test level, the effect sizes for the main effects and interactions that were significant (response mode-by-test mode preference, response mode, test mode preference) were all small. There were very few significant differences in the responses to the computer familiarity survey. Implications for practice and recommendations for research are provided.

Key words: comparability studies, computer-based tests, equivalency

Acknowledgements

There are many people I would like to acknowledge for their help during my masters work. First and foremost I must thank Dr. W. Todd Rogers for his sage advice, wisdom, insightful feedback, and never-ending support and encouragement. He is one of the most passionate teachers and mentors I have ever met, and I could not have completed this work without him.

Thank you to the members of my defense committee who generously gave their time and expertise to review my work and provide thoughtful input: Dr. George Buck and Dr. Florence Glanfield. I am also very grateful to Dr. Cheryl Poth for her support and encouragement along the way.

This thesis would not have been possible without the support of my friends and colleagues at the Assessment Sector, Alberta Education. Although there are too many to name, I sincerely thank them for their help and support.

To thank all of those people who have provided moral support and encouragement as I completed this degree would be an impossible task. To friends and family who had unwavering faith in me—thank you from the bottom of my heart. And last but not least, I would like to thank my husband and best friend, Peter, for not only providing moral support and encouragement, but most of all for never losing sight of what really matters.

Table of Contents

List of Tables

Chapter 1

Introduction

Two examinees of equal ability are writing a large-scale test. One is writing the test in a paper-and-pencil booklet while the other is writing the same test in a digital format. While the questions themselves are the same, are the scores of the two students the same?

Background

Many large-scale testing programs have either implemented computer-based tests or are exploring the possibility of implementing them. While the onset of computerized testing began in the 1970's (Drasgow, 2002), the use of computerized testing has increased markedly since the implementation of *No Child Left Behind* in the United States as states looked at new ways of measuring student performance more efficiently (Kim & Huynh, 2007; Way, Lin, & Kong, 2008). In the fall of 2010, the U.S. Department of Education announced that the SMARTER Balanced Assessment Consortium (SBAC) and the Partnership for the Assessment of Readiness for College and Careers (PARCC) were awarded Race to the Top grants to develop new digital assessments for the Common Core Standards (U.S. Department of Education, 2010). In Canada, provincial education departments are also looking at transitioning to computerized assessments. In particular, Alberta plans to begin piloting a new system for the digitization of diploma examinations in 2014/2015, with all diploma examination sessions being offered electronically by the fall of 2017 (Alberta Education, 2013a).

As more paper tests were converted to computerized versions, various professional testing organizations began publishing standards and guidelines for computer-based tests so that sound practices could be established. In 1986, the American Psychological Association (APA) published *The Guidelines for Computer-Based Tests* and in 1999 the *Standards for Educational*

and Psychological Testing were jointly published by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council of Measurement in Education (NCME). Guidelines for computer-based tests were also established by The International Test Commission (2005). All of these publications clearly address the importance of comparability between scores on a paper-and-pencil test and scores on a computer-based version of the same test; test delivery mode should not affect examinee performance.

Computerizing a paper-based test does not necessarily result in an equivalent test, so scores from the two versions may not be comparable (Arce-Ferrer & Guzman, 2009; Mazzeo & Harvey, 1988; Mead & Drasgow, 1993; Paek, 2005; Pommerich, 2004; Wallace & Clariana, 2005). Performance may be affected by computer experience, gender, and age (Collerton et al., 2007; Grignon et al., 2009; Parshall & Kromrey, 1993). Test-taking strategies may differ across modes (Murphy, Long, Holleran, & Esterly, 2000) or a different "response action" may be required to answer the item on the paper-and-pencil version than on the computer-based version (Sireci & Zenisky, 2006). Despite these findings, many states and provinces have found that not all schools have the necessary infrastructure and equipment to offer online testing to all students, so paper and online versions of the same test are typically offered side-by-side (Way, Lin, & Kong, 2008).

Potential Benefits of Computer-based Testing

The potential benefits of computer-based tests include cost savings related to printing and shipping the tests to schools, quicker turnaround of results, more flexibility for administration, and enhanced security (Drasgow, 2002; Pomplun, Frey, & Becker, 2002). Cost-benefit analyses have found that computer delivery is less costly and more efficient than paper-and-pencil

delivery (Zenisky & Sireci, 2002). Item and test development may be easier and more efficient (Bejar et al., 2003), and scoring constructed responses may be done more efficiently using distributive scoring (Way, Vickers & Nichols, 2008; Zhang, Powers, Wright, & Morgan, 2003) or computerized scoring (Williamson, Mislevy, & Bejar, 2006). The possible inclusion of innovative item formats made possible in a computer environment may more fully measure a construct (Way, Davis, & Fitzpatrick, 2006), which may improve reliability and enhance the validity of large-scale assessments (Zenisky & Sireci, 2002). For example, dynamic stimuli, such as audio, may more readily be incorporated into a computer-based test than into a paper-and-pencil test, thereby allowing for expanded measurement of constructs that are best presented auditorily (Bennett et al., 1999). Likewise, students' cognitive, psychomotor, and affective characteristics may be detected and recorded on a computer-based test but not as easily on a paper-and-pencil test (Csápo, Ainley, Bennett, Latour, & Law, 2012). Additionally, surveys have empirically shown that many students enjoy online testing, feel comfortable in a computer-based testing environment, and tend to prefer computer-based tests over paper-and-pencil based tests (Glassnapp et al., 2005; Way et al., 2006).

Issues Associated with Computer-Based Tests

Although, as mentioned in the preceding paragraph, there are benefits of computer-based testing, there are still issues that must be addressed and resolved before moving to only computer-based testing. Basic technological applications are available, but their effective application into everyday educational practice must be closely examined so that their features are consistent across applications, educationally optimized, and systematically introduced (Csápo, Ainley, Bennett, Latour, & Law, 2012). Innovative item formats have the potential to change the nature of the construct being measured, which affects comparability when measuring change

across years (Rowan, 2010), and further research must be done to determine how data collection with these new instruments affects reliability and validity (Csápo et al., 2012).

Although many believe that computer-based tests are less expensive and more secure than paper-and-pencil tests (Arce-Ferrer & Guzman, 2009; Drasgow, 2002; Higgins, Russell, & Hoffman, 2005; Pomplun & Custer, 2005; Pomplun et al., 2002; Wise & Plake, 1990; Zenisky & Sireci, 2002), there are in fact additional costs associated with security, test development, and maintenance costs (Foster, 2004; Maynes, 2009). Further, as on-demand computer-based tests gain in popularity, there is greater likelihood that test items may be stolen (Maynes, 2009).

Computer familiarity may influence student performance (Collerton et al., 2007; Hargreaves, Shorrocks-Taylor, Swinnerton, Tait, & Threlfall, 2004; Russell, Goldberg & O'Connor, 2003). The National Assessment of Educational Progress (NAEP) Math Online (MOL) study, which was the most comprehensive study on the role computer familiarity plays on student performance, revealed that while the number of students lacking computer familiarity was insignificant, familiarity with computers did affect performance on an eighth grade mathematics test (Bennett et al., 2008; Sandene, Bennett, Braswell, & Oranje, 2005). If computer familiarity is required to complete a computer-based test, but is not part of the construct being measured, it is recommended that testing programs develop online help, instructions, and tutorials so that construct irrelevant variance due to the lack of computer familiarity is not introduced (Parshall, Spray, Kalohn, & Davey, 2002).

The Alberta Context

Currently, the Assessment Sector, Alberta Education, develops and administers provincial achievement tests at grades 3, 6, and 9, and diploma examinations at the end of most

grade 12 courses. The majority of these provincial tests and examinations are currently offered in paper-and-pencil format.

In preparation for computerized testing, Alberta Education, in partnership with Respondus, Inc., developed Quest A+ as a production pilot to inform the design of future digital provincial assessment systems in Alberta. Since 2008, Quest A+ has been used on a limited basis to administer some machine-scored and written-response provincial achievement tests (PATs) as well as some Humanities written-response diploma examinations (Dan Karas, Exam Administration Director, Assessment Sector, Alberta Education, personal communication, November 27, 2012). In January and June 2013, over 5,400 secure provincial assessments—PATs and Humanities written-response diploma examinations—were written on Quest A+ (Alejandro Moreno, Systems Analyst, Assessment Sector, Alberta Education, personal communication, August 12, 2013). No machine-scored diploma examinations are currently offered on Quest A+, so all mathematics and science diploma examinations are only available in paper-and-pencil format.

In addition to providing a secure testing environment, Quest A+ also contains practice tests for most PATs and diploma examinations. Since January 2009, over 1.6 million practice tests have been requested via Quest A+ (Alejandro Moreno, Systems Analyst, Assessment Sector, Alberta Education, personal communication, personal communication, January 13, 2014).

Alberta Education is currently undertaking a *Curriculum Redesign* project that is aimed at ensuring the province's curriculum, which includes programs of study, assessment, and learning and teaching resources, remains responsive and relevant to students. This project includes plans to administer the diploma examinations electronically by the fall of 2017 (Alberta Education,

2013a). In preparation for this, all mathematics and science field tests will be administered digitally on Quest A+, starting in the fall of 2013 (Alberta Education, 2013b; Alberta Education, 2013c).

Prior to Quest A+, the Assessment Sector, Alberta Education, piloted an online environment for item development and administered an online field test for Applied Mathematics 30 and Pure Mathematics 30, both of which were diploma examination subjects, in 2004. Student experiences completing the field tests were examined, but no formal comparison of student performance between paper-and-pencil and computer-based versions of the field test for each subject was done. In 2007, a study to examine performance differences between a field test offered in paper-and-pencil format and an online format was conducted for Science 30, which is a diploma examination subject (Alberta Education, 2007). Classes were randomly split into two groups—one group of students in each class wrote the paper version and the other group wrote the online version. A total of 233 students participated, with 107 responding to the paper version and 126 responding to the online version. The difference in the mean performance for the total test was significantly higher ($p < 0.05$) for the paper-and-pencil version than for the online version. At the subtest level, there were no performance differences between the two modes on the machine-scored portion of the test, but there was a statistical difference on the written-response question ($p < 0.05$), favoring the paper-and-pencil version. At the item level, two multiple-choice questions were significantly more difficult on the online field test ($p < 0.05$). This study also considered the efficiencies, benefits, risks, and challenges of administering a field test online, and the authors of the study, the System Improvement Group of the Accountability and Reporting Division of Alberta Education, made recommendations about the feasibility of moving to an online environment for diploma examinations. The recommendations

were to continue field testing Science 30 multiple-choice, numerical-response, and written-response questions that did not require students to graph, draw, or provide tables online. However, once these features are available for students to use on the computer, another comparability study that compares student performance on written response questions should be conducted (Alberta Education, 2007). To date, this Science 30 study is the only published comparability study from Alberta Education.

Need for the Present Study

If officials of a testing program want to treat scores across different test delivery modes as being equivalent, then studies of the comparability of the test scores must first be undertaken (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; International Test Commission, 2005; Parshall, Spray, Kalohn, & Davey, 2002). One of the purposes of the Alberta diploma examination program is to certify student achievement, and the scores from these examinations count for 50% of students' final course marks for each subject area tested (Alberta Education, 2013b). Even though an overall mode effect may be small, a small effect can have significant consequences for individual examinees (Texas Education Agency, 2008).

A comparability study analyzes the effects of test delivery mode on student performance and can help address these issues. Mode effects cannot be predicted completely from previous research that focused on different tests than those to be used in the present situation or that focused on the same test that was administered to different students than the students in the present situation (Krőhne & Martens, 2011; Poggio, Glasnapp, Yang & Poggio, 2005). Furthermore, computerized assessment adds unknown elements to a complex assessment process, so a comparability study may help students, parents, teachers, and stakeholders accept

the results produced by a computer-based test, providing there is no difference between the scores on the paper-and-pencil test and the computer version (Csapó, Molnár, & Tóth, 2009). Therefore, if a high-stakes testing program is considering a transition from paper-and-pencil assessments to computer-based assessments, comparability studies are a necessity for each assessment. Diploma examinations in Alberta are high stakes.

Currently, machine-scored diploma examinations in Alberta are administered in a paper-and-pencil format. The majority of field testing is also paper-and-pencil, although all mathematics and science field tests will be offered digitally, starting in the fall of 2013 in preparation for the full implementation of computer-based diploma examinations in 2017 (Alberta Education, 2013a). Further, Alberta has a goal of making diploma examinations directly comparable from one administration to the next by using anchor items to equate the examinations (Alberta Education, 2013c). Therefore, comparability between paper-and-pencil and computer-based modes is crucial as the transition is made to a digital testing environment. Students must not be disadvantaged in any way by the mode of administration, and the interpretations of the test scores must be valid and not mode dependent. If it is assumed that scores are comparable when in fact they are not, then wrong decisions may be made, which directly violates fairness principles (van Lent, 2008) as well as Alberta Education's goal of maintaining consistent standards over time (Alberta Education, 2013c).

To date, the only published comparability study from Alberta Education has been for Science 30; no formal comparability studies have been done for Mathematics 30-2, which is part of the Alberta diploma examination program. Since "it is the responsibility of the test developers to show that a computer-based test and a paper-and-pencil test are equivalent" (Bugbee, 1996, p.

292), Alberta Education produced a Mathematics 30-2 field test that was offered in both paper-and-pencil form and in digital form in June 2013.

Purpose of the Study and Research Questions

Consequently, the purpose of the present study was to complete a secondary analysis of the data collected for the Mathematics 30-2 field test to examine the comparability of the psychometric properties and the students' scores for the computer-based and paper-and-pencil forms of the field test. The specific research questions addressed included:

1. Are the psychometric properties of the items comparable between the paper-and-pencil field test and the computer-based field test?
2. Does the reliability of the paper-and-pencil field test differ from the reliability of the computer-based field test?
3. Do students' scores differ significantly between the paper-and-pencil field test and the computer-based field test with at least a moderate effect size?
4. Do male and female students' scores differ significantly with at least a moderate effect size between the paper-and-pencil field test and the computer-based field test?
5. To what degree does student performance on a digital Mathematics 30-2 field test depend upon computer familiarity and prior online testing experience?

Definition of Terms

In order to ensure complete understanding of the terminology used in this study, the following definitions are provided.

Comparability. Comparability exists when there is a commonality of score meaning across testing conditions (Drasgow, Luecht, & Bennett, 2006), and the inferences made from

scores on a paper-and-pencil test are the same as the inferences made from the scores on the online version of the same test (Strader, 2012).

Computer familiarity. Computer familiarity is defined as having computer experience in a mathematical context.

Field tests. Field tests are developed and administered by Alberta Education at the end of each semester in a school year. They contain test questions that are being "tested" to determine their difficulty level and their appropriateness for use on future diploma examinations. Classical item analysis, along with teacher and student comments, are used to select only those field test questions that are clear, fair, valid, and reliable for future diploma examinations (Alberta Education, 2013b).

Mode effects. The effects of test administration mode on student performance.

Overall performance. The percent correct on a Mathematics 30-2 machine-scored field test.

Response mode. In this study, the Mathematics 30-2 field test was administered in two modes—questions were administered in a paper-and pencil test booklet and the same questions were administered digitally using Quest A+.

The Researcher's Dual Roles

In addition to being a master's student in Educational Psychology at the University of Alberta, the researcher is the Mathematics 30-2 Team Leader within the Assessment Sector of Alberta Education. Permission was granted by the Executive Director, Assessment Sector, Alberta Education, for the researcher to undertake this study. The data was collected by Alberta Education as part of its field testing program. Therefore, the present study will involve secondary data analyses designed to (a) determine the comparability of the machine-scored items in the

paper-and-pencil form and the computer-based form, and (b) compare the performance of the group of students administered these two forms of the field test. Prior to receiving the data for the study, all information that would identify a specific student was removed.

Delimitations of Study

This study was carried out on one field test for Mathematics 30-2 as part of the introduction of computer-based testing in Alberta. In light of previous studies of comparability between paper-and-pencil and computer-based test forms, which showed that sometimes there were no differences and other times there were, the results are delimited to the one field-test for Mathematics 30-2. Innovative items that are not possible on a paper-and-pencil test but are possible on a computer-based test were also beyond the scope of the study, as were tests that were only offered on a computer. However, the process followed to conduct the study is generalizable to other forms and subject areas.

A second delimitation was that it was not feasible to interview students. Consequently, it is not known if the students who responded to the paper-and-pencil field test and the students who responded to the computer-based field test used similar cognitive strategies, problem solving skills, and reasoning skills when responding to the items.

Organization of the Thesis

This thesis is organized in five chapters. Chapter 1 provided a background to the potential benefits as well as the issues with computer-based testing, described the Alberta context, and introduced the purpose of the study. Chapter 2 contains the literature review of mathematics comparability studies, and also gives an overview of how computer technology has affected assessment, describes the standards for computerizing paper-and-pencil tests, and outlines the general principles for examining the comparability of paper-and-pencil and

computer-based tests. Chapter 3 outlines the methods and statistical procedures that were used to compare student performance on the paper-and-pencil and computer-based modes, and also provides the preliminary analysis to establish if previous online test experience influences performance on the computer-based test. Comparative item analyses across modes are provided in Chapter 4, and a comparison of the two samples' field test performance on the two modes along with the survey results are provided in Chapter 5. Lastly, Chapter 6 contains a summary of the results, followed by a presentation of limitations, conclusions drawn in light of the limitations, implications for practice, and recommendations for future research.

Chapter 2

Literature Review

Chapter 2 contains a review of the relevant topics related to the comparability of paper-and-pencil test scores and computer-based test scores and is organized into four sections. The first section includes a brief history of how computer technology has affected assessment and standards for computerizing paper-and-pencil tests. Next, the importance of comparability, as well as some general principles for examining the comparability of paper-and-pencil scores and computer-based scores, is discussed. The third section discusses the results of comparability studies that have been done in mathematics. The final section includes a summary of the research findings.

