THE EFFECT OF REMOVING EXAMINEES WITH LOW MOTIVATION ON ITEM RESPONSE DATA CALIBRATION

by

Carlos Eduardo Zerpa

A dissertation submitted with the requirements for the degree of Doctor of Philosophy in Educational Studies

FACULTY OF EDUCATION
LAKEHEAD UNIVERSITY

Abstract

Many item-response models (IRM) used to estimate student abilities and test item parameters in large-scale assessments (LSA) do not account for the effect of student low motivation. This effect may pose a threat to the validity of test data interpretations. The first purpose of this study was to evaluate the effect of removing examinees with low motivation on the estimates of examinee abilities and test-item parameters calibrations using an item response theory model. The second purpose was to examine the significance of the relationship between students' motivation and their mathematics achievement, as measured by the LSA. Student motivation as defined by expectancy-value theory and self-efficacy theory was identified from self-report data using a principal component analysis. Two components scores, mathematics values and interest, were computed for each examinee to create two groups of examinees with high and low motivation. These groups were used to examine the effect of removing examinees with low motivation on the estimates of test item parameters and student abilities between a 3-parameter logistic (3PL) and modified 3PL IRM. The effect of student motivation on their academic achievement was examined using a hierarchical linear model (HLM). The results suggested that the modified 3PL IRM seemed to minimize the effect of low motivation on the estimates of student abilities and test item parameters when compared to the 3PL. The results from the HLM suggested that student mathematics values and interest are significant predictors of students' academic achievement. The outcome of this study builds on the research work of Swerdzewski, Harmes, and Finneys (2011) and supports the research work of Wise and DeMars (2006); Wolf and Smith (1995) in the use of IRM and measures of student motivation to provide more accurate interpretations of LSA data.

Acknowledgements

I would like to acknowledge and express my gratitude to my thesis committee members, Dr. Ruth Childs and Dr. Louis Volante for their support, suggestions, advice, and edits to my dissertation work. I thank Dr. Dwight Mazmanian for agreeing to be the internal examiner and for his edits and revisions to my dissertation work. I also thank Dr. Steve Wise for agreeing to be the external examiner and for his comments, advice, and suggestions that helped me improve my dissertation work.

I would like to express my deepest gratitude to my supervisor Dr. Christina van Barneveld for her support, encouragement, wisdom, guidance, advice, and knowledge throughout this difficult journey that has allowed me to grow and improve my knowledge as a young researcher. Dr. van Barneveld, I am indebted to you and I consider myself so fortunate to have you as my supervisor, thank you so much.

I thank the Educational Quality Accountability Office (EQAO) agency for allowing me to use the EQAO data to answer my research questions. I also would like to acknowledge that this research work was funded by the Social Science and Humanities Research council (SSHRC) and my supervisor Dr. Christina van Barneveld is the principal investigator of this grant.

To my mother and sisters, thank you for your encouragement and support on this journey of my life. Finally, to my wife Judith and my children Alyssa and Carlos (Jr), thank you for your patience, understanding, love and support. This PhD journey has been hard on all of us and we have made great sacrifices, but I promise I will make it up to you. Thank you for giving balance to my life and making me realize what truly matters. I love you all very much.

Table of Contents

List of Figures

List of Tables

Chapter One-Introduction

Item response theory (IRT) models have been used in conjunction with large-scale assessments (LSA) to examine the interactions between examinee abilities and test items. Both professionals and researchers use IRT models to design tests, assess educational programs, evaluate examinees' achievement and performance, and predict examinees' responses (Baker, 1992; Downing, 2003; Linn, 1989; Lord, 1990; National Council on Measurement in Education, 2009; van Barneveld, 2007).

In some situations, however, the item response theory models used to describe the pattern of examinee responses on a large-scale assessment do not account for unique characteristics that manifest as a result of interactions between examinee abilities and test items. These unique (or unaddressed) characteristics are referred to as "aberrant test-taking behaviours." Meijer and Sijtsma (1995) define these aberrant test-taking behaviours as "item score patterns that are unlikely, given that an item response model gives an adequate description of the data, or given the responses of the other persons in the group" (p.26).

Researchers in the field of educational measurement have developed models (referred to as person fit measurement techniques) to identify examinee test-taking behaviours that result in response patterns that are incongruent with an item response model. Some of these incongruent response patterns can be related to low motivation and in some situations, they pose a threat to the validity of large-scale assessment data interpretations (Ark, Emos, & Sijtsma, 2008; Meijer & Sijtsma, 2001; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999; Wright & Stone, 1979)**.** For example, if the examinee has low motivation because the exam does not hold personal consequences to his or her academic grade, then the examinee may not put forth the best effort in giving the correct answer to each item on the test. This lack of effort may result in the examinee guessing the correct answer, pattern marking, omitting items or quitting the test entirely.

Considering these test-taking behaviours in the item response model may produce a more accurate interpretation of the test data and, therefore, better estimates of student abilities.

Research Problem

Traditional item-response models used to validate and estimate student abilities and test item parameters on large-scale assessments do not include motivation as a parameter estimate in the item-response model. Some researchers, however, have included effort as a measure of student motivation when using a traditional item-response model to better estimate test item parameters and student abilities from large-scale assessment data (Wise & DeMars, 2006). For example, Wise and DeMars (2006) used response time as a measure of student effort in giving the correct response to a test item. The response time was dichotomized between solution behaviour and rapid guessing behaviour by selecting a threshold response time value. If the response time to a test item was below the threshold value, it was considered rapid guessing behaviour. If the item response time was above the threshold value, it was considered solution behaviour. This solution behaviour was considered a parameter estimate in the model and was used in conjunction with the other parameters of the item-response model to compute the test item parameter calibrations.

While Wise and DeMars' (2006) modeling technique is promising in including motivation as a parameter estimate in the item-response model by relating effort to motivation, a computer is required to measure examinee response time per item. Many large-scale assessments, however, are conducted using pencil and paper. Another approach to account for the effect of motivation in model parameter calibrations would be to modify existing item-response models to include measures of motivation from student self-report data.

Modifying an item-response model, but with measures of motivation obtained from student self-report questionnaires, may help provide more accurate estimates of examinee abilities and test item parameters on current large-scale assessments administered via pencil and paper. The motivation parameter to be included in the item response model can be obtained by identifying items related to student mathematics values and interest and link them to expectancy-value theory and self-efficacy theory of motivation. Expectancy-value theory links achievement performance, persistence, and choice directly to individuals' expectancy-related and task-value beliefs (Eccles & Wigfield, 2002). Self-efficacy refers to individual's perceived capabilities for learning or performing actions in relation to a task (Bandura, 1997). Interest relates to the extent to what an individual engages in a learning task based on the interactions of the individual with the activities and context he or she experiences (Mitchel, 1993; Renninger & Heidi, 2002; Singh, Grandville, & Dika, 2002; Wigfield & Cambria, 2010).

In this thesis, I modified an item response theory model to include measures of motivation obtained from student self-report data. The modified item response model used in this study was similar to a traditional item response model except that examinees with low motivation were not included in the calibration of test item parameters and examinee abilities. The measures of motivation included two constructs (math-values and interest) that I interpreted by drawing from expectancy-value theory and self-efficacy theory to address the effect of low motivation on the estimates of student mathematical abilities and test item parameters from the large-scale assessment data. This study builds on the work conducted by Swerdzewski, Harmes and, Finney (2011) and supports the work of Wise and DeMars (2006); Wise and Kong (2005); Wolf and Smith (1995); and Wise, Wise, and Bhola (2006).

Context of the Research Study

Every year, Grade-9 students in the province of Ontario are given a large-scale assessment in mathematics to monitor how well they are meeting the expectations of the mathematics curriculum (Klinger, Deluca, & Miller, 2008). The exams are administered by the Education Quality and Accountability Office (EQAO), which was established in 1996 by the government of Ontario to demonstrate the government's commitment to monitoring and accounting for students' mathematics achievement in our current educational system. The information obtained from EQAO assessments is used to inform schools, teachers, and parents about students' mathematics achievement in relation to a provincial standard (Volante, 2006).

EQAO's Grade 9 large-scale assessments of mathematics does not take into account the effect of motivation on students' test performance and academic achievement. If students do not value mathematics (Dweck &Elliot, 1983; Dweck & Grant, 2003; Singh, Grandville, & Dika, 2002) or if students know that test results do not count, it is possible that they may not place much importance and effort on a successful performance on the test (DeMars, 2000; Eccles & Wigfield, 2002). Students may randomly guess, omit questions, or display other low motivation test-taking behaviours that may pose a threat to the validity of the interpretation of the data in relation to their test performance and academic achievement (Cole, Bergin, & Whittaker, 2008; Eccles & Wigfield, 2002; Putwain, 2008).

Method

Goals of the study. The goals of this research study were:

1) To evaluate the effect of removing examinees with low motivation on the estimates of examinee abilities and test-item parameters calibrations using an item response theory model.
2) To examine the significance of the relationship between students' motivation and their mathematics achievement, as measured by the LSA, and how this relationship is influenced by school level variables.

I addressed the first goal of the study by examining the effect of removing examinees with low motivation on the estimates of examinee abilities and test item parameters computed by an item-response model (Baker, 1992). I accomplished this goal by focussing on motivation components as defined by the expectancy-value theory and self-efficacy theory and comparing two item-response models, which I calibrated by using student responses to multiple choice test items, and student self-report questionnaire data obtained from a Grade-9 large-scale assessment of mathematics, administered via the Education Quality Accountability Office (EQAO) in 2010.

To evaluate the effect of removing examinees with low motivation on test item parameters and examinee ability estimates, I conducted a principal component analysis of selected items from the EQAO student self-report questionnaire. This analysis helped me identify motivation components related to student mathematics value and interest as defined by expectancy-value and self-efficacy theory. I provided evidence of validity for the interpretation of the principal components as measures of motivation using three approaches; (1) comparing the principal component scores to responses from question 12, which asked about the use of the test in students' class marks, (2) analyzing and reanalyzing the components according to the literature (Singh, Grandville, & Dika, 2002; Pandura, 1997; Wigfield & Cambria, 2010), and (3) seeking the opinion of an expert on motivation theory.

In order to determine if removing examinees with low motivation affected the estimates of examinee abilities and test item parameters, I calibrated the data twice. The first time, I calibrated examinee multiple choice item responses by using a standard 3-parameter logistic (3PL) item response model. The second time, I calibrated examinee multiple choice item responses by using a modified 3PL item-response model, which included a motivation component obtained from students' self-report questionnaire data based on the expectancy-value and self-efficacy theory of motivation. It is important to clarify here that the modified 3PL item-response model is equivalent to the standard 3PL item response model, except that examinees with low motivation are coded as not administered in the modified 3PL model calibrations. After the calibration was completed, I compared the two models in terms of the bias, root mean square error (RMSE) and differential item functioning (DIF) techniques when estimating test item parameters and student abilities. The comparison of examinee abilities between the two models only pertained to examinees with high motivation that were included in both models calibrations. In the current study, bias is defined as the average difference of item parameter and ability estimates between the modified and standard models (Weiss, 1982). RMSE is defined as the standard deviation of item parameters and ability estimates of the standard model in reference to estimates of ability and item parameters of the modified model (Weiss, 1982). Differential item functioning (DIF) is defined as examinees in different groups in spite of their approximately equal knowledge and skills have a different probability to give a certain response to a test item (Bolt & Gield, 2006; Hambleton, Swaminathan, & Roger, 1991; Kim, Cohen, Alagoz, & Kim, 2007; Penfield & Algina, 2006; Penfield, 2007). Finally, I addressed the second goal of the study by conducting a hierarchical linear model (HLM) analysis to examine the relationship of student motivation and their mathematics achievement as measured by the EQAO test and how this relationship was influenced by school level variables.

The outcome of this study supports and builds on existing research (Swerdzewski, Harmes, & Finney, 2011; Wise & DeMars, 2006; Wolf & Smith, 1995). It also provides another avenue for researchers and measurement professionals to approach the problem of motivation in large-scale assessments by using a modified item response theory model in conjunction with expectancy-value and self-efficacy theory to provide more valid measures of student mathematical abilities and test item parameters. In addition, it may provide better information for teachers and educational agencies when planning and implementing educational policies in relation to students' test performance and academic achievement. Finally, it may suggest directions for expanding the scope of the research in relation to motivation theories and analyses techniques via item response models to better estimate student abilities and test item parameters.

Significance of the Study

This research study has theoretical and practical significance in the context of educational assessment. From the theoretical perspective, it supported existing research modeling techniques by Wise and DeMars (2006); Wise, Wise, and Bhola (2006) and built on the work of Swerdzewski, Harmes, and Finney (2011) to help explain the impact of low motivation on student academic achievement using a modified item-response model, which included motivation as one of the parameters of the model. In addition, it built on the work by Wolf and Smith (1995) on the application of expectancy-value theory of motivation in combination with measures of mathematics self-efficacy to examine student low motivation in relation to test items during large-scale assessments.

From the practical perspective, it offers another avenue through the use of self-report data based on subject domain values (i.e., general achievement motivation) to help explain the effect of student motivation on large-scale assessments so that more valid interpretations of student academic performance and achievement can be made from the large-scale assessment data (Dweck & Elliott, 1983; Eccles & Wigfield, 2002; Eklof, 2006; Ryan, Ryan, Arbuthnot, & Samuels, 2007). In addition, it suggests directions for expanding the scope of the research in relation to motivation analyses techniques when assessing student test performance and academic achievement via an item response theory model. Finally, it provides a different approach to evaluate LSA test item biases in relation to low motivation by using DIF techniques and item response theory models. This approach may also have implications for LSA test designs such as EQAO exams.

Chapter Two-Literature Review

This chapter provides an overview of relevant literature related to the importance of validity in large-scale assessments, evidence for validation, aberrant test-taking behaviours, construct-irrelevant variance, expectancy-value and self-efficacy theory of motivation, item response theory models, and statistical indices and techniques used to detect low motivation in large-scale assessments. In addition, in this chapter, I identify gaps in existing literature and provide information on a potential solution to address the research problem. The context of the research study is also addressed in the literature review.

Importance of Validity in Large-Scale Assessments

One commonly accepted definition of validity is offered by Messick (1989). He wrote, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment…. It is important to note that validity is a matter of degree, not all or none" (p.13). Messick addressed the word integrated, which in his concept of validity refers to the summation of many sources of information based on the existing evidence and the interpretation of test scores. Messick also stated that the theoretical rationale for the interpretation of test scores is not only based on the logical procedures for analyzing concepts, but also the rational procedures for addressing the conceptual problem based on the test scores. In other words, the validation process relies heavily on human judgement. More recently, Kane (2006) stated that the term validation tends "to have two distinct but closely related usages. In the first usage 'validation' involves the development of evidence to support the proposed interpretations and uses... In the second usage 'validation' is associated with an evaluation of the extent to which the proposed interpretation and uses are plausible and appropriate" (p.17).

The concept of validity is important in large-scale assessments because the interpretations of test results used to provide accountability to the public about specific schools, school based initiatives, programs, and teachers (Madaus & Kellaghan, 1992; McGehee & Griffith, 2001; Nagy, 2000; Volante, 2006) are subject to the inferences made from the test scores, which leaves test results open to evaluations and reinterpretations (Kane, 2006; Messick, 1989). In other words, to validate a proposed interpretation or use of test scores is to evaluate the rationale for the interpretations of the test scores (Kane, 2006).

Evidence for Validation of Test Score Interpretation in Large-Scale Assessments

In a number of ways, researchers have developed validity arguments to address the inferences made based on test-score interpretations from large-scale assessments (French & Oakes, 2004; Messick, 1989). Validity arguments may entail providing evidence to support the test score interpretations and their relevance to the proposed use. For instance, it may be possible that the test score interpretations gathered from large-scale examinations are based on providing arguments that test scores are a good indication of students' achievement on a specific skill, but do not support a wider set of skills (Kane, 2006; Schmidt, Le, & Llies, 2003). It may also be possible that test score interpretations are based on providing arguments to explain how different items on the test function differently for identifiable subgroups of examinees with similar overall ability (Standards for Educational and Psychological Testing, 1999).

There are various sources of evidence of validity that can be used to make a validity argument based on test score interpretations. These sources may include: evidence based on the internal structure of the test, evidence based on response processes, evidence based on test content and evidence based on relations to other variables (Standards for Educational and Psychological Testing, 1999).

Evidence based on the internal structure of the test is used to address the degree to which the test items relate to the construct being measured based on the test score interpretations. For example, claiming that the test is uni-dimensional. This claim can be supported by conducting a factor analysis to demonstrate that the score variability attributable to one major factor is higher than the score variability attributable to any other identified factors (Standards for Educational and Psychological Testing, 1999).

Evidence for validation based on examinee response processes is used to address concerns regarding the fit between a construct being measured and examinee level of engagement in responding to the test items addressing the construct (Standards for Educational and Psychological Testing, 1999). For example, in a large-scale examination, it may be relevant to know if the examinees are giving the best effort to answer each item on the test correctly or are just guessing or answer copying the responses to the test items. This behaviour can be examined by gathering documentation of other aspects of examinee performance like eye movement or response times.

Evidence for validation based on test content can be used to address the relationship between the test content (i.e., type of items, themes, and procedures) and the construct that is intended to be measured. This evidence can be obtained from an expert's judgement (Standards for Educational and Psychological Testing, 1999). For example, when developing a large-scale assessment in a specific area (e.g., mathematics), the major facets of a construct that is intended to be measured can be identified and an expert in the area can be asked to assign items to each of those facets on the test.

Evidence for validation based on relations to other variables can also be used to address the degree to which the construct being measured consistently relates to external variables

(Standards for Educational and Psychological Testing, 1999). For example, when administering a large-scale assessment, the researcher or practitioner may be interested in examining if student motivation to give the best performance on the test improves as a result of test consequences. This relationship may lead to a convergence of indicators that can be used to provide evidence for the validity of test score interpretations in large-scale assessments (Kane, 2006; Linn & Baker, 1996; Linn, Levine, Hastings, & Wardrop, 1981; Standards for Educational and Psychological Testing, 1999; Walter & Prescott, 1961).

Construct-Irrelevant Variance and Aberrant Behaviours in Large-Scale Assessments

When providing evidence for validation, construct-irrelevant variance is a key component of validity arguments because it poses a threat to test score interpretation from large-scale assessments (Messsick, 1989). Construct-irrelevant variance occurs when the assessment used to measure an educational or psychological construct includes measures that are not relevant to the intended construct, and cause scores to be different from what they should be (Messick, 1989; Standards for Educational and Psychological Testing, 1999).

There are several potential sources of construct-irrelevant variance that affect the validity of test score interpretations from large-scale assessments and some researchers have studied aberrant test-taking behaviours as a potential source of construct-irrelevant variance (Karabatsos, 2003; Meijer & Sijtsma, 1995; Wise, Wise, & Bhola, 2006; Wise & DeMars, 2005). Meijer and Sijtsma (1995) suggest that aberrant test-taking behaviours result in “item score patterns that are unlikely, given that an item-response model gives an adequate description of the data, or given the responses of the other persons in the group” (p. 26). Examples of aberrant test-taking behaviours are described in more detail in Table 1 (Meijer, 1996).

Table 1. Aberrant Test Taking Behaviours[1]

Aberrant Behaviour	Description
Sleeping Behaviour	An examinee fails to check the answers to some of the easier items in the test.
Guessing Behaviour	A person of low ability guesses the correct answer on medium difficulty items and the more difficult items.
Cheating Behaviour	A person of low ability copies the correct response to difficult items on a test from a more able examinee.
Plodding Behaviour	Examinee refuses to proceed to the next item until they have answered the current item.
Alignment Errors	Examinees enters the answer in the wrong position
Extremely Creative Examinees	Examinee finds easy items on a test too simple to be true and may give the incorrect answer to these items.
Deficiency of Abilities	An examinee that does not have the ability to answer easy items on a test, but has the ability to answer difficult items may generate many correct answers to the difficult items and many incorrect answers to the easy items.

Note: [1]This table was adapted from Meijer (1996, p.4-6)

Motivation, Test Stakes, Item Characteristics and Examinee Characteristics as a Source of Construct-Irrelevant Variance in Large-Scale Assessments

Motivation and test stakes. Test stakes (i.e., low or high stakes) are a source of construct-irrelevant variance because they can affect examinee level of motivation, which relates to aberrant test-taking behaviours (Wise, Wise, & Bhola, 2006). According to Wolf and Smith (1995), test stakes strongly affect examinees' level of motivation and have a modest, but significant impact on examinees' performance. The researchers also stated that when a test has personal consequences to an examinee, he or she may be more motivated to put forth a strong effort than when there are no personal consequences.

Motivation and item characteristics. The interactions between examinee motivation and test item characteristics can also be a source of construct-irrelevant variance (Standards for Educational and Psychological Testing, 1999; Wolf, Smith, & Birnbaum, 1995). These interactions may be attributed but not limited to item difficulty, mental taxation, and item position (Wolf, Smith, & Birnbaum, 1995). An interaction between examinee motivation and item difficulty relates to how likely an examinee is to get the correct response to an item if attempted (Pintrich, 1988; Wolf et al., 1995). For example, if the examinee has low motivation, the examinee may not apply his or her abilities to difficult items, opting instead to guess or omit these items. An interaction between examinee motivation and item mental taxation relates to how many steps or how much effort an examinee has to put forth to obtain a correct response to an item (Wolf et al., 1995). For example, if an examinee is motivated, he or she may put forth the best effort in giving the correct response to test items regardless of how many steps are needed to obtain the correct response. An interaction between examinee motivation and item position relates to the examinee's level of fatigue when attempting to give a correct response to a test item (Wolf et al., 1995). For example, an examinee may not persist when the test has no consequences. As a result, items at the end of the test may appear artificially more difficult because examinee motivation has waned. Interactions between examinee motivation and test item characteristics must be taken into account in large-scale assessments to minimize the effect of construct-irrelevant variance on test results (Wise, Wise, & Bhola, 2006; Wolf et al., 1995).

Some researchers have studied the interactions between examinee motivation and test item characteristics in large-scale assessments to assess the effect of these interactions as a source of construct-irrelevant variance for low and high stake conditions. For example, Wolf, Smith, and Birnbaum (1995) conducted a study to examine the interactions of examinee motivation and test item characteristics related to item difficulty, mental taxation, and item position for low and high stake conditions. Participants included 168 tenth graders and 133 eleventh graders from the same high school. For the eleventh graders, the results had no consequences; for the tenth graders, the results had consequences in terms of placement into remedial programs. The dependent variable measured was differential item functioning (DIF) with examinee groups determined by test stakes. The researchers used differential item functioning (DIF) because it provided an indication of the magnitude of the difference in performance between tenth and eleventh graders for each item on the test. The researchers conducted multiple regression analyses with DIF as the dependent variable and item characteristics (item difficulty, mental taxation, and item position) as independent variables. The researchers found that the interactions between examinee motivation and item characteristics (mental taxation, item difficulty, and item position) between the 10th and 11th graders were significantly related to the DIF index of the differences in performance between the two groups. The researchers indicated that test stakes influence test performance and this influence varies for easy or difficult items, for low or high mentally taxing items and the position of the item on the test. For instance, if there were no consequences, items related to non-consequential conditions although they were easy items, appeared more difficult, because they did not capture the complete effort of examinees. If there were no consequences, items that required multiple steps were more difficult for examinees than items that required one step. In addition, examinees experienced a higher level of fatigue toward the later items on the test under no consequential conditions. The researchers concluded that it is important to consider the conditions of the test (low and high stakes) and the interaction of examinee motivation with item characteristics before too much meaning is derived from the test scores.

Motivation and examinee characteristics. The relationship between examinees' characteristics (high or low ability) and aberrant test-taking behaviours that result from examinee low motivation is also a source of construct-irrelevant variance (Haladyna & Downing, 2004). For instance, Wolf and Smith (1995) found that the probability of guessing or omitting items on a test on the part of examinees increased as examinee ability estimates decreased. De Ayala, Plake and Impara (2001) investigated the effect on examinee ability estimates, when the examinee is presented an item and has ample time to answer, but decides not to respond to the item. The researchers found that the accuracy of examinee ability estimates decreases as the number of omissions increases. This effect has also been reported by Wise (1996a).

Expectancy-value theory of motivation. One motivation theory that researchers have used in combination with large-scale assessment data to address the effect of aberrant test-taking behaviours as a source of construct-irrelevant variance on student academic performance, is expectancy-value theory (Eccles & Wigfield, 2002; Printrich, 2004; Printrich, Smith, Garcia, & McKeachie, 1993; Printrich & Schunk, 1996; 2004; Wigfield & Cambria, 2010; Wolf & Smith, 1995). Expectancy-value theory of motivation links achievement performance, persistence, and choice directly to individuals' expectancy-related and task-value beliefs. Expectancy-related beliefs refer to individuals' beliefs about how well they will do on an upcoming task, either in the immediate or longer-term future (Eccles, 1993; Eccles & Wigfield, 2002). Task-value beliefs are defined by four components: (a) attainment value – the personal importance of doing well on a task, (b) intrinsic value – the enjoyment the individual gets from performing the task, (c) utility value – how well the task relates to current and future goals, such as career goals, and (d) cost – negative aspects of engaging in the task, such as fear of failure (Eccles & Wigfield, 2002).

As defined by expectancy-value theory, motivation to perform well on a large-scale assessment depends on the students' general ability-beliefs and task-value beliefs (Eccles, 1993; Eccles & Wigfield, 2002; Eklof, 2006; Pintrich, 1988;1989; McMillan, Simonetta, & Singh, 1994). General ability-beliefs relate to student beliefs about how well the student thinks he or she will do on the large-scale assessment (Pintrich, 1988; 1989). Task-value beliefs relate to how much importance the student places on a successful performance. That is, if the student values the outcome of the large-scale assessment, it is more likely he or she will be motivated, make an effort on tasks and engage with the tasks to the best of his or her ability (Eccles & Wigfield, 2002; Pintrich,1988;1989; Ryan, Ryan, Arbuthnot, & Samuels, 2007; Wigfield & Eccles, 2000).

Self-efficacy theory of motivation. Another motivation theory that researchers have used to address the effect of low motivation as a source of construct-irrelevant variance in large-scale assessments is self-efficacy (Pajares, 1996a, 1996b; Pajares & Graham, 1999; Pajares & Kranzler, 1995; Ryan, Ryan, Arbuthnot & Samuels, 2007). Self-efficacy refers to persons' judgement of their confidence to learn, perform academic tasks, or succeed in academic endeavours (Bandura, 1986). Research has revealed that self-efficacy influences individual's motivation, achievement, and self-regulation (Bandura, 1997; Stajkovic & Luthans, 1998). For example, in education, self-efficacy has been shown to affect students' choices of activities, effort expended, persistence, interest, and achievement (Pajares, 1996b; Schunk, 1995). That is, those students with high self-efficacy participate more readily, work harder, persist longer, show greater interest in learning, and achieve at higher levels (Bandura, 1997).

Achievement motivation and test-taking motivation. When using a motivation theory (e.g., expectancy-value, self-efficacy), students' achievement values, goal orientations, choices, persistence, and interest are considered motivation-related constructs that affect student behaviours to perform achievement activities, and these motivation-related constructs can be based on general achievement motivation or specific test-taking motivation (Bandura, 1997; Eklof, 2006; Wigfield & Eccles, 2000). General achievement motivation (i.e., motivation to learn a subject such as mathematics) is based on students' beliefs about their abilities to learn a subject, the value of school activities and endeavours, and the goals students brings to their classes and homework (Harter, 1982; Dweck & Elliot, 1983; Wigfield & Eccles, 2002). General achievement motivation is traditionally grounded on motivation theories (e.g., expectancy-value theory, self-efficacy theory, achievement goal theory) (Atkinson, 1964; Pintrich & Schunk, 2004) and more precisely on students' expectancy performance, perceived capabilities and perceived values of the task in relation to a goal (Bandura, 1997; Eccles & Wigfield, 2002). For example, student levels of engagement to perform mathematics tasks are based on the values that students bring to the mathematics task and these values are influenced by teachers and parents who play a critical role in the importance that students place in mathematics learning (Dweck & Elliot, 1983; Eccles & Harold, 1996; Stone, 2006).

Test-taking motivation (i.e., effort given to a test on the part of the student) can be thought of as the degree to which examinees exert effort in the attempt to provide the correct response to the test items (Swerdzewski, Harmes & Finney, 2011; Wise & DeMars, 2005; 2006; Wise & Kong, 2005).This definition relates to the influence that motivation has on test performance (Wolf & Smith, 1995) and it is grounded on Printrich's (1988;1989) modified model of expectancy-value theory and Eccles and Wigfield's (2002) model that relate individual characteristics to the nature of the task at hand. That is, when applying Pintrich's or Eccles and Wigfield's model to test-taking situations to measure student motivation, the researchers or professional must consider a) how well the student thinks he or she will do on the test, b) how important it is to do well on the test on the part of the student, and c) the reactions the student has to the test taking situation. The application of Pintrich's or Eccles and Wigfield's models to test-taking motivation has been supported by various studies that examined the effect of low student motivation on test-taking situations (Baumert & Demmrich, 2001; Sundre & Moore, 2002; Wolf & Smith, 1995; Wolf, Smith & Birnbaum, 1995, Wise & DeMars, 2005; 2006; Wise, Wise, & Bhola, 2006).

While the degree to which achievement motivation relates to test-taking motivation has not been extensively investigated, as these two types of motivation are considered distinct from one another, some evidence in the literature suggests that there is a relationship between them. For example, Eklof (2006) conducted a study to validate an instrument to measure student test-taking motivation by using the Trends in International Mathematics and Science (TIMSS) test. She used a sample of 350 Swedish participants from the eighth grade. Eklof compared measures of test-taking motivation to assess student perceived value of a good performance on the test (e.g., how motivated are you to do your best on TIMSS's mathematics items?), general attitudes toward mathematics and science that highlighted the value of learning and doing well on each subject (e.g., I enjoy learning mathematics, I like being in school), and performance-expectancy, that related to how well a student was expected to do on the test (e.g., how many of the mathematics items in TIMSS do you think you can answer correctly). The theoretical framework adopted in her research study was the expectancy-value theory (Pintrich & Schunck, 2004) and the achievement motivation model by Eccles and Wigfield (2000) and Wigfield and Eccles (2002). After conducting an exploratory factor analysis, the researcher found that the three constructs were distinct but related. That is, the test-taking motivation factor had a moderate positive correlation with the general attitude factor that highlighted values toward learning a subject ($r = 0.49$). The performance-expectancy factor was also moderately and positively correlated with the general attitudes factor ($r = 0.41$). The researcher also conducted a second analysis to examine the relationship of test-taking motivation, mathematics values and mathematics self-concept. The inter-factor correlation matrix revealed moderate positive correlations for all factors (*r* = 0.35 to 0.40). The researcher suggested that items asking for value perception and perceived feelings of test taking motivation could be used as a measure of student test-taking motivation. The researchers also suggested that test-taking motivation is related to but not the same "as general attitudes, self-concept or valuing of a subject" (p. 654). Swerdzewski, Harmes, and Finney (2011) also conducted a study to compare two approaches used to measure examinee motivation, self-report measures and response time effort (RTE). For the self-report measures, the researchers used the student opinion survey (SOS) by drawing from the research work of Sundre and Moore (2002) and Wise and DeMars (2005). For the measures of RTE, the researchers drew from the work of Wise and DeMars (2006). The researchers compared the two methods for filtered and unfiltered data in relation to cognitive and non-cognitive measures. For the non-cognitive measures such as attitudes toward learning (work avoidance), the researchers found that students who self-reported low test-taking motivation were also more work avoidance and academically amotivated about their attendance to university. The researchers suggested that "work avoidance and academic amotivation may cause low test-taking effort" (p.180). Some other researchers have also found that self-efficacy can have an effect on either student achievement in the classroom or in test-taking situations (Fast & Lewis, 2010; Pajares, 1996a, 1996b; Pajares & Graham, 1999; Pajares & Kranzler, 1995).