Computer Technology's Effect on Assessment

Personal computers did not exist prior to 1975 (Leeson, 2006). However, since then, the capabilities of personal computers and the percentage of people owning their own computer have increased exponentially (Lynch, 2000). The evolution of computer technology has also affected the field of assessment. For example, the use of high-speed scanners to score multiple-choice questions led to an increase in the use of multiple-choice items because of the quick turn-around (Parshall et al., 2002). Computerized testing began in the 1970's (Drasgow, 2002), and since then the increased affordability and computational ability of computers has led to a marked expansion of computerized testing. New item formats and item types are being developed (Zenisky & Sireci, 2002), including innovative, open-ended questions (Clariana & Wallace, 2002). Automatic item generation is being used for item development (Arendasy & Sommer, 2012; Gierl, Lai, & Turner, 2012). Computer scoring is more advanced (Clariana & Wallace, 2002), new scoring possibilities are being explored (Zenisky & Sireci, 2002), and in many cases

score reports are immediate (Kingston, 2009; Paek, 2005; Pomplun, Frey, & Becker, 2002). The use of fixed-form computer-based tests and computer-adaptive tests (CATs), where students respond to a subset of items dependent on their ability, is increasing with the more recent creation of the Smarter Balanced Assessment Consortium (SBAC) and Partnership of Readiness for College and Careers (PARCC) for the Common Core State Standards (CCSS) for mathematics and English language arts in the United States (Herman & Linn, 2013). These two consortia, which are composed of states, some of which belong to both consortia, are charged with developing comprehensive, technology-based assessment systems to measure students' performance on the CCSS. These assessment systems will serve both a formative and summative function, with the summative tasks being on demand and containing performance tasks for both mathematics and English language arts. SBAC will offer CATs for their summative assessments, whereas PARCC will offer fixed form computer-based assessments.

Standards for computerizing paper-and-pencil tests. Once computer-based tests started gaining in popularity, it became necessary to establish standards for developing and administering them to ensure that the scores from the computer-based tests could be validly interpreted in terms of what the student knew and could do. The American Psychological Association (APA) published *Guidelines for Computer-Based Tests and Interpretations* in 1986. According to the APA, the principles used for traditional paper-and-pencils should also be used for computer-based tests. While writing a computer-based test, examinees should be able to review their answers to items, answer items in any order, make changes to their responses, and receive timely feedback after they have completed the computer-based test. The guidelines also stated that a computer-based test should be comparable to a paper-and-pencil test in terms of means, standard deviations, and rankings. Further, "…when interpreting scores from the

computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cut scores obtained from conventional tests" (APA, 1986, p. 18). The *Standards for Educational and Psychological Testing* published by the American Educational Research Association (AERA), APA and the National Council of Measurement in Education (NCME; AERA et al., 1999) included for the first time standards that addressed the use of online versions of tests. For example, Standard 4.10 recommends that

> A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably….The specific rationale and the evidence required will depend in part on the intended use for which score equivalence is claimed. (AERA et al., 1999, p. 57)

Lastly, The International Test Commission (ITC) published the *International Guidelines on Computer-Based and Internet-Delivered Testing* in 2005 (ITC, 2005). Guideline 21c states "Where the computer-based test/Internet test has been developed from a paper-and-pencil version, ensure that there is evidence of equivalence" (ITC, 2005, p. 156). These documents clearly state that performance must not be affected by test delivery mode, and evidence of measurement equivalence must be reported.

In addition to the standards, researchers have called for the comparability of scores from the same versions of a paper-and-pencil and computer-based test. For example, Mead and Drasgow (1993) in 1993 indicated comparability across media is a prerequisite for equivalent construct validity across modes and that computerization should not affect the construct. To reinforce this need, Drasgow, Luecht, and Bennett (2006) stated in 2006 that comparability is necessary "when scores need to have common meaning with respect to one another, to some

reference group, or to a content standard" (p. 501). Thus, scores must be comparable when the same test is administered in a paper-and-pencil mode and in a computer-based mode at the same time.

Likewise, when measuring change from one year to the next, scores must be comparable if the first year's test form is administered in a paper-and-pencil format and the second year's test form is administered on a computer. If the scores are not comparable when a computer-based test and a paper-and-pencil test are offered simultaneously or when measuring change using different test modes, then decisions made based on one of these scores may be inaccurate, which directly violates fairness principles (van Lent, 2008). Equivalence evidence (or the lack of) must be reported so that test users can make informed decisions about how they are to interpret the scores from paper-and-pencil and online versions of the test (Pomplun & Custer, 2005; Russell et al., 2003).

Examining Comparability

Comparability can be examined on two levels—*score equivalence* and *construct equivalence* (Lottridge et al., 2008; Lottridge, Nicewander, & Mitzel, 2011). If the paper-and-pencil and computer-based tests produce similar score distributions, then there is score equivalence or, in other words, the *equal distribution property* has been met (Kolen, 2000). This means that there is no evidence to suggest that using a computer changed the construct being measured by the paper-and-pencil version (Lottridge et al., 2008). Although this satisfies Lord's (1980) definition of equity, which states that scale scores on Test 1 and Test 2 achieve equity if scores on the two tests have the same distribution for examinees of a given true level on the construct, the *Guidelines for Computer-Based Tests and Interpretations* (APA, 1986) also state that in order for scores to be equivalent across different modes, examinees must be rank ordered

in approximately the same way. Due to unreliability, the rank order of examinees' scores on any two tests of the same ability will usually differ; however, if the rankings are more dissimilar than would be expected based on the unreliability, then the two tests are not comparable (Mead & Drasgow, 1993). It is also important to remember that a high rank order correlation does not necessarily mean that the scores of the students across modes are the same. For example, two students may have the same rank, but the student who wrote the computer-based test received a score of 80% whereas the student who wrote the paper-and-pencil version received a score of 74%. These two scores may be viewed quite differently for post-secondary acceptance and scholarship opportunities. Ghiselli (1964; Ghiselli, Campbell & Zedeck, 1981) lists an additional requirement—two tests are parallel if they correlate to the exact same degree with the scores on another variable. Again though, this does not necessarily mean that the students' scores across modes of students of equal ability are the same. To confirm the scores across the two modes are the same, the mean absolute deviation or the root mean square difference should be used (Rogers, 2012).

The standards and guidelines identified above all outline methods for investigating comparability, but they do not specify criteria that indicate whether or not equivalence has been achieved (Lottridge et al., 2008). Comparisons of score distributions, comparisons of relationships with other criterion measures, and comparisons of relationships of scores across modes are called for, but judgment must still be used by the investigator when interpreting results.

Examinee characteristics and mode effects. Comparability studies analyze the effects of test delivery mode on student performance to determine if there is a *mode effect* (Paek, 2005). When exploring mode effects in a comparability study, there are 29 variables, human as well as

technological, that may contribute to differences between performance on a computer-based test and performance on a paper-and-pencil test (Muter, 1996). Mode effects may differ across different subject areas, item types, and examinee groups (Hamilton, Klein, & Lorié, 2000). These effects may be due to test questions, test conditions, scoring, or the examinees themselves (Kolen, 2000); and these factors may vary from one test administration to another (Texas Education Agency, 2008). The examinee characteristics that may affect comparability include age, race, gender, and computer familiarity (Collerton et al., 2007; Parshall & Kromrey, 1993; Wise & Plake, 1990). Therefore, in a comparability study, in addition to examining differences between item types, test conditions, and scoring, it is also important to examine differences between subgroups so that valid inferences can be made from examinees' scores (Rowan, 2010).

The examinee characteristics that were of interest in the present study are gender and computer familiarity, specifically as they relate to mathematics. If males and females of the same ability perform differently across modes, then a factor other than mathematics ability is being measured. However, only a few comparability studies for mathematics have examined gender differences, and the results have been mixed, as will be seen in the next section. Further, no comparability studies that examined gender have been conducted in a large-scale testing environment in which exit examinations are administered to high school students wishing to graduate. If computer familiarity differs, then construct irrelevant variance is a factor (Parshall et al., 2002). A majority of the research on the effect computer familiarity may have on performance has been done for writing. Several studies from the 1990's showed that students who had less experience with computers were disadvantaged on computer-based assessments for writing. Computer familiarity did play a role in writing performance on the 2002 Writing Online (WOL) study done by the National Assessment of Educational Progress (Horkay, Bennett, Allen

& Kaplan, 2005). However, other research has found that students performed better on computer-based assessments for writing regardless of their computer familiarity, which may be because students are more familiar with computers and there is increased motivation to do well on computer-based assessments (Hargreaves et al., 2004; Russell & Plati, 2002). Strader (2012) reported that lack of familiarity and interface issues did not affect the performance of Grade 5 or Grade 8 students on a high-stakes state science test, but recommended that future research should focus on familiarity and interface issues, studying all subgroups, grade levels, subjects, and interfaces in order to fully dismiss the notion that computer familiarity remains an issue. Very few studies have been done on the relationship between computer familiarity and performance on a computer-based test in mathematics.

Comparability studies on paper-and-pencil and computer-based tests have been conducted for more than 20 years, although they have been limited in Kindergarten to Grade 12, particularly in the context of high-stakes, large-scale assessments (Paek, 2005). The next section discusses the results of comparability studies that have been conducted in mathematics; a majority of these involved linear non-adaptive tests.

Mathematics Comparability Studies

Early research of mode effects on arithmetic reasoning tests administered in a paper-and-pencil format and also online produced mixed results, although many of the studies did not employ rigorous methodologies that would indicate true test equivalency (Russell et al., 2003). Initial meta-analysis of comparability studies for various subjects, including mathematics, also reported mixed results (Mazzeo & Harvey, 1988). Since many of the early studies investigated cases with discrete, single-screen items, Mead and Drasgow (1993) hypothesized that including graphical displays for items could introduce mode effects. However further studies have asserted

that primitive technology was to blame (Paek, 2005). Pommerich and Burden (2000) found that subtle item formatting differences could create mode effects and that examinee characteristics contributed to many of the observed mode effects.

Wang, Jiao, Young, Brooks, and Olson (2007) did a meta-analysis of comparability studies conducted between 1980 and 2005 for K-12 mathematics. To be selected for this meta-analysis, the minimum sample size had to be 25, the tests had to be in English, and the mean and standard deviation for both the paper-and-pencil test and the computer-based test had to be reported so that an effect size (ES) could be calculated. A repeated-measures design with either counter-balanced or random order was used in 25 of the 44 cases.

Despite the presence of 25 cases with repeated measures, Wang et al. (2007) estimated the ES for the repeated measures studies using means and standards deviations because the correlation was not reported in many of these studies. Thus, Wang et al. (2007) recommended that the overall results should be taken as the upper bound of the actual ES as they believed their estimation overestimated the ES. However, it is likely the correlation between the two modes would be positive, in which case the ES would be underestimated. Of the 44 ESs, 13 were significant ($p < 0.05$).

Wang et al. (2007) considered study design, grade level, sample size, type of test, computer delivery method, and computer practice as moderators in their analysis. The only moderator variable that was a significant predictor of ES was the computer delivery procedure (fixed versus adaptive). Wang et al. (2007) cautioned that sample size may influence the generalization of the results. However, Bennett et al. (2008) also pointed out that many of the studies used were unpublished, the samples were not necessarily representative, and all but 14 of the effects came from three investigations.

Kingston (2009) synthesized the results of 81 comparability studies for various subjects and grades performed between 1997 and 2007. Twenty-nine of the studies were for mathematics from grades 1 to 12. No sample sizes were reported for any of the studies. Meta-analysis was used to estimate overall ES attributable to administration mode, grade, and subject area. Since the variability among ES was greater than expected by chance, Kingston used the weighted and un-weighted random effects models, where a positive ES indicated that students performed better on the computer-based test. For just the mathematics studies, the mean of the un-weighted effect sizes was -0.05 and the weighted mean effect size was -0.06. While the mean effect size was small (Cohen, 1988), when it is applied uniformly across the underlying proficiency distribution to a group of 100 students, two to three additional students could be below proficient if every student used a computer-based test and the cut score was the median of the score distribution (Kingston, 2009). Also, the large range of effect sizes means that the absolute value of the effect size could be small one year, but then large the next year. Kingston did not disclose which formula for the pooled standard deviation was used when the same sample was administered both forms, so the ES for some of the studies in Kingston's study may be underestimated (Rogers, 2012).

One of the first quasi-controlled empirical investigations for comparability in a large-scale assessment program for mathematics was done in Kansas (Poggio, Glasnapp, & Yang, 2005). The purpose of the study was to see if it was necessary to offer a paper-and-pencil test when the move statewide to computer-based test was undertaken. On a voluntary basis, 48 schools agreed to allow testing of some or all of their Grade 7 students using a computer-based test as well as a paper-and-pencil test. Only two of these schools had previously done any formal online computerized testing. In total, 644 students were tested. Due to the voluntary

participation, a strictly randomized counterbalanced design was not possible. Schools could select which mode to administer first, with the majority choosing the computer-based test. While most schools administered parallel and equated forms, two schools administered identical test forms under the two modes. Although this was not the intended design and the order effect was not controlled for, these data allowed for the evaluation of the impact of repeating the same test under the two different modes. Preliminary analysis revealed that the sample of students was not an aberrant sample of students in Kansas.

The differences between the paper-and-pencil and the computer-based test at the test level were not statistically significant ($p > 0.01$), and the ESs were generally small ($\Delta = 0.20$; Cohen, 1988). This result was observed for all the student groups, regardless of which mode was administered first or whether they had the same tests or equivalent tests. The correlation between the total scores attained by students who took both the computer-based test and the paper-and-pencil was 0.96, so the majority of students maintained their rank position regardless of mode. No gender differences were observed, nor were there differences with respect to academic placement or socioeconomic status (SES). Analysis using item response theory (IRT) confirmed the results based on observed scores.

Nine of the 204 items were more difficult on the computer-based test than on the paper-and-pencil ($p < 0.01$). Poggio et al. (2005) hypothesized that the differences were likely due to the cognitive complexity of scrolling, although this was not confirmed. One of the few studies to examine differential item category functioning (DICF), no differences between the item choice distributions for the two modes were found. Based on the findings and given the assessment context, it was recommended that there was no need to simultaneously offer both modes since computer-based tests are a credible and comparable option to paper-and-pencils.

Kim and Huynh (2007) examined the comparability of scores for large-scale end-of-course examinations in algebra for 788 Grade 9 to Grade 12 students in a southeastern state. Scale scores, item parameter estimates, test characteristic curves, test information functions, Rasch ability estimates at the content domain level, and the equivalence of the construct were examined. A counter-balanced, repeated measures design was used to control for order effects. In order to control for motivation effects, student scores on the first test were not reported until the second test was completed; students were also told that they would be allowed to count the higher score in their final grade. The two alternate tests contained calibrated multiple-choice items and were pre-equated to be parallel forms.

Kim and Huynh (2007) reported that the results supported the comparability of the paper-and-pencil and the computer-based test at the item-level, sub-test level, and test-level. While there was a significant mode effect favoring the paper-and-pencil test ($p < 0.01$), the ES of 0.17 was small (Cohen, 1988). At the item level, overall differences in item parameter estimates and the average absolute difference were not statistically different. Similar patterns for the test characteristic curves and the test information functions supported the comparability of the two modes. However, Kim and Huynh (2007) pointed out that the convenience nature of the sampling, along with an underrepresentation of African-American and Hispanics, may limit the generalizability of their study. The researchers also noted that although the two alternate test forms were constructed to be parallel, the results of the study are only reasonable to the extent that the forms were properly equated.

When the Department of Education in Oregon initiated the Technology-Enhanced Student Assessment (TESA) for the state-wide assessment of mathematics in Grades 3, 4, 5, 6, 7, 8, and 10 in 2001, districts had a choice of using TESA or conventional paper-and-pencil

assessments. Several unpublished equating studies were conducted to determine whether the existing paper-and-pencil reporting scale could be used for the computer-based test. Although the studies suggested that scores were comparable when the administration mode changed, the decision was made to maintain separate item parameters for the test in the two different modes since slight mode effects were found at both the item and scale level. These studies also suggested that younger test takers who have less experience with computers may experience more difficulty on a computer-based test than a paper-and-pencil (Choi & Tinkler, 2002, as cited in Oregon Department of Education, 2007). No significance levels were reported.

Another comparability study involving the TESA was done in 2007 to ensure comparability of scores across the two modes and to generate "linking blocks" for future paper-and-pencil tests so that they could be equated to the TESA scale. A single group counterbalanced design (Kolen & Brennan, 2004) was used. The sample sizes for mathematics ranged from 156 students to 396 students across the grades. The correlation between TESA and paper-scale scores for mathematics ranged from 0.70 to 0.83, and it was reported that none of the mean scores differed significantly across the two modes, although no significance levels were reported. Rank orders may have been affected, particularly in the cases where the correlation was at the bottom of the range (Oregon Department of Education, 2007). No comparison at the item level was done in this study. Since 2007, paper-and-pencil tests have been used as an accommodation only.

Csapó, Molnár, and Tóth (2009) were one of the first to do a mathematics comparability study in Hungary. Their repeated-measures design study, with the paper-and-pencil administered first, was done in a low-stakes testing context, and achievement differences at the test, subtest, item, and subsample level across delivery modes were examined. Participants were

843 Grade 5 students involved in a longitudinal study. Inductive reasoning, which included number analogies, number series, and verbal analogies subtests, was assessed. Performances on the open-ended and multiple-choice items were separately compared, and gender differences were considered at the test and subtest level.

Although the correlation between the total scores of the paper-and-pencil and the computer-based test was 0.79, the mean scores of the two modes differed significantly ($p < 0.05$). On the subtests, students' achievement was statistically significantly higher ($p < 0.05$) on the paper-and-pencil than the computer-based test with one exception—verbal analogy, which consisted of only multiple-choice items. The greatest mode effect was on the open-ended questions from the number series unit.

Girls achieved significantly better ($p < 0.05$) on the paper-and-pencil than the computer-based test whereas the delivery mode had no impact on boys' achievement. At the subtest level, girls' achievement was significantly different ($p < 0.05$) between the two modes, with all but verbal analogy favoring the paper-and-pencil. Boys' achievement only differed significantly on number series, where the paper-and-pencil was favored, and verbal analogies, where the computer-based test was favored. No effect sizes were reported.

The Minnesota Department of Education has performed comparability studies for both the Mathematics Minnesota Comprehensive Assessment-Series III (MCA-III), which is taken by students from Grades 3 through 8 inclusive (Minnesota Department of Education, 2012), and the Graduation-Required Assessment for Diploma (GRAD) mathematics exam, which is taken by students in Grade 11 (Minnesota Department of Education, 2009). A matched samples comparability analysis (MSCA; Way et al., 2006) was used for the MCA-III. The matching variables were the student's mathematics scale score from the previous grade (with the exception

of Grade 3), their reading score from the current administration, and various demographic indicators, including free and reduced price lunch status (FRP) and ethnic group membership. However, no sample sizes were reported. Across Grades 3 through 7 inclusive for MCA-III, students who took the paper-and-pencil test scored significantly higher ($|z \text{ statistic}| \geq 2$) on the common items than those who took the computer-based test, although the effect sizes, $\Delta = -0.13$ to -0.02, were small (Cohen, 1988). The Department decided that a set of linking items not impacted by mode differences would be used to scale the paper mode-specific or unique items to the online scale to make the scores more comparable and to enhance fairness (Minnesota Department of Education, 2012).

The Minnesota GRAD mathematics comparability study used a randomized group design, with 1,036 students taking the online form of the test and 1,035 taking the paper-and-pencil form. During sampling, the subgroup of students for whom computer familiarity was most questionable—those who were eligible for FRP—was oversampled to increase the chances of detecting a mode effect for these students. The mean and proportion passing for this subgroup was higher on the paper-and-pencil form, and similar conclusions were drawn for the other subgroups in this study, including gender and school location. Rather than conducting a large number of independent statistical tests to determine whether these differences were significant, which inflates the probability of making a Type I error, a hierarchical modeling approach was taken whereby nested, linear mixed models with covariates were used. No evidence of a mode effect was found at the test level or subgroup level ($p > 0.10$ to $p > 0.50$, depending on the models being compared), and further analysis from a bootstrap IRT found that the scaling constants did not statistically differ from the identity scaling function. Thus, no statistical adjustment for scaling the GRAD Mathematics online test was made. However, the different

levels of significance that were used in this study means that the results should be interpreted with caution.

Johnson and Green (2006) did a comparability study in the United Kingdom with two sets of matched mathematics questions using a counter-balanced design and four experimental groups. *Facility values* **(**i.e., *p* values) were analyzed to explore the impact of mode on performance. In this study, 104 eleven-year old students who wrote in both modes were provided with "scratchwork space" so that the researchers could study the students' thinking processes and code their error types. Also, the provision of scratch work space enabled the researchers to further investigate Russell et al.'s (2003) conclusion that validity is threatened when students experience difficulty in accessing scratch paper to perform calculations. Affective responses of a sub-sample of students were also investigated.

Although no significance level was indicated, Johnson and Green (2006) reported that the empirical evidence indicated there was no significant order effect, and the differences between the means of the four groups were not statistically significant. Although the qualitative interview data revealed that five of the eight students interviewed actually thought that the computer-based test was easier, 11 of the 16 items were easier on the paper-and-pencil test than on the computer-based test, with three of these questions having a difference greater than the standard error. Johnson and Green (2006) stated that this reinforces the need for further investigation to explore how overall test level findings may mask individual question mode differences.

With respect to scratchwork space, Johnson and Green (2006) observed that there were some mode-related differences. For 9 of the 16 questions, more students showed their work on the scratchwork space for the computer-based test than for the paper-and-pencil test. However, for the three questions that were statistically easier on the paper-and-pencil test, computational

and mental calculation errors were more frequent on the computer-based test than on the paper-and-pencil test. Johnson and Green (2006) hypothesized that the increased errors on these questions on the computer-based test were perhaps due to a reliance on mental strategies. In contrast to Russell et al.'s (2003) findings, Johnson and Green (2006) suggested that difficulty accessing scratch paper was not a physical one but more likely a mental one.

For both modes, computation and mental calculation errors were the most frequent error types, and transcription errors were more frequent on the computer-based test than the paper-and-pencil test. It is interesting to note that for the students who showed work for both modes for at least one question, 47% changed their strategy according to mode. An important confounding issue with this study was the fact that the software used did not allow students to go back to earlier questions or view forthcoming questions until they had completed the current question (Johnson & Green, 2006), which directly violates one of the APA guidelines established in 1986. However, the researchers observed that while students were writing the paper-and-pencil test, they could review past questions to inform their strategies for new questions and also preview upcoming questions.

Overall, students left more questions unanswered on the paper-and-pencil test than on the computer-based test. A possible explanation for this is that the computer-based test environment may be less threatening (Gallagher, Bridgeman, & Cahalan, 2000) or that a "computer game schema" might influence students' perceptions of the true demand of a computer-based test (Johnson & Green, 2006). If this is true, it suggests that students may have a more positive attitude toward a computer-based test, which in turn may lead to greater motivation to complete a computer-based test versus a paper-and-pencil test. Wang, Young, and Brooks (2004) also reported that test takers' attitudes usually were more positive towards a computer-based test than

a paper-and-pencil test in their comparability study for the Stanford Diagnostic Reading and Mathematics Tests.

Threfall, Pool, Homer, and Swinnerton (2007) cloned 24 questions from the 2003 published key stage 2 and stage 3 mathematics assessments in England for a comparability study. Key stage 2 and stage 3 assessments are administered to 11 year olds and 14 year olds, respectively. A counter-balanced design with four test pairings was used, with approximately 400 students answering each question in each format. Threfall et al. (2007) reported that the overall performance of students at key stage 2 favored the computer-based test by 3%, whereas the overall performance of students at key stage 3 favored the paper-and-pencil test by 5%. Although the authors did not report whether or not these differences were statistically significant, they did state that "A difference of 5% or less in performance cannot be said to be indicative of an underlying effect" (Threfall et al., 2007, p. 340).

Seven items for which the *p* values on the two modes differed between 11.7% and 34.4% were further examined. Five of these items favored the computer-based test, four of which involved elements that needed to be arranged to find the solution. Threfall et al. (2007) proposed that the difference in performance on these items may be due to a "relative affordance" (Gaver, 1991) that was available on the computer-based test and not the paper-and-pencil test—the ability of the students to explore different arrangements interactively. The static paper-and-pencil environment meant that an increased cognitive load (Sweller, 1994) was required for these items, which may explain the poorer performance. Two of the seven items that had contrasting *p* values across the two modes involved items where scrap paper was needed. Although a "working booklet" was provided for the computer-based test, very few students chose to use it; in contrast, there was extensive evidence of students using the working booklet for the paper-

and-pencil test. This is in contrast to Johnson and Green's (2006) study, where it was reported that more than half of the students in their study actually showed more "scratchwork" for the computer-based test than for the paper-and-pencil test. These two items with contrasting *p* values illustrated cases where the paper-and-pencil test perhaps offered an *affordance* (Greeno, 1998) that the computer-based test did not (Threfall et al., 2007). Affordances can be thought of as "qualities of systems that can support interactions and therefore present possible interactions for an individual to participate in" (Greeno, 1998, p. 9). However, an affordance does not necessarily lead to increased validity. It may be that the affordance was not warranted, so the paper-and-pencil item was actually less valid than the computer-based test version; or it may be that the affordance of the computer-based test led to a less valid item version (Threfall et al., 2007).

The Texas Assessment of Knowledge and Skills (TAKS) has utilized computer-based testing since 2004. TAKS is offered in various subject areas, including mathematics, at Grades 7, 8, 9, 10, and the exit level. Each time a TAKS test form is offered in both computer-based and paper-based modes, a comparability study is conducted. The summary of comparability results for TAKS mathematics from 2005-2008 reveals that there have been test-level differences between the two modes 13 times, with no difference reported only once (Texas Education Agency, 2008). This is interesting, given that the maximum number of times any other subject area (reading, English language arts, science, or social studies) has had test-level differences is six times. However, it is not clear whether all of these differences have been statistically significant. Passing the TAKS at the exit level is a requirement for high school graduation, so the recommendation has been made that until such time as all schools have the technology infrastructure required to test all students on computer, comparability studies be continued to

help ensure the defensibility of the testing programs. Decisions about how test scores should be adjusted are made before test results are released and are based on the test-level comparability findings.

Way, Davis, and Fitzpatrick (2006) published the results of the 2005 comparability study for various TAKS subjects, including Grade 8 and Grade 11 mathematics. Participation in online testing for Grade 8 was voluntary, so a MSCA was done, with previous test scores on Grade 7 reading and mathematics TAKS tests used as matching criteria. Way et al. (2006) used the bootstrap procedure with 500 replications. For each bootstrap in the Grade 8 comparability study, 1,273 scores for the total test for each mode were used. For the Grade 11 comparability study, students who had previously failed one or more of the Grade 11 exit level assessments were randomly assigned to either an online version or paper version of the TAKS. There were 958 students who wrote the online version and 1,198 who wrote the paper version.

Way et al. (2006) used the Dorans and Lawrence (1990) criterion when comparing computer-based test and paper-and-pencil score conversions for TAKS—the ratio of the difference between the equating function and the identify function divided by the standard error of equating should be between plus or minus two. For Grade 8, Way et al. (2006) reported a difference of only 0.16 between the mean raw scores for the two modes favoring paper-and-pencil; however, at the upper raw score points, scaled score differences exceeded two standard errors of linking. The results for Grade 11 showed scale score differences within ±2 bootstrap standard errors, but there was a greater mode effect for Grade 11 mathematics than there was for Grade 8 mathematics, again favoring paper-and-pencil (Way et al., 2006). When Keng, McClarty, and Davis (2008) performed an item-level comparative analysis, they found significant differences ($p < 0.05$) between p values and different item response distributions for

four items on the Grade 8 TAKS, three favoring the paper-and-pencil test and one favoring the computer-based test. The three items that favored the paper-and-pencil test involved graphing and geometric manipulation. When a follow-up investigation on samples of actual test booklets was done, the authors found that two of these three items were the ones that were drawn on and labeled most frequently. No clear reason for the item that favored the computer-based test was evident. On the Grade 11 TAKS, three items had significant differences ($p < 0.05$) between p values and different item response distributions, two which favored the paper-and-pencil test and one which favored the computer-based test. The items that favored the paper-and-pencil test involved graphing and geometric manipulation; the item that favored the computer-based test had sizable diagrams in the item stimulus as well as the alternatives. Since the correct response to this item was B, Keng, McClarty, and Davis (2008) hypothesized that some students may not have realized that scrolling was required to see the other two alternatives (C and D).

The only Canadian mathematics comparability study found in the literature was done by Gaskill and Marshall (2006) for British Columbia's Grade 7 Foundation Skills Assessment (FSA) program. The FSA contains a multiple-choice component and a constructed response component. Starting in 2004, the multiple-choice component was offered electronically, but the constructed response was only available on paper. In the two parts of the study, mode effects for numeracy were examined with a focus on overall test score, gender, and ability grouping differences. Gaskill and Marshall (2006) hypothesized that the achievement on the computer-based test would be lower than that on the paper-and-pencil test and that ability grouping would be a moderator, with a greater difference for students of lower ability. They also hypothesized that when the constructed response was analyzed in relation to performance in paper mode years and computer mode years, there would be no difference.

In the first part of the study, six cohorts at the school level from 2001 to 2006 were compared with school, mode, and gender as fixed factors. Only schools that administered the FSA in both modes were included. The results of 13 schools with 2,836 students writing the paper-and-pencil test and 769 students writing the computer-based test during this time period were analyzed. For the multiple-choice component, the interaction of mode-by-school was significant ($p < 0.05$), as were the main effects mode and school ($p < 0.05$). The mode difference favored those who wrote this component on paper. For the constructed response component, the interaction of mode-by-school was significant ($p < 0.05$), as were the main effects mode, gender, and school ($p < 0.05$). All students wrote this component on paper, but the mode difference favored those students who wrote the multiple-choice component on computer, with the gender difference favoring females.

In the second part of the analysis, the goal was to determine if ability grouping was a moderator, so it was added as another fixed effect. Since the FSA is written in both Grade 4 and Grade 7, only the results of those students who wrote the Grade 7 FSA from 2004 – 2006 and had also previously written the Grade 4 FSA from 2001 – 2003 were used; there were 1,452 for the paper-and-pencil test and 637 for the computer-based test. These participants were assigned to one of three achievement categories based on their performance on the Grade 4 FSA. The gender-by-category interaction was significant ($p < 0.05$), with the greatest difference being at the high achievement level, favoring females. The gender-by-mode interaction was also significant ($p < 0.05$), favoring males on the paper test. Gaskill and Marshall (2006) hypothesized that transferring information back and forth from the screen to paper would have a greater negative impact on low-performing students than high-performing students, but this was not the case. The mode-by-achievement category interaction was significant ($p < 0.05$), but the

difference between modes was smaller for the low achievement level than the other two achievement levels. Given that the interaction mode-by-school was significant as was the main effect school ($p < 0.05$), Gaskill and Marshall (2006) suggested that future comparability studies also consider differences in teaching strategies, preparations for a computer-based test, and the use of computer-based assignments in the classroom.

The most extensive comparability studies at the elementary and secondary levels were undertaken by the National Assessment of Educational Progress (NAEP). In the first study, Russell and Haney (1997) investigated mode effects for mathematics, language arts, and science items from NAEP. For mathematics, Grade 8 students were randomly selected to either take a computer-based test or a paper-and-pencil test consisting of the same multiple-choice items. The study was part of a longitudinal study with matrix sampling, so only 50 students were selected to take the computer-based test and 70 were selected for the paper-and-pencil test. The overall performance difference between the two modes was not statistically significant ($p > 0.01$), and the effect size of 0.12 was small (Cohen, 1988). Regression analysis with a covariate, an open-ended assessment (OE) that all students wrote on paper before doing the multiple-choice questions, confirmed this result. There were no significant mode effects for any of the subtests. In a follow-up comparability study with just OE mathematics items, there was no statistical difference ($p > 0.05$) in performance on the two modes (Russell, 1999). For this study, only 54 students wrote the paper-and-pencil test form and only 56 wrote the computer-based test.

A third NAEP comparability study was done in 2001. Nationally representative samples of Grade 8 students were used for both the paper-and-pencil test and the computer-based test in the NAEP 2001 Math Online (MOL) study conducted by Sandene et al. (2005). In this study, 954 students wrote the paper-and-pencil test and 1,016 students wrote the computer-based test.

Multiple-choice and constructed-response questions were included. The mean of the computer-based test was statistically lower ($p < 0.05$) than the mean of the paper-and-pencil test, but only by about 0.14 standard deviations (Sandene et al., 2005; Bennett et al., 2008), which is less than the 0.20 minimum for small effects suggested by Cohen (1988). Bennett et al. (2008) reported that the standard deviation of the computer-based test was greater than that of the paper-and-pencil test, suggesting greater variability in the computer-based test scores, although no significance level was stated.

Item level analysis can help to determine whether any mode effects at the total score level are linked to uniform differences in item functioning or can be attributed to a few outliers (Bennett et al., 2008). All but 4 of the items were easier on the paper-and-pencil test. Taken across all 25 items, the range of the differences in the estimated item response theory (IRT) *b* parameters was -0.25 to 0.81, with the mean of the absolute value of the differences being 0.28 logits, suggesting minimal effects. However, the discrepancy was, on average, twice as large for the constructed-response questions than for the multiple-choice questions. The constructed-response questions required more adaptations to be rendered on screen than did the multiple-choice questions, which may have introduced the need for computer skills when responding or may have changed the nature of what is being measured (Bennett et al., 2008). When the differences in the discrimination estimates were compared, 16 of the 25 items appeared to be more discriminating on the paper-and-pencil test; however, across all the items, the mean absolute difference was 0.13, suggesting minimal effects.

The 2001 NAEP Math Online (MOL) study was the most comprehensive study of the role computer familiarity plays on student performance. Their definition of computer familiarity contained three components—computer experience, input accuracy, and input speed (Sandene et

al., 2005). Sandene et al. (2005) hypothesized that examinees should have a minimal level of proficiency in each of these components to effectively take an online test that contains constructed-response questions. While the number of students lacking familiarity was insignificant, familiarity with computers did affect performance on the Grade 8 mathematics test (Bennett et. al, 2008; Sandene et al., 2005). After controlling for performance on a paper-and-pencil block of 20 mathematics questions, the increment in variance accounted for in the MOL score was 8 percentage points, which was statistically significant ($p < 0.05$). Thus, some students performed better than their equally mathematically proficient peers because of their computer proficiency (Bennett et al., 2008). Thus, Drasgow et al. (2006) believed that the rank order was affected across the two modes in the MOL study, because the mix of skills (specifically related to computer familiarity) measured in each mode were different. Drasgow et al. (2006) also suggested that these results illustrate why judgments of comparability should never be based solely on score equivalence.

Summary of the Literature

Different properties related to mode effects have been studied in the comparability studies for mathematics. In addition to differences in overall scores, some researchers examined subtest differences (Csapó et al., 2009; Kim & Huynh, 2007; Russell & Haney, 1997; Sandene et al., 2005) or differences at the item level (Csapó et al., 2009; Johnson & Green, 2006; Kim & Huynh, 2007; Keng et al., 2006; Poggio et al., 2005; Sandene et al., 2005; Threfall et al, 2011). Some of the same and other researchers examined sub-group differences like gender and academic placement (Csapó et al., 2009; Gaskill & Marshall, 2006; Kim & Huynh, 2007; Minnesota Department of Education, 2009; Poggio et al., 2005). Very few studies examined the

relationship between computer familiarity and performance (Bennett et al., 2008; Choi & Tinkler, 2002, as cited in Oregon Department of Education, 2007; Sandene et al., 2005).

The tests in most studies contained only multiple-choice items, although a few also examined mode effects for open-ended or constructed response questions (Bennett et al., 2008; Csapó et al., 2009; Russell & Haney, 1997; Sandene et al., 2005). Only a few studies considered scratchwork (Johnson & Green, 2006; Keng et al., 2006; Threfall et al., 2011) or examined response distributions (Poggio et al., 2005; Keng et al., 2006). Very few studies were done at the high school level (Kim & Huynh, 2007; Minnesota Department of Education, 2009; Oregon Department of Education, 2007), although none were done at the grade 12 level. And only two comparability studies were identifiable as being a high-stakes large-scale testing environment (Minnesota Department of Education, 2009; Way et al., 2006).

Overall, the results of the mathematics comparability studies are inconclusive. Many of the studies found no statistical difference in performance on a paper-and-pencil test versus a computer-based test. A majority of the studies that did find statistically significant differences reported that the computer-based test was more difficult than the paper-and-pencil test, but often the effect sizes were small. These mixed results are not entirely surprising, given that tests can be nonequivalent when administered in different modes (Thissen, Reeve, Bjorner, & Chang, 2007), different settings produce different results (Clarianna & Wallace, 2002), and random assignment is often difficult to do, which may affect the interpretation of the results (Kingston, 2009; Way et al., 2008).

Chapter 3

Methods

The methods described in this chapter were selected to yield the results needed to answer the research questions first stated in Chapter 1, which are restated here for easy reference:

1. Are the psychometric properties of the items comparable between the paper-and-pencil field test and the computer-based field test?
2. Does the reliability of the paper-and-pencil field test differ from the reliability of the computer-based field test?
3. Do students' scores differ significantly between the paper-and-pencil field test and the computer-based field test with at least a moderate effect size?
4. Do male and female students' scores differ significantly with at least a moderate effect size between the paper-and-pencil field test and the computer-based field test?
5. To what degree does student performance on a digital Mathematics 30-2 field test depend upon computer familiarity and prior online testing experience?