Self-report questionnaires as measures of student motivation. While there are different techniques used to measure student motivation in large-scale assessment data (e.g., measures of response time using computer-based technology and statistical indices), some researchers have used self-report questionnaires to measure student motivation in large-scale assessments. For instance, the Motivated Strategies Learning Questionnaire (MSLQ) (Pintrich, Smith, Garcia, & McKeachie, 1993) and the Student Opinion Survey (Sundre & Kitsantas, 2004; Wolf and Smith, 1995) have been used by researchers to measure motivational beliefs and values of students ranging in age from late elementary to university. Marsh, Koller, Trautwein, and Baumert (2005) developed a learning survey to measure: how much students look forward to learning mathematics; how important mathematics is to them; the importance of being a good mathematician; and the enjoyment of learning mathematics, by drawing on the expectancy-value theory of motivation. O'Neil, Abedi, Miyoshi, and Mastergeorge (2005) used an adaptation of the State Thinking Questionnaire (O'Neil, Sugrue, Abedi, Baker, & Golan, 1997) to measure motivation in their study by using monetary incentives as a way to increase student effort and performance. Roderick and Engel (2001) used the Reynolds Adolescent Depression Scale (Reynolds, 1984) to cross check interview data of student descriptions of their motivation. Swerdzewski, Harmes, and Finney (2011) conducted a study to examine the interrelationships of self-report data based on the Student Opinion Scale questionnaire (SOS) and computerized testing based on response time measures to identify low motivated examinees in a low-stake assessment context. Swerdzewski et al. (2011) found that the two methods (self-report questionnaire and computerized based testing) were consistent in the degree to which they identified examinees with low and high motivation in low stake situations.

One of the strengths of using self-report questionnaires to measure student motivation in large-scale assessments is that they can be easily implemented using a pencil-and-paper method (Swerdzewski et al., 2011) as opposed to other measurement techniques that may require the use of computer-based technology (Wise & DeMars, 2006). In addition, self-reported questionnaires can be grounded in the expectancy-value theory of motivation or other motivation theories (e.g., attribution theory, achievement goal theory, and self-efficacy) to assess students' motivation and effort in relation to their academic achievement (Eccles &Wigfield, 2002; Pintrich, 2004; Pintrich et al., 1993; Pintrich & Schunk, 1996; Swerdzewski et al., 2011; Wigfield & Cambria, 2010).

One of the challenges, however, is to develop and use motivation self-report questionnaires that have a clear structure, high internal consistency, and strong evidence of validity. Another challenge is to create a motivation self-report questionnaire that clearly addresses factors and variables as they relate to test effort and performance (Cole et al., 2008; Eccles & Wigfield, 2002; Harlen & Crick, 2003; Putwain, 2007; 2008; Wigfield & Cambria, 2010). Also, motivation questionnaires are not typically part of large-scale assessment data collection plans.

Problem of Student Motivation in Large-Scale Assessments

Low motivation of students is a potential source of construct-irrelevant variance in large-scale assessments because it poses a threat to the validity of inferences based on test scores interpretations (DeMars, 2000; Haladyna & Downing, 2004). The effect of low motivation is that students may not give the best effort in providing the correct response to each item on the test, especially in situations when the test scores from the large-scale assessment do not have personal consequences to students (DeMars, 2000; Sundre & Wise, 2003; Wise, Wise, & Bhola, 2006; Wise & DeMars, 2005). As a result, low motivation may cause student abilities to be underestimated and test item parameters to be overestimated when using standard statistical modeling techniques to calibrate the large-scale assessment data (DeMars, 2000; van Barneveld, 2007; Wise & DeMars, 2006). The underestimation of examinee abilities and overestimation of item parameters can affect the decision making process in relation to curriculum changes, allocation of funding, and the implementation of policies in the educational system.

Effort to Address the Student Motivation Problem by other Researchers in LSA

Researchers have developed and use item response theory models and statistical indices to address this effect of low motivation as a potential source of construct-irrelevant variance on test score interpretations derived from large-scale assessments (Baker, 1992; Drasgow, Levine, & Williams, 1985; Fraire, Tideman, & Watts, 1997; Lord, 1990; Sotaridona & Meijer, 2003; Sotaridona, Linden, & Meijer, 2006; Wise & DeMars, 2006; Wolf & Smith, 1995). Part of the effectiveness of the item response models and statistical indices, however, depends on the psychometric considerations (i.e., low motivation parameter, response time effort parameter) that are used to minimize the effect of construct-irrelevant variance on the estimates of examinee abilities and test item parameters (Gierl, Henderson, Jodoin, & Klinger, 2001; Wise, Wise, & Bhola, 2006). Before I review some of the statistical techniques that are used in combination with item response theory models to examine the effect of low student motivation in large-scale assessments, I have provided a general review of item response theory models.

Item response models. Item response theory models provide an avenue to describe and estimate not only the item-by-item parameters on a test but also examinee-by-examinee parameter responses in a probabilistic manner (Baker, 1992; Lord, 1990). Thus, the performance of an examinee on an item can be expressed as a mathematical function, which specifies the probability of making a correct response to an item given the examinee's level of ability (Baker, 1992).

Local independence. A defining principle of item response theory models is local independence. The principle of local independence states that " in a subpopulation in which r latent traits $F_1 \ldots F_r$, take fixed values $f_1 \ldots f_r$, the responses are (conditionally) independent……That is , $F_1…F_r$ are latent traits if the item responses are independent in a subpopulation in which $F_1…F_r$ are fixed" (MacDonald, 1999, p. 255). This principle means that no relationship exists among examinee's responses to different items. Said differently, when examinees' "abilities influencing test performance are held constant, examinees' responses to any pair of items are statistically independent" (Hambleton, Swaminathan, & Roger, 1991; MacDonald, 1999, p.11).

Invariance. The keystone of item response theory models is the property of invariance, which "implies that the parameters that characterize an item do not depend on the ability distribution of the examinees and the parameter that characterizes an examinee does not depend on the set of test items" (Hambleton, Swaminathan, & Roger, 1991, p. 20). Said differently, if invariance holds, the parameter estimates should be the same regardless of the distribution of ability in the groups of examinees used to estimate the item parameters (Hambleton, Swaminathan, & Roger, 1991).

Item response models for dichotomous data. There are several item response models used in educational measurements which are considered appropriate for dichotomous item response data (e.g., true and false questions or multiple choice questions) in large-scale assessments. These models include: one, two, and three parameter logistic models (Baker, 1992; Hambleton, Swaminathan, & Roger, 1991; Lord, 1990). The one parameter logistic model estimates examinee probabilities of correct responses to test items based on examinee level of ability and the level of difficulty of each item on a test (Baker, 1992). The two parameter logistic model estimates examinee probabilities of correct responses to test items similarly to the one parameter logistic model; however, it includes two additional elements. These elements are: a) the factor D, which is used to scale the logistic function to a normal ogive function and b) the parameter a, which is called the discriminant parameter and it is proportional to the slope of the item characteristic curve. The discriminant parameter or slope can be used to separate examinees into different ability levels (Baker, 1992). The steeper the slope the better the item discriminates between lower and higher ability examinees (Baker, 1992). The three parameter logistic model is similar to the two parameter logistic model; however, it includes a pseudo-chance parameter to represent the probability of examinees with low ability answering an item correctly on a test (Baker, 1992).

Among item response models, the model that best fits the data after implementing statistical curve fitting techniques (i.e., Chi-square goodness of fit test, likelihood ratios) will be the one selected to estimate examinee abilities and item parameters from the test data (Baker, 1992). For example, if the model that best fits the data is a three parameter logistic item response model (3PL) as depicted in Equation 1, then this model is used to estimate examinee abilities and item parameters from the test data (Baker, 1992; Hambleton, Swaminathan, & Roger, 1991; Lord, 1990).

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}, \text{ i=1,2,....n.} \qquad (1)$$

Where:

$P_i(\theta)$: is the probability that a randomly chosen examinee with ability θ answers item *i* correctly.

b_i: is the item level of difficulty.

n: is the number of items in a test.

e: is a numerical constant whose value is 2.718.

a_i: is the discrimination parameter, which is proportional to the slope of the item.

c_i: is the pseudo-chance parameter.

D: is a scaling factor to make the logistic function as close as possible to the ogive function. The value of D is 1.702.

Maximum likelihood estimation of item parameters. Item response model parameters can be estimated using maximum likelihood functions and derivatives in the case where examinees' abilities are observed in conjunction with item responses on a test (Lord, 1990). Via maximum likelihood function and derivatives, it is possible to optimize the estimation of the parameters of a model based on observed data (Freund, 1999; Lord, 1990). In item response models, this means that mathematical expressions can be developed using the maximum likelihood criterion to estimate the parameters of a one, two, or three parameter logistic model and therefore, compute the probability of examinees' correct responses to each item on a test (Cohen, 1960; Gierl, Harwell, Baker, & Zwarts, 1998; Henderson, Jodoin, & Klinger, 2001).

In the case of the three-parameter logistic item response model and under the local independence assumption, success on one item is statistically independent of success on other items given one single dominant value of $\theta i,$ which represents examinee ability (Harwell, Baker, & Zwarts, 1998). Therefore, the probability of a response based on observed items is the joint distribution of success and failure, which under maximum likelihood functions and derivatives can be given by Equation 2.

$$L = \prod_{i=1}^{n}\prod_{j=1}^{J} P_j(\theta)^{y_{ij}} Q_j(\theta_i)^{1-y_{ij}} \qquad (2)$$

Where:

Yij: represents the response vector, which is conditional on the known value of θ and the item parameters (Harwell, Baker, & Zwarts, 1998). For convenience and simplicity, the log of the likelihood function is applied and it yields Equation 3.

$$\log L = \sum_{i=1}^{n}\sum_{j=1}^{J}[y_{ij} \log P_j(\theta_i) + (1 - y_{ij}) \log Q_j(\theta_i)] \tag{3}$$

The parameter items that maximize the likelihood of Equation 3 when estimating examinee abilities are found by setting the first derivatives of the likelihood equal to zero (Harwell, Baker, & Zwarts, 1998). This computation leads to Equations 4, 5, and 6.

$$\frac{\partial}{\partial a_j}(\log L) = 0 \tag{4}$$

$$\frac{\partial}{\partial b_j}(\log L) = 0 \tag{5}$$

$$\frac{\partial}{\partial c_j}(\log L) = 0 \tag{6}$$

Based on Lord (1990), after taking the first derivatives and solving the equations simultaneously to estimate the item parameters for a three parameter logistic model, likelihood Equations 7, 8, and 9 are produced for the estimation of the model parameters.

$$a_j : (1 - c_j)\sum_{i}^{n}[y_{ij} - P_j(\theta_i)](\theta_i - b_j)w_{ij} = 0 \tag{7}$$

$$b_j : (-a_j)(1 - c_j)\sum_{i}^{n}[y_{ij} - P_j(\theta_i)]w_{ij} = 0 \tag{8}$$

$$c_j : (1 - c_j)^{-1}\sum_{i}^{n}\frac{[y_{ij} - P_j(\theta_i)]}{P_j(\theta_i)} = 0 \tag{9}$$

Where :

$$w_{ij} = \frac{P_j^*(\theta_i)Q_j^*(\theta_i)}{P_j(\theta_i)Q_j(\theta_i)}$$

θ_i : is known

$P_j^*(\theta_i)Q_j^*(\theta_i)$: is the probability of success and failure of a prior joint distribution.

Statistical models and techniques used to detect low motivation. In this section, I review statistical models and techniques that have been developed and used by researchers to examine the effect of low motivation, which were more relevant to my research study. These statistical techniques include differential item functioning (DIF) and response time effort. These techniques provide an avenue to either examine or minimize the effect of low motivation as a source of construct-irrelevant variance on the inferences made from test-score interpretations derived from large-scale assessments (Wise, 1996a; Wise, Wise, & Bhola, 2006; Wise & DeMars, 2006; Wise & Kong, 2005; Wolf, Smith, & Birnbaum, 1995).

Differential item functioning (DIF). DIF is a technique that researchers use in combination with item response theory models to examine the effect of student low motivation on the estimates of test item parameters when using large-scale assessments (Wolf, Smith, & Birnbaum, 1995). DIF is defined as examinees in different groups in spite of their approximately equal knowledge and skills have a different probability to give a correct response to a test item (Bolt & Gield, 2006; Kim, Cohen, Alagoz, & Kim, 2007; Penfield, 2007; Penfield & Algina, 2006; Wainer, 1993). The presence of DIF on test items may pose a serious threat to the fairness of the test items and the validity of the interpretation of test scores (Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995).

One simple method used to detect DIF between two groups is by stating a null hypothesis that the test item parameters characteristics from the two groups are identical, that is, $H_o: b_1 = b_2; a_1 = a_2;\ c_1 = c_2$. A chi-square statistic is used to test this null hypothesis:

$$\chi^2 = \left(a_{diff} b_{diff} c_{diff}\right) \Sigma^{-1} (a_{diff} b_{diff} c_{diff}) \qquad (10)$$

where: $a_{diff} = a_2 - a_1$; $b_{diff} = b_2 - b_1$; $c_{diff} = c_2 - c_1$ (Hambleton, Swaminathan, & Roger, 1991; Lord, 1990).

Another method used to detect DIF is by computing the area in between two functions instead of the parameter differences (Hambleton, Swaminathan, & Roger, 1991). In order to compute the area, the parameter estimates need to be on a common scale. If after placing the parameter estimates on a common scale, the item characteristics curves (ICCs) are identical, then the area between them should be zero (Hambleton, Swaminathan, & Roger, 1991). A symbolic representation of this procedure may be expressed as:

$$Area = A_i = \int_{\theta=r}^{s} |P_{i1}\theta - P_{i2}\theta| d\theta \qquad (11)$$

and the exact expression as derived by Raju (1998) can be expressed as:

$$Area = (1 - c) \left| \frac{2(a_2 - a_1)}{D a_1 a_2} ln\left[1 + e^{D a_1 a_2 (b_2 - b_1)/(a_2 - a_1)}\right] - (b_2 - b_1) \right|. \qquad (12)$$

One disadvantage of the area computation method, however, is the need to find a cut-off value for the area statistics to decide whether or not DIF is present. A cut-off value for the area statistics can be very computationally intensive and may require a large examinee ability range, especially in situations when there is an interaction between examinee ability and group membership (Hambleton, Swaminathan, & Roger, 1991). Based on these concerns, some researchers choose to use the Mantel-Haenzel procedure (Camilli & Congdon, 1999; Camilli & Penfield, 1997; Holland & Thayer, 1988). This procedure is considered a practical, inexpensive, and powerful method to detect DIF (Holland & Thayer, 1988). With the Mantel-Haenzel approach, a focal and reference group are compared. The focal group is the group of interest (F) and the reference group (R) is the standard against which the focal group is compared. The problem with the Mantel- Haenzel approach, however, is that it does not have high power for non-uniform DIF (Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Uttaro & Millsap, 1994). Non-uniform DIF exists when there is an interaction between ability level and group membership. That is, the difference in the probabilities of a correct answer for the two groups is not the same at all ability levels as shown in Figure 1 (Mellenberg, 1982; Roussos, Schnipke & Pashley, 1999). Whereas, a uniform DIF exists when there is no interaction between ability level and group membership, that is, the probability of answering the item correctly is greater for one group than the other uniformly over all levels of ability as shown in Figure 2 (Mellenberg, 1982; Roussos, Schnipke, & Pashley, 1999)**.** One approach used to compute DIF, however, for non-uniform item characteristic curves (ICC) is via a logistic regression model technique, which is capable of identifying both non-uniform and uniform ICC (Roussos, Schnipke, & Pashley, 1999).

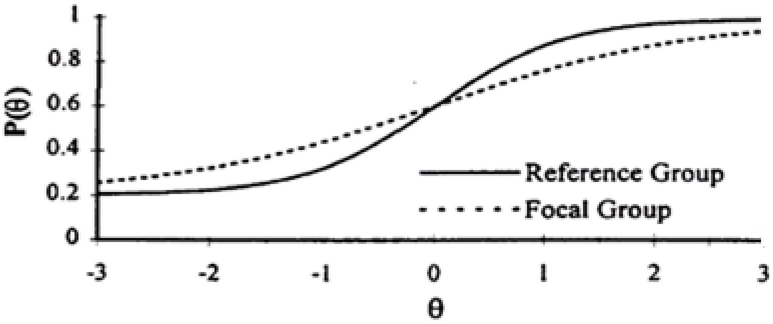


Figure 1: Nonuniform Differential Item Functioning

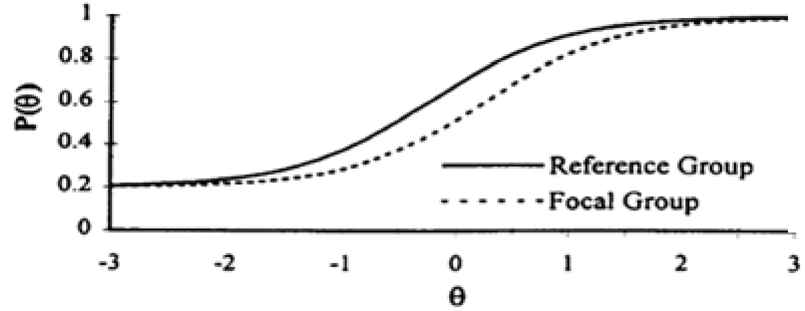


Figure 2: Uniform Differential Item Functioning

Thissen, Steinberg, and Wainer (1988; 1993) also proposed a method to compute DIF, which is based on the likelihood ratio model comparison. This procedure consists of two parts. First, a model is estimated in which the item parameters are freely estimated within the reference and focal groups, and a log-likelihood value is obtained. Second, a model is estimated in which the item parameters are constrained to be the same for each group and again, a log-likelihood value is obtained. The difference between the log likelihood values for the two models is computed and distributed as chi-square with degrees of freedom equal to the number of parameter differences between the two models. If the resulting chi-square is large, this will indicate that constraining the parameter results in a worse model and therefore, the items that were freely estimated contain DIF. Following this approach, Lord's chi-square statistic can be used as an index of DIF to examine which items contain DIF between the two groups (reference and focal).

Response time effort. While DIF methods have been used for detecting low motivation test-taking behaviours, other researchers have developed and used other methods of measuring student motivation such as response time effort (RTE) (Swerdzewski, Harmes, & Finney, 2011;Wise & Kong, 2005; Wise & DeMars, 2006). For instance, Wise and Kong (2005) developed a new technique to detect examinee test-taking behaviours by using response time effort. Response time effort is based on the hypothesis that when an item is administered, unmotivated examinees will have the tendency to answer it too quickly. Wise and Kong also stated that when measures of item response time effort are acquired on a test either for speeded or non-speeded exams, there are two types of examinees' behaviours manifested for each item on a test. The first one is called solution behaviour, in which examinees actively seek to determine the correct answer. The second one is called rapid-guessing behaviour, in which examinees rapidly respond to items in a generally random fashion. Wise and Kong explored these concepts further and hypothesized that when measures of item response time effort are acquired on a test, rapid-guessing behaviours reflect a lack of examinees' effort. In each encounter with an item on a test, the examinee makes a choice to engage in either solution-seeking or rapid-guessing behaviour. This choice is reflected in the time the examinee takes to respond to an item. The researchers stated that for each response time effort to an item on a test, there is a threshold, *Ti*, which represents the response time boundary between rapid-guessing behaviour and solution behaviour as stated in Equation 13. In Wise and Kong's study, 506 freshmen participated in an achievement information literacy test at the university level. The results showed that measurements of response time effort (RTE) provided information regarding examinees down to the level of individual items. The results also revealed that RTE scores seemed to be more accurate measures of examinees' efforts than self-reported efforts. Furthermore, RTE scores were found to be reliable and showed evidence of convergent and discriminant validity. Finally, the researchers concluded that RTE scores can be used as a motivation filter for item calibration during high stake examinations.

$$RTE = \frac{\sum SB_{ij}}{K}, SB_{ij} = \left\{ \begin{array}{l} 1 \rightarrow RT_{ij} \geq T_i \\ 0 \rightarrow otherwise \end{array} \right\} \tag{13}$$

Where:

RTE is the response time effort to answer a question.
SB is the solution behaviour to get the correct answer.
$1\text{-}SB$ is the rapid-guessing behaviour.
K is the number of items in the test.
RT is the response time per item measure with a computer-based test.
T_i is the response time boundary or threshold.

One of the challenges with RTE, however, relates to selecting the right threshold (T_i) to be able to differentiate between solution behaviour and rapid guessing behaviour. In order to develop a method to differentiate between solution behaviour and rapid guessing behaviour on a test item, Kong, Wise, and Bhola (2007) conducted a study on 524 undergraduate students. They compared four methods for establishing response time threshold. The methods included: a common threshold of 3 seconds per item; a threshold based on the amount of reading and scanning per item; visual inspection of response time distribution per item; and a mixture model by hypothesizing that each item should have a bimodal distribution. The researchers found that all the methods produced RTE scores showing very consistently positive results. In addition, the four methods were virtually identical in terms of internal consistency, discriminant, and motivation filtering.

Wise and DeMars (2006) built on work done by Wise and Kong (2005) and created an effort-moderated item response theory model to account for student motivation as depicted in Equation 14. According to Wise and DeMars, when encountering an item, the examinee will engage in either solution or rapid guessing behaviour. If the examinee engages in solution behaviour, the probability of giving the correct response to an item increases based on examinee proficiency and this probability can be effectively modeled under a traditional item response model. On the contrary, if the examinee engages in rapid guessing behaviour, the probability of a correct response to an item remains near the level expected by chance, which is a constant probability regardless of examinee proficiency. In Wise and DeMars' model as stated in Equation 14, the solution behaviour model is represented by a three parameter logistic model and guessing behaviour is represented by a constant probability. This constant probability is computed based on the reciprocal of the number of response options.

$$P_i(\theta) = \left(SB_{ij}\right) * (solution\ behaviour\ model) + \left(1 - SB_{ij}\right) * guessing\ behaviour \tag{14}$$

Wise and DeMars compared the effort-moderated model as stated in Equation 14 to a standard IRT model as stated in Equation 1 to examine the IRT psychometric characteristics of the effort-moderated item response model. These comparisons included model fit, item parameter estimation, test information and convergent validity. The researchers selected 524 mid-year sophomores from a medium-sized southern US university and administered a low stake 60 items computer-based literacy test. During the test administration, the researchers collected response time for each examinee encounter with an item. The researchers dichotomized the response time into solution behaviour and rapid guessing behaviour by using a threshold value. A dichotomized value of 1 represented solution behaviour and a value of 0 represented rapid guessing behaviour, similar to the approach of Wise and Kong (2005). The researchers found that under the model fit likelihood ratio technique, the item parameters from the effort-moderated model provided a better fit to examinees' response patterns than the item parameters from the standard model. In terms of item parameter estimation, the two models differed in item discrimination parameters with the standard model yielding higher discrimination parameter estimates. The models also differed in item level difficulty. The mean difficulty was substantially higher under the standard model for easy items, but there were no substantial differences between the two models for the more difficult items. In terms of reliability, the researchers found higher information functions for the standard model. The researchers stated the discrimination parameters could be spuriously high under the standard model due to the occurrence of low accuracy rapid guesses on the part of examinees. Finally, the researchers compared the two models in terms of convergent validity by correlating estimates of examinee proficiency to an external variable (i.e., grade point average) expected to correlate with proficiency. They found that the effort-moderated model had a significantly higher correlation to the external variable than the standard model. The researchers concluded that relative to the standard model, the effort-moderated model provided a better fit to examinees' responses and therefore more accurate estimates of examinees' proficiency.

Gaps in Existing Literature

While the modeling technique of Wise and DeMars (2006) is promising, this modeling technique requires the use of computer-based technology to measure examinee response time per item. Many large-scale assessments, however, are conducted using pencil and paper (Ferrando & Lorenzo, 2005; Swerdzewski et al., 2011) and therefore, there is a need to find other avenues to effectively filter examinees with low motivation from the test data by using self-report measures of student motivation in combination with item response theory modeling techniques.

Addressing the research problem. I addressed a portion of the motivation problem in the current study by using students' self-report questionnaires from large-scale assessment data. I used the research work by Wigfield and Cambria (2010) and experts' opinions to help me identify motivation items related to student math-values and interest from the self-report questionnaires. I used student math-values and interest as measures of their motivation in combination with a modified item response theory model that accounted for a motivation parameter when estimating examinee abilities and test item parameters form large-scale assessment data, similar to the modeling approach of Wise and DeMars (2006) and Wise, Wise and Bhola (2006).

In summary, the research work addressed in the current study is mostly concerned with modifying an item response theory model, to include motivation as a parameter estimate in the model and evaluate the effect of removing examinees with low motivation on item response data calibration. This modeling technique dichotomizes students' math-values and interest components into examinees with high and low motivation based on student self-report data. This approach may provide more valid measures of student academic performance because it accounts for the effect of low motivation on the estimates of examinee abilities and test item parameters when using a large-scale assessment conducted via pencil and paper such as EQAO. This study builds on work done by Swerdzewski, Harmes, and Finney (2011) and supports the work of Wise and DeMars (2006) and Wolf and Smith (1995).

Context of the Study

In 1996 the Education Quality and Accountability Office (EQAO) was established in Ontario as an agency to assist in improving the quality and accountability of Ontario's public education and to comply with recommendations made by a Royal Commission of Education in 1995 (Begin & Caplan, 1994). EQAO is responsible for developing, administering, collecting, and disseminating province-wide assessments. The EQAO results are used to provide accountability in the educational system about student academic performance in reading, writing, and mathematics (Education Quality and Accountability Office, 2004b).

EQAO mandate. The EQAO mandate is to design and implement a comprehensive program of student assessment, measure the quality of education in the province of Ontario, report the results to various stakeholder groups, lead the province in national and international assessments, promote research on best practices in assessment and accountability, and conduct quality reviews in consultation with school boards (Education Quality and Accountability Office, 2012). The information obtained from EQAO assessments is then used to better inform schools, teachers, and parents about students' mathematics achievement in relation to a provincial standard (Volante, 2006).

Administration of EQAO large-scale assessments in Ontario. EQAO large-scale assessments are administered in Grades 3, 6, 9, and 10 every year. The focus of this thesis is the Grade-9 assessment of mathematics. This assessment is administered twice (winter and spring semesters) during the year to students in applied and academic programs (Education Quality and Accountability Office, 2012). Students in academic programs are those who develop knowledge and skills through the study of theory and abstract problems. They also learn about practical applications where appropriate. Students in applied programs are those who develop knowledge and skills through practical applications and concrete examples (The Ontario Curriculum Grades 9 to 12, 2000). Students enrolled in first-semester mathematics courses write the test in January. Students enrolled in second-semester and full-year mathematics courses write the test in May/June (Education Quality and Accountability Office, 2012).

EQAO Grade-9 assessments of mathematics. EQAO Grade-9 assessments of mathematics were introduced in 2000-2001. The Grade-9 assessments measure how well students have met the provincial expectations of the Ontario Curriculum in relation to the knowledge and skills in mathematics that students are expected to have acquired from the first-semester to the second-semester in both academic and applied programs. The assessments are based on four mathematics strands: Number Sense and Algebra, Linear Relations, Analytic Geometry (academic program only), and Measurement and Geometry. Students enrolled in the applied mathematics program are given a different assessment from those in the academic mathematics program (Education Quality and Accountability Office, 2012).

In 2010, EQAO Grade-9 assessments of mathematics included: a) an exam that was composed of 24 multiple choice questions and 7 open response questions. Both multiple choice and open response questions were related to the four mathematics strands (Number Sense and Algebra, Linear Relations, Analytic Geometry “academic program only”, and Measurement and Geometry); b) a student self-report questionnaire , which was composed of 20 questions that were administered to both applied and academic students separately. This questionnaire gathered information about students’ backgrounds, attitudes toward mathematics, activities outside school, and expectations about their future; c) a teacher questionnaire, which was composed of 26 questions that were administered to teachers in applied and academic programs to gather information on the learning environment, the use of instructional resources, communication with parents, the use of EQAO resources in schools, and information on teachers’ backgrounds and their professional development; d) a principal questionnaire, which was composed of 19 questions to gather information on principals’ backgrounds, the learning environment, the use of EQAO resources, and parental engagement in schools (Education Quality and Accountability Office, 2011).

Item response theory models used by EQAO. EQAO uses a 3PL item response model to calibrate multiple-choice items for Grade-9 mathematics assessments (Kozlow, 2007). It is EQAO’s mandate to ensure that appropriate item response models are used to calibrate the data and accurately estimate examinee abilities (Kozlow, 2007). A study conducted by Kozlow (2007) on EQAO data revealed that the 3PL item response model with the pseudo-guessing parameter fixed at 0.20 was a better fit for the EQAO data as compared to the one or two parameter logistic item response models. The value of 0.20 was calculated based on 1/ (k+1) ratio, where k represented the number of response options. The technique of fixing the pseudo-chance parameter has been used in other research studies which used a 3PL response model to calibrate examinee responses to test items (Wise & DeMars, 2006; Wise & Kong, 2005).

Another item response model that EQAO uses to estimate examinee abilities and item parameters for constructed or open responses is the Generalized Partial Credit (GPC) model (Muraki, 1992; 1997). This model represents a family of mathematical models that deals with ordered polytomous data. Since the current research thesis only deals with EQAO multiple choice items, the emphasis and discussions are on the 3PL item response model.

Validity of EQAO assessments. EQAO test score interpretations are based on content-related and construct-related evidence of validity (Crundwell, 2005; Wolfe, Childs, & Elgie, 2004; Herbet, Dunn, & Luthra, 2004). While EQAO assessments provide content- and construct-related evidence of validity, some researchers have questioned the validity of EQAO assessments and stated that there is a need to provide evidence of validity beyond these conventional views (Crundwell, 2005).

According to Crundwell (2005) other types of validity evidence need to be taken into consideration to provide more accurate interpretations of EQAO test results. For example, EQAO assessments do not provide evidence of consequential validity, that is, the social consequences of the EQAO assessments to society (Crundwell, 2005; Volante, 2006) or evidence of validity that takes into account the effect of students' low motivation on their academic achievement and performance, especially in situations when the large-scale assessment does not hold personal consequences to students' academic grades (Wise, 2006; Zerpa, Hachey, van Barneveld, & Simon, 2011).

Wolfe, Childs, and Elgie (2004) also indicated in their review of the EQAO assessment instrument and procedures that to ensure valid interpretations and use of EQAO test scores, an active program of validity research needed to be initiated and supported based on "examination of the content, the process students use to respond to test items, marking and scoring, general external factors and consequences from EQAO assessments" (p. 68). Since then, EQAO has made great effort to implement procedures to: a) ensure validation of test items, b) continue to involve educational experts in the process of test design, c) include quality assurance so that the test are administered consistently and fairly across the province, and d) include scoring validity (Education Quality and Accountability Office, 2010). EQAO has not addressed, however, the effect of low motivation on test score interpretations and student academic performance.

Why is EQAO Grade-9 mathematics assessment appropriate for this study? The EQAO Grade-9 mathematics assessment is appropriate for the current study because this is the only grade where a portion of the EQAO test (0% to 30%) may count toward students' final grades (Education Quality and Accountability Office, 2011). In addition, students in Grade-9 are given a self-report questionnaire to gather information about their attitudes toward mathematics. These variations in test stake in conjunction with student self-report data provided me with the necessary information to answer my research questions.

Chapter Three-Method

Goals of the Study

There are two goals of this research study:

1) To evaluate the effect of removing examinees with low motivation on the estimates of examinee abilities and test-item parameters calibrations using an item response theory model.
2) To examine the significance of the relationship between students' motivation and their mathematics achievement and how this relationship is influenced by school level variables.

Research Questions

For the first goal, the research questions that guided this study were:

1) Can some items on the Student Questionnaire (SQ) for EQAO's Grade-9 Assessment of Mathematics, 2010, be used as a measure of student motivation?
2) What is the magnitude of bias and RMSE in item parameter and student ability estimates as a result of student low motivation on a Grade-9 large-scale assessment of mathematics when using multiple choice questions?
3) What items contribute to the bias (if any) in item parameter estimates on the Grade-9 mathematics EQAO test?
4) What changes in the proportions of examinees with high motivation affect the item parameter and ability estimates of the modified and standard IRM for academic and applied programs?

For the second goal, the research question that guided this study was:

5) What is the relationship of student motivation and academic achievement as measured by the EQAO Grade-9 mathematics assessment and how is this relationship influenced by school level variables?

For this study, I used quantitative methods based on item response theory and classical test theory. To answer the research questions, I used data from the Ontario's Education Quality and Accountability Office (EQAO) Grade-9 assessments of mathematics, 2010 spring administration for applied and academic programs. The data included self-reported questionnaires and test item response data files for English students in the academic and applied programs. The method I used to address each research question is described below.

Research question #1.

Participants. I used data from 63,783 Grade-9 students who took the English version of the test from the spring 2010 administration of the EQAO test. This sample included 43,308 students in the academic program and 20,475 students in the applied program. I recoded the missing data in both academic and applied program data samples by assigning a value of 99 to each missing value. There were 1221 cases with missing values for the academic program data and 708 cases with missing values for the applied program data. Only the missing values were excluded from the analysis.

Instruments. I used self-report questionnaires for Grade-9 English students in the academic and applied programs from the spring 2010 administration of the EQAO test. The questionnaires were composed of 20 questions and were administered to both applied and academic students separately. These questionnaires provided information about students' backgrounds, attitudes toward mathematics, activities outside school and expectations about their future.

Procedures.

Select items. I selected items from the student self-report questionnaires that may be related to student motivation based on research work by Wigfield and Cambria (2010), expert opinion, expectancy-value theory and self-efficacy theory. These items are listed in Equations 15 and 16. For example, self-report items such as: " I like math" I classified as intrinsic or interest value because it related to students' enjoyment from learning mathematics on the emotional side (Wigfield & Cambria, 2010, p.4); " The math I learn now is very useful for everyday life" and " I need to keep taking math for the kind of job I want after school" I classified as utility values because they reflected the importance of mathematics learning for student future plans (Wigfield & Cambria, 2010, p.4).; "I understand most of the mathematics I am taught"; "I am good in math" and "Mathematics is an easy subject" I classified as measures of self-efficacy because these items relate to student perceived capabilities for learning mathematics (Bandura, 1997). Finally, items such as: "How much time do you usually spend in math homework?," "How often have you been absent from your Grade 9 mathematics class this year?" and "How often have you been late for your Grade nine mathematics class this year?" I classified as student interest to engage in mathematics learning (Wigfield & Cambria, 2010, p.9) because academic tasks such as doing homework and regular attendance are documented in the literature to reflect student interest based on their engagement and motivation (Singh, Grandville, & Dika, 2002, p.324).

Assessing the component structure of the selected items. Similar to work done by Wolf, Smith and Birnbaum (1995), I assessed the component structures of the selected items from the student self-report questionnaire (SQ) by conducting a principal component analysis (PCA). Before conducting the PCA, however, the data were transformed from categorical to continuous data using optimal scaling techniques. In this transformation process, I used a value decomposition technique in which the ordinal points of the variables were used to create vectors using SPSS software. The spacing between the points in the vectors was computed and this spacing corresponded to the optimal quantification of the transformed variables (Bartholomew, 1980). After the variables were transformed, a PCA was conducted with varimax rotation to examine the component structure of the selected items for both academic and applied programs separately. I identified two components (math-values and interest) via the PCA and I computed component scores for each student using Equations 15 and 16.

Initially, I had identified two components from the PCA as task-values and effort based on the literature and expert's opinion. Taking in consideration, however, the revisions and suggestions made to this research study, the concept of validity, which states that the inferences made from data interpretations are open for reinterpretations to the extent to which the proposed interpretations and uses are plausible and appropriate (Kane, 2006; Messick, 1989), and the notion that the identification of items and the labelling of the components when conducting either a PCA or factor analysis are considered a subjective process that hinges on reflective researcher's judgement (Ford, MaCallum, & Tai, 1986; Henson, & Roberts, 2006), I decided to rename these two components as math-values and interest. I felt that these component labels aligned better with the definition of student values in mathematics learning. That is, mathematics values relate to student beliefs, importance and interest to engage in mathematics tasks (Boaler, 1999; Ernest, 1989). I reduced the subjectivity in the current study, however, by only considering as significant those items that loaded higher than 0.40 on their respective components, asking expert's opinion for a consensus judgement of the items as measures of motivation, cross-referencing the expert's judgement with the literature and reinterpreting the items and components based on reviewer's revisions and suggestions made to this research work.

$$\textbf{\textit{Math-Valu}}\textbf{es} = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 + \alpha_6 X_6 + \alpha_7 X_7 + \alpha_8 X_8 + \alpha_9 X_9 \quad (15)$$

$$\textbf{\textit{Interest}} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 \quad (16)$$

Where:

$\mathbf{X_1}$ = I like Math.

$\mathbf{X_2}$ = I am good in Math.

$\mathbf{X_3}$ = I understand most of the mathematics I am taught.

$\mathbf{X_4}$ = The mathematics I learn now is very useful for everyday life.

$\mathbf{X_5}$ = I need to keep taking mathematics for the kind of job I want after I leave school.

$\mathbf{X_6}$ = Mathematics is an easy subject.

$\mathbf{X_7}$ = How much time do you usually spend on mathematics homework (in or out of school) on any given day?

$\mathbf{X_8}$ = How often have you been absent from your Grade 9 mathematics class this year?

$\mathbf{X_9}$= How often have you been late for your Grade 9 mathematics class this year?

$\boldsymbol{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9}$ are the coefficients for the variables in the math-values component.

$\boldsymbol{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9}$ are the coefficients for the variables in the student interest component.

Evidence of validity. I collected some evidence of validity to confirm that the interpretations of responses to the selected items from EQAO student self-report questionnaires (math-values and interest components) related to test-taking motivation. To collect some convergent-related evidence of validity, I correlated the PCA components (math-values and interest) to an external motivation variable, SQ question 12, "Does counting the Grade-9

Assessment of Mathematics as part of your class mark motivate you to take the assessment more seriously?" Specifically, I selected a sample of examinees with low and high motivation based on the principal component scores (math-values and interest) using criteria described in Equation 17. Examinees with high motivation were defined as having a math-value score > 0.25 and interest score >0.25 standard deviations above the mean respectively. Examinees with low motivation were defined as having a math-value score < -0.25 and interest score < -0.25 below the mean respectively. I used this criteria based on the research of Wise and DeMars (2006), which stated that "motivated examinees tend to outperform their less motivated peers by 0.58 standard deviations" (p.19). Using chi-square statistics, I compared the principal component scores obtained from Equation 17 to "yes" and "no" responses from SQ question 12, "Does counting the Grade-9 Assessment of Mathematics as part of your class mark motivate you to take the assessment more seriously?" The final sample size for this analysis was 8,742 examinees from the academic program and 2,032 examinees from the applied program.

$$MB_{ij} = \begin{Bmatrix} 1\ if\ MV \geq \overline{MV} + \frac{1}{4}\delta_1\ and\ IN \geq \overline{IN} + \frac{1}{4}\delta_2 \\ 0\ \ if\ MV \leq \overline{MV} - \frac{1}{4}\delta_1 and\ IN \leq \overline{IN} - \frac{1}{4}\delta_2 \end{Bmatrix} \tag{17}$$

Where:

MB_{ij}= Motivation Behavior Parameter

MV= Math-Values Component

IN=Student Interest Component

$\overline{MV}$=Mean Math-Values

$\overline{IN}$= Mean Student Interest

δ_1 = Standard Deviation for Math-Values Component

δ_2 = Standard Deviation for Student Interest Component

To collect some content-related evidence of validity, I solicited expert opinion on the interpretations of the self-report items and principal components as measures of motivation. I also provided content-related evidence of validity by analyzing and re-analyzing the principal component items and component labels in relation to expectancy-value theory, self-efficacy theory of motivation, and mathematics values using Wigfield and Cambria (2010), Bandura (1997), Ernest (1989), Boaler (1999), and external reviewer's suggestions and revisions made to this research thesis.

Analysis. I used Chi-square statistics and Cramer's V effect size to determine if there was a relationship between question 12 and the dichotomized motivation variable (examinees with low and high motivation) obtained from the principal component scores (math-values and interest). This analysis was conducted separately for applied and academic programs.

Research question #2.

Participants. I used data from 43,308 English students in the academic program and 20,475 English students in the applied program. These samples resulted from the merge of student self-report questionnaire data with the multiple choice test item response data from the 2010 spring administration of the EQAO test.

Instruments. I used self-report questionnaires and multiple-choice item data for Grade-9 English students in the academic and applied programs from the spring 2010 administration of the EQAO test. I only used student self-report items related to math-values and interest as stated in Equations 15 and 16, as well as the 24 multiple choice item responses from the EQAO test to address the purposes of my study.

Procedures.

Data merge. I addressed the second research question by merging EQAO student self-report questionnaire data and multiple-choice item response data. I merged both data files (students' self-reported questionnaires and test item response data) by student record ID using a one to one relational database script via SPSS (see APPENDIX A1 for SPSS script).

Recoding item responses and treatment of missing data. When addressing the second research question, I implemented similar procedures to both academic and applied program samples; however, I analyzed both samples (student academic and applied) separately. Before calibrating the merged data with BILOG-MG software (Zimowski, Muraki, Mislevy, & Bock, 1996) and selecting the item response model that was a best fit for the data, I recoded the EQAO test item response data. I assigned a value of 1 to items with a correct score and I assigned a value of 0 to items with an incorrect score. For the cases with missing data, I assigned a value of 9 to the items with missing values in both academic and applied samples respectively (see APPENDIX A3 for SPSS script). According to the literature, there are several methods used to treat the missing data when calibrating data with an item response theory model. One method is to assign to the missing data a value, which is equal to the reciprocal of the number of response options so that the missing items are fractionally correct. A second approach is to treat the missing data as incorrect responses. Similarly to Emenogu and Childs (2005), I decided to use the second approach. I selected this option because there is evidence in literature to support that item response models do not distinguish between those examinees who omit items and those examinees that ignore the instructions and guess (Mislevy & Brock, 1990); therefore, treating the items as incorrect responses, it is possible to suppress the effect of guessing by using a biweight estimation technique (Mislevy & Brock, 1990). This method reduces the weight to unlikely correct responses and improves the accuracy of estimating ability in the presence of guessing in multiple choice items (Mislevy & Brock, 1990).

Model fit. I compared the one, two and three parameter logistic models to determine which one of these models was a better fit to the test items based on the chi-square values obtained from BILOG-MG software. This comparison was conducted separately for applied and academic programs data. Similarly to Kozlow (2007), I found that the three parameter item response model with the c-parameter fixed at a value of 0.20 was a better fit for the data (academic and applied programs). I followed a similar approach as Kozlow (2007) and Wise and DeMars (2006) by computing the c-parameter based on 1/ (k+1) ratio, where k represented the number of response options.

Selecting equal proportions of examinees with high and low motivation. From the merged data file (student self-report and item response data), I selected equal proportions of examinees with high and low motivation by using the principal component analysis scores related to math-values and interest, which were obtained from Equations 15 and 16. Each sample of data (academic and applied) was analyzed separately. Since both principal components (math-values and interest) were continuous variables normally distributed with a mean = 0 and *SD* = 1, it was possible for me to dichotomize the component scores into examinees with high and low motivation and therefore, create a motivation component as stated in Equation 17. This approached allowed me to select the top 40% of examinees with high motivation and the bottom 40% of examinees with low motivation if both conditions were met for math-values and interest. After implementing this approach, I obtained from the academic program data a sample of 9,123 examinees with high motivation and 9,123 examinees with low motivation from a total sample of

43,308 examinees. From the applied program data, I obtained 3,912 examinees with high motivation and 3,912 examinees with low motivation from a total sample of 20,475 examinees.

Formatting the data for BILOG-MG analysis. I configured in BILOG_MG format the newly created data file sample, which was composed of examinees with high and low motivation (see APPENDIX A3 for example of SPSS script). For each data sample (academic and applied), the BILOG_MG formatted file contained columns for examinee RECID, multiple choice item (1-24) responses and motivation parameter. I also created a key file for the missing data and I treated the missing data as incorrect responses using biweight estimation techniques facilitated by the BILOG-MG software (Mislevy & Brock, 1992). Since I recoded the multiple choice responses as 1 or 0 for each item per examinee, there was no need to create a key file for correct item responses because the BILOG_MG software can automatically differentiate between correct and incorrect responses coded as: 1 = correct and 0 = incorrect.

Item calibration using two models. Via the BILOG-scripts, I calibrated the data twice; once using a traditional 3PL item-response model as stated in Equation 1 and once using a modified 3PL item-response model (M3PL) as stated in Equation 18. This equation includes a motivation parameter in the model when estimating examinee abilities and test item parameters based on examinee responses to the test items. The motivation parameter was obtained from Equation 17.

$$P_i(\theta) = (MB_{ij})\left(c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}\right) + (1-MB_{ij})(g_i) \tag{18}$$

Where:

MB_{ij}= Motivation Behavior Parameter.

g_i = is the reciprocal of the number of response options for item *i*.

Under the modified item response model, each student has a motivation behaviour (MB) score which indicates that they exhibited either low motivation (0) or high motivation (1). If the examinee has low motivation, that is the MB = 0, then the probability estimate is fixed for all items, which corresponds to some estimated ability value given by the reciprocal of the number of response options. This implies that all students exhibiting low motivation will have the exact same estimated ability, which is not a measurement value of ability, but rather an expected score under random guessing. On the contrary, if the examinee has high motivation, that is MB = 1, then the probability of the examinee giving the correct response increases with examinee proficiency. This means that the M3PL is a motivation filter method that removes those examinees with low motivation from the test data when estimating examinee abilities and test item parameters. That is, those exhibiting high motivation get scores when using the M3PL and those examinees exhibiting low motivation do not get scores or the scores are unavailable. Since the modified IRT model is not included in the BILOG_MG software, I decided to use a similar approach as Wise and DeMars (2006), in which I considered the a, b, and c parameter estimates of the modified IRT model equivalent to those of the standard IRT model but with low motivated examinees removed from the test data. The reason why is because under low motivation behaviour, the probability of a correct response to an item is fixed and it is given by a constant g_i across all levels of θ. This constant does not influence where the maximum value of an item parameter occurs when computing the likelihood functions and equating the first derivative of the function to zero. In contrast, when calibrating the same data with the standard IRT model as stated in Equation 1, I considered the low motivation responses as valid meaning that those examinees exhibiting low motivation were included in the calibration (see APPENDIX A4 for example of BILOG script).

Scaling. After I calibrated the student item response data with each item-response model, I needed to put the two models (Standard and Modified) data calibration on the same scale. According to the literature, there are two possible ways of fixing the scale between two models or groups: (a) scaling the item difficulty values between the two models so that abilities estimates can be compared and (b) scaling the ability values between the two models so that item parameter estimates can be compared (Kolen & Brennan, 2004; Hambleton, Swaminathan, & Roger, 1991; Hanson & Beguin, 2002). Since I needed to compare both models in terms of ability and test item parameters, I implemented both scaling techniques. The first time, I scaled the item difficulty values between the two models by using the mean and sigma method (Hambleton, Swaminathan, & Roger, 1991). Via the mean and sigma method, the mean and standard deviation of the item difficulty parameters from the modified and standard models were used to estimate the α and β coefficients for Equation 19. I used this equation to put the standard model under the same scale as the modified model. This approach allowed me to compare examinee ability estimates for those examinees with high motivation only between these two models in terms of bias and RMSE (see APPENDIX A8 for example of SPSS script).

$$\hat{\theta} = \alpha \quad + \beta \tag{19}$$

Where:

$\hat{\theta}$ is the scaled ability estimates for the standard model

θ is the unscaled ability estimates for the standard model

$\alpha = \frac{SD_{bM}}{SD_{bS}}$

SD_{bM} is the standard deviation of the threshold estimates of the modified model

SD_{bS} is the standard deviation of the threshold estimates of the standard model

$\beta = \bar{b}_M - \alpha * \bar{b}_S$

$\bar{b}_M$ is the mean from the threshold estimates of the modified model

$\bar{b}_S$ is the mean from the threshold estimates of the standard model

The second time, similar to Wise and DeMars (2006), I scaled the test item parameters in terms of ability by using the empirical option in BILOG-MG so that the mean ability estimate was set to zero and standard deviation to 1 for each model calibration (Hambleton, Swaminathan, & Roger, 1991). Since examinee ability estimates were not exactly the same for each model calibration, an adjustment was necessary in order to compare the test item parameters of the standard to the modified model (Hambleton, Swaminathan, & Roger, 1991; Wise & DeMars, 2006). I accomplished this adjustment by using Equations 20 and 21 (see APPENDIX A10 for example of SPSS script).

$$\hat{b} = \alpha \mathrm{b} + \beta \tag{20}$$

Where:

$\hat{b}$ is the threshold estimate for the standard model after equating

b is the threshold estimate for the standard model before equating

α is the standard deviation of the ability estimate for the modified model

β is the mean of the ability estimate for the modified model

$$\hat{a} = \frac{a}{\alpha} \tag{21}$$

Where:

$\hat{a}$ is the slope estimate after equating

a is the slope estimate before equating

Analysis.

Bias and RMSE. After the data were scaled and adjusted, I analyzed the data by computing the bias and RMSE of item parameters and student ability estimates to examine the effect that examinees with low motivation had on the estimates of examinee abilities and test item parameters when comparing the two models (standard and modified) for applied and academic programs separately. Since low motivated examinee ability estimates were not available when using the M3PL, but they were available when using the standard model, the comparison between the two models (standard and modified) in terms of ability estimates was only accomplished for those examinees with high motivation that were included in both model calibrations (standard and modified). This approach allowed me to compare the same examinee ability estimate and the same number of examinees between the two models. I accomplished this comparison by merging the calibrated and scaled data in term of ability estimates for those examinees that were included in both models calibrations by using student RECID. This process permitted me to identify changes for equal proportions of low, middle and high ranges in examinee abilities estimates between the two models (standard and modified) for those examinees who exhibited high motivation only. That is, this process allowed me to evaluate the effect of removing examinees with low motivation on the estimates of examinee abilities with high motivation when using a standard model without low motivation filtering. I also examined parameter estimates for low, middle, and high ranges between the two models. The bias and RMSE for $\hat{a}$, $\hat{b}$ parameters and student ability, $\hat{\theta}$, were computed using the equations stated in Table 2 (see APPENDIX A11 for example of SPSS script).

Table 2. Bias and RMSE for the Standard and Modified Models

		Bias	**RMSE**
Parameter	Percent of low motivation	Standard and Modified IRT	Standard and Modified IRT
$\boldsymbol{\hat{a}}$	40	$\frac{\sum \hat{a}_j - \hat{a}_k}{n_1}$	$\sqrt{\sum \frac{(\hat{a}_j - \hat{a}_k)^2}{n_1}}$
$\boldsymbol{\hat{b}}$	40	$\frac{\sum \hat{b}_j - \hat{b}_k}{n_1}$	$\sqrt{\sum \frac{(\hat{b}_j - \hat{b}_k)^2}{n_1}}$
$\boldsymbol{\hat{\theta}}$	40	$\frac{\sum \hat{\theta}_j - \hat{\theta}_k}{n_2}$	$\sqrt{\sum \frac{(\hat{\theta}_j - \hat{\theta}_k)^2}{n_2}}$

Where:

$\hat{a}_j$ is the discriminant parameter estimate based on the standard IRM

$\widehat{a_k}$ is the discriminant parameter estimate based on the modified IRM

$\hat{\theta}_j$ is the student ability estimate based on the standard IRM

$\hat{\theta}_k$ is the student ability estimate based on the modified IRM

n_1 is the number of test items

n_2 is the number of examinees

Research question #3.

Participants. I used the same merged data sample from research question 2, which was composed of student self-report data and multiple choice items.

Instruments. I used the same self-report questionnaires and multiple-choice exams as research question 2. I only used, however, student self-report items related to math-values and interest as stated in Equations 15 and 16.

Procedures.

Selecting equal proportions of examinees with high and low motivation. I used the same procedure as in question 2 to select a data sample of examinees with 40 percent of low motivation and 40 percent of high motivation.

Matching the sample size between the two models for DIF analysis. Before conducting the DIF analyses to examine the effect of removing examinees with low motivation on test item parameter estimates, I needed to match the sample size between the two models for academic and applied programs data separately. Matching the sample size was necessary because the standard model included examinees with low motivation that were not administered to the modified model, resulting on a larger sample for the standard model, which could increase the chance of affecting the DIF detection rates when conducting a DIF analysis (Pankaja & Swaminathan, 1994).

I matched the sample sizes between the two models (modified and standard) by grouping examinees in either model based on their correct responses out of 24 items tried. For example, I computed a score distribution for each model (standard and modified) to see how many examinees got a score of 1 out of 24, 2 out of 24, three out of 24…, and 24 out of 24 in each model. If there was a difference in sample size between the modified and standard models for any grouping scores out of 24, then the one with the larger sample was reduced by randomly selecting examinees to match the size of the smaller sample. This approach allowed me to conduct a final adjustment to make sure that the groups were of equal size and the same or similar abilities.

Analysis. I conducted a differential item functioning (DIF) analysis technique to examine which items contributed to the bias when low motivation was present between two groups of examinees with approximately the same level of abilities. One group, however, contained examinees with low and high motivation and the other groups contained only examinees with high motivation. This approach is very similar to Wolf, Smith, and Birnbaum (1995), which examined DIF between two groups of examinees with approximately the same level of ability in relation to motivation, test item characteristics (item difficulty, mental taxation, and item position) and test stakes (high and low) situations.

DIF analysis method used. While there are different procedures and techniques used to compute DIF such as: Mantel Haenszel (MH), SIBTEST, standardization (STD), logistic regression, and item response theory (IRT) methods and there is no consensus on which method to use (Emenogu & Childs, 2005), I decided to use the Thissen, Steinberg, and Wainer (1988; 1993) method, which computes DIF based on the likelihood ratio model comparison. I used this method because it is included with the BILOG-MG software and the DIF distribution between the two models was uniform. Furthermore, the c-parameter was kept fixed and the slope was common for each item between the two models. Via the Thissen, Steinberg, and Wainer (1988; 1993) method, I created a BILOG-MG script (see APPENDIX A12) to calibrate the data for the two models (standard and modified). I calibrated the data twice. For the first calibration, I created a text file, which included the data for the two models as two separate sets by assuming that there was DIF. I computed the likelihood ratios and used the Thissen, Steinberg, and Wainer (1988; 1993) DIF technique to obtain the item parameters for each model. For the second calibration, I created a text file, which included the data for the two models as one set by assuming that there was not DIF. I calibrated the data as one model and computed the likelihood parameter ratio. Following this, I computed the difference between the two calibration likelihood ratios and used chi-square to determine if there was DIF present. The chi-square analysis revealed that the assumed DIF model from the first calibration was a better fit to the data, which indicated to me that there was DIF present in the item level difficulty parameter estimates between the two models (standard and modified).

Lord's chi-square statistics. I used Lord's chi-square statistic as an index of DIF to examine each item between the two models (standard and modified). Lord's chi-square statistics was computed as stated in Equation 22. I compared the chi-square index value to a critical chi-square value corresponding to an alpha level of 0.05 with one degree of freedom.

$$\mathcal{X}^2 = v_i \textstyle\sum_i^{-1} v_i \qquad (22)$$

Where:

v_i is a vector of the differences in the estimated item parameters for the ith item between the standard and modified models $\{b_{1i} - b_{2i}\}$.

$\sum_i$ is the asymptotic variance-covariance matrix for the differences in item parameter estimates (Lord, 1990).

DIF area between items. I also computed the DIF area between the item characteristic curves (ICCs) for the two models to examine which mathematics strands (numeracy, algebra, and geometry) might be more affected by examinee motivation. I accomplished these area computations by using Equation 12.

Research question #4.

Participants. I used the same merged dataset as in research questions 2 and 3 to address research question 4. I analyzed the data separately for academic and applied programs.

Instruments. I used the same self-report questionnaires and multiple-choice exams as in research question 2 and 3. I only used, however, student self-report items related to math-values and interest as stated in Equations 15 and 16.

Procedures.

Data calibration for different proportions of high motivation. I calibrated the academic program data first and the applied program data second by using the standard and modified IRT models. This approach allowed me to examine the effect of different proportions of high motivation on estimates of examinee abilities and test item parameters for each model. Similar to question 2, I conducted these calibrations by selecting data samples with different proportions (e.g., top 30% high and bottom 40% low, top 50% high and bottom 40% low) of examinees with low and high motivation from the merged data using the PCA scores from Equations 14 and 15. Next, I created text files for each proportion in BILOG_MG format via SPSS. In addition, I created three BILOG-MG scripts for each model (standard and modified) to calibrate the multiple choice items for each high motivation proportion.

Examining the effect of different proportions of high motivation. I examined the effect of different proportions of examinees with high motivation on item parameter and student ability estimates by first placing each proportional calibration comparison on the same scale using Equations 19, 20, and 21 and similar procedures as stated in the method for research question 2. I accomplished this by comparing the first calibration of the modified model (40% of high motivation), which was used as reference to each proportional calibration comparison (30% of high and 40% of low motivation, 50% high and 40% of low motivation) between the modified and standard models as depicted in Tables 3 and 4. To compare the same examinee ability estimates, however, between the reference (40% of examinees with high motivation) and each proportional calibration (30% of high and 40% of low motivation, 50% high and 40% of low motivation) for those examinees who exhibited high motivation when using the standard and modified models, I merged the calibrated data by student RECID using a SPSS script (see APPENDIX A7). This approach allowed me to compare the same examinee ability estimate and equal number of examinees between the reference and each proportion of high motivation when using the standard and modified models for those examinees who exhibited high motivation.

Analysis. I computed the magnitude of bias and RMSE for examinee ability estimates (examinees with high motivation only) and test items between the reference and each proportion of high motivation between the standard and modified models. The bias and RMSE equations are listed in Tables 3 and 4.

Table 3. Bias and RMSE for Item Parameter Estimates for the Standard and Modified IRM

		Bias		**RMSE**	
Parameter	Percent of low motivation	Standard IRM	Modified IRM	Standard IRM	Modified IRM
$\hat{\boldsymbol{a}}$	30	$\frac{\sum \hat{a}_{l40} - \hat{a}_{r30}}{n}$	$\frac{\sum \hat{a}_{l40} - \hat{a}_{l30}}{n}$	$\sqrt{\sum \frac{(\hat{a}_{l40} - \hat{a}_{r30})^2}{n}}$	$\sqrt{\sum \frac{(\hat{a}_{l40} - \hat{a}_{l30})^2}{n}}$
	50	$\frac{\sum \hat{a}_{l40} - \hat{a}_{r50}}{n}$	$\frac{\sum \hat{a}_{l40} - \hat{a}_{l50}}{n}$	$\sqrt{\sum \frac{(\hat{a}_{l40} - \hat{a}_{r50})^2}{n}}$	$\sqrt{\sum \frac{(\hat{a}_{l40} - \hat{a}_{l50})^2}{n}}$
$\hat{\boldsymbol{b}}$	30	$\frac{\sum \hat{b}_{l40} - \hat{b}_{r30}}{n}$	$\frac{\sum \hat{b}_{l40} - \hat{b}_{l30}}{n}$	$\sqrt{\sum \frac{(\hat{b}_{l40} - \hat{b}_{r30})^2}{n}}$	$\sqrt{\sum \frac{(\hat{b}_{l40} - \hat{b}_{l30})^2}{n}}$
	50	$\frac{\sum \hat{b}_{l40} - \hat{b}_{r50}}{n}$	$\frac{\sum \hat{b}_{l40} - \hat{b}_{l50}}{n}$	$\sqrt{\sum \frac{(\hat{b}_{l40} - \hat{b}_{r50})^2}{n}}$	$\sqrt{\sum \frac{(\hat{b}_{l40} - \hat{b}_{l50})^2}{n}}$

Where:

$\hat{a}_{l40}$ is the discriminant parameter estimate of a modified IRM for 40% of examinees with high motivation.

$\hat{a}_{r30}$ is the discriminant parameter estimate of a standard IRM for 30% of examinees with high motivation.

$\hat{a}_{r50}$ is the discriminant parameter estimate of a standard IRM for 50% of examinees with high motivation.

$\hat{a}_{l30}$ is the discriminant parameter estimate of a modified IRM for 30% of examinees with high motivation.

$\hat{a}_{l50}$ is the discriminant parameter estimate of a modified IRM for 50% of examinees with high motivation.

$\hat{b}_{l40}$ is the item level difficulty parameter estimate of a modified IRM for 40% of examinees with high motivation.

$\hat{b}_{r30}$ is the item level difficulty parameter estimate of a standard IRM for 30% of examinee with high motivation.

$\hat{b}_{r50}$ is the item level difficulty parameter estimate of a standard IRM for 50% of examinees with high motivation.

$\hat{b}_{l30}$ is the item level difficulty parameter estimate of a modified IRM for 30% of examinees with high motivation.

$\hat{b}_{l50}$ is the item level difficulty parameter estimate of a modified IRM for 50% of examinees with high motivation.