Since the data used were previously collected by Alberta Education as part of its regular field testing procedures, this study involved secondary data analysis. Approval to request data was submitted to and granted by the Executive Director, Assessment Sector, Alberta Education. Prior to receiving the data, Alberta Student Numbers (ASN) were removed and replaced with consecutive numbers beginning with 1 (one) to protect the identity of the students. Ethics approval for the study was granted by the Research Ethics Board at the University of Alberta.

The chapter is organized in six sections. First, a description of the Mathematics 30-2 Diploma Examination is provided. This is followed by an overview of the field test processes used by Alberta Education. Third, the research design employed by Alberta Education to

determine if performance is influenced by the mode of delivery is described. Fourth, information about the instruments used—the Mathematics 30-2 year-end field test used in the study and the computer familiarity survey administered to the students who wrote the paper-and-pencil form of the field test and the students who wrote the computer-based form of the field test—is described. Next, a description of the data collection procedures is provided. Lastly, the statistical analyses conducted to answer the research questions are provided. The preliminary analysis conducted to determine if previous on-line test experience influences performance on the computer-based field test is provided first. This is then followed in turn by a description of the item analyses procedures, the analyses conducted at the field test level, and the analyses of the survey data.

Mathematics 30-2

Mathematics 30-2 is a course intended for students planning to attend a university, college, or technical institute after high school, but who do not need calculus (Alberta Education, 2010). The Mathematics 30-2 Diploma Examination is the school exit examination for this course. Provincial implementation of the Mathematics 30-2 course began in the fall of 2012. Therefore, the first administration of the Mathematics 30-2 Diploma Examination was in January 2013 for students who took the course in the fall semester, and the second administration was in June 2013 for students who took the course in the spring semester or as a full year course. The blueprint for the Mathematics 30-2 Diploma Examination, which is criterion referenced, is provided in Table 1 (Alberta Education, 2013c).

Table 1

Mathematics 30-2 Blueprint

Question Format	Number of Questions	Emphasis
Multiple-Choice	28	70%
Numerical-Response	12	30%

Cognitive Levels	Emphasis
Conceptual	34%
Procedural	30%
Problem Solving	36%

Topics	Emphasis
Logic and Reasoning	17%
Probability	33%
Relations and Functions	50%

A different form of the diploma examination is administered at each examination administration. On each form, the emphasis of multiple-choice and numerical-response items in each topic area differs. Given that provincial implementation of Mathematics 30-2 began in the fall 2012 semester, the students who wrote the field tests and diploma examinations in January and in June, 2013, are not necessarily representative of the student population for whom the course is intended. Once this population stabilizes, secure common items to allow linking of forms to provide a measure of change will be included in each Mathematics 30-2 Diploma Examination.

As shown in Table 1, 28 multiple-choice and 12 numerical-response questions are used. The emphasis given to the three cognitive levels is more evenly distributed than the emphasis for topics: 34%, 30%, and 36% versus 17%, 33%, and 50%. The multiple-choice questions have four alternatives and the open-ended numerical response questions have four boxes and bubbles that students use to record their answers. Numerical-response questions in Mathematics 30-2 may involve a calculation, require students to correctly order a sequence, require students to rank

using provided criteria, or select from a list using provided criteria. Each multiple-choice question has one correct answer but some numerical-response questions have more than one correct answer. Both multiple-choice and numerical-response questions are scored dichotomously.

Alberta Education provides sample multiple-choice and numerical-response questions in the practice section online as well as in the *Mathematics 30-2 Assessment Standards and Exemplars* (Alberta Education, 2013d) on their web site. While the sample questions are not included in any diploma examination, they illustrate the formats and general types of questions that may be included on a Mathematics 30-2 Diploma Examination.

Field Testing in Alberta

Alberta Education regularly administers field tests toward the end of a semester (January and June) for each of its diploma examinations. The purposes of the field tests are to determine the psychometric characteristics of the field test questions and then to use the findings to select questions for future diploma examinations. Classical test score theory item analysis is used to ensure that only questions with good psychometric properties (difficulty values between 0.30 and 0.85, corrected point-biserial values greater than 0.20, and responses for incorrect options being chosen by at least 5% of the students) are selected for future diploma examinations. Teacher and student comments are also used to ensure the wording is clear and the contexts portrayed in the items are appropriate for the students.

Field test procedures. Once approval to participate in field-testing has been granted by the superintendent of a school district in the province and the principal of a school, teachers in these schools can request a field test for a specific date within the allowable field-testing period, which is approximately 3 weeks long and ends approximately 1 week before the administration

of the diploma examination. In order to obtain a representative sample of the province, the Examination Administration Unit, Assessment Sector within Alberta Education considers the requests received and then selects classes from across the province to obtain a somewhat representative sample of students for each field test.

The field tests are administered under conditions that are similar to the administration of diploma examinations. Paper-and-pencil field tests are brought to a teachers' class and administered by an Alberta Education field test supervisor (teachers other than the students' teacher administer the field test). While the supervisor proctors the field test, the teacher reviews the items on the field test and records comments they might have in the field test booklet. Computer-based field tests are delivered electronically using the computer platform Quest A+ and are usually administered under the supervision of the teacher who requested the field test. Teachers are given a supervisor code so that they have access to the field test online. They enter any comments they might have about items in an interface at the end of the test. Both paper-and-pencil field tests and computer-based field tests are kept secure before, during, and after administration.

Paper-and-pencil field tests are scored by the Alberta Education field test supervisor and a copy of the students' scores is left with the classroom teacher. The computer-based field tests are scored automatically and the classroom teacher receives a copy of the students' scores once the teacher submits the electronic security declaration form to Alberta Education. Teachers are encouraged to use the field test scores in a manner that motivates students to perform well, such as allowing students to use the field test score as a replacement score for a low classroom test or quiz score obtained during the school year.

Computer requirements for Quest A+. To use Quest A+, recent versions of Opera, Safari, Chrome, or Firefox are required systems at the school level. Flash Player 11.1 or newer is also required, along with a screen resolution of at least 800 x 600. Recommended flash settings are posted on Alberta Education's website, together with the system requirements for both a Windows environment and a Mac environment.

Quest A+ is the production pilot that Alberta Education is presently using to deliver the computer versions of selected field tests. Once students log in, the Quest A+ browser locks the computer so that students cannot access the internet, any files stored on the hard drive of the computer, other network files, and a printer. A mock secure test on the Quest A+ website can be used to ensure the locked browser is installed and functioning properly before a secure test or examination is administered.

Field test procedures using Quest A+. Teachers who request a field test on Quest A+ receive an electronic copy of the *Digital Field Test Instructions: Diploma Examination and Achievement Testing Programs* from the Examination Administration, Assessment Sector unit. This document describes how to set up a practice test run for students, the student login process, the teacher login process, what to do in the event of a power or computer failure, how to input teacher comments, and how to receive scores after the field test is written by students.

Research Design

One of the seven June 2013 Mathematics 30-2 field tests was administered in a paper-and-pencil format and in a computer-based format. Random assignment of classes to the paper-and-pencil form and to the computer-based form was not possible given the manner in which schools and classes are identified for the field tests. Consequently, the field test design was equivalent to a non-equivalent groups quasi-experimental design.

There were 91 teachers who indicated they would participate in the June 2013 paper-and-pencil field tests and 23 teachers who indicated they would participate in the computer-based field test. Fourteen teachers were selected for the paper-and-pencil field test in this study and, since there was only one computer-based field test in June 2013, all 23 teachers who requested a computer-based field test were selected for the online form of the same test. Once these selections were completed by the Examination Administration unit, these teachers were contacted by Alberta Education and asked if they were willing to participate in a comparability study. They were informed that by agreeing to participate, their students would complete a survey questionnaire that included questions on computer familiarity and questions related to their field test experience. Of the 14 teachers who requested a paper-and-pencil field test, 12 agreed to participate; of the 23 teachers who requested a computer-based field test, 17 agreed to participate. Examination of the schools of these 29 teachers by the Examination Administration unit revealed that the two subsamples were a somewhat representative sample of schools in the province.

Instruments

Mathematics 30-2 field tests. Each Mathematics 30-2 field test is built to the same examination specifications as the Mathematics 30-2 Diploma Examination (see Table 1), but with fewer questions. Although there is no statistical information for the items included in each field test, content related validity evidence for each field test is established by a field-test validation working group, which consists of classroom teachers and Alberta Education staff, prior to administering the field test.

As indicated above, seven field tests were administered for Mathematics 30-2 in June, 2013. These field tests were designed to be completed in 60 minutes, and each consisted of 13

multiple-choice and 5 numerical-response questions. One field test was offered in both a paper-and-pencil format and in a computer-based format to determine if there were performance differences due to administration mode. This field test contained six multiple-choice and three numerical-response items from Relations and Functions; three multiple-choice items from Logical Reasoning; and four multiple-choice and two numerical-response items from Probability. For the cognitive level emphasis, six items were conceptual, five were procedural, and seven were problem solving.

The paper-and-pencil test booklet for this field test contained an instruction page, scrap paper, and a tear-out formula sheet. Students were allowed to use an approved graphing calculator for the field test. They recorded their answers to the multiple-choice and numerical-response questions on a Scantron sheet provided by Alberta Education, and they could complete the 18 questions in any order and go back and review their answers.

The computer-based form of this field test was administered using Quest A+. An instruction page and a drop-down formula sheet were provided online, and students could choose to hide or show these resources using the respective button. Students could also navigate through the field test using the page navigation buttons. Both the resources and the question components had zoom controls that the students could use to control window size. One question at a time appeared on the screen. If students wished to mark a question for later review, they clicked the Review box. They could also choose to have the remaining time hidden or shown using the Time is Hidden or Time Left buttons. They were allowed to use an approved graphing calculator and, as with the paper-and-pencil field test, they could complete the 18 questions in any order and go back and review their answers. They were allowed to use scrap paper during the online field test, but they were not allowed to take it with them once they finished the test.

Before signing off of Quest A+, a checklist popped up to remind students of any unanswered questions, although they could choose to not answer these questions.

Computer familiarity survey. The researcher developed a paper-and-pencil questionnaire to collect information about students' use of computers in their mathematics classes, their computer use in general, their experience with online testing, and their experience in the response medium they wrote the field test in. Some of the questions on the questionnaire were adapted from the questions used by Sandene, Bennett, Braswell, and Oranje (2005) for the Math Online (MOL) study for the NAEP Technology-Based Assessment (TBA) project. The first eight questions were common to both samples of students as were question 13 for the students who responded to the computer-based field test and question 15 for the students who responded to the paper-and-pencil field test. The students who wrote the paper-and-pencil field test were branched after question 8 to question 15. The students who wrote the computer-based field test responded to questions 9 to 14. A copy of the questionnaire is provided in Appendix A.

Data Collection

Prior to data collection. As mentioned earlier, the data for this study was part of Alberta Education's field testing program. The research design and data collection were discussed with the Assessment Sector examination administration team and the scoring and reporting team. Once approval was granted by these personnel at Alberta Education, a special letter was sent out to Mathematics 30-2 teachers whose classes wrote the field test involved in this study. This letter outlined the scope of the study, timeline, nature of participation, and contained information on administering the computer familiarity survey. Teachers were asked to read this letter, and then check the yes box if they were willing to participate in the study and the

no box if they were not willing to participate. There were no apparent risks to either students or their teachers in this study and participation was purely voluntary.

In addition to receiving instructions on the standard procedures for administering Alberta Education field tests, the proctors for the paper-and-pencil field test were given instructions to administer the computer familiarity survey. Teachers whose students wrote the computer-based field test were given instructions for administering the online version of the field test and the computer familiarity survey. All data collected was secured by Alberta Education.

Actual data collection. The Mathematics 30-2 field test in this study was administered by Alberta Education from June 4 – June 11, 2013. Proctors supervised the paper-and-pencil field test; classroom teachers supervised the computer-based field test. The computer familiarity survey was administered in paper-and-pencil format to both groups of students by the proctors in the case of the paper-and-pencil administration and by the teachers in the case of the computer-based administration. If class time allowed, the survey was done immediately after students completed the field test; if there was not sufficient time, the survey was done at the beginning of the next Mathematics 30-2 class.

As part of the computer-based field test process, the students' Alberta Student Numbers (ASN) were collected when students logged on to Quest A+. Alberta Education exam administration staff affixed stickers with ASNs on the Scantron sheets for the paper-and-pencil version of the field test and on the computer familiarity survey forms. Use of the ASNs allowed identification of the gender of the student and were used to match students' field test forms and surveys.

Statistical Analysis

Data entry. The students' answers sheets for the paper-and-pencil form of the field test were scanned using Alberta Education's optical scanners. Using Alberta Education's business rules for field-test data, the data from the two field test forms excluded students who either left at least 16% of the machine-scored items blank (i.e., 3 items), at least 16% of the multiple-choice items blank (3 items), or at least 67% of the numerical-response items blank (3 items).

The responses to the computer familiarity survey were input by a contractor with 100% verification by the researcher. Prior to analyses, the missing responses were considered item by item, as shown in Table 2. As described further in Chapter 5, in some questions a blank or missing response was indicative of "no" or "never" so it was coded as such; in other questions,

Table 2

Handling of Missing Responses in Survey Questions

Survey Question	What Was Done with Missing Responses
Question 1	Not applicable as there were no missing responses
Question 2	Missing responses coded as average of actual responses
Question 3	Missing responses coded as "never"
Question 4	Missing responses coded as "never"
Question 5	Missing responses coded as "never"
Question 6	Missing responses coded as "no"
Question 7	Missing responses coded as average of actual responses
Question 8	Missing responses not recoded
Question 9	Missing responses not recoded
Question 10	Missing responses coded as average of actual responses
Question 11	Missing responses coded as average of actual responses
Question 12	Missing responses coded as average of actual responses
Question 13	Missing responses coded as average of actual responses
Question 14	Missing responses coded as average of actual responses
Question 15	Missing responses coded as average of actual responses

the missing responses were coded using the average of the actual responses so as to maintain as much data as possible in the analysis. Only two questions (8 and 9) did not have missing responses recoded because neither of the input procedures would have resulted in meaningful and interpretable results for these two questions. The survey data and the data from the two field test forms were then merged into one file.

Preliminary analysis. Possible differences in performance between students who had previous online test experience before writing the computer-based field test and students who had no previous online test experience before writing the computer-based field test was of concern (Research Question 5). Therefore, a preliminary analysis was conducted to determine if previous online testing experience affected performance on the computer-based version of the field test. For the students who wrote the field test online, the difference in performance between the students who answered "yes" and the students who answered "no" to the survey question "Have you taken a test online before today?" was tested.

The basic descriptive statistics for these two subsamples are reported in Table 3. Of the 258 students who wrote the computer-based field test, two did not respond to this survey question. As shown, the subsample sizes are not equal; of the 256 students, 163 indicated that they had taken a test previously online and 93 indicated they had not. Given the interaction

Table 3

Description of Student Subsamples' Responses to Survey Question 9

	yes	no
Number of examinees	163	93
Mean	7.88	8.23
Standard Deviation	3.12	2.66

between unequal sample sizes and lack of homogeneity of variance influences Student's *t*-test statistic for two independent groups, Levene's test was used to determine if there was lack of

homogeneity of variance. If there was lack of homogeneity of variance, then Welch's correction to Student's *t*-test statistic would be used to test the significance of the difference between the two means. The results of Levene's test ($F = 4.24$, $df = 254$, $p < 0.05$) revealed that there was not homogeneity of variance and Welch's correction was needed. The value of Welch's correction to Student's *t*- test statistic was not significantly different from zero ($t' = -0.09$; $df = 123$; $p > 0.05$). The effect size, which was -0.19, was small (Cohen, 1988). Therefore, the two computer-based field test subsamples were combined for all subsequent analyses.

Item analyses. Using classical test score theory item analysis, the difficulty (*p*-value), corrected point-biserial correlation coefficient (*CRPB*), and alternative functionality were compared between the paper-and-pencil field test (PP) and the computer-based field test (CB) for each item. The *p*-value and the *CRPB* were calculated using Alberta Education's item analysis software (Ping Yang, Psychometrician, Assessment Sector, Alberta Education, personal communication, November 12, 2013). To compare the *p*-values across modes, the odds ratio was calculated and then Cox's (1970) Index was used for the effect size:

$$OR = \frac{p_{cb_i}(1 - p_{pp_i})}{p_{pp_i}(1 - p_{cb_i})}$$

and

$$\text{Cox Index} = \frac{ln(OR)}{1.65}$$

where *OR* is the odds ratio, p_{pp_i} is the difficulty of item *i*, PP, and p_{cb_i} is the difficulty of item *i*, CB.

To compare the *CRPB*s across mode, the values were first transformed to Fisher's (1958) *Z*, given correlations are on the closed interval and do not follow a normal distribution for values other than zero:

$$Z = \tanh^{-1} r$$

where *r* is the correlation coefficient. Then the *z*-test statistic for the test of the difference between two point-biserial correlation coefficients for the correct option *c*, item *i*, PP and CB was used:

$$|z_i| = \frac{z_{cPP_i} - z_{cCB_i}}{\sqrt{\frac{1}{n_{PP_i} - 3} + \frac{1}{n_{CB_i} - 3}}}$$

where $z_{cPP_i}$is Fisher's *Z* for the correct option, item *i*, PP, and $z_{cCB_i}$is Fisher's *Z* for the correct option for item *i*, CB. The effect size, *q*, (Heinrich Heine Univeristät Düsseldorf, n.d.) was determined by:

$$q = z_{cCB_i} - z_{cPP_i}$$

The functionality of each alternative for the multiple choice questions across modes was compared using the Chi-square test statistic goodness of fit test. The proportions of students who responded to the alternatives for each computer-based field test multiple-choice item were fit to the proportions of students who responded to the corresponding alternatives for the same paper-and-pencil field test item, given the paper-and-pencil field test was previously the only form used. For the numerical-response items, the Chi-square test statistic goodness of fit test was also used to compare the proportions of students who determined the correct response and the proportion of students who determined the incorrect response and responded to the computer-based field test and the corresponding proportions of students who responded to the paper-and-pencil field test.

Given failing to indicate that the difference between the two *p*-values, the two *CRPB*s, and the fit of the alternatives was a more critical error (Type II error) than saying the difference was due to chance (Type I error), the comparative analyses were completed at the 0.05 level of significance. All comparative analyses were conducted using Version 21, IBM SPSS Statistics (IBM Corporation, 2012).