Table 4. Bias and RMSE for Student Ability Estimates for the Standard and Modified IRM

		Bias		**RMSE**	
Parameter	Percent of low motivation	Standard IRM	Modified IRM	Standard IRM	Modified IRM
$\boldsymbol{\hat{\theta}}$	30	$\frac{\sum \hat{\theta}_{l40} - \hat{\theta}_{r30}}{n}$	$\frac{\sum \hat{\theta}_{l40} - \hat{\theta}_{l30}}{n}$	$\sqrt{\sum \frac{(\hat{\theta}_{l40} - \hat{\theta}_{r30})^2}{n}}$	$\sqrt{\sum \frac{(\hat{\theta}_{l40} - \hat{\theta}_{l30})^2}{n}}$
	50	$\frac{\sum \hat{\theta}_{l40} - \hat{\theta}_{r50}}{n}$	$\frac{\sum \hat{\theta}_{l40} - \hat{\theta}_{l50}}{n}$	$\sqrt{\sum \frac{(\hat{\theta}_{l40} - \theta_{r50})^2}{n}}$	$\sqrt{\sum \frac{(\hat{\theta}_{l40} - \hat{\theta}_{l50})^2}{n}}$

Where:

$\hat{\theta}_{l40}$ is the student ability estimate of a modified IRM for 40% of examinees with low motivation.

$\hat{\theta}_{r30}$ is the student ability estimate of a standard IRM for 30% of examinees with high motivation.

$\hat{\theta}_{r50}$ is the student ability estimate of a standard IRM for 50% of examinees with high motivation.

$\hat{\theta}_{l30}$ is the student ability estimate of a modified IRM for 30% of examinees with high motivation.

$\hat{\theta}_{l50}$ is the student ability estimate of a modified IRM for 50% of examinees with high motivation.

Research question #5.

Participants. I used the same data as in research question 1 to address research question 5. This sample included 43,308 students in the academic program and 20,475 students in the applied program.

Instruments. I used the same self-report questionnaires as in research question 1. I only used, however, items related to math-values and interest as stated in Equations 15 and 16.

Procedures.

Select items. I used the same procedure to select items from student self-report questionnaires as in research question 1. This procedure was conducted separately for applied and academic programs.

Assessing component structure. I assessed the component structures of the selected items from student self-report questionnaires (SQ) using the same procedure as stated in research question 1.

Data analysis. I used a two level hierarchical linear model (HLM) as stated in Equations 23, 24, and 25 - students nested in schools. The HLM allowed me to determine the significance of the relationship between students' motivation and their mathematics achievement. The mathematic achievement scores were based on student overall outcome level. The first level of the HLM contained a fixed-model effect, which I used to determine how significant the component scores (math-values and interest) were in relation to students' academic achievement. The second level contained a random model effect, which I used to determine the impact of different schools on students' academic achievement at random. This analysis was conducted separately for academic and applied programs.

Level 1 or Fixed Effect Model: $\boldsymbol{y} = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ (23)

Level 2 or Random Effect Model: $\boldsymbol{\alpha_0} = \beta_0 + \beta_1 z_1$ (24)

Combined: $\boldsymbol{y} = \beta_0 + \beta_1 z_1 + \alpha_1 x_1 + \alpha_2 x_2 + r$ (25)

Where:

y= Students' academic achievement
x_1= math-values
x_2= student interest
r= fixed effect model residual variance
α_0= mean student academic achievement for a given school
β_0= intercept which represents the grand mean due to schools
β_1= variance of the intercept due to schools
α_1, α_2= coefficient for the fixed factors of level 1
z_1 = school categories

Chapter Four-Results

This chapter contains the results for the five research questions. The results for each of the research questions are summarized by means of figures and tables.

Research Question#1

Evidence of validity for the academic and applied program data. The results of the principal component analysis for students in the academic and applied program suggested that the students' self-reported EQAO questionnaire variables clustered into two components. I interpreted these two components as math-values and interest that relate to measures of student motivation as defined by expectancy-value theory and self-efficacy theory using Bandura (1997), Boaler (1999), Wigfield and Cambria (2010), and expert's opinion. For the academic program, the math-values component accounted for 54.60% and the interest component accounted for 22.84% of the total variance. For the applied program, the math-values component accounted for 59.50% and the interest component accounted for 27.49% of the total variance. Since varimax rotation was implemented, the two components (math-values and interest) were orthogonal. The outcome of the PCA analysis in combination with expert's opinion, reviewers' suggestions and revisions made to this research work, and the literature provides some content-related evidence of validity for the use of selected items from student self-report questionnaire as measures of motivation for the academic and applied program data. See Table 5 and 7 for results of the principal component analysis for academic and applied program data.

Table 5. Extracted Components from the Principal Component Analysis for Academic Students

	Components	
Variables	**Math-values**	**Interest**
I like math	**.807**	-.001
I am good in math	**.836**	-.209
I understand most of the mathematics I am taught	**.805**	-.146
The mathematics I learn now is very useful for everyday life	**.556**	.288
I need to keep taking mathematics for the kind of job I want after I leave school	**.503**	.221
Mathematics is an easy subject	**.760**	-.303
How much time do you usually spend on mathematics homework (in or out of school) on any given day?	.022	**.666**
How often have you been absent from your Grade-9 mathematics class this year?	.203	**.543**
How often have you been late for your Grade-9 mathematics this year?	.222	**.570**

Note. Variables with a component loading equal or higher than .40 were considered to have high loadings on a respective component. The last two variables from this table were reversed coded before conducting the PCA analysis.

The results from the chi-square analysis as depicted in Table 6 show a significant relationship between examinee motivation (“high and low” obtained from principal component scores using Equation 17) and SQ question 12, “Does counting the Grade-9 Assessment of Mathematics as part of your class mark motivate you to take the assessment more seriously?” In addition, I found a moderate effect size based on Cramer’s effect size calculations, $\chi 2$ (1, 8742) = 627, p<0.005, V = 0.30. This result provides some convergent-related evidence of validity for the academic program data.

Table 6. Observed Frequencies for Academic Students

		Examinee Motivation Level		**Total**
		Low	High	
SQ question 12 **“Does counting the Grade-9 Assessment of Mathematics as part of your class mark motivate you to take the assessment more seriously?”**	Yes	2332	5104	7436
	No	883	423	1306
Total		3215	5527	8742

Table 7. Extracted Components from the Principal Component Analysis for Applied Students

	Components	
Variables	**Math-values**	**Interest**
I like math	**.794**	-.053
I am good in math	**.804**	-.258
I understand most of the mathematics I am taught	**.749**	-.189
The mathematics I learn now is very useful for everyday life	**.537**	.293
I need to keep taking mathematics for the kind of job I want after I leave school	**.464**	.234
Mathematics is an easy subject	**.728**	-.362
How much time do you usually spend on mathematics homework (in or out of school) on any given day?	.162	**.524**
How often have you been absent from your Grade-9 mathematics class this year?	.235	**.615**
How often have you been late for your Grade-9 mathematics this year?	.263	**.605**

Note. Variables with a component loading equal or higher than .40 were considered to have high loadings in a respective component. The last two variables from this table were reversed coded before conducting the PCA analysis.

The results from the chi-square analysis as depicted in Table 8 show a significant relationship between examinee motivation (“high and low” obtained from principal component scores using Equation 17) and SQ question 12, “Does counting the Grade-9 Assessment of Mathematics as part of your class mark motivate you to take the assessment more seriously?”

Similar to the academic program data, I found a moderate effect size based on Cramer's effect size calculations, $\chi 2$ (1, 2032) = 253, p<0.005, V = 0.353. This result provides some convergent-related evidence of validity for the applied program data.

Table 8. Observed Frequencies for Applied Students

		Examinee Motivation Level		**Total**
		Low	High	
SQ question 12 **"Does counting the Grade-9 Assessment of Mathematics as part of your class mark motivate you to take the assessment more seriously?"**	Yes	458	1276	1734
	No	219	79	298
Total		677	1355	2032

Research Question#2

Bias and RMSE ability estimates between the two models (standard and modified) for academic and applied programs data, which contained 40 Percent of High Motivation for the modified model and 40 Percent of High and 40% of Low Motivation for the standard model. The total magnitude of bias and RMSE results stated in Table 9 revealed that examinee ability estimates for the academic program data for those examinees that exhibited high motivation were positively biased toward the modified model, meaning that examinee ability estimates appeared to be underestimated by the standard model for those examinees with high motivation when compared to the modified model. For middle ability estimates, however, the magnitude of bias and RMSE decreased and there was more agreement between the two models. As ability estimates decreased or increased from middle ability estimates for those examinees with high motivation, the models became more in disagreement with the standard model underestimating examinees abilities when compared to the modified model.

The results in Table 9 also revealed that the total average bias and RMSE for examinee ability estimates in the applied program for those examinees exhibiting high motivation were positively biased toward the modified model, meaning that examinee ability estimates appeared to be underestimated by the standard model when compared to the modified model. After examining the ability estimates in terms of low, middle, and high ability for those examinees exhibiting high motivation, the magnitude of bias and RMSE indicated that the two models were more in disagreement for low ability examinees and the standard model underestimated examinee abilities. As ability estimates increased, the models became more in agreement. In addition, it can be noticed that the total magnitude of bias and RMSE for examinee ability estimates between the standard and modified models for those examinees that exhibited high motivation were larger for examinees in the applied program than examinees in the academic program. Note that examinees in the academic program were considered to have higher ability than examinees in the applied program. This result indicates that including examinee with low motivation in the data calibration when using the standard model seemed to have a larger effect on examinee ability estimates for those examinees that exhibited high motivation but had lower ability.

Table 9. Ability Estimates between Modified and Standard Models

	Parameter	Group	Number of Examinees	BIAS	RMSE
Academic Data					
	Θ=ability	Low	3041	0.113	0.136
	Θ=ability	Middle	3041	0.083	0.107
	Θ=ability	High	3041	0.119	0.130
	Θ=ability	Total	9123	0.105	0.125
Applied Data					
	Θ=ability	Low	1304	0.187	0.195
	Θ=ability	Middle	1304	0.129	0.133
	Θ=ability	High	1304	0.085	0.097
	Θ=ability	Total	3912	0.134	0.147

Bias and RMSE for item parameter estimates between the two models (standard and modified) for academic and applied programs, which contained 40 Percent of High and 40 Percent of Low Motivation. The results in Table 10 revealed that for the academic program data, the magnitude of bias and RMSE for the discriminant (slope) and item level difficulty (threshold) parameters were negatively biased toward the standard model, meaning that the items appeared more difficult and more discriminating under the standard model when compared to the modified model. When comparing the model in terms of low, middle and high discriminant items, the results revealed that the two models were more in agreement for low discriminant items. For high discriminant items, however, the magnitude of bias and RMSE increased and the items appeared more discriminating under the standard model.

When comparing the models in terms of low, middle and high item level difficulty (threshold), the magnitude of bias and RMSE increased between the two models for the easy items, implying that the item level difficulty parameter appeared to be overestimated for easy items by the standard model when compared to the modified model. There was more agreement, however, between the two models for the more difficult items.

The results in Table 10 also revealed that for the applied program data, the magnitude of bias and RMSE for the discriminant (slope) and item level difficulty (threshold) parameters were negatively biased toward the standard model. This outcome implies that the parameter estimates were more discriminating and more difficult under the standard model when compared to the modified model. When comparing the item parameters in terms of low, middle and high discriminant items, as well as, item level difficulty, the models strongly agree for very difficult items and less discriminant items. As the items became easier or more discriminating, the magnitude of bias and RMSE increased. Said differently, the discriminant (slope) and item level difficulty (threshold) parameters seemed to be overestimated under the standard model for easy and more discriminating items.

Table 10. Item Parameter Estimates between Modified and Standard Models for Academic and Applied Programs Data

	Parameter	Item Group	Number of Items	BIAS	RMSE
Academic Data					
	a=slope	Low	8	-0.274	0.307
	a=slope	Middle	8	-0.443	0.477
	a=slope	High	8	-0.598	0.650
	a=slope	Total	24	-0.438	0.478
	b=Threshold	Low	8	-0.790	0.790
	b=Threshold	Middle	8	-0.582	0.603
	b=Threshold	High	8	-0.347	0.367
	b=Threshold	Total	24	-0.573	0.586
Applied Data					
	a=slope	Low	8	-0.176	0.190
	a=slope	Middle	8	-0.154	0.171
	a=slope	High	8	-0.207	0.273
	a=slope	Total	24	-0.179	0.211
	b=Threshold	Low	8	-0.622	0.709
	b=Threshold	Middle	8	-0.229	0.255
	b=Threshold	High	8	-0.050	0.109
	b=Threshold	Total	24	-0.323	0.357

Research Question#3

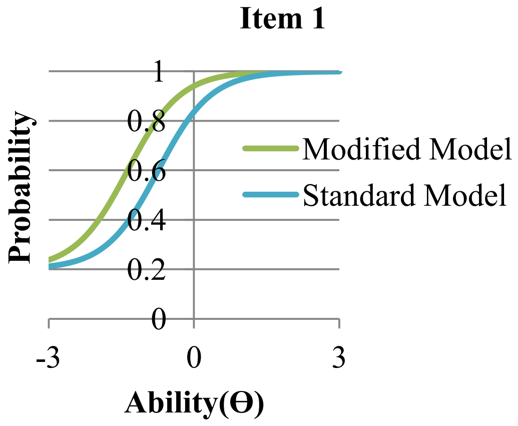
DIF model fit analysis for academic program data. The results for the academic program data indicated that under NON DIF assumption, the 2 log likelihood value obtained from the full calibration cycle of both models' data sets (standard and modified) combined as a single model was 45382. Under DIF assumption, however, the 2 log likelihood value from the full calibration cycle was 45139. The difference between the 2 log likelihood values for DIF and NON DIF was 243. This result represented a significant chi-square value distributed over four degrees of freedom, meaning that there was DIF present on the item level difficulty parameter estimates between the two models data sets (standard and modified) and that the DIF model was a better fit.

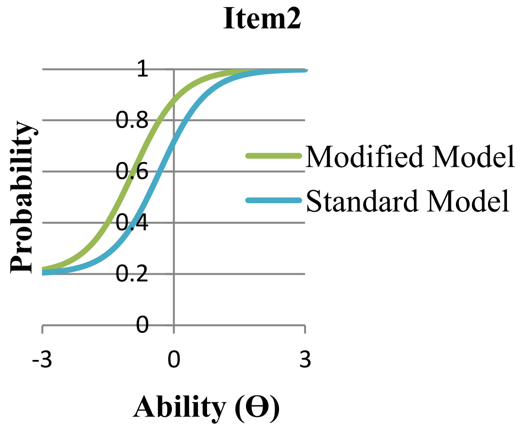
DIF item bias for academic program data. The results for DIF item bias in relation to item level difficulty (threshold) were obtained by using Lord (1990) chi-square statistics as stated in Equation 22. The rejection criteria value to identify the items that contained DIF based on Lord (1990) chi-square statistics was $\mathcal{X}^2_{(1,0.05)}$= 3.84. The c-parameter (pseudo-chance parameter) was held constant at a value of 0.20 for all items. The outcome of the data, as depicted in Table 11 and Figure 3, reveals that the two models shared a uniform DIF distribution with the same slope per item and a constant pseudo-chance parameter for all items. The probability estimates of correct responses per item as depicted in Figure 3 seem to be lower under the standard model than the modified model, that is, the items appeared to be more difficult under the standard model. In term of significant DIF per item, the chi-square values for items 1,2,3,8,10,11,12, 14, 15,16,17,18, 19, 21, 22, 23, and 24 indicated significant DIF between the two models distributed over one degree of freedom. The chi-square values for items 4,5,6,7,9,13 and 20, however, were lower than the rejection criteria and therefore, non-significant DIF was found for these items between the two models.

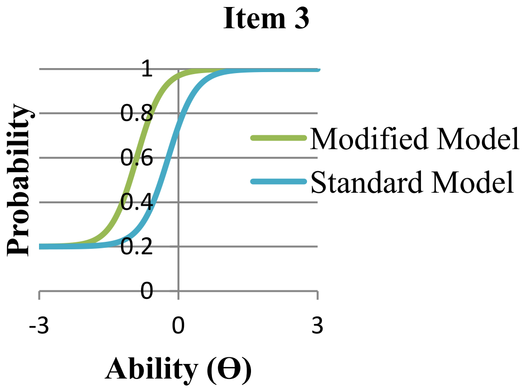
Table 11. Differential Item Functioning for Academic Program

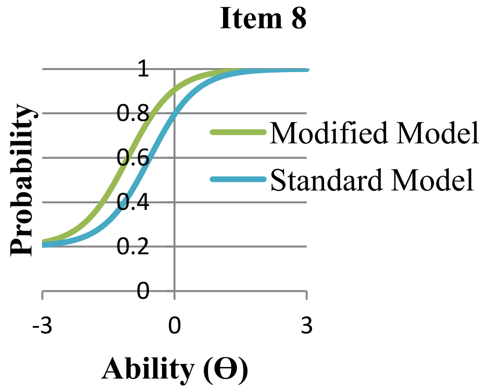
Item	a	Area	bS	bM	Vector Diff(b)	var(b1)+var(b2)	$\mathcal{X}^2_{(1,0.05)}$	Strand	DIF
1	1.08	0.12	-0.74	-1.37	-0.150	0.043	12.08	AG	yes
2	1.09	0.10	-0.32	-0.92	-0.125	0.036	12.06	NA	yes
3	2.00	0.16	-0.22	-0.90	-0.203	0.023	75.61	AG	yes
4	0.93	0.05	0.61	0.20	0.066	0.038	3.02	LR	no
5	1.44	0.03	0.33	-0.09	0.044	0.026	2.86	NA	no
6	1.79	0.00	-0.29	-0.29	-0.005	0.024	0.05	MG	no
7	0.93	0.04	0.39	-0.03	0.050	0.038	1.83	NA	no
8	1.12	0.01	-0.56	-1.06	-0.220	0.037	35.35	MG	yes
9	0.86	0.12	-2.24	-2.56	0.155	0.098	0.098	LR	no
10	1.12	0.04	-0.55	-0.98	0.051	0.038	13.57	AG	yes
11	1.32	0.17	-0.57	-1.27	-0.181	0.037	23.93	NA	yes
12	1.29	0.22	-1.04	-1.84	-0.280	0.046	37.05	NA	yes
13	1.82	0.01	-1.52	-1.55	0.013	0.046	0.09	MG	no
14	1.08	0.08	-0.65	-1.02	0.103	0.038	7.35	MG	yes
15	1.22	0.06	-0.25	-0.66	-0.076	0.031	6.01	LR	yes
16	1.30	0.05	0.71	0.29	0.060	0.033	4.00	AG	yes
17	0.79	0.11	-0.39	-0.68	0.194	0.040	23.52	AG	yes
18	1.69	0.06	0.05	-0.50	-0.084	0.024	12.25	AG	yes
19	0.77	0.17	-0.93	-1.19	0.222	0.050	19.36	LR	yes
20	1.10	0.01	-1.51	-1.97	0.016	0.060	0.070	MG	no
21	1.30	0.06	-0.10	-0.63	-0.077	0.030	6.59	MG	yes
22	0.71	0.16	-0.59	-0.87	0.200	0.051	15.38	LR	yes
23	1.58	0.04	0.32	-0.10	0.049	0.024	4.17	AG	yes
24	1.53	0.09	-0.73	-1.32	-0.117	0.030	15.21	LR	yes

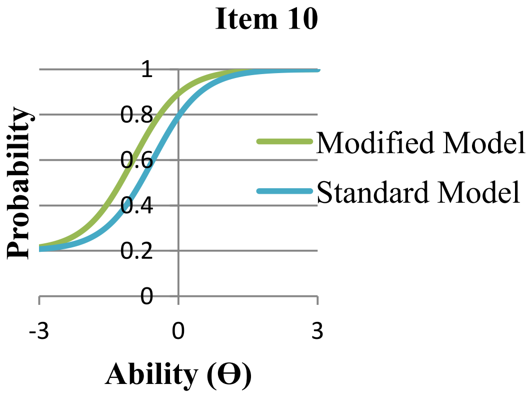
Note. a: slope; Area: area difference of item level difficulty between the two models; b:threshold; bS: item level difficulty standard model; bM: item level difficulty modified model, Vector diff(b): vector difference between the two models; var(b1)+var(b2): total variance of the difference; $\mathcal{X}^2$: Chi-square value given by vector diff(b)/(var(b1)+var(b2); Strand: (AG: Analytical Geometry; NA: Number Sense and Algebra; LR: Linear Relations; MG: Measurement and Geometry); DIF: differential item functioning significance.

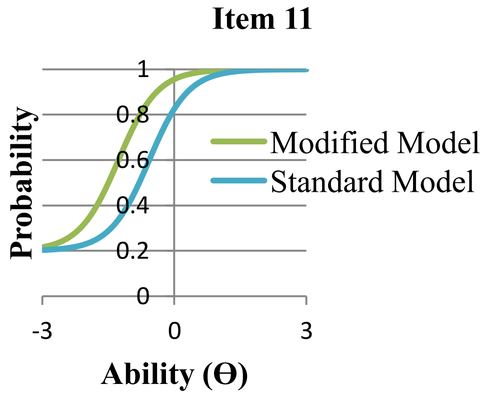

Item 1

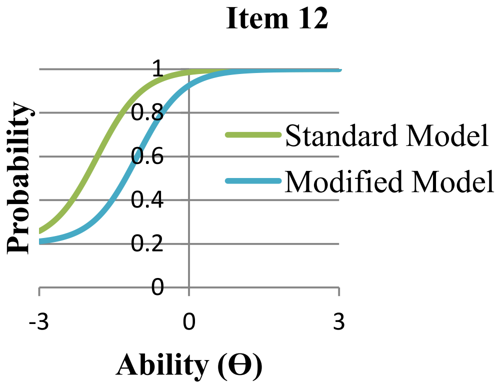

Item2

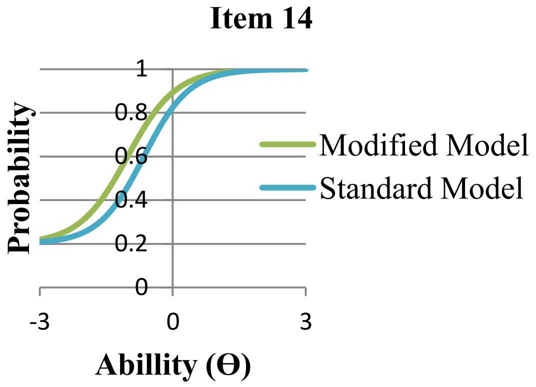

Item 3

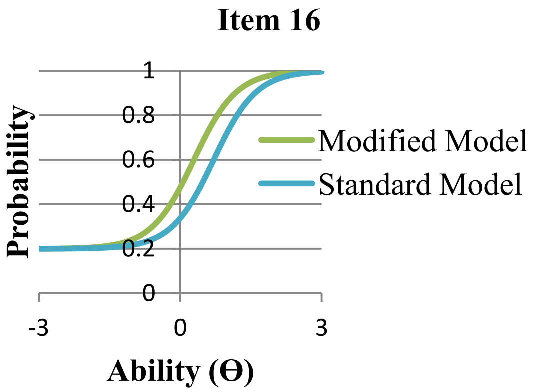

Item 8

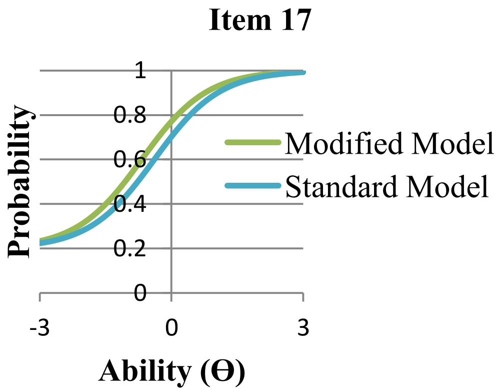

Item 10

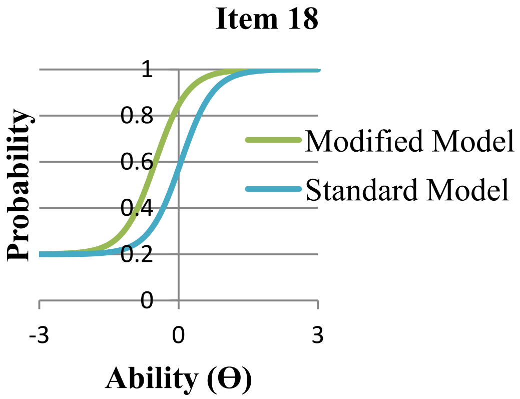

Item 11

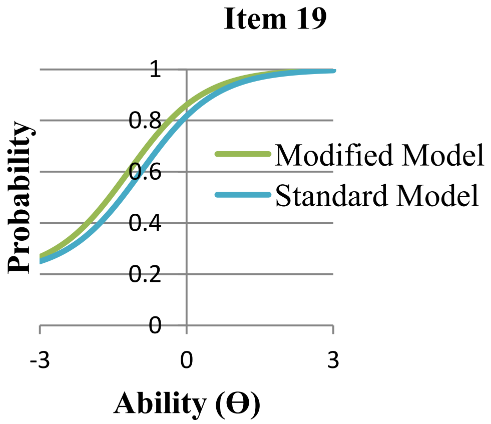

Item 12
Probability
1
0.8
0.6
0.4
0.2
0
-3
0
3
Ability (Θ)
Standard Model
Modified Model


Item 14
Probability
1
0.8
0.6
0.4
0.2
0
-3
0
3
Abillity (Θ)
Modified Model
Standard Model


Item 16
Probability
1
0.8
0.6
0.4
0.2
0
-3
0
3
Ability (Θ)
Modified Model
Standard Model


Item 17
Probability
1
0.8
0.6
0.4
0.2
0
-3
0
3
Ability (Θ)
Modified Model
Standard Model


Item 18
Probability
1
0.8
0.6
0.4
0.2
0
-3
0
3
Ability (Θ)
Modified Model
Standard Model


Item 19
Probability
1
0.8
0.6
0.4
0.2
0
-3
0
3
Ability (Θ)
Modified Model
Standard Model

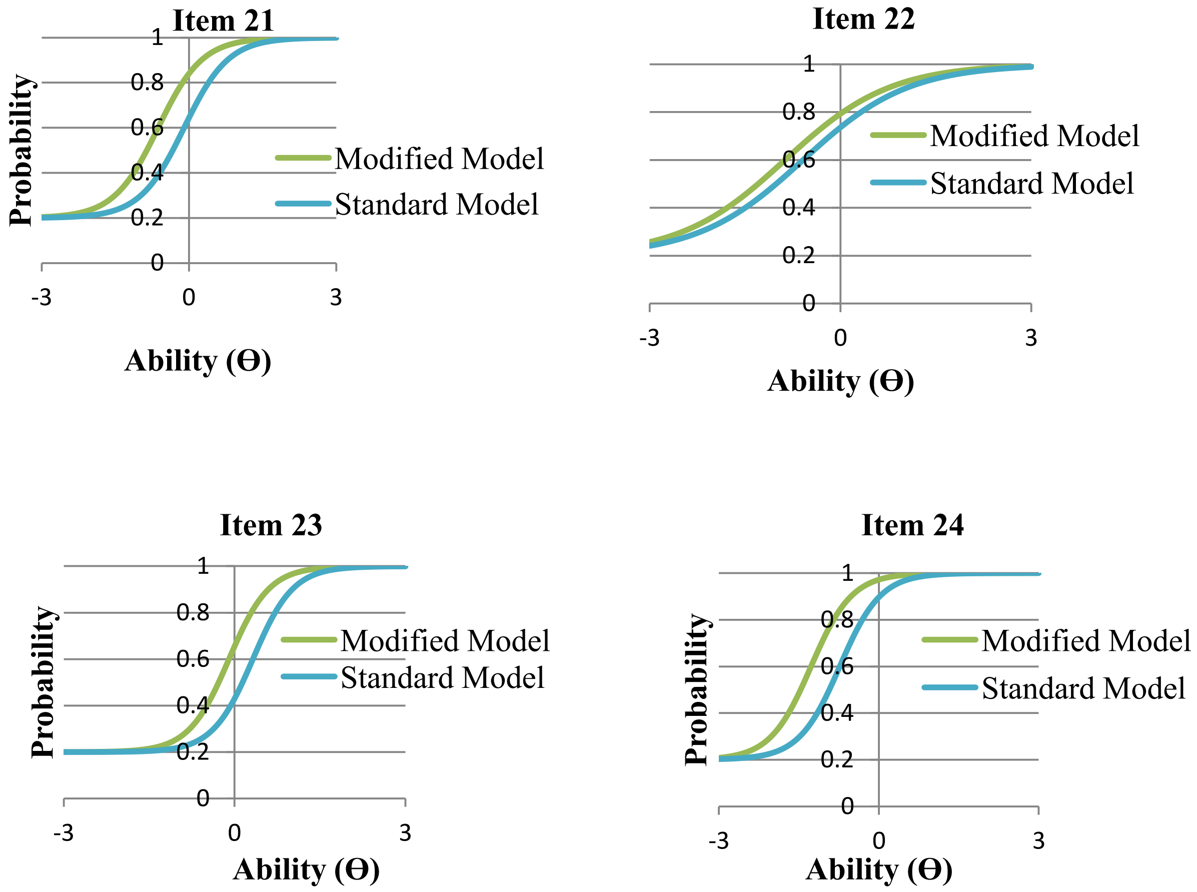


Figure 3: Differential Item Functioning for Academic Program Data

Note. The Y-axis represents probability of correct responses to a test item; X-axis represents examinee ability estimates; PM(Θ) represents probability of correct responses using the modified model; PS(Θ) represents probability of correct responses using the standard model.

DIF model fit analysis for the applied program data. The results for the applied program data indicated that under NON DIF assumption, the 2 log likelihood value from the full calibration cycle for both models data sets (standard and modified) combined as a single model was 312851. Under DIF assumption, however, the 2 log likelihood value from the full calibration cycle was 312551. The difference between the 2 log likelihood values for DIF and NON DIF was 300. This result represented a significant chi-square value distributed on four degrees of freedom, which indicated that there was DIF present on the item level difficulty parameter estimates between the two models (standard and modified) and that the DIF model was a better fit.

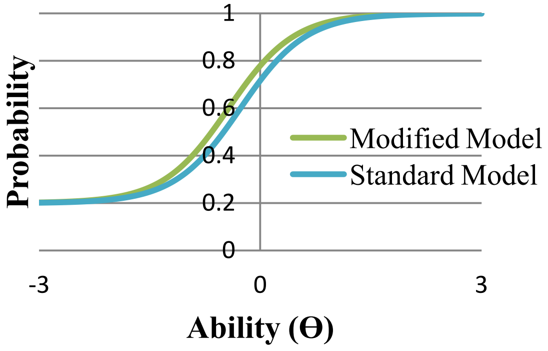
DIF item bias for applied program data. Similar to the academic data, the results for DIF item bias in relation to item level difficulty (threshold) were obtained by using Lord (1990) chi-square statistics as stated in Equation 22. The rejection criteria value to identify the items that contained DIF based on Lord (1990) chi-square statistics was $\mathcal{X}^2_{(1,0.05)}$= 3.84. The c-parameter (pseudo-chance parameter) was held constant at a value of 0.20 for all items. The results as in Table 12, Figure 4 indicate that the two models shared a uniform DIF distribution with the same slope per item and a constant pseudo chance parameter for all items. The probability estimates of correct responses per item as depicted in Figure 4 were lower under the standard model than the modified model, meaning that the items appeared to be more difficult under the standard model. I used Lord (1990) chi-square statistics to examine which items contained DIF between the two models as stated in Table 12. Significant DIF was found for items 1,9,10,12,15,16,20,22,23, and 24 in relation to item level difficulty as depicted in Figure 4.