Test level analyses. Since random assignment of the students to the mode of administration was not possible, the original plan for this study was to use the June 2013 Mathematics 30-2 Diploma Examination as a covariate to account for possible differences in the mathematical abilities of the students in the two samples at the test level. However, there was severe flooding in the southern part of the province in June 2013, and a large number of students who wrote the field test and completed the computer familiarity survey were unable to write the Mathematics 30-2 Diploma Examination. Removing these students from the samples decreased the sample size considerably and also impacted the representativeness of the samples, so the covariate was not used.

Instead, a 2 x 2 x 3 (response mode-by-gender-by-test mode preference) fully crossed fixed effects ANOVA was conducted to determine the significance of the interactions between gender, mode, and test mode preference; each of the two way interactions (e.g., gender-by-test mode preference); and each of the main effects (e.g., gender). The responses to the test mode preference survey question (Question 8) were used to measure test mode preference ("If I had a choice, I would prefer taking a math test a. online, b. using paper and pencil, and c. either online or using paper and pencil"). This ANOVA was used in light of the way the field test samples were selected, which while not random, was systematic. Both samples were selected to represent the province.

Computer familiarity survey analyses. The analysis of the computer familiarity survey was completed at the item level. First, for the common questions (2 through 8, and 15 (paper-and-pencil field test) with 13 (computer-based field test)), the difference between the proportions of responses of the students who responded to the paper-and-pencil field test and the students who responded to the computer-based field test were compared using the Chi-square test statistic. For the remaining questions that were answered by only the students who did the computer-based field test, the frequencies were computed and examined.

With respect to fairness to students, it was believed that a Type II error had more serious consequences than a Type I error, so all analyses were completed at the 0.05 level of significance.

Chapter 4

Results at the Item Level

The results for the analysis at the item level are presented in this chapter in two sections. First, the results of a comparative analysis of the psychometric properties of each item across response modes is provided and discussed. However, since the field test is secure, specific details about the items cannot be disclosed. Second, as will be seen, since there were no differences between the psychometric properties of the items in the two test forms, the two samples were combined and an item analysis for the full sample was conducted. The results of this analysis, which is provided in the second section, will be used to select items for future Mathematics 30-2 Diploma Examinations.

Comparative Analyses across Response Modes at the Item Level

The psychometric properties of each item were examined using classical test theory item analysis. For each item, the alternative functionality across response modes, the difficulty (*p*-value), and the discrimination (corrected point biserial correlation coefficient) were compared for each item across the paper-and-pencil field test and the computer-based field test.

Alternative functionality. The functionality of each alternative for the multiple-choice items across response modes was compared using the Chi-square test statistic goodness of fit test. The proportions of students who responded to each alternative for each of the computer-based field test (CB) items were fit to the proportions of students who responded to the corresponding alternative for each of the paper-and-pencil field test (PP) items, given the paper-and-pencil field test was the only form used in previous years. The results of the analysis for each multiple-choice item are shown in Table 4, with the correct answers indicated by the letter

c. The items are grouped by topic area: multiple-choice items 1 through 4, 12 and 13 are from Relations and Functions; multiple-choice items 5 through 7 are from Logical Reasoning;

Table 4

Chi-square Test Statistic Comparison between the Multiple-Choice Items on the Paper-and-Pencil Field Test (n = 252) and Computer-Based Field Test (n = 378)

		Alternative/Proportion Within Response Mode				
Topic/Item Number	Response Mode	A	B	C	D	Sig.*
Relations & Functions						
MC 1	PP	0.26	0.02	0.14	0.58c	No
	CB	0.26	0.04	0.14	0.57	
MC 2	PP	0.07	0.10	0.69c	0.14	No
	CB	0.05	0.11	0.74	0.10	
MC 3	PP	0.34	0.53c	0.11	0.02	No
	CB	0.32	0.58	0.07	0.03	
MC 4	PP	0.28	0.56c	0.10	0.06	No
	CB	0.28	0.59	0.09	0.04	
MC 12	PP	0.43	0.29c	0.22	0.06	No
	CB	0.46	0.29	0.17	0.08	
MC 13	PP	0.31	0.35	0.14	0.20c	No
	CB	0.27	0.38	0.16	0.19	
Logical Reasoning						
MC 5	PP	0.77c	0.03	0.07	0.13	No
	CB	0.85	0.01	0.04	0.10	
MC 6	PP	0.15	0.09	0.20c	0.56	No
	CB	0.16	0.07	0.16	0.61	
MC 7	PP	0.09	0.13	0.06	0.72c	No
	CB	0.07	0.17	0.04	0.72	
Probability						
MC 8	PP	0.43c	0.21	0.27	0.09	No
	CB	0.50	0.17	0.24	0.09	
MC 9	PP	0.06	0.30	0.17	0.47c	No
	CB	0.03	0.28	0.15	0.54	
MC 10	PP	0.51c	0.36	0.09	0.04	No
	CB	0.47	0.41	0.09	0.03	
MC 11	PP	0.21	0.36	0.15	0.28c	No
	CB	0.18	0.39	0.12	0.31	

Note: MC = multiple-choice.
*$p < 0.05$; c indicates the correct answer

and multiple-choice items 8 through 11 are from Probability. IBM SPSS Statistics (IBM Corporation, 2012), which was used to generate these data, recommends that the Chi-square test statistic not be used unless the minimum expected frequency in each cell is at least 5. However, this study followed the recommendations made by Roscoe and Byars (1971), Convor (1974), and Camilli and Hopkins (1978, 1979) that an average expected frequency of 2 is sufficient.

As shown in Table 4, none of the Chi-square test statistic values were significant ($p > 0.05$) for any of the 13 multiple-choice items. Therefore, the distribution of proportions of students choosing the four alternatives for each multiple-choice item on the computer-based field test was not significantly different from the distribution of proportions of students choosing the four alternatives for the same multiple-choice item on the paper-and-pencil field test.

The proportion of students who choose the correct response should be greater than the proportion that chooses an incorrect alternative. If this is not the case, it may be that the item was miskeyed, the item is flawed in some way, or too many students are drawn to the misconception that the alternative is based on. This is the case in multiple-choice items 12, 13, 6, and 11. Multiple-choice item 12 was a word problem that required careful reading. Multiple-choice item 13 required students to use the context to identify an appropriate domain and range; in this question, two alternatives were more popular than the correct response and in both cases, the domain was not appropriate. Multiple-choice item 6 required students to properly apply set notation symbols, which are identified on the formula sheet, to two regions of a diagram; the alternative that was more popular than the keyed answer used the wrong symbols for the second region. Multiple-choice item 11 was a probability word problem; students who missed a key word in the problem choose the wrong alternative. These items were not miskeyed, and the

members of the teacher validation committee felt that these items were fair and that the alternatives were appropriate.

Numerical-response items. For the numerical-response items, the proportion of students who responded correctly and the proportion who responded incorrectly on the computer-based field test were fit to the corresponding proportion of students on the paper-and-pencil field test. The results of the analysis for each numerical-response item are shown in Table 5. The items are grouped by topic area: numerical-response items 1, 2, and 3 are from Relations and Functions; numerical-response items 4 and 5 are from Probability. Logical Reasoning contained no numerical-response items.

Table 5

Chi-square Test Statistic Comparison between the Numerical-Response Items on the Paper-and-Pencil Field Test (n = 252) and Computer-Based Field Test (n = 378)

Topic/Item Number	Response Mode	Correct Response	Incorrect Response	Sig.*
Relations & Functions				
NR 1	PP	0.39	0.61	No
	CB	0.41	0.59	
NR 2	PP	0.32	0.68	No
	CB	0.38	0.62	
NR 3	PP	0.08	0.92	Yes
	CB	0.18	0.82	
Probability				
NR 4	PP	0.15	0.85	No
	CB	0.11	0.89	
NR 5	PP	0.20	0.80	No
	CB	0.22	0.78	

Note: NR = numerical-response
*$p < 0.05$

As shown in Table 5, only one numerical-response item had a significant Chi-square test statistic value ($p < 0.05$). While the proportion of students who answered item 3 are low (0.08 for the paper-and-pencil field test and 0.18 for the computer-based field test), the proportion of students who wrote the computer-based field test and determined the correct answer was

significantly greater than the proportion of students who wrote the paper-and-pencil field test and determined the correct answer. There is no apparent reason for this difference. Further, for the remaining four numerical-response items, an incorrect response was more popular than the correct response for both response modes, particularly so for the last two items. On this field test, all of the correct answers were verified, and teacher validation committees felt that each of the five numerical-response items was not flawed and was fair. Further, students typically find numerical-response items more difficult than multiple-choice items.

Difficulty and discrimination. The difficulty index or *p*-value for an item is the proportion of examinees who answer the item correctly. The corrected point biserial correlation coefficient (*CRPB*) is the correlation between the responses to the alternatives of an item and the total score minus that item, and it is used to indicate how well the item discriminates between students who performed well on the test and students who performed poorly. Alberta Education recommends that the *CRPB* be used over the point biserial correlation coefficient (*RPB*) because of the small sample sizes that are typically used for field tests.

Table 6 provides the *p*-value and the *CRPB* for the correct option for each item for the students who responded to the paper-and-pencil field test and for the students who responded to the computer-based field test, and it also indicates whether or not the differences between the two *p*-values and the two *CRPB*s differ significantly or not. The results for the difficulty are in columns 2, 3, and 4, and the results for discrimination are in columns 5, 6 and 7. The items are grouped by topic area: multiple-choice items 1 through 4, 12 and 13, and numerical-response items 1 through 3 are from Relations and Functions; multiple-choice items 5 through 7 are from Logical Reasoning; and multiple-choice items 8 through 11 and numerical-response items 4 and 5 are from Probability.

As shown in Table 6, with three exceptions, the differences between the two *p*-values and the two *CRPB* values did not statistically differ. The *p*-values of numerical-response item 3 and multiple-choice item 5, and the *CRPB* values for numerical-response item 3 differed significantly

Table 6

Comparison of Difficulty and Discrimination between Paper-and-Pencil Field Test (n = 252) and Computer-Based Field Test (n = 378)

	Difficulty			Discrimination		
Topic/Item Number	PP	CB	Sig.*	PP	CB	Sig.*
Relations & Functions						
MC 1	0.58	0.57	No	0.32	0.24	No
MC 2	0.69	0.74	No	0.21	0.28	No
MC 3	0.53	0.58	No	0.06	0.08	No
MC 4	0.56	0.59	No	0.23	0.25	No
MC 12	0.29	0.29	No	0.15	0.13	No
MC 13	0.20	0.19	No	0.03	0.04	No
NR 1	0.39	0.41	No	0.35	0.40	No
NR 2	0.32	0.38	No	0.30	0.27	No
NR 3	0.08	0.18	Yes	0.23	0.38	Yes
Logical Reasoning						
MC 5	0.77	0.85	Yes	0.10	0.20	No
MC 6	0.20	0.16	No	0.17	0.11	No
MC 7	0.72	0.72	No	0.20	0.09	No
Probability						
MC 8	0.43	0.50	No	0.25	0.18	No
MC 9	0.47	0.54	No	0.24	0.31	No
MC10	0.51	0.47	No	0.27	0.34	No
MC 11	0.28	0.31	No	0.20	0.25	No
NR 4	0.15	0.11	No	0.30	0.30	No
NR 5	0.20	0.22	No	0.27	0.41	No

Note: MC = multiple-choice. NR = numerical-response
*$p < 0.05$

($p < 0.05$). The effect sizes for numerical-response item 3 were 0.62 for difficulty and 0.16 for discrimination; for multiple-choice item 5, the effect size was 0.30 for difficulty. Cohen's (1988) guidelines for interpreting the effect size are 0.20 small, 0.50 moderate, and 0.80 large. For the purposes of this study, the mid-point between the suggested cut-scores was used to

differentiate small from moderate (0.35) and moderate from large (0.65). Therefore, the effect size for numerical-response item 3 for difficulty was moderate, and the other two effect sizes, numerical-response item 3 for discrimination and multiple-choice item 5 for difficulty, were small. Consequently, it was concluded that, with the exception of numerical-response item 3, the item difficulties of the items did not differ between the paper-and-pencil field test and the computer-based field test, and the discrimination of the items did not differ between the paper-and-pencil field test and the computer-based field test either. The results for numerical-response item 3 with respect to difficulty are consistent with the results of the Chi-square test statistic, which was also significant, as shown in Table 5.

As can be seen from the results presented in Tables 4, 5 and 6, none of the multiple-choice items differed in the response distributions for the alternatives, and only one numerical response item had a significant difference in response distributions across response modes, with the correct response favoring the computer-based field test. Two items had a significant response mode effect for their *p*-values, both favoring the computer-based field test, but only one of these items had a medium effect size. Only one item had a significant response mode effect for discrimination, favoring the computer-based field test, but the effect size was small. Therefore, it can be concluded that, with the exception of numerical-response item 3, the students who responded to the computer-based field test and the students who responded to the paper-and-pencil field test were not disadvantaged at the item level across response modes for the field test in this study. Further, as indicated above, the proportions of students who responded correctly to numerical-response item 3 were both small and less than the desired value of 0.30 (see below). Therefore, the results of the two samples were combined to produce item

analysis results that would be more stable due to the larger sample size (n= 630); these results are provided in the next section.

Item Level Analysis for the Combined Sample

Table 7 provides the *p*-value and the *CRPB* for each item for the combined sample of 630 students who responded to the field test. As before, the items are grouped by topic area: multiple-choice items 1 through 4, 12 and 13, and numerical-response items 1 through 3 are from

Table 7

Difficulty and Discrimination for Combined Sample of Paper-and-Pencil and Computer-Based Field Tests (n = 630)

Position on Field Test	Topic/Item Number	Difficulty	Discrimination
	Relations & Functions		
1	MC 1	0.57	0.26
2	NR 1	0.40	0.38
3	MC 2	0.72	0.25
4	MC 3	0.56	0.08
5	NR 2	0.36	0.28
6	MC 4	0.57	0.24
7	NR 3	0.14	0.33
17	MC 12	0.29	0.13
18	MC 13	0.20	0.04
	Logical Reasoning		
8	MC 5	0.82	0.16
9	MC 6	0.18	0.13
10	MC 7	0.72	0.13
	Probability		
11	MC 8	0.48	0.21
12	MC 9	0.51	0.29
13	NR 4	0.12	0.30
14	MC10	0.48	0.30
15	MC 11	0.30	0.23
16	NR 5	0.21	0.36

Note: MC = multiple-choice. NR = numerical- response
*$p < 0.05$

Relations and Functions; multiple-choice items 5 through 7 are from Logical Reasoning; and multiple-choice items 8 through 11 and numerical-response items 4 and 5 are from Probability.

The position of the item on the field test is also provided in Table 7. These results will be used to select items for future Mathematics 30-2 examinations.

The minimum and maximum *p*-values for field tested items that can be used for future diploma examinations, according to Alberta Education's branch standards, are 0.30 and 0.85, respectively. As seen in Table 7, multiple-choice items 6, 12, and 13 failed to meet the minimum criterion for difficulty (i.e., 0.18, 0.29, and 0.20, respectively) as did numerical-response items 3, 4, and 5 (i.e., 0.14, 0.12, and 0.21, respectively). None of the *p*-values of the items exceeded the maximum.

In terms of topic, two of the three multiple-choice items (items 12 and 13) for which the *p*-value was less than 0.30 measured Relations and Functions and the third (item 6) measured Logical Reasoning. Two of the three numerical-response items (items 4 and 5) for which the *p*-value was less than 0.30 measured Probability, and the third (item 3) measured Relations and Functions.

Three of the six items for which the *p*-value was less than 0.30 measured Relations and Functions, with two of the three items measuring the same sub topic on Relations and Functions (i.e., sinusoidal functions). For the two items from Probability with a *p*-value of less than 0.30, one was from the sub topic permutations and the second was a probability question.

Further, as shown in Table 7, it appears that item position did not have an effect on the *p*-values, but, as mentioned earlier, item type did. While the *p*-values of the last two multiple-choice items for Relations & Functions, which were the last two items on the test, are low, as mentioned in the previous paragraph, both measured a difficult concept. There are two other multiple-choice items (items 6 and 11) with low *p*-values, but it is not clear that the values are

due to their position in the test. However, the *p*-values for the numerical-response items tend to be lower than the *p*-values for the multiple-choice items referenced to the same subdomain.

The *CRPB* for the correct answer must be positive and should be greater than or equal to 0.20, according to Alberta Education's branch standards. The value of the *CRPB* is influenced by the *p*-value (Gulliksen, 1950; Lord & Novick, 1968). Items that are very difficult and very easy for a group of students usually have substantially lower *CRPB*s than do items of medium difficulty. As shown in Table 7, all of the items had a positive *CRPB*. However multiple-choice items 3, 5, 6, 7, 12, and 13 have a *CRPB* below 0.20. Three of these six items (items 6, 12, and 13) were more difficult items, while the remaining three were of moderate difficulty. But not all of the difficult items had low CRPBs. Numerical-response items 3, 4 and 5 were difficult, but the values of the *CRPB* for these items were 0.30 or greater.

The *CRPB* for three of the six multiple-choice items that measured Relations and Functions (items 3, 12, and 13) and all three items that measured Logical Reasoning were below the minimum value. Three of the items measured Logical Reasoning, which suggests that students learned this topic to the same level, as measured by these three items.

Based only on the difficulty and discrimination values reported in Table 7, the following items from the field test in this study that would be considered for inclusion in a Mathematics 30-2 Diploma Examination are multiple-choice items 1, 2 and 4 and numerical-response items 1 and 2 for Relations and Functions; and multiple-choice items 8, 9, 10, and 11 for Probability. No numerical-response items would be selected for Probability and no items would be selected for Logical Reasoning.

Chapter 5

Results at the Test Level and Survey Results

This chapter presents the results for the analysis of the field test at the test level and the results of the analysis of the computer familiarity survey at the item level. This chapter is organized in two major sections. In the first major section, the psychometric properties of the paper-and-pencil field test and the computer-based field test are provided and discussed, followed in turn by the results of a comparative analysis of the students who responded to the computer-based field test and completed the computer familiarity survey and the students who responded to the computer-based field test and did not complete the survey, and, lastly, the results of a comparative analysis at the test level across response mode, gender, and test preference. In the second major section, the computer familiarity survey results are presented.

Psychometric Properties of the Paper-and-Pencil Field Test and the Computer-based Field Test

The psychometric properties of the paper-and-pencil field test (PP) and the computer-based field test (CB) are reported in Table 8. The field test contained 18 items in total, all of which were dichotomously scored.