Table 12. Differential Item Functioning for Applied Program

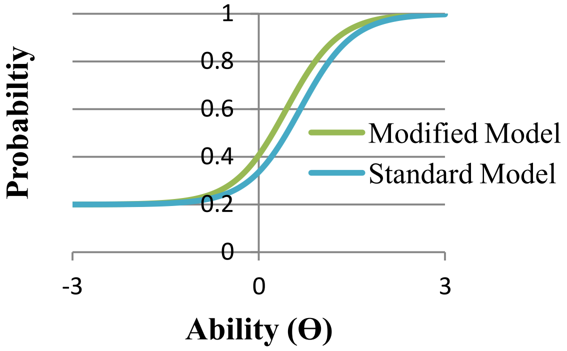
					Vector			Strand	DIF
Item	**a**	**Area**	**bS**	**bM**	**Diff(b)**	**var(b1)+var(b2)**	$\mathcal{X}^2_{(1,0.05)}$		
1	1.32	0.13	-0.27	-0.43	-0.16	0.045	12.64	NA	yes
2	1.33	0.03	-2.74	-3.01	-0.037	0.037	0.04	LR	no
3	2.11	0.06	0.44	0.13	-0.075	0.063	1.42	LR	no
4	0.90	0.03	0.72	0.53	0.041	0.063	0.42	LR	no
5	0.66	0.10	1.77	1.66	0.137	0.137	1.00	NA	no
6	1.33	0.05	0.36	0.22	0.071	0.043	2.73	MG	no
7	1.04	0.02	-1.85	-2.12	-0.035	0.108	0.11	MG	no
8	1.59	0.01	1.57	1.32	-0.010	0.068	0.02	LR	no
9	1.39	0.16	0.67	0.44	0.210	0.043	23.85	NA	yes
10	0.83	0.11	1.10	0.72	-0.140	0.070	4.00	NA	yes
11	0.62	0.04	-2.55	-2.72	0.062	0.200	0.10	LR	no
12	1.28	0.05	-1.17	-1.77	-0.124	0.057	4.73	LR	yes
13	0.91	0.00	1.15	0.92	0.002	0.075	0.00	NA	no
14	0.67	0.07	-2.52	-2.85	-0.090	0.091	0.22	LR	no
15	0.47	0.21	-0.84	-1.35	-0.274	0.130	4.44	NA	yes
16	1.22	0.11	0.35	0.26	0.146	0.044	11.01	MG	yes
17	1.05	0.04	-0.51	-0.68	0.061	0.050	1.49	MG	no
18	1.42	0.04	0.19	-0.10	-0.059	0.040	2.18	MG	no
19	0.99	0.03	1.32	1.13	0.039	0.078	0.25	LR	no
20	1.24	0.17	-0.34	-0.35	0.219	0.045	23.68	LR	yes
21	1.04	0.04	0.42	0.12	-0.061	0.050	1.49	LR	no
22	1.50	0.09	0.05	-0.30	0.121	0.039	9.63	MG	yes
23	0.87	0.13	-0.88	-1.28	-0.168	0.077	4.76	NA	yes
24	0.88	0.14	-0.83	-1.88	0.185	0.069	7.19	LR	yes

Note: a: slope; Area: area difference of item level difficulty between the two models; b:threshold; bS: item level difficulty standard model; bM: item level difficulty modified model, Vector diff(b): vector difference between the two models; var(b1)+var(b2): total variance of the difference; $\mathcal{X}^2$: Chi-square value given by vector diff(b)/(var(b1)+var(b2); Strand: (NA: Number Sense and Algebra; LR: Linear Relations; MG: Measurement and Geometry); DIF: differential item functioning significance.

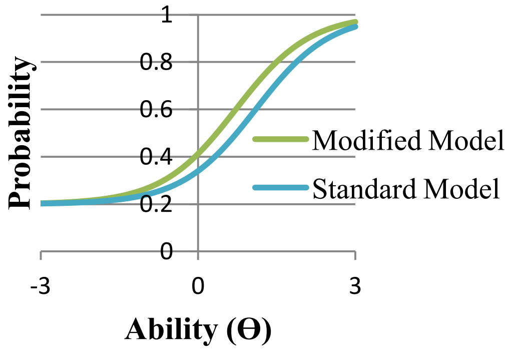
Item 1



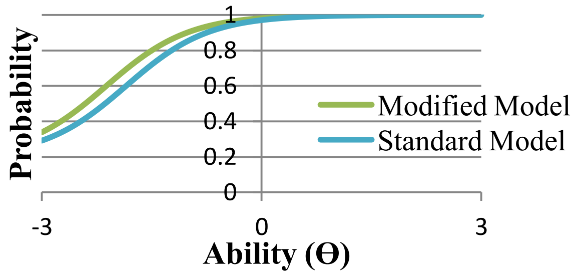
Item 9



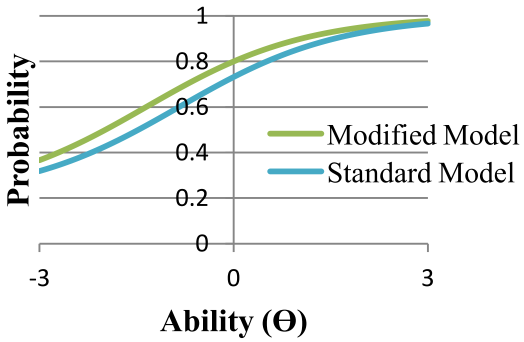
Item 10



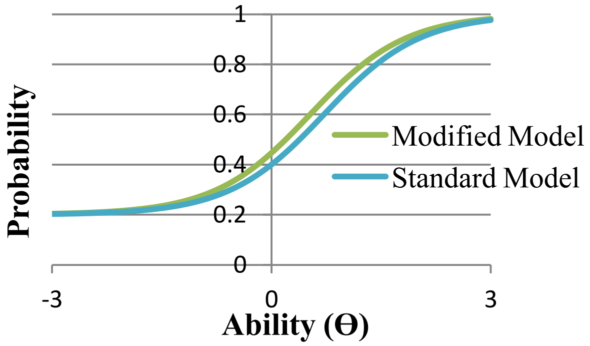
Item 12

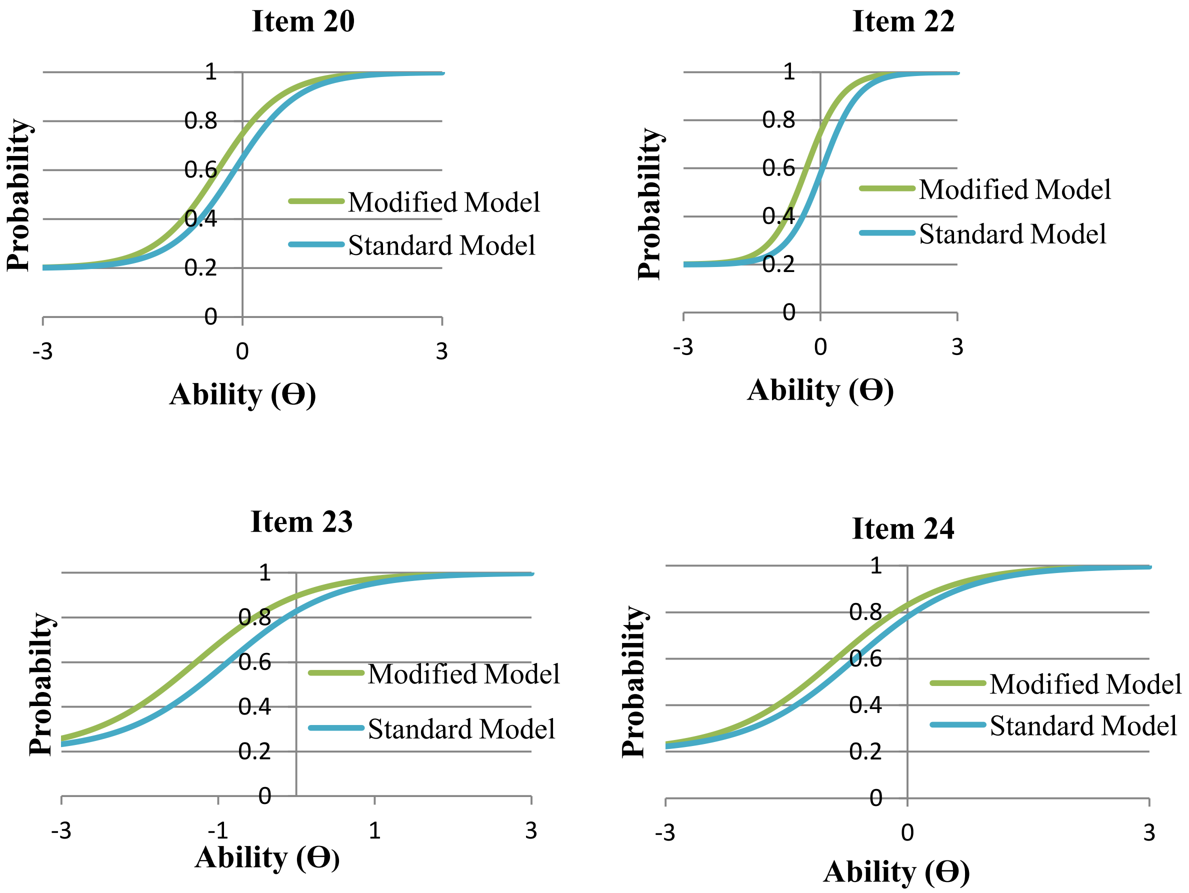


Item 15



Item 16





Item 20

Item 22

Item 23

Item 24

Figure 4: Differential Item Functioning for Applied Program Data

Note. The Y-axis represents probability of correct responses to a test item; X-axis represents examinee ability estimates; the green colour represents the probability of correct responses using the modified model; the blue colour represents the probability of correct responses using the standard model.

Most of the items with significant DIF clustered under the analytic geometry strand for the academic program data and under the number sense and algebra strand for applied program data. The area difference between the two models (standard and modified) was also computed for all items related to academic and applied programs data. These results revealed that the area difference between the two models under the analytical geometry strand for the academic program data and number sense and algebra for the applied program data was smaller as the items got more difficult, meaning that based on DIF area calculations, there was more agreement between the two models for difficult items and more disagreement for easy items.

Research Question#4

Comparing the estimates of ability between the two models for 30% and 50% of high motivation for academic and applied programs. The total magnitude of bias and RMSE in Tables 13 indicate that for those examinees who exhibited high motivation, examinees' estimates of ability using the standard model (calibrated with 30% of high and 40% of low motivation; 50% of high and 40% of low motivation) for either academic or applied programs data were lower than the reference, (calibrated only with 40% of high motivation using the modified model). As the proportion of examinees with high motivation increased, the standard model was more in agreement with the reference when estimating examinee abilities for those examinees that exhibited high motivation.

On the contrary, the total bias and RMSE results in Tables 13 indicate that examinees' estimates of ability calibrated with 30% and 50% of high motivation using the modified model were very similar to the estimates of examinee abilities using the reference for either academic or applied programs data. Note that this comparison is based on the same examinees with high motivation that were included in each proportional calibration (30% and 50%) and the reference. This result means that regardless of the proportion of examinee with high motivation present in the data, examinees' ability estimates appeared to be very consistent when using the modified model for either academic or applied programs data. Said differently, when low motivation was removed from the data using the modified model as a motivation filter, the magnitude of bias and RMSE decreased, that is, more accurate estimates of examinee abilities were obtained.

I also examined the magnitude of bias and RMSE for examinee ability estimates in terms of low, medium, and high ability when the two models (standard and modified) were compared to the reference for examinees that exhibited high motivation only; the results as stated in Table 13 revealed that when comparing the standard model (calibrated with 30% of high and 40% of low motivation; 50% of high and 40% of low motivation) to the reference (calibrated only with 40% of high motivation under the modified model), there was disagreement between the two calibration comparisons across all examinee ability levels for examinees that exhibited high motivation. Higher magnitudes of bias and RMSE were found, however, for those examinees with low ability in the applied program data. Conversely, when comparing the modified model calibration (30% and 50% of high motivation) to the reference (40% of high motivation under the modified model) for the same examinees included in each proportion of high motivation (30% and 50%) and the reference in terms of low, middle, and high examinees' abilities estimates, the calibrations were consistent across all levels of examinee ability estimates. Said differently, there seemed to be no differences between the two proportional calibrations (30% high motivation and 50% high motivation) and the reference in terms of examinee ability estimates for different levels of ability (low, middle, and high) when using the modified model for either academic or applied programs data. This result means that the modified model behaved as a low motivation filter across all levels of examinee abilities producing more accurate estimates of examinee abilities when compared to the reference.

Comparing the estimates of item parameters between the two models for 30% and 50% of high motivation for academic and applied programs. The magnitude of bias and RMSE results in Table 14 indicate that the parameter estimates (item level difficulty and slope) under the standard model (calibrated with 30% of high and 40% of low motivation; 50% of high and 40% of low motivation) were higher than the parameter estimates of the reference (calibrated only with 40% of high motivation under the modified model) for either academic or applied programs data. This means that the item parameters appeared to be more difficult and more discriminating under the standard model. As the proportions of high motivation increased, there was more agreement between the standard model and the reference.

On the contrary, when comparing the item parameter estimates (threshold and slope) under the modified model (calibrated with 30% and 50% of high motivation) to the parameter estimates of the reference, the magnitude of bias and RMSE decreased. That is, there seemed to be more consistency in item parameter estimates (threshold and slope) when using the modified model, which neglected the effect of low motivation in the calibration process.

Table 13. Bias and RMSE of Examinee Abilities for Different Proportions of High Motivation (30% and 50%) in the Standard and Modified Models Compared to 40 Percent of High Motivation in the Modified Model for Academic and Applied Programs

				BIAS		RMSE	
	Percent of High Motivation	Ability Group (Θ)	Number of Examinees	Standard	Modified	Standard	Modified
Academic							
	30	Low	1635	0.106	0.076	0.133	0.078
		Middle	1635	0.128	0.068	0.145	0.070
		High	1635	0.160	0.054	0.168	0.055
		Total	4095	0.131	0.066	0.150	0.068
	50	Low	3041	0.105	0.012	0.124	0.063
		Middle	3041	0.067	0.016	0.089	0.063
		High	3041	0.098	0.029	0.108	0.031
		Total	9123	0.090	0.019	0.108	0.077
Applied							
	30	Low	738	0.234	0.008	0.243	0.024
		Middle	738	0.169	0.019	0.175	0.030
		High	738	0.123	0.031	0.134	0.037
		Total	2214	0.175	0.019	0.189	0.030
	50	Low	1304	0.220	0.031	0.228	0.034
		Middle	1304	0.111	0.030	0.118	0.034
		High	1304	0.018	0.027	0.060	0.031
		Total	3912	0.116	0.029	0.135	0.033

Table 14. Item Bias and RMSE for Standard and Modified Models for Different Proportions of High Motivation Compared to 40 Percent of High Motivation in the Modified Model for Academic program

				BIAS		RMSE	
	Percent of High Motivation	Item Parameter	Number of Items	Standard	Modified	Standard	Modified
Academic							
	30	a	24	-0.174	0.005	0.280	0.034
	30	b	24	-0.713	0.100	0.750	0.112
	50	a	24	-0.149	-0.013	0.225	0.044
	50	b	24	-0.514	-0.088	0.555	0.100
Applied							
	30	a	24	0.010	-0.015	0.131	0.068
	30	b	24	-0.392	0.010	0.449	0.127
	50	a	24	0.010	0.015	0.090	0.044
	50	b	24	-0.264	-0.038	0.320	0.080

Research Question#5

Student motivation and academic achievement using academic program data. The results from the hierarchical linear model (HLM) conducted on the academic program data as depicted in Table 15 suggest that the math-values and interest components were significant predictors of students' academic achievement for the academic program. Although significant in the first-level analysis (fixed analysis effect model) of the HLM, these predictors (math-values and interest) only accounted for 18% of the variance. This result means that 82 % of the variance for the academic program data was left unexplained in level 1. The second level of the HLM analyses (random model effect) as depicted in Table 16, accounted for 8.9% of the variance. The intra-class correlation coefficient computation showed that from the total variance accounted for at the second level (8.9%), there was 9.3% of between school variance and 90.6% of within school variance that affected students' academic achievement. In addition, the total coefficient of determination (R^2 = 26.90%) between the observed and predicted students' achievement data computed for the entire HLM model for academic data suggested that the motivation components (math-values, student interest, and score variability within and between schools) accounted for 26.90% of the variance in relation to students' academic achievement. This means that 73.10% of the variance was unaccounted for by the HLM model and this variance might be related to other factors besides motivation and student score variability within and between schools that affected students' academic achievement.

Table 15. HLM Fixed Analysis Effect Model for Academic Program

Source	Numerator df	Denominator df	F	Sig.	R square
Intercept	1	592.83	127058.00	**.000**	
Math-Values	1	52396.05	11624.79	**.000**	**.18**
Interest	1	52566.96	148.00	**.000**	

Note: Dependent Variable- Academic Achievement

Table 16. HLM Random Analysis Effect Model for Academic Program

Parameter	Estimate	Std.Error	Wald Z	Sig
Residual	**.293**	.001	161.20	**.000**
Intercept/Variance Subject=School	**.036**	.002	15.11	**.000**

Note: Dependent Variable- Academic Achievement

Student motivation and academic achievement using applied program data. The results from the hierarchical linear model (HLM) conducted on the applied program data as depicted in Table 17 suggest that the math-values and interest components were significant predictors of students' academic achievement for the applied program. Although significant in the first-level analysis (fixed analysis effect model) of the HLM, these predictors (math-values and interest) only accounted for 14% of the variance. This result means that 86 % of the variance for the applied program data was left unexplained in Level 1. The second level of the HLM analyses (random model effect) as depicted in Table 18, accounted for 12.9% of the variance. The intra-class correlation coefficient computation showed that from the total variance accounted for at the second level (12.9%), there was 12.6% of between school variance and 87.3% of within school variance that affected students' academic achievement. In addition, the total coefficient of determination (R^2 = 24.90%) between the observed and predicted students' achievement data computed for the entire HLM model for applied data suggested that the motivation components (math-values, student interest, and score variability within and between schools) accounted for 24.90% of the variance in relation to students' academic achievement. This result means that 75.10% of the variance was unaccounted for by the HLM model and this variance might be related to other factors besides motivation and student score variability within and between schools that affected students' academic achievement in the applied program.

Table 17. HLM Fixed Analysis Effect Model for Applied Program Data

Source	Numerator df	Denominator df	F	Sig.	R square
Intercept	1	611.91	26530.21	**.000**	
Math-Values	1	22301.86	3576.36	**.000**	**.14**
Interest	1	41549.10	394.56	**.000**	

Note: Dependent Variable- Academic Achievement

Table 18. HLM Random Analysis Effect Model for Applied Program Data

Parameter	Estimate	Std.Error	Wald Z	Sig
Residual	**.69**	.006	104.68	**.000**
Intercept/Variance Subject=School	**.10**	.007	14.33	**.000**

Note: Dependent Variable- Academic Achievement

Chapter Five-Discussion

This chapter is divided into nine sections. The first five sections contain a discussion of the results. The remaining four sections contain discussions about limitations of the study, conclusion, recommendations, and future directions.

Research Question#1

Although EQAO student self-report questionnaires were not designed as measures of student motivation, the results from the expert's opinion, the principal component analysis, and the literature provided some content-related evidence of validity that it is possible to measure student motivation using these self-report data (Bandura, 1997; Boaler, 1999; Cole, Bergin, & Whittaker, 2008; Eccles & Wigfield, 2002; Putwain, 2008; Wigfield & Cambria, 2010). By relating the component scores of math-values and interest to an external variable such as question 12, which asked about student motivation and test stakes, a significant relationship with a medium effect size was found through the chi-square analysis. This result provided some convergent-related evidence of validity. While the literature shows that one of the challenges when using self-report questionnaires is to select a questionnaire that has strong evidence of validity measures (Cole, Bergin, & Whittaker, 2008; Eccles & Wigfield, 2002; Harlen & Crick, 2003; Putwain, 2008; Wigfield & Cambria, 2010), the outcome of this study showed some content-related and convergent-related evidence of validity when using math-values and interest components as measures of student motivation from a large-scale assessment.

Ideally, a researcher would validate the principal components and motivation behaviour scores by administering the nine self-report items used in this study as measures of motivation in conjunction with an established motivation scale such as the student opinion survey (SOS) (Sundre & Moore, 2002; Sundre & Wise, 2003) immediately after the test administration to provide concurrent-related evidence of validity. Since the current study is based on secondary data analysis, I used the existing nine self-report items to create a motivation measure, although these items were not specifically designed for that purpose. I was able, however, to relate the self-report items to general achievement motivation but not to test-taking motivation. There is some evidence in the literature to support that test-taking motivation and motivation to learn a subject (as it relates to general achievement motivation) are distinct but related (Eklof, 2006). I felt that there was some evidence of validity from the principal component analysis, expert's opinion, question 12, and the literature grounded on expectancy-value theory and self-efficacy theory to used the PCA components (math-values and interest) as a proxy to measure motivation. The results of the analysis of research question 1 allowed me to build on the research work of Swerdzewski, Harmes, and Finney (2011) by dichotomizing motivation components (math-values and interest) into examinees with high and low motivation using student self-report questionnaires from a large-scale assessment. The results of research question 1 also provided me with an avenue that can be used to build on the literature (Wise & DeMars, 2005, 2006; Wolf & Smith, 1995) to further examine the effect of removing examinees with low motivation on the estimates of students' abilities and test item parameters when using an item response theory model in combination with measures of student motivation based on self-report data.

Research Question#2

Bias and RMSE in ability estimates for academic and applied programs data. By including a motivation parameter in an item response model to estimate student abilities, it was possible for me to examine the effect of removing examinees with low motivation on the estimates of student abilities that exhibited high motivation and who were included in both models (standard and modified) calibrations for academic and applied programs based on EQAO

Grade-9 assessments of mathematics. The bias and RMSE results revealed that for both programs (academic and applied), student ability estimates for examinees that exhibited high motivation and were included in both models calibrations seemed to be underestimated by the standard model due to the effect of low motivation. This outcome can be explained under the rationale that including examinees with low motivation in the data calibrations when using an item response model creates a source of construct-irrelevant variance that affects the estimates of examinee abilities and the validity of test data interpretations (Cole, Bergin, & Whittaker, 2008; DeMars, 2000; Wise, Wise, & Bhola, 2006).

In the case of the academic program data, when examining the standard and modified models in terms of bias and RMSE for low, middle, and high abilities estimates for examinees with high motivation only, the models appeared to be more in disagreement for low and high ability estimates. These findings are consistent with previous research, that is, examinee with low motivation who are either low or high ability may guess or omit items and including them in the model calibrations causes an underestimation of examinee abilities (Cole, Bergin, & Whittaker, 2008; De Ayala, Plake, & Impara, 2001).

In the case of the applied program, which contained examinees that were considered to have lower ability estimates than those examinees in the academic program, the standard and modified models appeared to agree more for high ability estimates for those examinees with high motivation who were included in both models calibrations as measured by the bias and RMSE. This outcome suggests that including examinees with low motivation in an item response model calibration seems to have a lower effect on ability estimates for those examinees with high motivation but lower ability. These findings seem to support the literature under the rationale that high motivation even under consequential testing does not contribute to test scores above or beyond students' knowledge and self-regulatory strategies (Cole et al., 2008), that is, higher motivation does not have much of an impact on examinee ability estimates far and beyond their topic knowledge level (DeMars, 2000).

In summary, using a standard item response theory model in the presence of low motivation, that is, including examinees with low motivation in the data calibration to estimate student abilities for EQAO Grade-9 assessments of mathematics seems to produce a source of construct-irrelevant variance that may pose a threat to the validity of the interpretation of the results. It may be possible that by not accounting for the effect of low motivation in the estimate of examinee abilities, educational agencies, policy makers, school boards and teachers can misunderstand student ability estimates and therefore, make the wrong decision in the allocation of funding and curriculum changes. As Palomba and Banta (1999) stated, this low motivation effect raises a concern of whether data collected from large-scale assessments are valid measures of student achievement. Using a modified item response model minimizes the effect of low motivation as a source of construct-irrelevant variance to more accurately estimate student mathematical abilities (Wise & DeMars, 2006) because it can function as a low motivation filter. That is, responses from low motivated examinees are systematically removed from the test data (Swerdzewski, Harmes, & Finney, 2011; Wise, Wise, & Bhola, 2006).

Bias and RMSE in parameter estimates for academic and applied programs. Although I used a different method to measure student motivation, the results of this study seem to be consistent with the findings of Wise and DeMars (2006). The results were that the threshold and slope parameter estimates were negatively biased when using the modified model as reference and the standard model as focal. The items appeared to be more difficult and more discriminating under the standard model for both applied and academic programs.

For the item level difficulty, similar to Wise and DeMars (2006), there was more agreement between the two models in terms of threshold parameter estimates for more difficult items and more disagreement for easier items. More disagreement for easier items can be explained by the notion that under the effect of low motivation, examinees, especially those with high ability, may not provide a correct response to the test items (Wise & DeMars, 2005; Wise & Kong, 2005; Wise, Wise, & Bhola, 2006) and may opt to guess or omit easy items, which causes an overestimate of item difficulty parameters on the test (De Ayala et al., 2001; Wise & DeMars, 2006; Wolf & Smith, 1995; Zhang & Walker, 2008). On the contrary, as the items became more challenging, the magnitude of bias and RMSE between the modified and standard model seemed to indicate that there was a lower effect when including examinees with low motivation in the model calibration. While the modeling techniques used in this study do not account for a change on examinee level of motivation per item like item response time modeling techniques, one explanation for this outcome may be that examinees found some of these difficult items to be engaging, especially those examinees with low motivation and higher ability that were included in the standard model calibration and as a result generated more correct responses to these difficult items. As Printrich and Schunk, (2004) stated students' engagement increases when they are given tasks that are challenging but that can be accomplished based on student ability level.

In the case of the discriminant parameter, similar to Wise and DeMars (2006), the items appeared to be more discriminating under the standard model. One explanation for this, as stated in Wise and DeMars (2006), may be that the discriminant parameters from the standard model were spuriously high due to the present of examinees with low motivation, which caused the discriminant parameters to be more discriminating in the standard model.

Overall, my results indicate that both examinee ability and test item parameters seemed to have been influenced by the effect of student low motivation in both academic and applied programs. Examinee ability estimates for those examinees with high motivation seemed to be underestimated and test item parameter estimates seemed to be overestimated when using a standard item response model. This outcome seems to indicate that including examinee with low motivation in current item response model calibrations without motivation filtering may pose a potential threat to the validity of test score interpretations from large-scale assessments such as EQAO.

The outcome of this study in relation to question 2 builds on the research work of Swerdzewski's et al. (2011) and supports the work of Wise and DeMars (2006) and Wise, Wise, and Bhola (2006) because it used student self-report data to account for a motivation parameter in an item response model to minimize the effect of low motivation as a source of construct-irrelevant variance. This approach may provide another avenue for test administration agencies such as EQAO to account for the effect of low motivation when administering pencil and paper exams and more accurately estimate student abilities and test item parameters, from large-scale assessment data.

Research Question#3

Low motivation not only affected the parameter estimates but also it appeared to influence how examinees responded to different kinds of items on the test (Wolf & Smith, 1995). For students in the academic and applied programs, there was significant DIF between the groups of examinees calibrated using the standard model and modified model for some items. For the academic program data, the majority of the items (MC01, MC03, MC10, MC16, MC17, MC18 and MC23) that manifested DIF grouped in analytical geometry. For the applied program data, however, the majority of the items (MC01, MC09, MC10, MC15 and MC23) that manifested DIF grouped in number sense and algebra. It is important to clarify here that analytical geometry for Grade-9 EQAO exams only pertained to students in the academic program. The reason for the majority of the items with DIF clustering in the analytical geometry strand for the academic program and number sense and algebra for the applied program may be that examinees found these tasks to be more mentally taxing. For instance, analytical geometry requires students to use spatial visualization and computation for 2 or 3 dimensional objects. Spatial visualization varies based on the individual's ability to perform cognitive tasks related to geometric shapes in 2 or 3 dimensional space (Salthouse & Mitchell, 1990) that can result in a mentally taxing process. Number sense and algebra also require the use of a number of steps involving proportional reasoning and algebra manipulation that can also be mentally taxing, especially for those individuals with lower ability (Ontario Mathematics Curriculum Grade 9, 2011; Wolf & Smith, 1995). To further support the previous claims, it is stated in the literature (Wolf et al., 1995) that students perform fairly well on items that are not mentally taxing, but have more difficulty with more complex items or items with multiple steps. In addition, mentally taxing items are found to have the strongest relation to DIF (Wolf et al., 1995). This finding may also explain why the majority of the items related to analytical geometry in the case of academic program and the majority of items related to number sense and algebra in the case of applied program were more susceptible to manifesting DIF under the effect of low motivation.

In the case of the academic program and under the DIF analysis, the area differences between the two models were smaller as the items got more difficult for the analytical geometry strand. In the case of the applied program, the area differences were also smaller as the items got more difficult for the number sense and algebra strand. A similar trend was found in Wise and

DeMars' (2006) research and the results of research question 2 in the current study, that is, mean differences, biases, and RMSE between the two models were smaller as the items got more difficult. One of the reasons for these outcomes may be that students found more difficult items to be less taxing or perhaps more engaging and therefore, generated more correct responses to these items (Wolf & Smith, 1995).

The DIF results in the current study reinforce the praxis of DIF (Miller, Chahine, & Childs, 2010; Zumbo, 2007) as an avenue to examine the effect of low motivation on test item bias. They also support the literature under expectancy-value and self-efficacy theory of motivation, which states that when students encounter an item, they determine how likely they are to get the item right if attempted, which underlines item difficulty or how much work they will have to put forth to reach a correct answer, which underlines mental taxation (Printrich, 1988; 1989; Printrich & Schunk, 2004).

The DIF results support the need for low motivation filtering in large-scale assessments such as EQAO to provide more accurate estimates of test item parameters (i.e., item level difficulty parameter). As the DIF results in the current study indicate, under the effect of low motivation, the item level difficulty parameters appeared to be more difficult than when the low motivation effect was removed for both applied and academic programs. Some items, however, were more affected by low motivation than others, especially items that had the tendency to be more mentally taxing (i.e., analytical geometry and number sense and algebra).

Finally, the DIF results seem to provide another avenue for researchers and practitioners to use hybrid models (i.e., the effort moderated model or the modified model) in combination with DIF modeling techniques to identify items that are affected by the effect of low motivation.

This can have implications for test designs and data interpretations (Eklof, 2006; Wise, Wise, & Bhola, 2006; Wise & DeMars, 2006) when using large-scale assessments such as EQAO.