As shown in Table 8, the minimum and maximum scores for the group of students who responded to the paper-and-pencil form and the group of students who responded to the computer-based form were identical. The means and the standard deviations are similar. The scores for both groups are positively skewed by about the same amount, which, as seen in the item level results reported in Chapter 4, indicates that the field test was difficult for the students in both samples. Further, the values for kurtosis suggest that the score distribution of the

Table 8

Psychometric Characteristics of the Paper-and-Pencil Field Test and the Computer-Based Field Test

	Field Test Form	
	PP	CB
Number of examinees	252	378
Minimum	1	1
Maximum	16	16
Mean	7.37	7.78
Standard Deviation	2.90	2.99
Skewness	0.30	0.35
Kurtosis	-0.11	-0.26
Internal Consistency[a]	0.60	0.61
Standard Error of Measurement	0.18	0.15

[a] Cronbach's Alpha

computer-based field test sample is somewhat more flat than the score distribution of the paper-and-pencil field test sample. The internal consistency of the paper-and-pencil field test and the computer-based field test are comparable, as are the standard errors of measurement. The low values of the internal consistencies of the two forms are likely due to the small number of items and the fact that the numerical response scores were essentially constant.

Issue of nonresponse to the survey. Two variables of interest were students' gender and preference for response mode. While the gender of all students was known, not all students completed the computer familiarity survey, either because there was not enough time following completion of the field test or the teacher did not administer the survey to the students during the next class period. Seven (2.8%) students in the paper-and-pencil sample did not complete the survey; in contrast, 120 (31.7%) students in the computer-based sample did not complete the survey. To determine the influence of this non-response on performance, Students' *t*-test for independent groups was conducted to test the difference in performance between the students who wrote the computer-based field test and completed the survey and the students who wrote the computer-based field test and did not complete the survey. The results of this analysis are

reported in Table 9. As shown, the mean of the students who completed the survey is not significantly different from the mean of the students who did not complete the survey. The effect size, 0.26, is small (Cohen, 1988). Thus the results for the students who responded to the computer-based field test and completed the survey are generalizable to the students who responded to the computer-based field-test but did not complete the survey.

Table 9

Comparison of Performance of Students who Responded to the Computer-Based Field Test: Completed Survey versus did not Complete Survey

	n	$\overline{X}_{\cdot}$	$\hat{\sigma}_X$	*t*	*p*
Completed	258	7.97	2.95	0.67	> 0.05
Not Completed	120	7.22	3.03		

Influence of response mode, gender, and test mode preference on student performance. A 2 x 2 x 3 (response mode-by-gender-by-test mode preference survey question) fully crossed fixed effects ANOVA was conducted to determine if there was a significant interaction between gender, response mode, and test mode preference; a significant interaction between response mode and gender, response mode and test mode preference, and gender and test mode preference; and a main effect due to response mode, gender, and test mode preference. In total, there were five students who did not respond to this survey question—four students who completed the paper-and-pencil field test and one student who completed the computer-based field test. Since using the mean to replace these missing responses may not have been a true representation of their actual test mode preference, these five students were excluded from the analysis. The number of students, mean, and standard deviation in the cells of this analysis are reported in Table 10 and the results of the ANOVA are presented in Table 11.

Levene's test of equality of error variances was not significant, ($F(11,486) = 0.78$, $p > 0.05$), so the assumption of homogeneity of variance was met. Given the unbalanced design and

Table 10

Description of Subsamples' Performance Across Response Mode, Gender, and Test Mode Preference

		PP			CB		
Test Preference	Gender	*n*	$\bar{X}_.$	$\hat{\sigma}_X$	*n*	$\bar{X}_.$	$\hat{\sigma}_X$
PP	Males	82	7.38	2.76	49	7.88	3.05
	Females	101	8.03	3.05	88	8.03	3.08
Either	Males	22	6.77	2.65	33	8.09	2.83
	Females	14	6.57	2.65	48	8.33	3.19
CB	Males	15	5.47	2.50	17	7.09	2.15
	Females	7	6.00	2.45	22	8.41	2.45

Table 11

ANOVA Results for the Mathematics 30-2 Field Test Across Response Mode, Gender, and Test Mode Preference

	n	$\bar{X}_.$	$\hat{\sigma}_X$	*F*	*p*
Response Mode				12.78	0.00
PP	241	6.70	2.92		
CB	257	7.97	2.96		
Gender				1.61	0.20
Males	223	7.11	2.83		
Females	275	7.56	3.02		
Test Mode Preference				3.26	0.04
PP	320	7.83	2.99		
Either PP or CB	117	7.44	2.99		
CB	61	6.74	2.58		
Response Mode by Gender				0.12	0.73
Response Mode by Test Mode Preference				3.17	0.04
Gender by Test Mode Preference				0.41	0.67
Response Mode by Gender by Test Mode Preference				0.42	0.66

the fact that the three main effects—response mode, gender, and test mode preference—were considered to be equally important, Type III sums of squares were used.

As shown in Table 10, the majority of students who responded to the paper-and-pencil field test indicated that they preferred to respond to a test presented in a paper-and-pencil format. Likewise the majority of students who responded to the computer-based field test indicated that they preferred to respond to a test presented in a paper-and-pencil format. For both the paper-and-pencil sample and the computer-based sample, the next greatest number indicated either format, and the fewest students indicated that they preferred to respond to a computer-based test. As shown in Table 11, the three way interaction among response mode, gender, and test mode preference was not significant. Hence, the differences among the cell means reported in Table 10 are not significant. Whereas the two way interaction between response mode and test mode preference was significant, the two way interactions between response mode and gender and gender and test mode preference were not. Lastly, whereas the main effects for response mode and for test mode preference were significant, the main effect for gender was not (see Table 11).

However, the effect sizes were all small (partial $\eta^2 = 0.013$ for the interaction between response mode and test mode preference, partial $\eta^2 = 0.026$ for response mode, and partial $\eta^2 = 0.013$ for test mode preference; Cohen, 1988). Therefore, students were not disadvantaged due to response mode, their gender, or their test mode preference on the field test considered in this study.

Computer Familiarity Survey Results

The results for the computer familiarity survey are presented and discussed in two subsections. The first subsection contains the analysis of the differences in survey responses for

the questions that were asked of both samples. The second subsection examines the responses of the students in the computer-based field test sample to the items asked only of those students.

Results for survey questions asked of both samples. The questions on the computer familiarity survey that were completed by students who responded to the paper-and-pencil field test and the students who responded to the computer-based field test were questions 2 through 8; also, question 15 (computer-based field test) and question 13 (paper-and-pencil) were related. The Chi-square test statistic goodness of fit test was used to compare the fit of the proportions of students in the computer-based field test sample who responded to each alternative to the proportions of students in the paper-and-pencil field test sample who responded to the corresponding alternative, given the paper-and-pencil field test was the only form used in previous years. As discussed in Chapter 4, this study followed the recommendations made by Roscoe and Byars (1971), Convor (1974), and Camilli and Hopkins (1978, 1979) with regards to expected cell counts and the Chi-square test statistic. For example, in question 7, "strongly disagree" and "disagree" were combined as were "agree" and "strongly agree." However, in some cases where the expected cell count was less than 2, it was not logical to combine alternatives. For example, in question 4, it would not be logical to combine "once every few months" with "never." In these cases where it was not logical to combine alternatives, an asterisk is included next to the proportions where the expected cell count is less than 2 and the Chi-square test statistic should be interpreted with caution.

The results for the common survey items are presented in Tables 12 through 19. In some cases, parts of questions had to be shortened due to space limitations. The sample size for the paper-and-pencil field test was 245 and the sample size for the computer-based field test was 258.

As shown in Table 12, only small proportions of students who responded to the paper-and-pencil field test and students who responded to the computer-based field test rated their experience as weak or poor for a desktop, laptop, tablet, and smartphone. Of the remaining

Table 12

Chi-square Test Statistic Comparison for Question 2: *How would you rate your computer experience on a*

		Alternative/Proportion Within Response Mode				
	Response Mode	Weak	Poor	Moderate	Excellent	Sig.*
desktop	PP	0.02	0.04	0.55	0.39	No
	CB	0.00	0.04	0.53	0.43	
laptop	PP	0.02	0.06	0.49	0.43	Yes
	CB	0.00	0.03	0.61	0.36	
tablet	PP	0.04	0.16	0.53	0.27	Yes
	CB	0.06	0.10	0.68	0.16	
smartphone	PP	0.07	0.06	0.40	0.47	Yes
	CB	0.08	0.08	0.61	0.23	

* $p < 0.05$.

students, the larger majority of students in both samples rated their experience as moderate rather than excellent on each of the devices. However, the Chi-square test statistic was significant for three of the four devices because of the differences in the distributions for “moderate” and “excellent” between the students who responded to the paper-and-pencil field test and the students who responded to the computer-based field test. For these three devices, whereas a larger proportion of students who responded to the computer-based field test than students who responded to the paper-and-pencil field test rated their experience as moderate, a smaller proportion of students who responded to the computer-based field test than students who responded to the paper-and-pencil field test rated their experience as excellent. It is also interesting to note the relatively small proportion of students in the computer-based field test

sample who rated themselves as excellent on the smartphone, given that a majority of students own a smartphone, as shown in Table 13.

Table 13

Chi-square Test Statistic Comparison for Question 3: Which of the following digital devices do you have access to at home?

	Response Mode	Alternative/Proportion Within Response Mode			
		Do not own	Access but do not own	Own	Sig*
desktop	PP	0.14	0.42	0.44	No
	CB	0.17	0.43	0.40	
laptop	PP	0.06	0.22	0.72	No
	CB	0.08	0.16	0.76	
tablet	PP	0.25	0.39	0.36	Yes
	CB	0.36	0.31	0.33	
smartphone	PP	0.05	0.05	0.90	No
	CB	0.10	0.06	0.84	

* $p < 0.05$.

On the actual survey, students were instructed to leave the row blank in question 3 if they did not have access to or own each specific device. In total, only 1% of all the students left all the rows blank, meaning that they do not have access to or own any of these devices. The Chi-square test statistic was significant for only the tablet. Whereas a smaller percentage of students who responded to the paper-and-pencil field test indicated they did not own a tablet than students who responded to the computer-based field test, a greater percentage indicated they had access to or owned a tablet.

As shown in Table 14, there were no significant differences in the responses patterns for each of the items dealing with the students' use of a computer outside of school. The dominant uses are for social purposes and to search the internet (over 90% at least two times per week). Students do download music slightly more often than they do homework on the computer (68% versus 55 to 58%). Lastly, they do not play games as often as the other activities.

Table 14

Chi-square Test Statistic Comparison for Question 4: In general, how often do you use a computer outside school to do each of the following?

		Alternative/Proportion Within Response Mode					
	Response Mode	Never	Once every few months	2-3 times/mo.	2-3 times/wk.	Almost every day	Sig*
Play games	PP	0.36	0.21	0.08	0.13	0.22	No
	CB	0.34	0.22	0.11	0.19	0.14	
Chat electronically	PP	0.05	0.04	0.07	0.14	0.70	No
	CB	0.06	0.05	0.06	0.14	0.69	
Use social media	PP	0.03	0.02	0.04	0.08	0.83	No
	CB	0.04	0.02	0.03	0.16	0.75	
Download music, videos, etc.	PP	0.04	0.06	0.22	0.30	0.38	No
	CB	0.02	0.07	0.23	0.33	0.35	
Search the internet**	PP	0.03	0.00	0.02	0.13	0.82	No
	CB	0.01	0.00	0.02	0.19	0.78	
Do homework	PP	0.07	0.08	0.30	0.35	0.20	No
	CB	0.05	0.11	0.26	0.38	0.20	

* $p < 0.05$; ** indicates at least one cell with an expected count less than 2.

As shown in Table 15, the students use computers more for subjects other than mathematics in school. Further, whereas a greater proportion of the students who responded to the paper-and-pencil field test than students who responded to the computer-based field test indicated they never use computers at school to do mathematics (52% versus 42%), a smaller percentage of students who responded to the paper-and-pencil field test than who responded to

Table 15

Chi-square Test Statistic Comparison for Question 5: In general, how often do you use a computer at school for any of the following?

		Alternative/Proportion Within Response Mode					
	Response Mode	Never	Once every few months	2-3 times/mo.	2-3 times/wk.	Almost every day	Sig*
For mathematics	PP	0.52	0.20	0.16	0.10	0.02	Yes
	CB	0.42	0.35	0.12	0.09	0.02	
For any other subject	PP	0.03	0.20	0.36	0.29	0.12	No
	CB	0.07	0.17	0.37	0.31	0.08	

* $p < 0.05$

the computer-based field test indicated they did mathematics on the computer once every few months (20% versus 35%). That is, only about a quarter of the students in both groups use the computer for mathematics at least 2 or 3 times per month. This may be because students in Mathematics 30-2 are required to use a graphing calculator for their diploma examination, and these calculators can do many of the same mathematical functions that a computer can do. In contrast, there is greater use of the computer for the other subjects; more than 75% of the students used the computer more than once every few months for other subjects. The greater use may be due to the need for a program or application that the computer has that is pertinent for the subject (such as a word processing program or to search the internet for a particular topic).

As shown in Table 16, there were no significant differences between the response patterns of the students who responded to the paper-and-pencil field test and the students who responded to the computer-based field test for the five mathematics activities that students might do on the computer as part of their math homework. While perhaps not surprising given the students have their own graphing calculators, between 5 and 7 students out of 10 responded "no"

Table 16

Chi-square Test Statistic Comparison for Question 6: Do you use a computer to do any of the following activities when you are doing math homework?

			Alternative/Proportion Within Response Mode			
	Response Mode	No	Only at School	Only at Home	At Home and School	Sig*
Practice items on Quest A+	PP	0.61	0.21	0.10	0.08	No
	CB	0.57	0.24	0.11	0.08	
Look up math information	PP	0.54	0.06	0.26	0.14	No
	CB	0.56	0.07	0.20	0.17	
Go to math teacher's website	PP	0.71	0.04	0.13	0.12	No
	CB	0.65	0.08	0.10	0.17	
Perform calculations	PP	0.60	0.06	0.09	0.25	No
	CB	0.64	0.05	0.06	0.25	
Go to other math websites	PP	0.65	0.04	0.23	0.08	No
	CB	0.64	0.06	0.23	0.07	

* $p < 0.05$

for each activity. In contrast, what is surprising is the low proportion of students who wrote the computer-based field test yet indicated that they had not done the practice items on Quest A+. As indicated in Chapter 3, their teachers received the *Digital Field Test Instructions: Diploma Examination and Achievement Testing Programs*, which gives instructions on how to set up a practice test run for students on Quest A+. Further, these practice items are accessible on Alberta Education's public website.

As shown in Table 17, the Chi-square test statistic was significant for the categories pertaining to ease of reading the formula sheet and the font size. Whereas a greater proportion of the students who responded to the paper-and-pencil field test than students who responded to the computer-based field test agreed/strongly agreed that it was easy to use the formula sheet (0.82 versus 0.70), a smaller proportion of students who responded to the paper-and-pencil field test than students who responded to the computer-based field test strongly disagreed that the formula

Table 17

Chi-square Test Statistic Comparison for Question 7: Thinking about the field test you just wrote, please indicate whether you agree with the following statements.

		Alternative/Proportion Within Response Mode			
	Response Mode	Strongly Disagree/ Disagree	Neither Agree nor Disagree	Agree/ Strongly Agree	Sig *
Instructions were clear	PP	0.01	0.05	0.94	No
	CB	0.01	0.03	0.96	
Easy to read formula sheet	PP	0.04	0.14	0.82	Yes
	CB	0.19	0.11	0.70	
Easy to enter answers**	PP	0.01	0.05	0.94	No
	CB	0.01	0.05	0.94	
Easy to transfer information to scrap	PP	0.03	0.14	0.83	No
	CB	0.06	0.12	0.82	
Easy to read the font	PP	0.00	0.06	0.94	Yes
	CB	0.06	0.08	0.86	
Easy to read the graphics	PP	0.01	0.05	0.94	No
	CB	0.03	0.07	0.90	

* $p < 0.05$; ** indicates at least one cell with an expected count less than 2.

sheet was easy to read (0.04 versus 0.10). The formula sheet for the paper-and-pencil field test was a tear-out page at the back of the test booklet that the students could place to the side for easy reference; the formula sheet for the computer-based field test was in a drop-down menu that the students had to click on each time they wanted to use it. While a similar pattern of responses occurred for the second item with statistical significance, approximately 90% of the students in both groups agreed/strongly agreed that it was easy to read the font. The similarities in responses across the two response modes to the category "easy to transfer information to scrap paper" is somewhat surprising, given that students who wrote the computer-based field test would have had to transfer information from the computer screen to their scrap paper, whereas students who wrote the paper-and-pencil field test would not necessarily have to transfer information since they are allowed to write in the booklet. Indeed, the students in both response modes did not have any trouble with transferring information to their scrap paper.

Earlier in this chapter, the ANOVA revealed that there was a significant interaction between response mode and test mode preference (question 8 on the computer familiarity survey), although the effect size was small. As shown in Table 18, the Chi-square test statistic is again significant ($p < 0.05$). While slightly more than three-quarters of the students who responded to the paper-and-pencil field test indicated they preferred paper-and-pencil mathematics tests, slightly more than half of the students who responded to the computer-based field test indicated so. In contrast, approximately one third of the students who responded to the computer based field test indicated that the format did not matter, and about one sixth of the students who responded to the paper-and-pencil field test indicated that the format did not matter. Lastly, the proportions of students in both samples who indicated that they would prefer to take a mathematics test online were small (0.09 and 0.15).

Table 18

Chi-square Test Statistic Comparison for Question 8[a]: If I had a choice, I would prefer taking a math test

Response Mode	using paper and pencil	either online or using paper and pencil	online
PP	0.76	0.15	0.09
CB	0.53	0.32	0.15

Note: [a]Pearson Chi-Square χ^2 (3) = 30.12, $p < 0.05$

As can be seen in Table 19, the response distributions for the survey question on students' perceptions of the mark they would have received if they had written the field test in the other response mode were significantly different across the two response modes. While a greater proportion of students who wrote the computer-based field test than students who responded to the paper-and-pencil field test indicated that they would receive the same mark

Table 19

Chi-square Test Statistic Comparison for Questions 13(CB) and 15(PP) [a]: If I completed the field test on paper/online, I believe I would have received

Response Mode	a lower mark	the same mark	a better mark
PP	0.38	0.58	0.04
CB	0.08	0.74	0.18

[a]Pearson Chi-Square χ^2 (2) = 76.15, $p < 0.05$

had they had written the field test in the other response mode (0.74 vs. 0.58), and a greater proportion of students who responded to the computer-based field test than students who responded to the paper-and-pencil field test indicated that they would receive a higher mark if they had written the field test in the other response mode (0.18 versus 0.04), a greater proportion of students who wrote the paper-and-pencil field test than students who wrote the computer-based field test indicated they would receive a lower mark had they written in the other response mode.