Research Question#4

By changing the proportions of high motivation and keeping the low motivation proportion constant in both academic and applied programs data, it was possible to examine the effect of different levels of motivation filtering in the estimates of examinee abilities (examinees who exhibited high motivation only) and test item parameters for the standard and modified models. The results revealed that the bias and RMSE results of the standard and modified models when compared to the reference (calibrated with 40% of high motivation) were consistent with those found in research questions 2, 3 and previous research (DeMars, 2000; Wise & DeMars, 2006). That is, examinee ability estimates seem to be underestimated for those examinees who exhibited high motivation and test item parameters (item level difficulty and discriminant parameters) overestimated by the standard model when compared to the reference for either proportion (30% and 50%) of high motivation. There seem to be more agreement, however, between the standard model and reference when estimating examinee abilities (examinees who exhibited high motivation only) and test item parameters for the 50% proportion of high motivation. This means that it is possible to minimize the effect of low motivation by increasing the proportion of examinees with high motivation. This result is consistent with the literature (Swerdzewski et al., 2011 ; Wise,Wise, & Bhola, 2006) and it seems to provide an avenue for researchers and practitioners who may be concerned with decreasing the sample size under low motivation filtering, but are still interested in providing more accurate interpretations of the test results from the large-scale assessment data during low stake situations.

Examinee ability estimates (for examinees with high motivation only) and item parameter estimates under the modified model were very consistent to the reference for either proportion (30% and 50%) of high motivation. This indicates to me that the modified model exhibited more desirable psychometric characteristics than the standard model. For example invariance was preserved, when comparing the modified model calibrations (30% and 50%) to the reference. This was not the case, however, when comparing the standard model calibrations (30% and 50%) to the reference. These findings support research questions 2 and 3, and are consistent with the literature (Wise & DeMars, 2006; Wise & Kong, 2005; Zumbo, 2007).

Research Question#5

The results of the first level (fixed model effect) of the HLM(s) statistical analyses conducted on the Grade-9 students' self-reported questionnaire EQAO data suggest that students' math-values and interest are significant predictors of their academic achievement on the EQAO test for both academic and applied programs data. There was 18% of the variance accounted for by the first level of the HLM analysis for the academic program data and 14% of the variance accounted for by the first level of the HLM analysis for the applied program data that affected students' academic achievement in relation to their motivation levels (math-values and interest). This result left 82% of the variance for the academic program data and 86% of the variance for the applied program data unexplained within the first level, which may be related to other factors (i.e., test anxiety) not accounted for in this model.

The outcomes of the HLM first level reinforce the need to explain the effect of motivation on student academic performance on large-scale assessments and the need to include motivation as a parameter estimate when using item response theory models to estimate examinee abilities and test item parameters (Linn & Baker, 1996; Maehr & Meyer, 1997; Meijer, 1996; Meijer & Sijtsma, 2001; Schmitt, Sacco, McFarland, & Jennings, 1999, Wise,Wise, & Bhola, 2006). This information may be relevant for teachers and educational agencies such as EQAO to help them account for the effect of motivation on student academic achievement to make more valid interpretations of large-scale assessment data when assessing student academic performance.

The second level of the HLM analysis suggests that the variance within and between schools is also a significant predictor of student academic achievement for EQAO Grade-9 mathematics assessments for both academic and applied programs. There is 90.6% of score variability within schools and 9.3% of score variability between schools based on 8.9% of the total variance accounted for at this level for the academic program data. Similarly, there is 87.3% of score variability within schools and 12.6% of score variability between schools based on 10.9% of the total variance accounted for at this level for applied program data. The variance accounted for at the second level for both academic and applied programs data, however, may be related to specific school level variables that were not measured in student EQAO self-report data. According to Ma and Klinger (2000), school level variables may entail the disciplinary climate of the schools and academic press. Disciplinary climate relates to the rules of the school to monitor student behaviours in class and within the school (Ma & Klinger, 2000). Academic press relates to student attendance in class and how important it is to do well in schools through good effort (Ma & Klinger, 2000). While the current study did not address these variables, it may be an explanation for the variability within and between schools that affected student academic achievement, since academic press and disciplinary climate are predictors of academic achievement (Ma & Klinger, 2000).

The total coefficient of determination between the observed and predicted values was computed separately for both HLM models (academic and applied). The results indicate that the HLM models accounted for 26.90% of the variance in the academic program data and 24.90% of the variance in the applied program data when combining the first and second level of the HLM models. These outcomes indicate that student math-values, interest and school level variables seem to be significant predictors of academic achievement for EQAO data.

Limitations

Uni-dimensionality. The assumption of uni-dimensionality of the item-response models under the principle of local independence, as the complete latent space representing an examinee's mathematical performance was assumed to consist of a single ability (Hambleton, Swaminathan, & Roger, 1991; MacDonald, 1999). This notion was tested, however, by using an exploratory factor analysis for English academic item response data and an exploratory factor analysis for English applied item response data. From a cluster of 24 items, one component was identified to account for a large proportion of the variance. This proportion was 76% for the English academic program data and 75% for the English applied program data. Since in both programs (academic and applied), one component accounted for a large proportion of the variance, this outcome indicated that there was one single ability representing examinee mathematics performance (Sheng, 2007).

Self-report measures of motivation. The student motivation components (math-values and interest) on EQAO Grade-9 mathematics exams were based on student self-reported questionnaires, which were not originally designed as a measure of motivation. In addition, the items used to measure motivation relate to general achievement motivation and not test-taking motivation. While the degree to what these two types of motivation are related (general achievement or motivation to learn a subject and test-taking motivation) has not been extensively investigated, there is some evidence in the literature to support that although distinct, they are related (Eklof, 2006). I collected some evidence of validity for the use of the EQAO self-report items as measure of motivation, by referring to research work done by Wigfield and Cambria (2010), Bandura (1997), soliciting expert's opinion, reinterpreting the results based on the reviewers' suggestions and revisions made to this research study and by examining the motivation component results (math-values and interest) in relation to question 12 from student self-report questionnaire data that asked about student test-taking motivation and test stakes. It should also be noted that not all the motivation values as defined by the expectancy-value theory could be identified from EQAO student self-report questionnaire data. For example I could not identify cost values, which can affect the measures of students' motivation in relation to their reactions to mathematics.

Question 12 as a measure of test-taking motivation. I debated on whether to remove or include questions 12, "Does counting the Grade-9 Assessment of Mathematics as part of your class mark motivate you to take the assessment more seriously?"as a validity criterion. I thought about removing it because question 12 is inherently ambiguous. For example, a student who would be highly motivated either with or without the class mark consequence would likely not agree with this statement. In contrast, a student who would be highly unmotivated regardless of the class mark consequence would probably also not agree with the statement. This ambiguity minimizes the usefulness of question 12 as a validity criterion as it is hard to know if the n is small for highly motivated and unmotivated examinees. I decided to include question 12, however, because moderate positive correlations were found between question 12 and the PCA component scores (math-values and interest) that supported the relationship between test-taking motivation and general achievement motivation for mathematics learning as stated in the literature (Eklof, 2006).

Motivation filtering. An optimal level of motivation filtering was not determined to have a more accurate reference when examining the effect of different proportions of high motivation in relation to the standard and modified model calibrations. Using optimal motivation filtering helps preserve a larger sample size because it minimizes the use of stringent filters, which may be unnecessary. The results of the current study, however, revealed that the modified model produced more accurate and consistent results than the standard model when estimating test item parameters and examinee abilities for different proportions of motivation filtering.

Examinee effort-moderated per item versus filtering.

The modified model used in this study does not measure examinee effort per item and it is different from the effort-moderated model used by Wise and DeMars (2006). Wise and DeMars' effort-moderated model is able to differentiate between examinees with high or low effort per item based on examinee response time. That is, low effort examinee ability estimates under the effort-moderated model can be salvaged by basing examinee ability estimation on a portion of the test event that reflected high effort. On the contrary, the modified model used in this study is a motivation filtering method in which those examinees exhibiting high motivation get a score and those examinees exhibiting low motivation do not get a score and therefore, the ability estimates for those examinees under the effect of low motivation are unavailable. The modified model does not measure examinee effort per item, but rather the motivation levels that examinees may bring to the test-taking situation. Motivation was a treated as a constant throughout the test because it was related to the subject domain, generally, and not to specific items.

Credible calibration results. Although the modified model shows no differences when comparing the model across different proportions of high motivation (30% and 50%) to a reference calibrated with 40 percent of high motivation, a minimum percent of examinees with high motivation is required to obtain credible results. For example if the percent of high motivation is set to be 40 percent, any examinee with a motivation value lower than 40 percent will not receive a valid score. Under the modified model, this means that there will not be sufficient information for the estimates of examinee ability to be credible.

Conclusion

This study had two purposes: a) to evaluate the effect of removing examinees with low motivation on the estimates of examinee abilities and test-item parameters calibrations using an item response theory model and b) to examine the significance of the relationship between students' motivation and their mathematics achievement and how this relationship is influenced by school level variables. The findings indicate that using a modified model that removes examinees with low motivation from the test data may provide more accurate estimates of examinee abilities and test item parameters than those obtained with a standard IRT model. For example, using the modified model to calibrate Grade-9 EQAO large-scale assessments of mathematics may reduce bias and provide more valid interpretation of test results for educational agencies, teachers and policy makers to better assess student mathematics abilities and academic performance and therefore, make better curriculum decision changes in our current educational system.

This study supports previous research work by Wise and DeMars (2006) and Wise and Kong (2005) in the use of hybrid models to account for the effect of student low motivation when calibrating large-scale assessment data to estimate examinee abilities and test item parameters. While the modeling approach of Wise and DeMars (2006) is promising because it accounts for the effect of low motivation and it should be considered to monitor the effort that examinees give toward the test when calibrating the data from a large-scale assessment, the technique requires the use of computer-based technology. This study, however, builds on the work of Swerdzewski, Harmes, and Finney (2011) by providing another avenue to include a motivation component in an item response model using a similar approach as Wise, Wise, and Bhola's (2006) but with measures of motivation obtained from self-report data related to achievement motivation as in mathematics learning as opposed to test-taking motivation in mathematics.

All methods used to address the first and second purpose in the study suggest that the modified model minimizes the effect of low motivation when estimating student abilities (for examinees with high motivation) and test item parameters for the EQAO large-scale assessment data from both academic and applied programs. Testing organizations who administer pencil and paper tests should consider using a modified item response model (hybrid model) that accounts for the effect of student motivation in combination with item response theory models to better explain the results from large-scale assessment data.

The modified model examined in this study is a motivation filtering method in which those exhibiting high motivation get a score and those exhibiting low motivation do not get a score. That is, examinees with low motivation get filtered from the data. The model is compatible with commonly used IRT software. The model can be implemented using BILOG software or any software that allows examinee cases to be not administered under low motivation. This modeling approach behaves as a low motivation filter. Under low motivation, the probability estimate is a fixed value for all examinees, which corresponds to some estimated ability value obtained from random guessing but it is not considered a valid score. Adding this constant value to the likelihood function, however, does not influence where the maximum or critical value occurs in the estimates of examinees' abilities and test item parameters because the derivative of a constant is zero.

There is a unique approach, however, with the methods used in the current study to conduct the bias, RMSE, and DIF analyses, which are different from the expectancy-value model proposed by Pintrich (1988;1989) and used by Wise and DeMars (2005; 2006). That is, the current study used a modified framework that draws directly from expectancy-value and self-efficacy theory by using subject domain questions (i.e., I like mathematics, the mathematics I learn now is very useful for everyday life, I am good in math, how much time do you usually spend in mathematics homework?...) as measures of general achievement motivation to address student intrinsic values, utility values, self-efficacy, and interest in mathematics learning. Printrich's expectancy-value model used by Wise and DeMars (2005; 2006), specifies that student effort is a function of a) how well students feel they will do on the test, b) how much effort it will take to complete the test, c) how important they perceive the test to be, and d) their affective emotional reactions to the individual test items. Although, I used a different method to measure motivation, the outcomes are similar to previous research work (Wise & DeMars, 2006). That is, the results seem to indicate that the mathematics values that students bring to the test-taking situation based on subject domain achievement motivation appear to be related to test-taking motivation and these mathematics values appear to serve as a proxy to test-taking motivation in large scale assessments.

Besides using a modified item response model as a low motivation filter and examining its usefulness via bias, RMSE, and DIF methods, the current study also examined the effect of students' motivation in their academic achievement. It was found that student math-values and interest were significant predictors of academic achievement. There are other factors affecting student academic achievement within and between schools that may or may not be connected to differences in disciplinary climate and academic press across schools. There is no evidence, however, to say that this is the case because these school level variables were not measured in the current study.

Recommendations

From the practical perspective, the current study may have implications for teachers, school boards, educational agencies, policy makers, students, and researchers because it offers an avenue to better estimate students' abilities (for examinees with high motivation) and test item parameters to provide more valid interpretations of test scores from large-scale assessments such as EQAO. These low motivation filtering techniques, however, can have either a trustworthy or an untrustworthy implication for policy makers. As a trustworthy implication, it may be possible for educational agencies and policy makers to use this method as an avenue to gather more accurate information when implementing curriculum changes, educational policies, and allocating funding to different educational organizations, to improve student academic performance in our current educational system. As an untrustworthy implication, policy makers may feel reluctant to use this method to make decisions in curriculum changes and educational policies because they may be concerned that the sample is not quite representative of the population when providing accountability for each student academic achievement. One recommendation that a researcher can make to policy makers, however, is to use this method as an avenue to flag the effect of low motivation when estimating examinee abilities and test item parameters so that test scores from the large-scale assessment are interpreted carefully. Another recommendation that a researcher can make to policy makers is to use this motivation filtering method to inform teachers of the effect of low motivation in low stake large-scale assessment data calibrations. This information may help teachers implement better methods at the classroom level to engage students in mathematics learning to further develop students' mathematics values and improve test performance.

From the theoretical perspective, the current research provides an extension to Swerdzewski, Harmes, and Finney (2011), and supports Wise and DeMars (2005; 2006) and Wolf and Smith (1995) in the use of expectancy-value theory, and self-efficacy theory in combination with item response theory models to account for a motivation component to better estimate students' ability and test item parameters when using self-report data. It also highlights the application of a statistical technique based on item response models, expectancy-value theory and self-efficacy theory to create a low motivation filter to minimize threats to the validity of the test results due to the effect of examinees with low motivation in test data calibrations. A combination of these theories (expectancy-value theory, self-efficacy theory, and item response theory) might be particularly effective for researchers to build on and measurement professionals to use.

Future Directions

This study illustrates the use a modified IRT model to better estimate students' abilities and test item parameters from EQAO large-scale assessment data. While the results are promising and highlight the effect of low motivation on student ability estimates and test item parameters when using an item response theory model, future research should focus on developing student self-report questionnaires that will encompass all the task-values as defined by expectancy-value theory, as well as, items that assess students' well developed interest in mathematics tasks as measures of student effort. This may provide stronger measures of student motivation to examine the effect of low motivation on the estimates of examinee abilities and test item parameters when comparing the standard and modified item response models.

Future research should aim at administering the EQAO student self-report items along with an established test-taking motivation scale such as Student Opinion Survey (SOS) immediately after the test administration and use confirmatory factor analysis to assess the degree to which the self-report items used as measures of achievement motivation in the current study form factors with the established SOS items. This approach will help provide stronger evidence of validity. That is, the degree to which achievement motivation relates to test-taking motivation in mathematics. It may also open another window of research opportunities to examine the effect of low motivation as it relates to achievement motivation and test-taking motivation in conjunction with item response models calibrations as an avenue to provide better estimates of examinee abilities and test item parameter from large-scale assessment data.

A simulation study may provide further evidence of the use of the modified model to minimize the effect of low motivation and more accurately estimate examinee abilities and test item parameters. Simulated data with different proportions of examinees with low motivation may be generated by using item parameters and proficiency estimates from the current study. The estimates of examinee abilities and test item parameters from the simulated study (calibrated with the modified and standard item response models) may be compared to known values obtained from the modified model calibrations in the current study.

Future research should also focus on examining the effect of low motivation on examinee abilities and test item parameters estimates from large-scale assessment data by comparing the effort-moderated model of Wise and DeMars to the modified model used in this thesis for the same group of examinees. This approach may provide further evidences of validity when using either pencil and paper or computer-based testing. It may also provide other avenues to better understand the relationship between general achievement motivation and test-taking motivation to differentiate low and high motivated examinees from the data distribution.

In addition, research should aim at the development of new digital technologies that are in compliance with teaching tools (pencil and paper) used in the classroom. For example, the development of a digital pen to measure response time and record student scores will allow the combination of these measures with a hybrid item response model, similar the effort-moderated model by Wise and DeMars (2006). This approach will permit students to be tested with the same instruments (pencil and paper) used in the classroom to minimize any threats to the validity of the measurements due to examinee possible unfamiliarity with the testing tool. Yet, it will facilitate the collection of supplemental data (i.e., response time and pen tracking) that will aid in filtering aberrant behaviours that result from student low motivation.

References

Ark, A., Emos, W., & Sijtsma, K. (2008). Detecting answer copying using alternate test forms and seat locations in small-scale examinations. *Journal of Educational Measurement*, *45*(2), 99-117.

Atkinson, J. W. (1964). *An introduction to motivation*. Princeton, NJ: Van Norstrand.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological review*, *84*, 1991-215.

Baker, F. (1992). *Item Response Theory: Parameter estimation technique.* New York: Marcel Dekker Inc.

Bartholomew, D. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society*, *42*(3), 293-321.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441-462.

Begin, M., & Caplan, G. (1994). *For the love of learning: Report of the Royal Commission on Learning*, Publications, Ontario.

Boaler, J. (1999). Participation, knowledge and beliefs: A community perspective on mathematics learning. *Educational Studies in Mathematics*, 40, 259-281.

Bolt, D., & Gield, J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement, 4*(43), 313-333.

Camilli, G., & Penfield, D. (1997). Variance estimation for differential test functioning based on Mantel and Haenszel Statistics. *Journal of Educational Measurement*, *34*, 123-139.

Camilli, G., & Congdon, P. (1999). Application of method of estimating DIF for polytomous test items. *Journal of Educational and Behavioural Statistics, 24*(4), 323-341.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cole, J., Bergin, D., & Whittaker, T. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*, 609-624.

Crundwell, R. M. (2005). Alternative strategies for large scale student sssessment in canada. *Canadian Journal of Education Administration Policy, 1*(41), 1-21.

De Ayala, R., Plake, B., & Impara, J. (2001). The impact of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement*, *38*(3), 212-234.

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 109-132.

Downing, S. (2003). Item response theory: Applications of modern test theory in medical education. *Medical Education, 37*, 739-745.

Drasgow, F., Levine, M., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86.

Dweck, C. S., & Elliot, E. S. (1983). Achievement motivation. In E. M. Hetherington (Ed.) & P. H. Mussen (Series Ed.), *Handbook of child psychology: Vol. 4. Social and personality development* (pp. 643-691). New York: Wiley.

Dweck, C. S., & Grant, H. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology, 3*(85), 541-553.

Eccles, J. (1993). Expectancies, values, and academic behaviour. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75-137). New York: Freeman.

Eccles, J., & Harold, R. (1996). Family involvement in children and adolescents' schooling. In A. Booth & J. F. Dunn (Eds.), *Family-school links: How do they affect educational outcomes* (pp. 3-33). Mahwah, NJ: Erlbaum.

Eccles, J., & Wigfield, A. (2002). Motivational beliefs, values and goals. *Annual Reviews, 53*, 109-132.

Education Quality and Accountability Office (2004b). *Ensuring quality assessments: Building on strengths: Refining the program*. Toronto: EQAO.

Education Quality and Accountability Office (2010). *EQAO's Technical Report for the 2009-2010 Assessments*. Retrieved from http://www.eqao.com/pdf_e/11/Ctr_2009-10Report_ne_0511_web.pdf

Education Quality and Accountability Office (2011). *EQAO Assessments of reading, writing and mathematics*. Retrieved from http://www.eqao.com/results/results.aspx?grade=36,9, 10&year=2010&submit=View+Results&Lang=E

Education Quality and Accountability Office (2012). *EQAO mandate.* Retrieved from http://www.eqao.com/AboutEQAO/Mandate.aspx?Lang=E

Eklof, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*(4), 1-15.

Emenogu, B. C., & Childs, R. A. (2005). Curriculum, translation, and differential functioning of measurement and geometry items. *Canadian Journal of Education, 28*(1,2), 128-146.

Ernest, P. (1989). The knowledge, beliefs and attitudes of the mathematics teacher: A model. *Journal of Education for Teaching*, *15*(1), 13-33.

Fast, L. A., & Lewis, J. L. (2010). Does math self-efficacy mediate the effect of the perceived classroom environment on standardized math test performance? J*ournal of Educational Psychology*, *3*(102), 729-740.

Ferrando, P. J., & Lorenzo-Seva, U. (2005). IRT-related factor analytic procedure for testing the equivalence of paper-and-pencil and internet-administered questionnaires. *Psychological Methods, 10*(2), 193-205.

Ford, K., MacCallum, R., & Tai, M. (1986). The application of exploratory factor analysis in applied Psychology: A critical review and analysis. *Personnel Psychology, 39*, 291-314.

Fraire, R., Tideman, T., & Watts, T. (1997). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics*, *6*(2), 152-165.

French, B., & Oakes, W. (2004). Reliability and validity evidence of the institutional integration scale. *Educational and Psychological Measurement, 64*(1), 88-98.

Freund, J. E. (1999). *Mathematical statistics*. New Jersey: Prentice-Hall.

Gierl, M., Henderson, D., Jodoin, M., & Klinger, D. (2001). Minimizing the influence of item parameter estimation error in test development: A comparison of three selection procedures. *Journal of Experimental Education*, *69,* 261-279.

Haladyna, T., & Downing, S. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hambleton, R., Swaminathan, H., & Roger, J. (1991). *Fundamentals of item response theory.* London, New Delhi: Sage Publications.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3-24.

Harlen, W., & Crick, R. D. (2003). Testing and motivation for learning. *Assessment in Education*, *10*, 169-207.

Harter, S. (1982). The perceived competence scale for children. *Child Development*, *53*(1), 87-97.

Harwell, M., Baker, F., & Zwarts, M. (1998). Item parameter estimation via marginal maximum likelihood and an EM algorithm: A didactic. *Journal of Educational Statistics*, *13*(3), 243-271.

Henson, R., & Roberts, K.(2006). Use of exploratory factor analysis in published research. *Educational and Psychological Measurement*, *66*(3), 393-416.

Herbet, M., Dunn, J., & Luthra, V. (2004). *The validation of psychometric procedures: Teachers' perceptions and opinions of the grade 9 assessement of mathematics.* Research report for the EQAO Assessment Review Project.

Holland, P., & Thayer, D. (1988). Differential item peformance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds). *Test validity* (pp. 129-145). Hillsdale, NJ:Erlbaum.

Kane, M. (2006). Validation. In R. L. Brennan (Eds), *Educational measurements* (4th ed., pp. 17-64). New York: American Council of Education.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298.

Kim, C., Cohen, A., Alagoz, C., & Kim, S. (2007). Dif detection and effect size measures for polytomous scored items. *Journal of Educational Measurement, 2*(44), 93-116.

Klinger, D., Deluca, C., & Miller, T. (2008). The evolving culture of large-scale assessment in Canadian education. *Canadian Journal of Education Administration and Policy, 76*, 1-34.

Kozlow, M. (2007). *Model selection for analysis of EQAO assessment data* (Bulletin No. 1). Toronto, Ontario: EQAO.

Kolen, M., & Brennan, R. (2004). *Test Equating, scaling, and linking: Methods and practices*. New York: Springer.

Kong, X., Wise, S., & Bhola, D. (2007). Setting the response time parameter to differentiate solution behaviour from rapid-guessing behaviour. *Educational and Psychological Measurement*, *67*(4), 606-619.

Linn, R & Baker, E. (1996). *Assessing the validity of the National Assessment of Educational Progress: NAEP technical review panel white paper.* Washington, DC: National Center for Research on Evaluation, Standards, and Student Testing/National Center for Education Statistics.

Linn, R. L. (1989). *Educational measurement* (3 ed.). New York: National Council on Measurement on Education and American Council on Education.

Linn, R., Levine, M., Hastings, C., & Wardrop, J. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*, 159-173.

Linn, R. & Baker, E. (1996). *Assessing the validity of the National Assessment of Educational Progress: NAEP technical review panel white paper.* Washington, DC: National Center for Research on Evaluation, Standards, and Student Testing/National Center for Education Statistics.

Lord, F. (1990). *Applications of item response theory to practical testing problems.* Hillsdale, New Jersey: Lawrence.

Ma, X., & Klinger, D. (2000). Hierarchical linear modelling of student and school effect on academic achievement. *Canadian Journal of Education, 25*(1), 41-55.

MacDonald, R. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Maehr, M.L., & Meyer, H. (1997). Understanding motivation and schooling. Where we've been, where we are, and where we need to go. *Educational Psychology Review*, *9*, 371-409.

Madaus, G., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson, *Handbook of research curriculum* (pp. 119-154). New York: Macmillan.

Marsh, W., Koller, O., Trautwein, U., Ludtke, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effect models of casual ordering. *Child Development,76*, 376-416.

McMillan, J., Simonetta, L., & Singh, J. (1994). Student opinion survey: Development of measures of student motivation. *Educational and Psychological Measurement*, *54*, 498-505.

Mellenberg, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, *7*, 105-108.

McGehee, J. J., & Griffith, L. K. (2001). Large-scale assessments combined with curriculum alignment: Agents of change. *Theory Into Practice, 40*(2), 137-144.

Meijer, R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8.

Meijer, R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*, 261-272.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135.

Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, *85*, 424–436.

Miller, T., Chahine, S., & Childs, R. (2010). Detecting differential item functioning and differential step functioning due to differences that should matter. *Practical Assessment Research and Evaluation*, *15*(10), 1-13.

Mislevy, R. J., & Bock, R. D. (1990). PC-BIOLOG-Item analysis and test scoring with binary logistic models [computer software]. Mooresville, IN: Scientific Software.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council of Education.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16,* 159-176.

Muraki, E. (1997). A generalized partial credit model. In W.J. van der Liden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York: Springer.

Nagy, P. (2000). The three roles of assessment:gatekeeping, accountabilityand instructional diagnosis. *Canadian Journal of Education, 25*(2), 262-279.

NCME. (2009, January). *National Council on Measurement in Education.* Retrieved from http://www.ncme.org.

O'Neil, H. F., Abedi, J., Miyoshi, J. & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*(3), 185-208.

O'Neil, H. F., Jr., Sugrue, B., Abedi, J., Baker, E. L., & Golan, S. (1997). *Final report of experimental studies on motivation and NAEP test performance* (CSE Tech. Report No. 427). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Pajares, F. (1996a). Role of self-efficacy beliefs in mathematical problem-solving of gifted students. *Contemporary Educational Psychology*, *21*, 325-344.

Pajares, F. (1996b). Self-efficacy beliefs in achievement settings. *Review of Educational Research*, *66*, 543-578.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology, 24,* 124-136.

Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology, 26,* 426-443.

Palomba, C. A., & Banta, T. W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education.* San Francisco: Jossey-Bass.

Pankaja, N., & Swaminathan, H. (1994). Performance of Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*(4), 315-328.

Penfield, R. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education, 20*(3), 334-355.

Penfield, R., & Algina, J. (2006). A generalized DIF effect variance estimator for measuring unsigned differential test functioning in mixed format tests. *Journal of Educational Measurement, 43*(4), 295-312.

Pintrich, P. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, *16*(4), 385-407.

Pintrich, P. R. (1988). A process-oriented view of student motivation and cognition. In J. S. Stark & R. Mets (Eds), *Improving teaching and learning through research* (pp. 55-70). San Francisco: Jossey-Bass.

Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In C. Ames & M. Maehr (Eds), *Advances in achievement and motivation* (Vol. 6, pp. 117-160). Greenwich. CT: JAI Press.

Pintrich, P. R., & Schunk, D. H. (1996). *Motivation in education: Theory, research, and applications*. Englewood Cliffs, NJ: Merrill–Prentice Hall.

Pintrich, P. R., & Schunk, D. H. (2004). *Motivation in education: Theory, research, and applications*. (2nd ed.). Upper Saddle, NJ: Merrill–Prentice Hall.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*, 801-813.

Putwain, D. (2007). Test anxiety in UK school children: Prevalence and demographic patterns, *The British Psychological Society*, *77*, 579-593.

Putwain, D. (2008). Do examinations stakes moderate the test anxiety-examination performance relationship? *Educational Psychology, 28*(2), 109-118.

Raju, N. S. (1998). The area between the two item characteristic curves. *Psychometrika*, *53*, 495-502.

Renninger, K. A., & Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 173–195). San Diego: Academic Press.

Reynolds, W. M. (1984). Depression in children and adolescents: Phenomenology, evaluation and treatment. *School Psychology Review*, *13*, 171–182.

Ryan, K., Ryan, A., Arbuthnot, K., & Samuels, M. (2007). Students' motivation for standardized math exams, *Educational Researcher*, *36*(1), 5-13.

Roderick, M., & Engel, M. (2001). The grasshopper and the ant: Motivational responses of low-achieving students to high-stakes testing. *Educational Evaluation and Policy Analysis*, *23*, 197-227.

Rogers, H.J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2)*,* 105-116.

Roussos, L., Schnipke, D., & Pashley, P. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 293-322.

Salthouse, T., & Mitchell, D. (1990). Effects of age and naturally occurring experience on spatial visualization performance. *Developmental Psychology, 5*(26), 845-854.

Schmidt, F., Le, H., & Llies, R. (2003). Beyond alpha: An empirical examination of the effect of different sources on measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*, 206-224.

Schmitt, N., Chan, D., Sacco, J., McFarland, L., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*, 41-53.

Schunk, D. (1995). Self-efficacy and education and instruction. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 281-303). New York: Plenum.

Sheng, Y. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 67*(6), 899-919.

Singh, K., Grandville, M., & Dika, S. (2002). Mathematics and science achievement: Effect of motivation, interest and academic engagement. *The Journal of Educational Researcher*, 95(6), 323-332.

Sotaridona, L., & Meijer, R. (2003). Two new statistics to detect answer copying. *Journal of Educational Measurement*, *40*(1), 53-69.

Sotaridona, L., Linden, W., & Meijer, R. (2006). Detecting answer copying using kappa statistics. *Applied Psychological Measurement*, *30*(5), 412-431.

Stajkovic, A., Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis, *Psychological Bulletin*, *124*, 240-261.

Standards for Educational and Psychological Testing. (1999). *American Educational Research Association, American Psychological Association, & National Council on Measurement in Education*. Washington, DC: American Psychological Association.

Sundre, D. L., & Moore, D.L. (2002). The student opinion scale. A measure to examine motivation. *Assessment Update*, 14, 8-9.

Sundre, D., & Wise, S. (2003, April). *Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests.* Paper presented at the National Council on Measurement in Education, Chicago, II.

Sundre, D., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology, 29*(1), 6-26.

Stone, S. (2006). Correlates of change in student reported parent involvement in schooling: A new look at the National Education Study of 1988. *Journal of Orthopsychiatry*, 76, 518-530.

Swaminathan, H., & Rogers, H. (1991). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Swerdzewski, P., Harmes, C., & Finney, S. (2011). Two approaches for indentifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*, 162-188.

The Ontario Curriculum Grades 9 to 12 (2000). *Program planning and assessment*. Retrieved from http://www.edu.gov.on.ca/eng/curriculum/secondary/progplan912curr.pdf.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study group differences in trace lines. In H. Wainer & H. I. Braun (Eds), *Test validity* (pp.147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds), *Differential item functioning* (pp.67-113). Hillsdale, NJ: Erlbaum.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. *Applied Psychological Measurement, 18*, 15-25.

van Barneveld, C. (2007). The effect of examinee motivation on test construction within an IRT framework. *Applied Psychological Measurement, 31*(1), 31-46.

Volante, L. (2006). An alternative vision for large-scale assessment in Canada. *Journal of Teaching and Learning,4*(1), 1-14.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, *30*, 1-21.

Walter, D., & Prescott, G. (1961). *Essential of measurements for teachers* (1st ed). New York: Harcourt, Brace & World, Inc.

Weiss, D. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473-492.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review, 30*, 1-35.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68-81.

Wise, S. (2006). An investigation of the differential effort received by items on a low-stakes computer based test. *Applied Measurement in Education*, *19*(2), 95-114.

Wise, V., Wise, S., & Bhola, D. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65-83.

Wise, S., & DeMars, C. (2005). Low examinee effort in low stake-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 11-17.

Wise, S., & DeMars, C. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurements, 43*(1), 19-38.

Wise, S., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(3), 163-183.

Wise, L. (1996a). *A persistent model of motivation and test performance*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Wolf, L., & Smith, J. (1995). The consequence of consequence: Motivation, anxiety and test performance. *Applied Measurement in Education, 8*(3), 227-242.

Wolf, L., Smith, J., & Birnbaum, M. (1995). Consequences of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*(4), 341-351.

Wolfe, R., Childs, R., & Elgie, S. (2004). *Final Report of the External Evaluation of EQAO'S Assessment Process.* Toronto: EQAO. Retrieved from http://www.ontla.on.ca/library/repository/mon/8000/244807.pdf.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: Mesa Press.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class models* (ETS Research Report RR-89-41). Princeton, NJ: Educational Testing Service (ERIC Document Reproduction Service No. ED 310161.

Zerpa, C., Hachey, K., van Barneveld., & Simon, M. (2011). Modeling student motivation and students' ability estimates from a large-scale assessment of mathematics. *Sage Open, 1*(2), 1-9.

Zhang, B., & Walker, C. (2008). Impact of missing data on person-model fit and person trait estimation. *Applied Psychological Measurement*, *32*(6), 466-479.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multigroup IRT analysis and test maintenance for binary items.* Chicago: Scientific Software.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now and where it is going. *Language Assessment Quartely*, *4*(2), 223-233.

Appendices

Note: Examples of SPSS and BILOG_MG scripts are included in the appendices. These scripts were used to merge and calibrate the data to compute examinee abilities and test item parameters. The scripts were also used to scale the data to be able to compute bias, RMSE and DIF results between the modified and standard models

Appendix A1

Example of SPSS script used to merge item responses and student questionnaire.

```
********Reading item response data from EQAO 2010*******************************

GET DATA  /TYPE = TXT
 /FILE = 'C:\EQAO2010\G9 2010 Data Merge\selfreportanditemresponses'+
 ' Merge\G9_2010_ItemResponses.csv'
 /DELCASE = LINE
 /DELIMITERS = ","
 /QUALIFIER = '"'
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 2
 /IMPORTCASE = ALL
 /VARIABLES =
 RecID A8
 StudentID A9
 ClassID A7
 SchoolID A6
 Language F1.0
 Program F1.0
 Assessment F1.0
 MC01 F2.0
 MC02 F2.0
 MC03 F2.0
 MC04 F2.0
 MC05 F2.0
 MC06 F2.0
 MC07 F2.0
 MC08 F2.0
 MC09 F2.0
 MC10 F2.0
 MC11 F2.0
 MC12 F2.0
 MC13 F2.0
 MC14 F2.0
 MC15 F2.0
 MC16 F2.0
 MC17 F2.0
 MC18 F2.0
 MC19 F2.0
 MC20 F2.0
 MC21 F2.0
 MC22 F2.0
 MC23 F2.0
 MC24 F2.0
 OR01 F1.0
 OR02 F1.0
 OR03 F2.0
 OR04 F2.0
 OR05 F1.0
 OR06 F1.0
 OR07 F2.0
 .
CACHE.
EXECUTE.
CROSSTABS
  /TABLES=Language BY Program BY Assessment
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.
*Drop cases with invalid StudentIDs.
SUMMARIZE
  /TABLES=StudentID
  /FORMAT=NOLIST
  /TITLE='Case Summaries'
  /CELLS=COUNT MIN MAX .
FILTER OFF.
USE ALL.
SELECT IF(StudentID ne "000000000").
EXECUTE .

SUMMARIZE
  /TABLES=StudentID
  /FORMAT=NOLIST
  /TITLE='Case Summaries'
  /CELLS=COUNT MIN MAX .

CROSSTABS
  /TABLES=Language BY Program BY Assessment
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.

*Sort cases.

SORT CASES BY
  RecID (A).

*Save Student Item Response File.
```

```
SAVE OUTFILE='C:\EQAO2010\G9 2010 Data Merge\selfreportanditemresponses'+
 ' Merge\G9_2010_ItemResponsesScript.sav'
 /COMPRESSED.
.
**********Reading student self-report data

GET DATA  /TYPE = TXT
 /FILE = 'C:\EQAO2010\G9 2010 Data Merge\selfreportanditemresponses'+
 ' Merge\G9_2010_ISD_SQd.csv'
 /DELCASE = LINE
 /DELIMITERS = ","
 /QUALIFIER = ""
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 2
 /IMPORTCASE = ALL
 /VARIABLES =
 RecID A8
 StudentID A9
 ClassID A7
 SchoolID A6
 Language F1.0
 Program F1.0
 MathClassWhen F1.0
 OverallOutcomeLevel A1
 OverallRawLevel_Dot F5.1
 OverallLeftBarEnd F6.2
 OverallRightBarEnd F6.2
 NSAOutcome F2.0
 LROutcome F2.0
 AGOutcome F2.0
 MGOutcome F2.0
 KUOutcome F2.0
 APOutcome F2.0
 PSOutcome F2.0
 AssessmentTotalNonBlanks F2.0
 TotalNumOfItems F2.0
 Prior_G6_MOverallLevel A2
 Prior_G3_MOverallLevel A2
 Gender F1.0
 StudentType F1.0
 ESLELD_ALFPDF F1.0
 SIF_IEP F1.0
 SIF_IPRCBehaviour F1.0
 SIF_IPRCAutism F1.0
 SIF_IPRCDeaf F1.0
```

```
SIF_IPRCBlind F1.0
SIF_IPRCGifted F1.0
SIF_IPRCIntellectual F1.0
SIF_IPRCDevelopmental F1.0
SIF_IPRCMultiple F1.0
SIF_IPRCPhysical F1.0
SIF_IPRCSpeech F1.0
SIF_IPRCLanguage F1.0
SIF_IPRCLearning F1.0
SIF_SpecialProvisionsSetting F1.0
SIF_SpecialProvisionsTime F1.0
SIF_SpecialProvisionsBreaks F1.0
SIF_SpecialProvisionsInstructions F1.0
SIF_AccommodationSetting F1.0
SIF_AccommodationSeating F1.0
SIF_AccommodationDevicesSetting F1.0
SIF_AccommodationPrompts F1.0
SIF_AccommodationTime F1.0
SIF_AccommodationBreaks F1.0
SIF_AccommodationSigned F1.0
SIF_AccommodationBraille F1.0
SIF_AccommodationLargePrint F1.0
SIF_AccommodationColoured F1.0
SIF_AccommodationLPColoured F1.0
SIF_AccommodationAudio F1.0
SIF_AccommodationDevicesPresentation F1.0
SIF_AccommodationInstructions F1.0
SIF_AccommodationRecording F1.0
SIF_AccommodationScribing F1.0
SIF_AccommodationComputer F1.0
SIF_AccommodationDevicesResponse F1.0
SIF_SpPermTempInjury F1.0
SIF_SpPermNewStudent F1.0
SQ1aLikeMath F2.0
SQ1bGoodMath F2.0
SQ1cUnderstandMath F2.0
SQ1dUsefulMath F2.0
SQ1eKeepMath F2.0
SQ1fBoringMath F2.0
SQ1gEasyMath F2.0
SQ2aNumberSense F2.0
SQ2bAlgebra F2.0
SQ2cRelations F2.0
SQ2AnalyticGeometry F2.0
SQ2Measurement F2.0
SQ2Geometry F2.0
```

```
 SQ3aComputerAtHome F2.0
 SQ3bCalculatorScientificAtHome F2.0
 SQ3cCalculatorGraphingAtHome F2.0
 SQ4aMathHomework F2.0
 SQ4bMathHomeworkComplete F2.0
 SQ5LanguageAtHome F2.0
 SQ6Age F2.0
 SQ7AbsentMath F2.0
 SQ8LateMath F2.0
 SQ9HowManySchools F2.0
 SQ10ClassMark F2.0
 SQ11aClassMarkTold F2.0
 SQ11bClassMarkHowMuch F2.0
 SQ12ClassMarkMotivate F2.0
 DateCreatedUpdated A10
 .
CACHE.
EXECUTE.
DATASET NAME DataSet2 WINDOW=FRONT.

*Drop cases with invalid StudentIDs.

SUMMARIZE
  /TABLES=StudentID
  /FORMAT=NOLIST
  /TITLE='Case Summaries'
  /CELLS=COUNT MIN MAX .

FILTER OFF.
USE ALL.
SELECT IF(StudentID ne "000000000").
EXECUTE .

SUMMARIZE
  /TABLES=StudentID
  /FORMAT=NOLIST
  /TITLE='Case Summaries'
  /CELLS=COUNT MIN MAX .

*Sort cases.

SORT CASES BY
  RecID (A).

*Save Student Questionnaire Response File.
```

```
SAVE OUTFILE='C:\EQAO2010\G9 2010 Data Merge\selfreportanditemresponses'+
 ' Merge\G9_2010_StudentQuestionnaire.sav'
 /COMPRESSED.

**********Merging the self-report data with item response data**************

MATCH FILES
 /TABLE='C:\EQAO2010\G9 2010 Data Merge\selfreportanditemresponses'+
 ' Merge\G9_2010_ItemResponsesScript.sav'
 /FILE='C:\EQAO2010\G9 2010 Data Merge\selfreportanditemresponses'+
 ' Merge\G9_2010_StudentQuestionnaire.sav'
 /RENAME (ClassID Language Program SchoolID StudentID = ClassID_SQ Language_SQ
         Program_SQ SchoolID_SQ StudentID_SQ)
 /BY RecID.
EXECUTE.

FILTER OFF.
USE ALL.
SELECT IF(Assessment gt 0).
EXECUTE .


SAVE OUTFILE='C:\EQAO2010\G9 2010 Data Merge\selfreportanditemresponses'+
 ' Merge\G9_2010_MergestudentData.sav'
 /COMPRESSED.


**********Delete cases that are not applicable (-99), excluded  (-97), and dropped from
         assessment (-96)

FILTER OFF.
USE ALL.
SELECT IF NOT (MC01 = -99 OR MC01=-97 OR MC01=-96 ).
SELECT IF NOT (MC02 = -99 OR MC02=-97 OR MC02=-96 ).
SELECT IF NOT (MC03 = -99 OR MC02=-97 OR MC03=-96 ).
SELECT IF NOT (MC04 = -99 OR MC04=-97 OR MC04=-96 ).
SELECT IF NOT (MC05 = -99 OR MC05=-97 OR MC05=-96 ).

SELECT IF NOT (MC06 = -99 OR MC06=-97 OR MC06=-96 ).
SELECT IF NOT (MC07 = -99 OR MC07=-97 OR MC07=-96 ).
SELECT IF NOT (MC08 = -99 OR MC08=-97 OR MC08=-96 ).
SELECT IF NOT (MC09 = -99 OR MC09=-97 OR MC09=-96 ).
SELECT IF NOT (MC10 = -99 OR MC10=-97 OR MC10=-96 ).

SELECT IF NOT (MC11 = -99 OR MC11=-97 OR MC11=-96 ).
SELECT IF NOT (MC12 = -99 OR MC12=-97 OR MC12=-96 ).
SELECT IF NOT (MC13 = -99 OR MC13=-97 OR MC13=-96 ).
SELECT IF NOT (MC14 = -99 OR MC14=-97 OR MC14=-96 ).
SELECT IF NOT (MC15 = -99 OR MC15=-97 OR MC15=-96 ).

SELECT IF NOT (MC16 = -99 OR MC16=-97 OR MC16=-96 ).
SELECT IF NOT (MC17 = -99 OR MC17=-97 OR MC17=-96 ).
SELECT IF NOT (MC18 = -99 OR MC18=-97 OR MC18=-96 ).
SELECT IF NOT (MC19 = -99 OR MC19=-97 OR MC19=-96 ).
SELECT IF NOT (MC20 = -99 OR MC20=-97 OR MC20=-96 ).

SELECT IF NOT (MC21 = -99 OR MC21=-97 OR MC21=-96 ).
SELECT IF NOT (MC22 = -99 OR MC22=-97 OR MC22=-96 ).
SELECT IF NOT (MC23 = -99 OR MC23=-97 OR MC23=-96 ).
SELECT IF NOT (MC24 = -99 OR MC24=-97 OR MC24=-96 ).

EXECUTE .

****Recoding the data for missing values for item responses

IF (MC01=-9 OR MC01=-6) MC01 = 9.
IF (MC02=-9 OR MC02=-6) MC02 = 9.
IF (MC03=-9 OR MC03=-6) MC03 = 9.
IF (MC04=-9 OR MC04=-6) MC04 = 9.
IF (MC05=-9 OR MC05=-6) MC05 = 9.
IF (MC06=-9 OR MC06=-6) MC06 = 9.
IF (MC07=-9 OR MC07=-6) MC07 = 9.
IF (MC08=-9 OR MC08=-6) MC08 = 9.
IF (MC09=-9 OR MC09=-6) MC09 = 9.
IF (MC10=-9 OR MC10=-6) MC10 = 9.
IF (MC11=-9 OR MC11=-6) MC11 = 9.
IF (MC12=-9 OR MC12=-6) MC12 = 9.
IF (MC13=-9 OR MC13=-6) MC13 = 9.
IF (MC14=-9 OR MC14=-6) MC14 = 9.
IF (MC15=-9 OR MC15=-6) MC15 = 9.
IF (MC16=-9 OR MC16=-6) MC16 = 9.
IF (MC17=-9 OR MC17=-6) MC17 = 9.
IF (MC18=-9 OR MC18=-6) MC18 = 9.
IF (MC19=-9 OR MC19=-6) MC19 = 9.
IF (MC20=-9 OR MC20=-6) MC20 = 9.
IF (MC21=-9 OR MC21=-6) MC21 = 9.
IF (MC22=-9 OR MC22=-6) MC22 = 9.
IF (MC23=-9 OR MC23=-6) MC23 = 9.
IF (MC24=-9 OR MC24=-6) MC24 = 9.


******Recoding the data for missing values for self-report
```

```
IF (SQ1aLikeMath = 0 OR SQ1aLikeMath = -1 ) SQ1aLikeMath = 99.
IF (SQ1bGoodMath=0 OR SQ1bGoodMath=-1 ) SQ1bGoodMath=99.
IF (SQ1cUnderstandMath=0 OR SQ1cUnderstandMath=-1 ) SQ1cUnderstandMath=99.
IF (SQ1dUsefulMath=0 OR SQ1dUsefulMath=-1 ) SQ1dUsefulMath=99.
IF (SQ1eKeepMath=0 OR SQ1eKeepMath=-1) SQ1eKeepMath=99.
IF (SQ1fBoringMath=0 OR SQ1fBoringMath=-1) SQ1fBoringMath=99.
IF (SQ1gEasyMath=0 OR SQ1gEasyMath=-1 ) SQ1gEasyMath=99.
IF (SQ2aNumberSense=0 OR SQ2aNumberSense=-1) SQ2aNumberSense=99.
IF (SQ2bAlgebra=0 OR SQ2bAlgebra=-1) SQ2bAlgebra=99.
IF (SQ2cRelations=0 OR SQ2cRelations=-1 ) SQ2cRelations=99.
IF (SQ2AnalyticGeometry=0 OR SQ2AnalyticGeometry=-1) SQ2AnalyticGeometry=99.
IF (SQ2Measurement=0 OR SQ2Measurement=-1 ) SQ2Measurement=99.
IF (SQ2Geometry=0 OR SQ2Geometry=-1 ) SQ2Geometry=99.
IF (SQ3aComputerAtHome=0 OR SQ3aComputerAtHome=-1) SQ3aComputerAtHome=99.
IF (SQ4aMathHomework=0 OR SQ4aMathHomework=-1) SQ4aMathHomework=99.
IF (SQ4bMathHomeworkcomplete=0 OR SQ4bMathHomeworkcomplete=-1 )
        SQ4bMathHomeworkcomplete=99.
IF (SQ7AbsentMath=0 OR SQ7AbsentMath=-1) SQ7AbsentMath=99.
IF (SQ8LateMath=0 OR SQ8LateMath=-1) SQ8LateMath=99.

RECODE SQ1aLikeMath   (SYSMIS=99).
RECODE SQ1bGoodMath  (SYSMIS=99).
RECODE SQ1cUnderstandMath (SYSMIS=99).
RECODE SQ1dUsefulMath (SYSMIS=99).
RECODE SQ1eKeepMath  (SYSMIS=99).
RECODE SQ1fBoringMath (SYSMIS=99).
RECODE SQ1gEasyMath  (SYSMIS=99).
RECODE SQ2aNumberSense (SYSMIS=99).
RECODE SQ2bAlgebra (SYSMIS=99).
RECODE SQ2cRelations (SYSMIS=99).
RECODE SQ2AnalyticGeometry (SYSMIS=99).
RECODE SQ2Measurement (SYSMIS=99).
RECODE SQ2Geometry (SYSMIS=99).
RECODE SQ3aComputerAtHome (SYSMIS=99).
RECODE SQ4aMathHomework (SYSMIS=99).
RECODE SQ4bMathHomeworkcomplete (SYSMIS=99).
RECODE SQ7AbsentMath (SYSMIS=99).
RECODE SQ8LateMath (SYSMIS=99).
RECODE OverallOutcomeLevel (SYSMIS=99).
RECODE Math_Value (SYSMIS=99).
RECODE Interest(SYSMIS=99).
EXECUTE.
```

```
**********Delete missing data from self-report data****************
FILTER OFF.
USE ALL.
SELECT IF NOT (SQ1aLikeMath = 99 ).
SELECT IF NOT(SQ1bGoodMath = 99 ).
SELECT IF NOT (SQ1cUnderstandMath = 99 ).
SELECT IF NOT (SQ1dUsefulMath = 99 ).
SELECT IF NOT (SQ1eKeepMath = 99 ).
SELECT IF NOT (SQ1fBoringMath = 99 ).
SELECT IF NOT (SQ1gEasyMath = 99 ).
SELECT IF NOT (SQ4aMathHomework  = 99 ).
SELECT IF NOT (SQ4bMathHomeworkcomplete = 99 ).
SELECT IF NOT (SQ7AbsentMath = 99 ).
SELECT IF NOT (SQ8LateMath = 99 ).

EXECUTE .
```

Appendix A2

Example of SPSS script used to rescale student self-report data.

```
*******This script will rescale the self-report data********************
IF (SQ1aLikeMath=1) SQ1aLikeMath1 = 5 .
IF (SQ1aLikeMath=2) SQ1aLikeMath1 = 4 .
IF (SQ1aLikeMath=3) SQ1aLikeMath1 = 3 .
IF (SQ1aLikeMath=4) SQ1aLikeMath1 = 2 .
IF (SQ1aLikeMath=5) SQ1aLikeMath1 = 1 .

IF (SQ1bGoodMath=1) SQ1bGoodMath1 = 5 .
IF (SQ1bGoodMath=2) SQ1bGoodMath1 = 4 .
IF (SQ1bGoodMath=3) SQ1bGoodMath1 = 3 .
IF (SQ1bGoodMath=4) SQ1bGoodMath1 = 2 .
IF (SQ1bGoodMath=5) SQ1bGoodMath1 = 1 .

IF (SQ1cUnderstandMath=1) SQ1cUnderstandMath1 = 5 .
IF (SQ1cUnderstandMath=2) SQ1cUnderstandMath1 = 4 .
IF (SQ1cUnderstandMath=3) SQ1cUnderstandMath1 = 3 .
IF (SQ1cUnderstandMath=4) SQ1cUnderstandMath1 = 2 .
IF (SQ1cUnderstandMath=5) SQ1cUnderstandMath1 = 1 .

IF (SQ1dUsefulMath=1) SQ1dUsefulMath1 = 5 .
IF (SQ1dUsefulMath=2) SQ1dUsefulMath1 = 4 .
IF (SQ1dUsefulMath=3) SQ1dUsefulMath1 = 3 .
IF (SQ1dUsefulMath=4) SQ1dUsefulMath1 = 2 .
IF (SQ1dUsefulMath=5) SQ1dUsefulMath1 = 1 .
IF (SQ1eKeepMath=1) SQ1eKeepMath1 = 5 .
IF (SQ1eKeepMath=2) SQ1eKeepMath1 = 4 .
IF (SQ1eKeepMath=3) SQ1eKeepMath1 = 3 .
IF (SQ1eKeepMath=4) SQ1eKeepMath1 = 2 .
IF (SQ1eKeepMath=5) SQ1eKeepMath1 = 1 .

IF (SQ1fBoringMath=1) SQ1fBoringMath1 = 5 .
IF (SQ1fBoringMath=2) SQ1fBoringMath1 = 4 .
IF (SQ1fBoringMath=3) SQ1fBoringMath1 = 3 .
IF (SQ1fBoringMath=4) SQ1fBoringMath1 = 4 .
IF (SQ1fBoringMath=5) SQ1fBoringMath1 = 1 .

IF (SQ1gEasyMath=1) SQ1gEasyMath1 = 5 .
IF (SQ1gEasyMath=2) SQ1gEasyMath1 = 4 .
IF (SQ1gEasyMath=3) SQ1gEasyMath1 = 3 .
IF (SQ1gEasyMath=4) SQ1gEasyMath1 = 2 .
IF (SQ1gEasyMath=5) SQ1gEasyMath1 = 1 .

IF (SQ4aMathHomework=1) SQ4aMathHomework1 = 4 .
IF (SQ4aMathHomework=2) SQ4aMathHomework1 = 3 .
IF (SQ4aMathHomework=3) SQ4aMathHomework1 = 2 .
IF (SQ4aMathHomework=4) SQ4aMathHomework1 = 1 .

IF (SQ7AbsentMath=1) SQ7AbsentMath1 = 4 .
IF (SQ7AbsentMath=2) SQ7AbsentMath1 = 3 .
IF (SQ7AbsentMath=3) SQ7AbsentMath1 = 2 .
IF (SQ7AbsentMath=4) SQ7AbsentMath1 = 1 .

IF (SQ8LateMath=1) SQ8LateMath1 = 4 .
IF (SQ8LateMath=2) SQ8LateMath1 = 3 .
IF (SQ8LateMath=3) SQ8LateMath1 = 2 .
IF (SQ8LateMath=4) SQ8LateMath1 = 1 .
EXECUTE .
```

Appendix A3

Example of SPSS script used to create BILOG-MG text file.

```
IF (MC01=-1  | MC01=-2 | MC01=-3  |  MC01=-4 ) MC01 = 0 .
IF (MC01=1  | MC01=2 | MC01=3  |  MC01=4 ) MC01 = 1.
IF (MC01=-99  | MC01=-97 | MC01=-96  |  MC01=-9 | MC01=-6) MC01 = 9.

IF (MC02=-1  | MC02=-2 | MC02=-3  |  MC02=-4 ) MC02 = 0 .
IF (MC02=1  | MC02=2 | MC02=3  |  MC02=4 ) MC02 = 1.
IF (MC02=-99  | MC02=-97 | MC02=-96  |  MC02=-9 | MC02=-6) MC02 = 9.

IF (MC03=-1  | MC03=-2 | MC03=-3  |  MC03=-4 ) MC03 = 0 .
IF (MC03=1  | MC03=2 | MC03=3  |  MC03=4 ) MC03 = 1.
IF (MC03=-99  | MC03=-97 | MC03=-96  |  MC03=-9 | MC03=-6) MC03 = 9.

IF (MC04=-1  | MC04=-2 | MC04=-3  |  MC04=-4 ) MC04 = 0 .
IF (MC04=1  | MC04=2 | MC04=3  |  MC04=4 ) MC04 = 1.
IF (MC04=-99  | MC04=-97 | MC04=-96  |  MC04=-9 | MC04=-6) MC04 = 9.

IF (MC05=-1  | MC05=-2 | MC05=-3  |  MC05=-4 ) MC05 = 0 .
IF (MC05=1  | MC05=2 | MC05=3  |  MC05=4 ) MC05 = 1.
IF (MC05=-99  | MC05=-97 | MC05=-96  |  MC05=-9 | MC05=-6) MC05 = 9.

IF (MC06=-1  | MC06=-2 | MC06=-3  |  MC06=-4 ) MC06 = 0 .
IF (MC06=1  | MC06=2 | MC06=3  |  MC06=4 ) MC06 = 1.
IF (MC06=-99  | MC06=-97 | MC06=-96  |  MC06=-9 | MC06=-6) MC06 = 9.

IF (MC07=-1  | MC07=-2 | MC07=-3  |  MC07=-4 ) MC07 = 0 .
IF (MC07=1  | MC07=2 | MC07=3  |  MC07=4 ) MC07 = 1.
IF (MC07=-99  | MC07=-97 | MC07=-96  |  MC07=-9 | MC07=-6) MC07 = 9.

IF (MC08=-1  | MC08=-2 | MC08=-3  |  MC08=-4 ) MC08 = 0 .
IF (MC08=1  | MC08=2 | MC08=3  |  MC08=4 ) MC08 = 1.
IF (MC08=-99  | MC08=-97 | MC08=-96  |  MC08=-9 | MC08=-6) MC08 = 9.

IF (MC09=-1  | MC09=-2 | MC09=-3  |  MC09=-4 ) MC09 = 0 .
IF (MC09=1  | MC09=2 | MC09=3  |  MC09=4 ) MC09 = 1.
IF (MC09=-99  | MC09=-97 | MC09=-96  |  MC09=-9 | MC09=-6) MC09 = 9.

IF (MC10=-1  | MC10=-2 | MC10=-3  |  MC10=-4 ) MC10 = 0 .
IF (MC10=1  | MC10=2 | MC10=3  |  MC10=4 ) MC10 = 1.
IF (MC10=-99  | MC10=-97 | MC10=-96  |  MC10=-9 | MC01=-6) MC10 = 9.

IF (MC11=-1  | MC11=-2 | MC11=-3  |  MC11=-4 ) MC11 = 0 .
IF (MC11=1  | MC11=2 | MC11=3  |  MC11=4 ) MC11 = 1.
IF (MC11=-99  | MC11=-97 | MC11=-96  |  MC11=-9 | MC11=-6) MC11 = 9.

IF (MC12=-1  | MC12=-2 | MC12=-3  |  MC12=-4 ) MC12 = 0 .
IF (MC12=1  | MC12=2 | MC12=3  |  MC12=4 ) MC12 = 1.
IF (MC12=-99  | MC12=-97 | MC12=-96  |  MC12=-9 | MC12=-6) MC12 = 9.

IF (MC13=-1  | MC13=-2 | MC13=-3  |  MC13=-4 ) MC13 = 0 .
IF (MC13=1  | MC13=2 | MC13=3  |  MC13=4 ) MC13 = 1.
IF (MC13=-99  | MC13=-97 | MC13=-96  |  MC13=-9 | MC13=-6) MC13 = 9.

IF (MC14=-1  | MC14=-2 | MC14=-3  |  MC14=-4 ) MC14 = 0 .
IF (MC14=1  | MC14=2 | MC14=3  |  MC14=4 ) MC14 = 1.
IF (MC14=-99  | MC14=-97 | MC14=-96  |  MC14=-9 | MC14=-6) MC14 = 9.

IF (MC15=-1  | MC15=-2 | MC15=-3  |  MC15=-4 ) MC15 = 0 .
IF (MC15=1  | MC15=2 | MC15=3  |  MC15=4 ) MC15 = 1.
IF (MC15=-99  | MC15=-97 | MC15=-96  |  MC15=-9 | MC15=-6) MC15 = 9.

IF (MC16=-1  | MC16=-2 | MC16=-3  |  MC16=-4 ) MC16 = 0 .
IF (MC16=1  | MC16=2 | MC16=3  |  MC16=4 ) MC16 = 1.
IF (MC16=-99  | MC16=-97 | MC16=-96  |  MC16=-9 | MC16=-6) MC16 = 9.

IF (MC17=-1  | MC17=-2 | MC17=-3  |  MC17=-4 ) MC17 = 0 .
IF (MC17=1  | MC17=2 | MC17=3  |  MC17=4 ) MC17 = 1.
IF (MC17=-99  | MC17=-97 | MC17=-96  |  MC17=-9 | MC17=-6) MC17 = 9.

IF (MC18=-1  | MC18=-2 | MC18=-3  |  MC18=-4 ) MC18 = 0 .
IF (MC18=1  | MC18=2 | MC18=3  |  MC18=4 ) MC18 = 1.
IF (MC18=-99  | MC18=-97 | MC18=-96  |  MC18=-9 | MC18=-6) MC18 = 9.

IF (MC19=-1  | MC19=-2 | MC19=-3  |  MC19=-4 ) MC19 = 0 .
IF (MC19=1  | MC19=2 | MC19=3  |  MC19=4 ) MC19 = 1.
IF (MC19=-99  | MC19=-97 | MC19=-96  |  MC19=-9 | MC19=-6) MC19 = 9.

IF (MC20=-1  | MC20=-2 | MC20=-3  |  MC20=-4 ) MC20 = 0 .
IF (MC20=1  | MC20=2 | MC20=3  |  MC20=4 ) MC20 = 1.
IF (MC20=-99  | MC20=-97 | MC20=-96  |  MC20=-9 | MC20=-6) MC20 = 9.

IF (MC21=-1  | MC21=-2 | MC21=-3  |  MC21=-4 ) MC21 = 0 .
IF (MC21=1  | MC21=2 | MC21=3  |  MC21=4 ) MC21 = 1.
IF (MC21=-99  | MC21=-97 | MC21=-96  |  MC21=-9 | MC21=-6) MC21 = 9.

IF (MC22=-1  | MC22=-2 | MC22=-3  |  MC22=-4 ) MC22 = 0 .
IF (MC22=1  | MC22=2 | MC22=3  |  MC22=4 ) MC22 = 1.
IF (MC22=-99  | MC22=-97 | MC22=-96  |  MC22=-9 | MC22=-6) MC22 = 9.
IF (MC23=-1  | MC23=-2 | MC23=-3  |  MC23=-4 ) MC23 = 0 .
IF (MC23=1  | MC23=2 | MC23=3  |  MC23=4 ) MC23 = 1.
IF (MC23=-99  | MC23=-97 | MC23=-96  |  MC23=-9 | MC23=-6) MC23 = 9.

IF (MC24=-1  | MC24=-2 | MC24=-3  |  MC24=-4 ) MC24 = 0 .
IF (MC24=1  | MC24=2 | MC24=3  |  MC24=4 ) MC24 = 1.
IF (MC24=-99  | MC24=-97 | MC24=-96  |  MC24=-9 | MC24=-6) MC24 = 9.

EXECUTE .
```