Although the common survey questions were designed to explain any possible differences between the two samples, as can be seen from the results presented above, there were

very few statistical differences between the responses by the two samples. The only significant differences were in response to laptop experience, tablet experience, and smartphone experience (question 2), tablet access (question 3), the use of computers for math at school (question 5), the use of the formula sheet and the ease of reading the font size (question 7), test preference (question 8), and mark prediction for other response mode (question 13, CB; question 15, PP).

Results of survey questions asked of only of students who responded to the computer-based field test. Questions 9 through 12 and question 14 on the survey questionnaire were completed only by students who wrote the computer-based field test.

The proportions of students who responded to each alternative in questions 9 through 11 are presented in Table 20. Approximately 64% of the students who wrote the computer-based field test had previously taken a test online. If they responded "yes" to question 9, students were asked to specify when and where in the space provided below the question. Although many students who answered yes left this space blank, for those who did respond, the majority indicated that their previous experience was either a classroom assessment or an Alberta Education field test. All but 2% of the students wrote the computer-based field test on the school's computers. However, one third of these students did not feel confident writing the field test online. This may be related to the fact that not many students who wrote the computer-based

Table 20

Questions 9 - 11

	Alternative/Proportion	
Question 9 – Have you taken a test online before?	No	Yes
	0.36	0.64
Question 10 – I wrote the field test	on the school's computer	on my own laptop
	0.98	0.02
Question 11 – I felt confident writing the field test online.	No	Yes
	0.33	0.67

field test had previous experience with the practice items for Mathematics 30-2 on Quest A+ (survey question 6).

As indicated in Table 21, when asked about the ease of taking a computer-based test in comparison to taking a paper-and-pencil test, nearly half (0.47) indicated taking either was about the same; slightly more than a third (0.36) felt taking a computer-based test was more difficult, although about a third of these students indicated things got easier as they progressed through the computer-based field test; and slightly less than a fifth (0.17) indicated that taking a computer-based test would be easier than taking a paper-and-pencil test.

Table 21

Question 12: Taking the field test on Quest A+ was

	Proportion
more difficult than taking a paper-and-pencil test	0.24
more difficult at first, but became easier as I got used to it	0.12
about the same as taking a paper-and-pencil test	0.47
easier than taking a paper-and-pencil test	0.17

Given that the mean for the students who wrote the computer-based field test was not statistically different than the mean for the students who wrote the paper-and-pencil field test, it is interesting that approximately 36% felt that taking the computer-based field test was more difficult than a paper-and-pencil field test in some way, as shown in Table 21. Yet in Table 19, only 18% of those who wrote the computer-based field test felt they would have received a higher mark on the paper-and-pencil field test. It is also interesting to note that although 47% of the students in the computer-based field test sample thought that taking a field test on Quest A+

was about the same as taking one on paper, a greater proportion (74%) in Table 19 indicated that they felt that would have received the same mark, had they taken the paper-and-pencil field test.

Although approximately 70% of the students who wrote the computer-based field test agreed or strongly agreed that it was easy to use the formula sheet on the field test they had just written (question 7), as shown in Table 22, 49% believe that it is easier to use the formula sheet on the paper-and-pencil field test. Likewise, approximately 82% either agreed or strongly agreed that it was easy to transfer information to the scrap paper (question 7), whereas the responses to question 14 indicate that the students feel that it is easier to use scrap paper with a paper-and-pencil field test. It is interesting to note that many felt that it was quicker to complete the computer-based field test (44%), although very few felt that it was easier to read the computer-based field test (16%).

Table 22

Question 14: Compare your experience writing paper-and-pencil tests to completing the field test on Quest A+. Which is

	PP	they are the same	CB
easier to read?	0.27	0.57	0.16
less tiring?	0.21	0.45	0.34
quicker to complete?	0.18	0.38	0.44
easier to use if you need to refer to the formula sheet?	0.49	0.34	0.17
easier to use if you need to use scrap paper?	0.44	0.41	0.15

Chapter 6

Discussion and Conclusions

Chapter 6 is organized in seven sections. In the first section, the purpose of the study and a summary of the methods are provided. The key findings are provided in the second section. The findings are then discussed in the third section. The limitations of the study are provided in the fourth section, followed by the conclusions drawn in light of the limitations in the fifth section. Implications for practice and recommendations for future research are presented in the last two sections.

Purpose and Summary of Methods

The purpose of the present study was to complete a secondary analysis of the data collected for the Mathematics 30-2 field test that was administered in a paper-and-pencil format and a computer-based format, and to examine the comparability of the psychometric properties and the students' scores for the two response modes. The research questions addressed included:

1. Are the psychometric properties of the items comparable between the paper-and-pencil field test and the computer-based field test?
2. Does the reliability of the paper-and-pencil field test differ from the reliability of the computer-based field test?
3. Do students' scores differ significantly between the paper-and-pencil field test and the computer-based field test with at least a moderate effect size?
4. Do male and female students' scores differ significantly with at least a moderate effect size between the paper-and-pencil field test and the computer-based field test?
5. To what degree does student performance on a digital Mathematics 30-2 field test depend upon computer familiarity and prior online testing experience?

To address research question 1, the psychometric properties of each item were examined using classical test theory item analysis. First, the distributions of responses for the paper-and-pencil field test and for the computer-based field test across the alternatives of the items were compared using the Chi-square goodness of fit statistic. Then, the difficulty and discrimination for the correct option were specifically compared between the two forms using Cox's Index for difficulty and Fisher's Z and the z-test statistic for discrimination. The remaining research questions involved working with the paper-and-pencil and computer-based samples. First, the psychometric properties, including reliability, of the two forms were compared (Research Question 2). Next, a 2 x 2 x 3 (response mode-by gender-by-test mode preference survey question) fully crossed fixed effects ANOVA was conducted to address research questions 3 and 4. Given the cells sizes were not equal, Type III sums of squares in which each main and interaction effect is adjusted for each other was used. Lastly, to address the fifth research question, a series of analyses were conducted. First, the difference in performance between the students in the computer-based field test sample who answered "yes" and the students in the computer-based field test sample who answered "no" to the survey question "Have you taken a test online before today?" was tested using Welch's correction to Student's t-test statistic for two independent groups, given unequal sample size and lack of homogeneity of variance. Second, the responses to the individual common survey questions were compared between the students in the paper-and-pencil and computer-based samples using the Chi-square goodness of fit test. For the remaining survey questions that were answered by only the students who responded to the computer-based field test, the frequencies were computed and examined.

Summary of Findings

The results at the item level were as follows:

- except for numerical-response item 3, the distribution of the proportions of students choosing the alternatives for each item on the computer-based field test form was not significantly different from the distribution of the proportions of students choosing the alternatives for each item on the paper-and-pencil field test. The proportion of students who wrote the computer-based field test and determined the correct answer to numerical-response item 3 was significantly greater than the proportion of students who wrote the paper-and-pencil field test and determined the correct answer, although both proportions were low (0.08 for paper-and-pencil and 0.18 for computer-based); and
- the *p*-values of numerical-response item 3 and multiple-choice item 5, and the *CRPB* values for numerical-response item 3 differed significantly ($p < 0.05$). However, while the effect size for the difference in *p*-values for multiple-choice item 5 and the effect size for the difference in *CRPB*s for numerical-response item 3 were both small, the effect size for the difference in *p*-values for numerical-response item 3 was moderate (Cohen, 1988).

The results at the test level were as follows:

- students' performance on the computer-based field test was not affected by prior online testing experience, and the two subsamples were combined for all subsequent analysis;
- the reliability (Cronbach's alpha) of the paper-and-pencil field test was 0.60, and the reliability of the computer-based field test was 0.61;
- whereas the three way interaction between response mode, gender, and test mode preference, the two way interactions between response mode and gender and

gender and test mode preference, and the main effect of gender were not significant at the 0.05 level of significance, the two way interaction between response mode and test mode preference and the main effects of response mode and test mode preference were. However, the effect sizes were all small (Cohen, 1988).

The results of the computer familiarity survey were as follows:

- the students who responded to the computer-based field test and to the paper-and pencil field test rated their experience with a desktop in the same way, with 94% and 96% indicating at least "moderate" on the four point scale. In contrast, while a greater percentage of students who responded to the computer-based field test rated their experience with a laptop, tablet, and smartphone as moderate, a greater percentage of students who responded to the paper-and-pencil field test rated their experience with a laptop, tablet, and smartphone as excellent.
- the majority of students in both samples owned or had access to a desktop, laptop, and a cellphone. However, a smaller percentage of students who responded to the paper-and-pencil field test indicated they did not own a tablet (25% compared to 36% for the computer-based field test) whereas a smaller percentage of students who responded to the computer-based field test indicated they had access to or owned a tablet (64% compared to 75% for the paper-and-pencil field test).
- nearly a quarter of all students in both samples used the computer for mathematics at least 2 or 3 times per month. A greater proportion who wrote the paper-and-pencil field test than the computer-based field test (0.52 versus 0.42) indicated they never use a computer at school for mathematics but a smaller proportion

who wrote the paper-and-pencil field test than the computer-based field test (0.20 versus 0.35) indicated they use a computer for mathematics once every few months. In contrast, 97% of the paper-and-pencil field test and 93% of the computer-based field test used computers at school for other subjects.

- the proportion in both samples who strongly agreed/agreed the instructions were clear, it was easy to enter answers, and it was easy to transfer information were similar. A greater proportion of students who wrote the paper-and-pencil field test than the computer-based field test strongly agreed/agreed it was easier to read the formula sheet and it was easier to read the font; however, a smaller proportion of students who wrote the paper-and-pencil field test than the computer-based field test strongly disagreed/disagreed in those same two categories.
- while slightly more than three-quarters of the students who responded to the paper-and-pencil field test indicated they preferred paper-and-pencil mathematics tests, slightly more than half of the students who responded to the computer-based field test indicated so. In contrast, approximately one sixth of the students who responded to the paper-and-pencil field test indicated that the format did not matter, whereas about one third of the students who responded to the computer-based field test indicated that the format did not matter. Lastly, the proportions of students in both samples who indicated that they would prefer to take a mathematics test online were small (0.09 and 0.15). The results for this survey question are in agreement with the results of the ANOVA, where there was a significant interaction between response mode and test mode preference, although the effect size was small.

- a greater proportion of students who wrote the computer-based field test than who wrote the paper-and-pencil field test indicated they would receive the same mark or a better mark if they wrote in the other mode; in contrast, a greater proportion of students who wrote the paper-and-pencil field test than who wrote the computer-based field test indicated they would receive a lower mark if they wrote in the other mode.

Discussion

Currently, the diploma examinations in Alberta are administered in a paper-and-pencil format. However, there are plans to move to computer-based diploma examinations by the fall of 2017 (Alberta Education, 2013a). However, as in other jurisdictions, it is likely that not all students will be able to respond online due to the lack of needed computer technology in some schools. Further, there is a need to measure progress from one year to another (Alberta Education, 2013c). Consequently, it is necessary that the paper-and-pencil form of a diploma examination and a computer-based form of a diploma examination be interchangeable so that students are not disadvantaged in any way by the response mode, and so that the interpretations of the test scores are equally valid. If it is assumed that scores are comparable when in fact they are not, wrong decisions may be made, which directly violates fairness principles (American Educational Research Association et al., 1999; *Principles for Fair Student Assessment Practices for Education in Canada*, 1993; van Lent, 2008) as well as Alberta Education's goal of maintaining consistent standards over time (Alberta Education, 2013c).

The results for the field test considered in this study suggest that there will essentially be no difference in performance between students who respond to a paper-and-pencil form and students who respond to a computer-based form, and that factors such as gender, test preference,

online experience, and ease in completing the field test will not affect performance. This finding agrees with most other comparability studies for mathematics in the literature. At the test level, only two of the ten studies reported statistically significant differences, although neither reported an effect size. Seven studies in the literature examined differences at the item level; five of these studies reported statistical differences, some favoring paper-and-pencil and some favoring computer-based tests. However, where effect sizes were reported, they were small (Cohen, 1988). Only three comparability studies considered gender, but only one reported a statistical difference, although no effect size was given. And only two comparability studies considered computer familiarity, with the authors in both cases reporting that familiarity did affect performance, but no effect sizes were stated. As for the meta-analyses completed by Wang et al. (2007) and Kingston (2009), the majority of comparisons were not significant or, if significant, had small effect sizes (Cohen, 1988).

It was somewhat surprising that ease of transferring information between the field test and scrap paper was not influential. Russell et al. (2003) concluded that validity is threatened when students experience difficulty in accessing scratch paper to perform calculations. Kingston (2009) stated that switching between scratch paper and the computer is spatially much larger and utilizes different planes than working out a problem on paper, yet over 80% of the students in the present study agreed or strongly agreed that it was easy to transfer information between the field test and scrap paper regardless of mode.

Limitations of the Study

Unfortunately, not all students who responded to the computer-based field test completed the computer familiarity survey. While there was no difference in the performance of the students who responded to the computer-based field test and completed the survey questionnaire

and the performance of the students who responded to the computer-based field test but did not complete the questionnaire, there may have been some differences on the items of the questionnaire had all the students completed the questionnaire.

As mentioned in Chapter 3, although the planned analysis included the Mathematics 30-2 Diploma Examination as a covariate, the severe flooding in southern Alberta led to a number of school closures. Consequently, the Mathematics 30-2 Diploma Examination scores for the students in the sample schools that were forced to close were not available. Removing these students' field test scores and survey results led to an overall reduction of 40% of the sample, and the representativeness of the paper-and-pencil and computer-based samples was adversely impacted. Therefore, given the care given to selecting the schools for field tests, the covariate was dropped to allow analyses of the full field test sample.

Conclusions

The field test for Mathematics 30-2 considered in this study was administered in paper-and-pencil format and in a computer based format. There was no discernible difference in the performance of the students who responded to the paper-and-pencil form and the students who responded to the computer based form. This finding agrees with the majority of comparability studies conducted to date. However, the findings cannot simply be generalized to other testing situations, including Alberta Education field tests and diploma examinations not only for Mathematics 30-2, but for all other subject areas' diploma examinations as well as the provincial achievement tests. As indicated in the literature, not all studies lead to finding that paper-and-pencil and online forms of the same test are comparable. That is, comparability is unique to each testing situation, and a comparability study must be conducted anytime a test is offered in two

different response modes. However, the methodology outlined and implemented in this study can be used as a model for future comparability studies.

Implications for Practice

As just stated, while the process used in this study is generalizable, the findings are not. It is strongly recommended that a comparability study be conducted whenever the same test is administered in a paper-and-pencil form and an online form. Only by doing so can the testing agency be sure that the forms are or are not comparable. If comparable, then the scores for the students who wrote either form can be validly interpreted in terms of the same construct. If the forms are not comparable, then testing agencies should be prepared to make adjustments in the two score scales so that they are on a common scale. Toward this latter end, it is recommended that testing agencies have a procedure ready to use to make the adjustment when needed.

A second implication concerns the manner in which field testing is conducted. In the present case, the field tests are administered prior to the diploma examinations. The finding of low performance found for both the paper-and-pencil and the computer-based field test, particularly for the numerical-response items, raises questions about the motivation of the students. This finding, taken with the low internal consistency of both forms of the field test, leads to scores that cannot be validly interpreted in terms of what students actually know. It is therefore recommended that the field testing procedure used be changed, if logistically possible, so that the items are embedded in the diploma examinations. This likely will require several forms, each with the same set of operational items to be scored but a different set of field test items so that a sufficient number of items are field tested.

Recommendations for Future Research

Although, with one exception at the item level, there were no differences at the item and test level in this study, it cannot be surmised that the students who responded to the paper-and-pencil field test and the computer-based field test used the same cognitive strategies, problem solving skills, and reasoning skills while they were responding to each item. Therefore it is recommended that future research should probe the links between student's thinking, their test-taking behavior, and the assessment mode (Johnson & Green, 2006; Sawacki, 2001) by utilizing think aloud interviews followed by protocol analysis (Ericson & Simon, 1993). The researcher could then examine whether students used similar cognitive strategies, problem solving skills, and reasoning skills on the two response modes. At the same time, given that it is also possible that some errors on a computer-based test may be transcription errors instead of conceptual errors, errors such as this could be addressed in the interviews (Johnson & Green, 2006; Russell et al., 2003).

References

Alberta Education (2007). *Science 30 Online Testing*. Edmonton: Accountability and Reporting Division, Alberta Education

Alberta Education (2013a). *Curriculum redesign: June 2013 update.* Retrieved from http://www.education.alberta.ca/media/7065028/june%202013%20curriculum%20redesign%20update%20-%20june%203%202013%20-%20final.pdf

Alberta Education (2013b). *General information bulletin*. Retrieved from http://education.alberta.ca/admin/testing/diplomaexams/diplomabulletin.aspx

Alberta Education (2013c). *Information bulletin: Mathematics 30-2, 2012-2013 diploma examinations program* Retrieved from http://education.alberta.ca/media/6738245/18%20math30-2%20bulletin%202012-13%20signoff.pdf

Alberta Education (2013d). *Mathematics 30-2 Assessment Standards and Exemplars*. Retrieved from http://www.education.alberta.ca/media/6758262/16%20math30-2%20standardsexemp2012-13signoff.pdf

American Psychological Association Committee on Professional Standards and Committee on Psychological Tests and Assessment (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: APA

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences*, 22(1), 112-117. doi: 10.1016/j.lindif.2011.11.005

Arce-Ferrer, A.J. & Guzman, E. M. (2009). Studying the equivalence of computer-delivered and paper-based administrations of the Raven Standard Progressive Matrices Test. *Educational and Psychological Measurement, 69*, 855-867

Bejar, I.I, Lawless, R.R., Morley, M.E., Wagner, M.E., Bennett, R.E. & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning and Assessment, 2(3)*. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1663

Bennett, R.E., Goodman, M., Hessinger, J., Ligget, J., Marshall, G., Kahn, H., & Zack, J. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behaviour, 15*, 283-294

Bennett, R.E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *The Journal of Technology, Learning, and Assessment*, 6, 1-38. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/index

Bugbee, A.C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28*, 282-299

Camilli, G., and Hopkins, K.D., (1978). Applicability of chi-square 2 x 2 contingency tables with small expected frequencies. *Psychological Bulletin*, 85, 163-167.