Appendix A4

Example of BILOG-MG script used to compute test item parameters.

```
IRT Model for English Academic Spring term for a Standard Model 2010
*
>GLOBAL DFName = 'G:\BilogEQAO\academic2\verify\SM.dat',
       NPArm = 3,
       LOGistic,
       OMIts,
       SAVe;
>SAVE CALib = 'G:\BilogEQAO\academic2\verify\SM.CAL',
     PARm = 'G:\BilogEQAO\academic2\verify\SM.PAR',
     SCOre = 'G:\BilogEQAO\academic2\verify\SM.SCO';
>LENGTH NITems = (24);
>INPUT NTOtal = 24,
      NALt = 5,
      NIDchar = 14,
      OFName = 'G:\BilogEQAO\academic2\verify\ASCII.OMT';
>ITEMS INAmes = (ITEM1(1)ITEM9, ITEM10(1)ITEM24);
>TEST1 TNAme = 'TEST1',
      INUmber = (1(1)24),
      GUEss = (0.200(0)24),
      FIX = (1(0)24);
(14A1, 1X, 24A1)
>CALIB NEWton = 5,
      PLOt = 0.0100,
      ACCel = 1.0000,
      CYCLES =25,
      FIXED,
      TPRIOR,
      SPRIOR,
      CRIT=0.001,
      CHIsquare = (24,5);
>SCORE METhod = 2,
      RSCTYPE=4,
      BIWeight;
```

Appendix A5

Example of BILOG-MG script used to compute examinee abilities.

```
IRT Model for English Academic Spring term for a Modified Model 2010
*
>GLOBAL DFName = 'G:\BilogEQAO\academic2\MotivationModified\40PERCENT\MM.dat',
       NPArm = 3,
       LOGistic,
       OMIts,
       SAVe;
>SAVE CALib = 'G:\BilogEQAO\academic2\MotivationModified\40PERCENT\MM.CAL',
     PARm = 'G:\BilogEQAO\academic2\MotivationModified\40PERCENT\MM.PAR',
     SCOre = 'G:\BilogEQAO\academic2\MotivationModified\40PERCENT\MM.SCO';
>LENGTH NITems = (24);
>INPUT NTOtal = 24,
      NALt = 5,
      NIDchar = 14,
      OFName = 'G:\BilogEQAO\academic2\MotivationModified\40PERCENT\ASCII.OMT';
>ITEMS INAmes = (ITEM1(1)ITEM9, ITEM10(1)ITEM24);
>TEST1 TNAme = 'TEST0001',
      INUmber = (1(1)24),
      GUEss = (0.200(0)24),
      FIX = (1(0)24);
(14A1, 1X, 24A1)
>CALIB NEWton = 5,
      PLOt = 0.0100,
      ACCel = 1.0000,
      CYCLES =25,
      CRIT=0.001,
      CHIsquare = (24,5);
>SCORE METhod = 2,
      BIWeight;
```

Appendix A6

Example of SPSS script used to read ability estimates from BILOG-MG file.

```
*********ABILITY 30 PERCENT STANDARD MODEL*************

GET DATA  /TYPE = TXT
 /FILE = 'G:\BilogEQAO\academic2\MotivationStandard\30PERCENT\SM.SCO'
 /DELCASE = LINE
 /DELIMITERS = " "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 3
 /IMPORTCASE = ALL
 /VARIABLES =
V1 F4.2
 RECID A14
 TRIED F2.0
 RIGHT F2.0
 PERCENT F5.2
 ABILITY F9.6
 SE F8.6
 V8 F8.6
 PROB F8.6
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

delete variables V1,V8.

SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\academic2\MotivationStandard\30PERCENT\ability.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.

********NOTE: USE EXCEL HERE TO ALIGN THE DATA AND SAVE IT AS
ability1.cvs****************

*********BRING THE DATA BACK FROM EXCEL TEXT FILE

GET DATA  /TYPE = TXT
 /FILE = 'G:\BilogEQAO\academic2\MotivationStandard\30PERCENT\ability1.csv'
 /DELCASE = LINE
 /DELIMITERS = ", "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 2
 /IMPORTCASE = ALL
 /VARIABLES =
 RECID A14
 TRIED F2.0
 RIGHT F2.0
 PERCENT F5.2
 ABILITY F9.6
 SE F8.6
 PROB F8.6
 V8 F1.0
 V9 F1.0
 V10 F1.0
 V11 F1.0
 V12 F1.0
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

FILTER OFF.
USE ALL.
SELECT IF(TRIED= 16).
EXECUTE .

DELETE VARIABLES V8,V9,V10,V11,V12.

SAVE OUTFILE='G:\BilogEQAO\academic2\RESULTS\30PERCENT\30ABILITYSM.sav'
 /COMPRESSED.

************ABILITY 40 PERCENT STANDARD MODEL*****************

GET DATA  /TYPE = TXT
 /FILE = 'G:\BilogEQAO\academic2\MotivationStandard\40PERCENT\SM.SCO'
 /DELCASE = LINE
 /DELIMITERS = " "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 3
 /IMPORTCASE = ALL
 /VARIABLES =
V1 F4.2
 RECID A14
 TRIED F2.0
 RIGHT F2.0
 PERCENT F5.2
 ABILITY F9.6
 SE F8.6
 V8 F8.6
 PROB F8.6
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

delete variables V1,V8.

SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\academic2\MotivationStandard\40PERCENT\ability.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.

********NOTE: USE EXCEL HERE TO ALIGN THE DATA AND SAVE IT AS
ability1.cvs****************

*********BRING THE DATA BACK FROM EXCEL TEXT FILE

GET DATA  /TYPE = TXT
 /FILE = 'G:\BilogEQAO\academic2\MotivationStandard\40PERCENT\ability1.csv'
 /DELCASE = LINE
 /DELIMITERS = ", "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 2
 /IMPORTCASE = ALL
 /VARIABLES =
 RECID A14
 TRIED F2.0
 RIGHT F2.0
 PERCENT F5.2
 ABILITY F9.6
 SE F8.6
 PROB F8.6
 V8 F1.0
 V9 F1.0
 V10 F1.0
 V11 F1.0
 V12 F1.0
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

FILTER OFF.
USE ALL.
```

```
SELECT IF(TRIED= 24).
EXECUTE .

DELETE VARIABLES V8,V9,V10,V11,V12.

SAVE OUTFILE='G:\BilogEQAO\academic2\RESULTS\40PERCENT\40ABILITYSM.sav'
 /COMPRESSED.


************ABILITY 50 PERCENT STANDARD********************

GET DATA  /TYPE = TXT
 /FILE = 'G:\BilogEQAO\academic2\MotivationStandard\50PERCENT\SM.SCO'
 /DELCASE = LINE
 /DELIMITERS = " "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 3
 /IMPORTCASE = ALL
 /VARIABLES =
V1 F4.2
 RECID A14
 TRIED F2.0
 RIGHT F2.0
 PERCENT F5.2
 ABILITY F9.6
 SE F8.6
 V8 F8.6
 PROB F8.6
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

delete variables V1,V8.

SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\academic2\MotivationStandard\50PERCENT\ability.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.

********NOTE: USE EXCEL HERE TO ALIGN THE DATA AND SAVE IT AS
ability1.cvs****************

*********BRING THE DATA BACK FROM EXCEL TEXT FILE

GET DATA  /TYPE = TXT
```

```
 /FILE = 'G:\BilogEQAO\academic2\MotivationStandard\50PERCENT\ability1.csv'
 /DELCASE = LINE
 /DELIMITERS = ", "
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 2
 /IMPORTCASE = ALL
 /VARIABLES =
 RECID A14
 TRIED F2.0
 RIGHT F2.0
 PERCENT F5.2
 ABILITY F9.6
 SE F8.6
 PROB F8.6
 V8 F1.0
 V9 F1.0
 V10 F1.0
 V11 F1.0
 V12 F1.0
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

FILTER OFF.
USE ALL.
SELECT IF(TRIED= 24).
EXECUTE .

DELETE VARIABLES V8,V9,V10,V11,V12.

SAVE OUTFILE='G:\BilogEQAO\academic2\RESULTS\50PERCENT\50ABILITYSM.sav'
 /COMPRESSED.
```

Appendix A7

Example of SPSS script used to merge ability estimates between the two models for the same examinee.

```
GET
  FILE='G:\BilogEQAO\Academic\MotivationStandard\30PECENT\ABILITYSM.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.

delete variables TRIED, RIGHT, PERCENT, SE, PROB.

RENAME  VARIABLES (ABILITY=ABILITYSM).

SORT CASES BY
  RECID (A) .

SAVE OUTFILE='G:\BilogEQAO\Academic\ABILITYMERGE\ABILITYSM.sav'
 /COMPRESSED.
GET
  FILE='G:\BilogEQAO\Academic\MotivationModified\ABILITYMM.sav'.
DATASET NAME DataSet2 WINDOW=FRONT.

RENAME VARIABLES (ABILITY=ABILITYMM).
.
delete varibles TRIED, RIGHT, PERCENT, SE, PROB.
SORT CASES BY
  RECID (A) .
SAVE
OUTFILE='G:\BilogEQAO\Academic\ABILITYMERGE\30PERCENT\ABILITYMM.sav'
 /COMPRESSED.
MATCH FILES  /TABLE='G:\BilogEQAO\Academic\ABILITYMERGE\ABILITYSM.sav'
 /FILE='G:\BilogEQAO\Academic\ABILITYMERGE\ABILITYMM.sav'
 / BY RECID.
EXECUTE.
SAVE OUTFILE='G:\BilogEQAO\Academic\ABILITYMERGE\MATCHEDABILITIES.sav'
 /COMPRESSED.

SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\Academic\ABILITYMERGE\MATCHEDABILITIES.csv'
TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.

GRAPH
  /SCATTERPLOT(BIVAR)=ABILITYSM WITH ABILITYMM
  /MISSING=LISTWISE
  /TITLE= 'MEASURES OF ABILITY STANDARD VS MODIFIED'.
```

Appendix A8

Example of SPSS script used to scale ability estimates via mean and sigma method; bias and MSE in reference to the modified model are computed.

```
***This script will scale 30 percent of high motivation SM to 40 percent of high motivation MM

GET

FILE='G:\BilogEQAO\academic2\ABILITYMERGE\30PERCENT\30ABILITYMATCHSM.sa
v'.
DATASET NAME DataSet2 WINDOW=FRONT.
COMPUTE scaledAbility30SM = 1.10364*ABILITY30SM-0.694380 .
EXECUTE .

COMPUTE scaledbias = ABILITY40MM - scaledAbility30SM .
EXECUTE .

COMPUTE scaledSE = scaledbias ** 2 .
EXECUTE .

**********Total mean and standard Deviation******************

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.
*********LOW BIAS AND MSE
FILTER OFF.
use  1 thru  1635  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.

*********MIDDLE BIAS AND MSE
FILTER OFF.
use  1636 thru  3270  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.


*********HIGH BIAS AND MSE
```

```
FILTER OFF.
use  3271 thru  4906  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.

OUTPUT SAVE NAME=Document1

OUTFILE='G:\BilogEQAO\Academic2\RESULTS\biasMSE_ResultsAbility1\Scaled40MM30S
MAbility.spo'.

*************This script will scale 30 percent of high motivation MM to 40 percent of high
motivation MM

GET

FILE='G:\BilogEQAO\academic2\ABILITYMERGE\30PERCENT\30ABILITYMATCHMM.sa
v'.
DATASET NAME DataSet2 WINDOW=FRONT.

COMPUTE scaledAbility30MM = 0.86*ABILITY30MM-0.0710 .
EXECUTE .

COMPUTE scaledbias = ABILITY40MM - scaledAbility30MM .
EXECUTE .

COMPUTE scaledSE = scaledbias ** 2 .
EXECUTE .

**********Total mean and standard Deviation******************

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.
*********LOW BIAS AND MSE
FILTER OFF.
use  1 thru  1635  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.

*********MIDDLE BIAS AND MSE
```

```
FILTER OFF.
use  1636 thru  3270  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.

*********HIGH BIAS AND MSE
FILTER OFF.
use  3271 thru  4906  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.

OUTPUT SAVE NAME=Document1

OUTFILE='G:\BilogEQAO\Academic2\RESULTS\biasMSE_ResultsAbility1\Scaled40MM30
MMAbility.spo'.

*****************scaled 50% SM and 40 percent MM************

GET
FILE='G:\BilogEQAO\academic2\ABILITYMERGE\50PERCENT\50ABILITYMATCHSM.sa
v'.
DATASET NAME DataSet2 WINDOW=FRONT.

COMPUTE scaledAbility50SM = 1.104322*ABILITY50SM-0.4750 .
EXECUTE .

COMPUTE scaledbias = ABILITY40MM - scaledAbility50SM .
EXECUTE .

COMPUTE scaledSE = scaledbias ** 2 .
EXECUTE .

**********Total mean and standard Deviation******************

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .
*********LOW BIAS AND MSE
FILTER OFF.
use  1 thru  3041 .
```

```
EXECUTE .

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .

*********MIDDLE BIAS AND MSE
FILTER OFF.
use  3042 thru  6082  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .

*********HIGH BIAS AND MSE
FILTER OFF.
use  6083 thru  9123  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .

OUTPUT SAVE NAME=Document1

OUTFILE='G:\BilogEQAO\Academic2\RESULTS\biasMSE_ResultsAbility1\Scaled40MM50S
MAbility.spo'.

*****************scaled 50% MM and 40 percent MM************

GET

FILE='G:\BilogEQAO\academic2\ABILITYMERGE\50PERCENT\50ABILITYMATCHMM.sa
v'.
DATASET NAME DataSet2 WINDOW=FRONT.

COMPUTE scaledAbility50MM = 1.0062*ABILITY50MM-0.08351 .
EXECUTE .

COMPUTE scaledbias = ABILITY40MM - scaledAbility50MM .
EXECUTE .

COMPUTE scaledSE = scaledbias ** 2 .
EXECUTE .
```

```
**********Total mean and standard Deviation*******************

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .

*********LOW BIAS AND MSE
FILTER OFF.
use  1 thru  3041 .
EXECUTE .

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .

*********MIDDLE BIAS AND MSE
FILTER OFF.
use  3042 thru  6082  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .

*********HIGH BIAS AND MSE
FILTER OFF.
use  6083 thru  9123  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN .

OUTPUT SAVE NAME=Document1

OUTFILE='G:\BilogEQAO\Academic2\RESULTS\biasMSE_ResultsAbility1\Scaled40MM50
MMAbility.spo' .

************This script will scale 40 percent of low motivation SM to 40 percent of high
motivation MM

GET

FILE='G:\BilogEQAO\academic2\ABILITYMERGE\40PERCENT\40ABILITYMATCHSM.sa
v'.
DATASET NAME DataSet2 WINDOW=FRONT.
```

```
COMPUTE scaledAbility40SM = 1.1123*ABILITY40SM-0.55063.
EXECUTE .

COMPUTE scaledbias = ABILITY40MM - scaledAbility40SM .
EXECUTE .

COMPUTE scaledSE = scaledbias ** 2 .
EXECUTE .

**********Total mean and standard Deviation*******************

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.
*********LOW BIAS AND MSE
FILTER OFF.
use  1 thru  3014  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.

*********MIDDLE BIAS AND MSE
FILTER OFF.
use  3015 thru  6082  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.

*********HIGH BIAS AND MSE
FILTER OFF.
use  6083 thru  9123  .
EXECUTE .

DESCRIPTIVES
 VARIABLES=scaledbias scaledSE
  /STATISTICS=MEAN STDDEV.
OUTPUT SAVE NAME=Document1

OUTFILE='G:\BilogEQAO\Academic2\RESULTS\biasMSE_ResultsAbility1\Scaled40MM40S
MAbility.spo'.
```

Appendix A9

Example of SPSS script used to read Bilog-MG data for item parameter estimates.

```
*********BEGINS FILE FOR 40 PERCENT Modified (ideal situation)

GET DATA  /TYPE = TXT
 /FILE =
  'G:\BilogEQAO\Academic2\MotivationModified\40PERCENT\MM.PAR'
 /FIXCASE = 1
 /ARRANGEMENT = FIXED
 /FIRSTCASE = 5
 /IMPORTCASE = ALL
 /VARIABLES =
 /1  V1 0-7 A8
 V2 8-15 A8
 V3 16-25 F10.3
 V4 26-35 F10.3
 V5 36-45 F10.3
 V6 46-55 F10.3
 V7 56-65 F10.3
 V8 66-75 F10.3
 V9 76-85 F10.3
 V10 86-95 F10.3
 V11 96-105 F10.3
 V12 106-115 F10.3
 V13 116-125 F10.3
 V14 126-135 F10.3
 V15 136-151 F16.3
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

********DELETING VARIABLES THAT ARE NOT NEEDED

DELETE VARIABLES  V2,V3,V4,V6,V8,V9,V10,V12,V13,V14,V15.

RENAME VARIABLES V1=ITEM V5=SLOPEMM40, V7=TRESHOLDMM40,
V11=ASYMPTOTEMM40.

SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\Academic2\RESULTS\40PERCENT\PARAMETERS40MM.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.
```

```
*************LOAD FILE FOR 15 PERCENT STANDARD

GET DATA  /TYPE = TXT
 /FILE =
  'G:\BilogEQAO\Academic2\MotivationStandard\15PERCENT\SM.PAR'
 /FIXCASE = 1
 /ARRANGEMENT = FIXED
 /FIRSTCASE = 5
 /IMPORTCASE = ALL
 /VARIABLES =
 /1  V1 0-7 A8
 V2 8-15 A8
 V3 16-25 F10.3
 V4 26-35 F10.3
 V5 36-45 F10.3
 V6 46-55 F10.3
 V7 56-65 F10.3
 V8 66-75 F10.3
 V9 76-85 F10.3
 V10 86-95 F10.3
 V11 96-105 F10.3
 V12 106-115 F10.3
 V13 116-125 F10.3
 V14 126-135 F10.3
 V15 136-151 F16.3
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

********DELETING VARIABLES THAT ARE NOT NEEDED

DELETE VARIABLES  V2,V3,V4,V6,V8,V9,V10,V12,V13,V14,V15.

RENAME VARIABLES V1=ITEM V5=SLOPESM15, V7=TRESHOLDSM15,
V11=ASYMPTOTESM15.

SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\Academic2\RESULTS\15PERCENT\PARAMETERS15SM.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.

************LOAD 30 PERCENT STANDARD MODEL

GET DATA  /TYPE = TXT
```

```
 /FILE =
  'G:\BilogEQAO\Academic2\MotivationStandard\30PERCENT\SM.PAR'
 /FIXCASE = 1
 /ARRANGEMENT = FIXED
 /FIRSTCASE = 5
 /IMPORTCASE = ALL
 /VARIABLES =
 /1  V1 0-7 A8
 V2 8-15 A8
 V3 16-25 F10.3
 V4 26-35 F10.3
 V5 36-45 F10.3
 V6 46-55 F10.3
 V7 56-65 F10.3
 V8 66-75 F10.3
 V9 76-85 F10.3
 V10 86-95 F10.3
 V11 96-105 F10.3
 V12 106-115 F10.3
 V13 116-125 F10.3
 V14 126-135 F10.3
 V15 136-151 F16.3
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

********DELETING VARIABLES THAT ARE NOT NEEDED

DELETE VARIABLES  V2,V3,V4,V6,V8,V9,V10,V12,V13,V14,V15.

RENAME VARIABLES V1=ITEM V5=SLOPESM30, V7=TRESHOLDSM30,
V11=ASYMPTOTESM30.
SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\Academic2\RESULTS\30PERCENT\PARAMETERS30SM.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.

**********LOAD 40 PERCENT STANDARD MODEL

GET DATA  /TYPE = TXT
 /FILE =
  'G:\BilogEQAO\Academic2\MotivationStandard\40PERCENT\SM.PAR'
 /FIXCASE = 1
 /ARRANGEMENT = FIXED
 /FIRSTCASE = 5
 /IMPORTCASE = ALL
 /VARIABLES =
 /1  V1 0-7 A8
 V2 8-15 A8
 V3 16-25 F10.3
 V4 26-35 F10.3
 V5 36-45 F10.3
 V6 46-55 F10.3
 V7 56-65 F10.3
 V8 66-75 F10.3
 V9 76-85 F10.3
 V10 86-95 F10.3
 V11 96-105 F10.3
 V12 106-115 F10.3
 V13 116-125 F10.3
 V14 126-135 F10.3
 V15 136-151 F16.3
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

********DELETING VARIABLES THAT ARE NOT NEEDED

DELETE VARIABLES  V2,V3,V4,V6,V8,V9,V10,V12,V13,V14,V15.

RENAME VARIABLES V1=ITEM V5=SLOPESM40, V7=TRESHOLDSM40,
V11=ASYMPTOTESM40.
SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\Academic2\RESULTS\40PERCENT\PARAMETERS40SM.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.

**********LOAD 50PERCENT STANDARD MODEL

GET DATA  /TYPE = TXT
 /FILE =
  'G:\BilogEQAO\Academic2\MotivationStandard\50PERCENT\SM.PAR'
 /FIXCASE = 1
 /ARRANGEMENT = FIXED
 /FIRSTCASE = 5
 /IMPORTCASE = ALL
 /VARIABLES =
 /1  V1 0-7 A8
 V2 8-15 A8
 V3 16-25 F10.3
 V4 26-35 F10.3
 V5 36-45 F10.3
 V6 46-55 F10.3
 V7 56-65 F10.3
 V8 66-75 F10.3
 V9 76-85 F10.3
 V10 86-95 F10.3
 V11 96-105 F10.3
 V12 106-115 F10.3
 V13 116-125 F10.3
 V14 126-135 F10.3
 V15 136-151 F16.3
 .
CACHE.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

********DELETING VARIABLES THAT ARE NOT NEEDED

DELETE VARIABLES  V2,V3,V4,V6,V8,V9,V10,V12,V13,V14,V15.

RENAME VARIABLES V1=ITEM V5=SLOPESM50, V7=TRESHOLDSM50,
V11=ASYMPTOTESM50.
SAVE TRANSLATE
  OUTFILE='G:\BilogEQAO\Academic2\RESULTS\50PERCENT\PARAMETERS50SM.csv'
  /TYPE=CSV /MAP /REPLACE /FIELDNAMES
  /CELLS=VALUES.
```

Appendix A10

Example of SPSS script used to equate parameter estimates.

```
************Equating 30percent SM and 40 percent MM*****************

GET
  FILE='G:\BilogEQAO\academic2\RESULTS\biasMSE_ResultsItems\40MM30SM.sav'.
DATASET NAME DataSet3 WINDOW=FRONT.

COMPUTE scaledThresholdSM = 0.82950*TRESHOLDSM30-0.0304 .
EXECUTE .

COMPUTE ScaledBiasThreshold = TRESHOLDMM40 - scaledThresholdSM .
EXECUTE .

COMPUTE ScaledSEThreshold = ScaledBiasThreshold ** 2 .
EXECUTE .

SORT CASES BY
  TRESHOLDMM40 (A) .

DESCRIPTIVES
  VARIABLES=ScaledBiasThreshold ScaledSEThreshold
  /STATISTICS=MEAN .

*********LOW BIAS AND MSE (TRESHOLD)
FILTER OFF.
use  1 thru  8  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=ScaledBiasThreshold ScaledSEThreshold
  /STATISTICS=MEAN .

*********MEDIUM BIAS AND MSE (TRESHOLD)
FILTER OFF.
use  9 thru  16  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=ScaledBiasThreshold ScaledSEThreshold
  /STATISTICS=MEAN .

*********HIGH BIAS AND MSE (TRESHOLD)
FILTER OFF.
```

```
use  17 thru  24  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=ScaledBiasThreshold ScaledSEThreshold
  /STATISTICS=MEAN .

*********SAVE OUTPUT************

OUTPUT SAVE NAME=Document1

OUTFILE='G:\BilogEQAO\Academic2\RESULTS\biasMSE\Scaled40MM40SMBIAS_MSE.sp
o'.
```

Appendix A11

Example of SPSS script used to compute bias and MSE for test item parameters.

```
***********THIS SCRIPT WILL COMPUTE BIAS AND MEAN SQUARE
ERROR*****************

**********THIS ANALYSIS WILL BE USED FOR THE SECOND QUESTION- 40
PERCENT SM AND 40 PERCENT MM

******OPEN 40 PERCENT MM *******************

GET DATA  /TYPE = TXT
 /FILE =
'G:\BILOGEQAO\ACADEMIC2\RESULTS\40PERCENT\PARAMETERS40MM.CSV'
 /DELCASE = LINE
 /DELIMITERS = ","
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 2
 /IMPORTCASE = ALL
 /VARIABLES =
 ITEM A6
 SLOPEMM40 F7.5
 TRESHOLDMM40 F8.5
 ASYMPTOTEMM40 F2.1
 .
CACHE.
EXECUTE.
DATASET NAME DATASET1 WINDOW=FRONT.

SORT CASES BY
  ITEM (A) .

SAVE
OUTFILE='G:\BILOGEQAO\ACADEMIC2\RESULTS\40PERCENT\40PERCENTMM.SAV'
 /COMPRESSED.

**********OPEN 40 PERCENT SM

GET DATA  /TYPE = TXT
 /FILE =
'G:\BILOGEQAO\ACADEMIC2\RESULTS\40PERCENT\PARAMETERS40SM.CSV'
 /DELCASE = LINE
 /DELIMITERS = ","
 /ARRANGEMENT = DELIMITED
 /FIRSTCASE = 2
 /IMPORTCASE = ALL
 /VARIABLES =
 ITEM A6
 SLOPESM40 F7.5
 TRESHOLDSM40 F8.5
 ASYMPTOTE40 F2.1
 .
CACHE.
EXECUTE.
DATASET NAME DATASET2 WINDOW=FRONT.

SORT CASES BY
  ITEM (A) .

SAVE
OUTFILE='G:\BILOGEQAO\ACADEMIC2\RESULTS\40PERCENT\40PERCENTSM.SAV'
 /COMPRESSED.

************MERGE 40MM PERCENT AND 40 PERCENT SM

MATCH FILES
/TABLE='G:\BILOGEQAO\ACADEMIC2\RESULTS\40PERCENT\40PERCENTMM.SAV'
 /FILE='G:\BILOGEQAO\ACADEMIC2\RESULTS\40PERCENT\40PERCENTSM.SAV'
 / BY ITEM.
EXECUTE.


*******COMPUTE BIAS AND MSE FOR SLOPE AND THRESHOLD

COMPUTE BIASSLOPE = SLOPEMM40- SLOPESM40 .
EXECUTE .

COMPUTE SESLOPE = BIASSLOPE ** 2 .
EXECUTE .

COMPUTE BIASTHRESHOLD = TRESHOLDMM40 - TRESHOLDSM40 .
EXECUTE .

COMPUTE SETHRESHOLD = BIASTHRESHOLD ** 2 .
EXECUTE .

DESCRIPTIVES
  VARIABLES=BIASSLOPE SESLOPE BIASTHRESHOLD SETHRESHOLD
  /STATISTICS=MEAN STDDEV MIN MAX .
```

```
*********SAVE MERGED RESPONSES

SAVE OUTFILE='G:\BILOGEQAO\ACADEMIC2\RESULTS\BIASMSE\40MM40SM.SAV'
 /COMPRESSED.

SORT CASES BY
  SLOPEMM40 (A) .

GRAPH
  /SCATTERPLOT(BIVAR)=SLOPEMM40 WITH SLOPESM40
  /MISSING=LISTWISE .

*********************DIFFERENT SLOPE LEVEL****************************

*********LOW BIAS AND MSE (SLOPE)
FILTER OFF.
USE  1 THRU  8  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=BIASSLOPE SESLOPE
  /STATISTICS=MEAN STDDEV MIN MAX .

*********MEDIUM BIAS AND MSE (SLOPE)
FILTER OFF.
USE  9 THRU  16  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=BIASSLOPE SESLOPE
  /STATISTICS=MEAN STDDEV MIN MAX .

*********HIGH BIAS AND MSE (SLOPE)
FILTER OFF.
USE  17 THRU  24  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=BIASSLOPE SESLOPE
  /STATISTICS=MEAN STDDEV MIN MAX .

***********************DIFFERENT THRESHOLD
LEVELS************************
SORT CASES BY
  TRESHOLDMM40 (A) .
```

```
GRAPH
  /SCATTERPLOT(BIVAR)=TRESHOLDMM40 WITH TRESHOLDSM40
  /MISSING=LISTWISE

*********LOW BIAS AND MSE (TRESHOLD)
FILTER OFF.
USE  1 THRU  8  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=BIASTHRESHOLD SETHRESHOLD
  /STATISTICS=MEAN STDDEV MIN MAX .

*********MEDIUM BIAS AND MSE (TRESHOLD)
FILTER OFF.
USE  9 THRU  16  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=BIASTHRESHOLD SETHRESHOLD
  /STATISTICS=MEAN STDDEV MIN MAX .

*********HIGH BIAS AND MSE (TRESHOLD)
FILTER OFF.
USE  17 THRU  24  .
EXECUTE .

DESCRIPTIVES
  VARIABLES=BIASTHRESHOLD SETHRESHOLD
  /STATISTICS=MEAN STDDEV MIN MAX .

*********SAVE OUTPUT************

OUTPUT SAVE NAME=DOCUMENT1

OUTFILE='G:\BILOGEQAO\ACADEMIC2\RESULTS\BIASMSE\40MM40SMBIAS_MSE.S
PO'.
```

Appendix A12

Example of BILOG-MG script used to compute DIF.

```
DIF IRT MODEL FOR ENGLISH ACADEMIC SPRING TERM (2010)
*
>GLOBAL DFNAME = 'G:\BILOGEQAO\ACADEMIC2\DIF\M12.DAT',
       NPARM = 3,
       LOGISTIC,
       OMITS,
       SAVE;
>SAVE CALIB = 'G:\BILOGEQAO\ACADEMIC2\DIF\M12.CAL',
     PARM = 'G:\BILOGEQAO\ACADEMIC2\DIF\M12.PAR',
     DIF =  'G:\BILOGEQAO\ACADEMIC2\DIF\M12.DIF'
     SCORE = 'G:\BILOGEQAO\ACADEMIC2\DIF\M12.SCO';
>LENGTH NITEMS = (24);
>INPUT NTOTAL = 24,
      NIDCHAR = 14,
      NGROUP=2,
      DIF,
      OFNAME = 'G:\BILOGEQAO\ACADEMIC2\DIF\ASCII.OMT';
>ITEMS INAMES = (ITEM01(1)ITEM09, ITEM10(1)ITEM24);
>TEST1 TNAME = 'TEST1',
      INUMBER = (1(1)24);
>GROUP1 GNAME = 'MODEL1',
      LENGTH = 24,
      INUMBER = (1(1)24);
>GROUP2 GNAME = 'MODEL2',
      LENGTH = 24,
      INUMBER= (1(1)24);
(14A1, 26X, I1, T16, 24A1)
>CALIB CYCLES = 25,
      NEWTON = 5,
      NQPT = 30,
      CRIT = 0.010,
      PLOT = 1,
      ACCEL = 1.0000,
      FIXED,
      COMMON,
      SPRIOR,
      REFERENCE=2,
      CHISQUARE = (24, 5);
```