Camilli, G., and Hopkins, K.D., (1978). Testing for association in 2 x 2 contingency tables with very small sample sizes. *Psychological Bulletin*, 86, 1011-1014.

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

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd edition). L. Erlbaum Associates (pp. 531–535)

Collerton, J., Collerton, D., Arai, Y., Barrass, K., Eccles, M., Jagger, C., McKeith, L., Saxby, B.K., and Kirkwood, T. (2007). A comparison of computerized and pencil-and-paper tasks in assessing cognitive function in community-dwelling older people in the Newcastle 85+ pilot study. *Journal of American Geriatric Society, 55*, 1630-1635.

Conovor, W. J. (1974). Some reasons for not using Yates' Continuity Correction on 2 x 2 contingency tables. *Journal of the American Statistical Association*, 69, 374-382.

Cox, D.R. (1970). *Analysis of binary data*. New York: Chapman & Hal/CRC.

Csápo, B., Ainley, J., Bennett, R.E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McCaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143-230). London, New York: Springer Dordrecht Heidelberg, Publishers

Csapó, B., Molnár, G., & Tóth, K. (2009). Comparing paper-and-pencil and online assessment of reasoning skills: a pilot study for introducing TAO in large-scale assessment in Hungary. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: new approaches to skills assessment and implications for large-scale testing* (pp. 120-126) Luxembourg: Office for Official Publications of the European Community. Retrieved from http://www.gesci.org/assets/files/reporttransition.pdf

Drasgow, F. (2002). The work ahead: A psychometric infrastructure for computerized adaptive tests. In C.N. Mills, M.T. Potenza, J. J. Fremer, & W. C. Ward (Eds.) *Computer-based testing: Building the foundation for future assessments* (pp 1 – 35). Mahwah, NJ: Lawrence Erlbaum Associates

Drasgow, F., Luecht, R.M., & Bennett, R. E. (2006). Technology and Testing. In Brennan, R. L. (Ed), *Educational Measurement, 4th edition.* (pp. 471–515)

Dorans, N.J., & Lawrence, I.M. (1990). Checking the statistical equivalence of nearly identical test forms. *Applied Measurement in Education*, 3, 245-254.

Ericson, K. A. & Simon, H.A. (1993). *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT Press

Fisher, R. A. (1958). *Statistical methods and scientific inference* (2nd ed.). New York: Hafner

Foster, D. (2004). *Testing technology: Take one giant step backward*. Retrieved from Caveon website: www.caveon.com/articles/df_article11.htm

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

Gaskill, J. & Marshall, M. (2006). *Comparisons between paper- and computer-based tests: foundations skills assessment – 2001 to 2006 data*. Kelowna, BC: Society for the Advancement of Excellence in Education. Retrieved from http://www.maxbell.org/sites/default/files/038.pdf

Gaver, W.W. (1991). *Technology affordances*. In Proceedings of the special interest group on computer-human interaction (SOGHCI) conference on human factors in computing systems (pp. 79-84). New Orleans

Gierl, M., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8), 757-765. doi:10.1111/j.1365-2923.2012.04289.x

Glasnapp, D.R., Poggio, J., Poggio, A., & Yang, X. (2005). *Student attitudes and perceptions regarding computerized assessment and the relationship to performance in large-scale assessment programs.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, CA.

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York: McGraw Hill

Ghiselli, E.E., Campbell, J.P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: W. H. Freeman and Co.

Greeno, J. G. (1998). The situativity of knowing, learning and research. *American Psychologist*, 53, 5–26.

Grignon, S., Gregoire, C.A., Durand, M., Mury, M., Elie, D., & Chianetta, J.M. (2009). Age-dependent discrepancies between computerized and paper cognitive testing in patients with schizophrenia. *Social Psychiatry and Psychiatric Epidemiology*, *44*, 73-77.

Gulliksen, H. (1950). Item Analysis. In *Theory of mental tests* (pp. 363-395). Hoboken, NJ: John Wiley & Sons.

Hamilton, L.S., Klein, S.P., & Lorié, W. (2000). *Using web-based testing for large-scale assessment.* Santa Monica: RAND Education

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

Heinrich Heine Univeristät Düsseldorf (n.d.). *Correlations: two independent Pearson r's.* Retrieved from http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/user-guide-by-distribution/z/correlations_two_independent_pearson_rs

Herman, J.L, & Linn, R.L., (2013). *On the Road to Assessing Deeper Learning: The Status of Smarter Balanced and PARCC Assessment Consortia* (CRESST Report 823). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://ampersand.gseis.ucla.edu/assets/CRESSTReport8232.pdf

Horkay,N., Bennett, R.E., Allen, N., & Kaplan, B. (2005). Online assessment in writing. In B. Sandene, N. Horkay, R.E. Bennett, N. Allen, J. Braswell, B. Kaplan, et al. (Ed.), *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project* (NCES 2005-475). Washington, D.C.: National Center for Education Statistics, US Department of Education

Higgins, J., Russell, M., & Hoffman, T. (2005). Examining the effect of compute-based passage presentation on reading test performance. *The Journal of Technology, Learning, and Assessment, 3*, 3-35.

IBM Corporation (2012). IBM SPSS Statistics Version 21

International Test Commission (2005). *International guidelines on computer-based and internet delivered testing*. Retrieved from http://www.intestcom.org/itc_projects.htm#ITC Guidelines on Computer-Based and Internet Delivered Testing. Granada, Spain: International Test Commission

Johnson, M. & Green, S. (2006). On-line mathematics assessment: the impact of mode on performance and question answering strategies. *The Journal of Technology, Learning,*

and Assessment, 4(5), 1-35. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/index

Keng, L., McClarty, K. L, & Davis, L.L. (2008). Item-level comparative analysis of online and paper administrations of the Texas assessment of knowledge and skills. *Applied Measurement in Education, 21*, 207-226. doi: 10.1080/08957340802161774

Kim, D. & Huynh, H. (2007). Comparability of computer and paper-and-pencil versions of algebra and biology assessments. *The Journal of Technology, Learning, and Assessment, 6(4)*,1-31. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/index

Kingston, N.M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: a synthesis. *Applied Measurement in Education, 22*, 22-37. doi: 10.1080/08957340802558326

Kolen, M.J. (2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, 6(2), 73-96.

Kolen, M. J. and Brennan, R. L. (2004) *Test equating, scaling, and linking: Methods and practices*. (2nd Ed.). New York: Springer-Verlag

Krőhne, U., & Martens, T. (2011). Computer-based competence tests in the national educational panel study: the challenge of mode effects. *Z Erziehungswiss, 14*, 169-186. doi: 10.1007/s11618-011-0185-4

Leeson, H.V. (2006). The mode effect: a literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*, 1-24

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley Publishing Company.

Lottridge, S., Nicewander, A., & Mitzel, H. (2011). A comparison of paper and online tests using a within-subjects design and propensity score matching study. *Multivariate Behavioral Research, 46,* 544-566. doi: 10.1080/002737171.2011.569408

Lottridge, S., Nicewander, A., Schulz, M. & Mitzel, H. (2008). *Comparability of paper-based and computer-based tests: a review of the methodology*. Monterey, CA: CCSSO Technical Issues in Large Scale Assessment Comparability Research Group. Retrieved from http://www.pacificmetrics.com/white-papers/Comparability_of_Paper-based_and_Computer-based_Tests.pdf

Lynch, R. (2000). Computer-based testing: the test of English as a foreign language (TOEFL). *The Source*, Fall 2000. Retrieved from http://www.usc.edu/dept/education/

Mazzeo, J. & Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: a review of the literature* (ETS RR No. 88-21). New York: College Entrance Examination Board

Maynes, D. (2009). Caveon speaks out on IT exam security: The last five years. Retrieved from Caveon website: http://www.caveon.com/articles/it_exam_security.htm

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

Minnesota Department of Education (2009). *Graduation-required assessment for diploma (GRAD) mathematics comparability study report*. Roseville, MN: Minnesota Department of Education. Retrieved from http://education.state.mn.us/search?q=Mode+comparability+study+report&output=

xml_no_dtd&oe=UTF-8&ie=UTF-8&client=New_frontend&proxystylesheet=New_frontend&site=default_collection

Minnesota Department of Education (2012). *Mathematics Minnesota comprehensive assessment-series III (MCA-III) mode comparability study report*. Roseville, MN: Minnesota Department of Education. Retrieved from http://education.state.mn.us/search?q=Mode+comparability+study+report&output=xml_no_dtd&oe=UTF-8&ie=UTF-8&client=New_frontend&proxystylesheet=New_frontend&site=default_collection

Murphy, P.K., Long, J., Holleran, T., & Esterly, E. (2003). Persuasion online or on paper. *Learning and Instruction, 13,* 511-532

Muter, P. (1996). Interface design and optimization of reading of continuous text. In H. van Oostendorp & S. De Mul (Eds.), *Cognitive aspects of electronic text processing* (p. 161-180). Norwood, NJ: Ablex.

Oregon Department of Education (2007). *Comparability of student scores obtained from paper and computer administrations.* Salem, OR: Oregon Department of Education. Retrieved from http://www.ode.state.or.us/teachlearn/testing/manuals/2007/doc4.1comparabilitytestatopandp.pdf

Paek, P. (2005). *Recent trends in comparability studies*. (Pearson Educational Measurement Research Report 05-05). Retrieved from http://www.pearsonedmeasurement.com/research/research.htm

Parshall, C.G., & Kromrey, J. D. (1993). *Computer testing versus paper-and-pencil testing: An analysis of examinee characteristics with mode effect.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag

Poggio, J., Glasnapp, D.R., Yang, X. , & Poggio, A.J. (2005). A comparative evaluation of score results from computerized and paper & pencil mathematics testing in a large scale state assessment program. *The Journal of Technology, Learning, and Assessment, 3(6),* 4-30. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/index

Pommerirch, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning, and Assessment, 2(6),*1-45. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/index

Pommerich, M. & Burden, T. (2000). *From simulation to application: examinees react to computerized testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Pomplun, M., & Custer, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3 Reading tests. *Educational Computing Research, 32*, 153-166

Pomplun, M., Frey, S., & Becker, D. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement, 62*, 337-354

Rogers, W.T. (2012). *Review of literature: computer-based assessments.* Unpublished manuscript, Department of Educational Psychology, University of Alberta, Edmonton, Canada

Roscoe, J.T., and Byars, J.A. (1971). An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association*, 66, 755-759.

Rowan, B. E. (2010). *Comparability of paper-and-pencil and computer-based cognitive and non-cognitive measures in a low-stakes testing environment.* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3403048)

Russell, M. (1999). Testing on computers: a follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7.* Retrieved from http://epaa.asu.edu/epaa/v7n20

Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives.* Retrieved from http://epaa.asu.edu/epaa/v5n3.html

Russell, M. & Plati, T. (2002). Does it matter with what I write? Comparing performance on paper, computer and portable writing devices. *Current Issues in Education, 5(4),* 24. Retrieved from http://cie.asu.edu/volume5/number4/

Russell, M., Goldberg, A., & O'Connor, K. (2003). Computer-based testing and validity: a look back into the future. *Assessment in Education, 10,* 279-294

Sandene, B., Bennett, R.E., Braswell, J., & Oranje, A. (2005). Online assessment in mathematics. In B. Sandene, N. Horkay, R.E. Bennett, N. Allen, J. Braswell, B. Kaplan et al. (Eds), *Online assessment in mathematics and writing: Reports from NAEP technology-based assessment project (NCES 2005-457).* (pp. v – 67). Washington, DC: U.S. Department of Education, National Centre for Education Statistics

Sireci, S. G. & Zenisky, A.L. (2006). Innovative item formats in computer-based testing: in pursuit of improved construct representation. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp.329-347). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Strader, D.A. (2012). *Comparability of computer delivered versus traditional paper and pencil testing.* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3511477)

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4,* 295-312.

Texas Education Agency (2008). A review of literature on the comparability of scores obtained from examinees on computer-based and paper-based tests. Unpublished Paper, Austin, TX: Author

Thissen, D., Reeve, B.B., Bjorner, J.B., & Chang, C.H. (2007). Methodological issues for building item banks and computerized adaptive scales. *Quality of Life Research, 16,* 109-119

Threfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66,* 335-348. doi: 10.1007/210649-006-9078-5

U.S. Department of Education (2010, September 2). *U.S. Secretary of Education Duncan Announces Winners of Competition to Improve Student Assessments.* Retrieved from http://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse

van Lent, G. (2008). Important considerations in e-assessment. In Scheuermann, F., & Bjornsson (Eds.). *Towards a Research Agenda on Computer-Based Assessment* (pp. 97-103)

Wallace, P., & Clariana, R.B. (2005). Test mode familiarity and performance – gender and race comparisons of test scores among computer-literate students in advanced information systems courses. *Journal of Information Systems Education, 16,* 177-182

Wang, S., Young, M.J., & Brooks, T.E. (2004). *Administration mode comparability study for Stanford diagnostic reading and mathematics tests* (Research Report). San Antonio, TX: Harcourt Assessment.

Wang, S., Jiao, H., Young, M. J., Brooks, T. & Olson, J. (2007). A meta-analysis of testing mode effects in grade K-12 mathematics tests. *Educational and Psychological Measurement, 67(2), 219-237*. doi: 10.1107/0013164406288166

Way, W.D., Davis, L.L., & Fitzpatrick, S. (2006). *Score comparability of online and paper administrations of the Texas assessment of knowledge and skills*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Way, W.D., Lin, C., & Kong, J. (2008). *Maintaining score equivalence as tests transition online: issues, approaches and trends.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Way, W.D., Vickers, D., & Nichols, P. (2008). *Effects of Different Training and Scoring Approaches on Human Constructed Response Scoring.* Paper presented at the meeting of the National Council on Measurement, New York City

Williamson, D. M., Mislevy, R. J., & Bejar, I.I. (2006). Automated scoring of complex tasks in computer-based testing: an introduction. In D. M. Williamson, R.J. Mislevy & I.I. Bejar (Eds), *Automated scoring of complex tasks in computer-based testing* (pp. 1-13). Mahwah, NJ: Erlbaum

Wise, S.L. & Plake, B.S. (1990). Computer-based testing in higher education. *Measurement and Evaluation in Counseling and Development, 23*, 3-10.

Zenisky, A.L. & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362.

Zhang, Y., Powers, D.E., Wright, W., & Morgan, R. (2003). Applying the online scoring network (OLN) to advanced placement program (AP) tests (RM-03-12). Princeton: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-03-12-Zhang.pdf

Appendix A

Name: _____ School: _____

Mathematics 30-2 Field Test Comparability Study Survey

Alberta Education is collecting feedback to assess students' use of computers in mathematics class as well as their computer use in general. Your feedback is greatly appreciated, and your responses will be kept strictly confidential.

1. I wrote the field test ____on Quest A+ ____ on paper
2. How would you rate your computer experience on each of the following?

	Excellent	Moderate	Poor	Weak
desktop				
laptop				
tablet				
smartphone				

The next two questions relate to your computer use at home and outside school.

3. Which of the following digital devices do you have access to **at home**? Please check all that are applicable; if you do not have access to an item, please leave that row blank.

	I have access to but do not own a	I own my own
desktop		
laptop		
tablet		
smartphone		

For the remaining questions, "computer" includes any of these digital devices.

4. In general, how often do you use a computer (desktop, laptop, tablet, or smartphone) outside school to do each of the following?

	Almost every day	2-3 times a week	2-3 times a month	Once every few months	Never
Play computer games					
Chat electronically					
Use social media					
Download music, videos, etc.					
Search the Internet					
Do homework assignments using a word processor, spreadsheet, drawing program, etc.					

5. *Now think about your computer use at school*. In general, how often do you use a computer (desktop, laptop, tablet, or smartphone) at school for each of the following? This can be anywhere in the school and at any time during the school day.

	Almost every day	2-3 times a week	2-3 times a month	Once every few months	Never
For mathematics					
For any other subject					

6. *Now think about your computer use at home or at school when you are doing math homework.* Do you use a computer (desktop, laptop, tablet, or smartphone) to do any of the following activities when you are doing math homework?

	Yes, but only at home	Yes, but only at school	Yes, at home and at school	No
Do practice items on Alberta Education's Quest A+ website				
Look up math information on the Internet				
Go to my math teacher's school website				
Perform calculations				
Go to other math websites to try practice items or get math help				

7. Thinking about the field test you just wrote, please indicate whether you agree with the following statements:

	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
The instructions on how to complete the field test were clear.					
It was easy to use the formula sheet.					
It was easy to enter my answers.					
It was easy to transfer information between the field test and my scrap paper.					
The size of the font was easy to read.					
The graphics were easy to read.					

8. If I had a choice, I would prefer taking a math test
    a. online
    b. using paper and pencil
    c. either online or using paper and pencil

This section is **ONLY** for students who wrote the field test on **Quest A+.** If you wrote the **paper-and-pencil** test, please go to **Question 15**.

9. Have you taken a test online before today?
    a. Yes
    b. No

If you answered "yes", please specify when and where in the space below.

10. I wrote the field test
    a. on the school's computer
    b. on my own laptop

11. I felt confident writing the field test online.
    a. Yes
    b. No

Think about a time you took a paper-and-pencil test.

12. Taking the field test on Quest A+ was…. (circle one)
    a. easier than taking a paper-and-pencil test
    b. about the same as taking a paper-and-pencil test
    c. more difficult than taking a paper-and-pencil test
    d. more difficult at first, but became easier as I got used to it

13. If I had completed the field test on paper, I believe I would have received… (circle one)
    a. a better mark
    b. the same mark
    c. a lower mark

Please turn the page and complete the survey.

14. Compare your experience writing paper-and-pencil tests to completing the field test on Quest A+. Please check the most appropriate box for each question.

Which is:	Paper-and-pencil	Quest A+	They are the same
easier to read?			
less tiring?			
quicker to complete?			
easier to use if you need to refer to the formula sheet?			
easier to use if you need to use scrap paper?			

If you wrote the test online, you have now finished the survey. Thanks for your feedback! Please return your completed survey to the field test proctor or your teacher.

This section is **ONLY** for students who wrote the **paper-and-pencil field test**.

15. If I had completed the test online, I believe would have received …. (pick one)
    a. a better mark
    b. the same mark
    c. a lower mark

Thanks for your feedback! Please return your completed survey to the field test proctor or your teacher.