EVIDENTIARY VALIDITY OF THE EDUCATION QUALITY AND ACCOUNTABILITY OFFICE'S MATHEMATICS ASSESSMENT

By

Nazli Uygun Emil

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods — Doctor of Philosophy

2020

ABSTRACT

EVIDENTIARY VALIDITY OF THE EDUCATION QUALITY AND ACCOUNTABILITY OFFICE'S MATHEMATICS ASSESSMENT

By

Nazli Uygun Emil

Validity of a measurement refers to appropriate test score meanings, uses, and interpretations (Messick, 1989; Kane, 1992). There are different approaches to validity: an evidentiary aspect of validity is one requiring gathering statistical evidence to evaluate test score meaning. A common approach to validation is comparisons of test score equity across different population groups. This research examines the evidentiary validity of mathematics test administrations by the Education Quality and Accountability Office in Ontario, Canada. Using factorial invariance and differential item functioning analyses, score validity was investigated across both achievement level and gender groups. Validity evidence was provided via a four-step measurement invariance procedure and differential item functioning analyses. However, items causing invariance problems and/or functioning in favor or against a certain population group are identified through the analyses. These violations are not necessarily a threat to construct validity but provide guidance for revisiting the test framework and revising some of the items. Content strands functioning differently for gender groups are mostly consistent with conclusions in the existing literature. Practical suggestions for measurement equity are discussed after reporting statistical findings. To reduce achievement gaps between student groups, future studies are needed to identify items causing invariance obstacles, presenting partial invariance solutions, and revising items with differential item functioning.

To Kemal & his father Huseyin,
my nephew Uras Ata Uygun,
my grandparents Halil & Fikriye,
and our cats Fistik & Hoshaf.

ACKNOWLEDGEMENTS

I would like to present a sincere appreciation to my advisor Dr. Mark Reckase, for all his supports during my doctoral education. This research would not be completed without his invaluable and continuous supports. I would like to send my gratitude to my committee members Drs. Corey Drake, Richard Houang, and Kimberly Kelly as well, for their contributions to this research study. I also wish to acknowledge all faculty members of the Measurement and Quantitative Methods for providing me necessary knowledge, skills, and resources to earn this doctorate degree at Michigan State University, staff in our college, Brooke Stodyk and other staff in international office for helping me understanding requirements of being an international student. Thanks to William Sullivan as well, for providing grammar edits and suggestions.

Many thanks to EQAO, Gönülden Gönüle Anadolu (Heart2Heart Anatolia), TACAM, and TUBITAK for their technical or financial supports during my school years in the United States.

Although it is not possible to mention each person who contributed my academic and life goals, I would like to thank everyone in my family, all my friends; here in the United States and in the Republic of Turkey, to my previous advisors at University of Florida and at Middle East Technical University, and to everyone at MSU, who have all contributed my learning years as a lifelong student.

Heartful thanks to my spouse Mustafa Kemal Emil, who has followed me to overseas and has supported me at each step in this life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

CI:	Confidence Interval
CFI:	Comparative Fit Index
CTT:	Classical Test Theory
DIF:	Differential Item Functioning
EQAO:	Education Quality and Accountability Office
ICC:	Item Characteristic Curve
IRT:	Item Response Theory
NELS: 88:	National Educational Longitudinal Study of 1988
OECD:	Organization for Economic Cooperation and Development
PISA:	Programme for International Student Assessment
RMSEA:	Root Mean Square Root of Error
SA:	Signed Area
SAT:	Scholastic Aptitude Test
SEM:	Structural Equation Modeling
SES:	Socio-economic Status
STEM:	Science Technology Engineering and Mathematics
TLI:	Tucker-Lewis Index
UNESCO:	United Nations Educational, Scientific and Cultural Organization
US:	Unsigned Area
WASL:	Washington Assessment of Student Learning
3PL:	Three Parameter Logistic Model

Chapter 1: Introduction of the Research Questions

In this chapter, research questions to be investigated through this dissertation study are introduced after providing a brief statement of purpose and significance of the research inquiries.

1.1.Purpose of the Research

The intended purpose of this study is to investigate evidentiary aspects of validity by use of item-level response data obtained from a 9th grade mathematics assessment administered by the Education Quality and Accountability Office (EQAO) in the province of Ontario, Canada. An extensive definition of validity in the context of standards-based achievement testing is provided by Messick (1989) as the "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." Although there are different aspects of validity with their own significance, the main focus of the present research will be evidential aspects by means of group comparisons. Evaluation of validity can be directly made via comparisons of examinee groups' test score results (Messick, 1989). Mainly, cross-sectional test level and item level analyses will be conducted to investigate the existence of differences across examinee groups based on item responses of academic and applied math tests administered in 2015. As for particular statistical procedures: factorial measurement invariance analyses with structural equation modeling, and differential item functioning for bias evaluation will be conducted to provide statistical information for evidentiary aspects of construct validity.

After describing the unique requirements for investigations of these aspects of validity, the mathematical constructs assessed by EQAO are described. In addition, the EQAO Provincial Assessment system, Ontario's Mathematical Curriculum, and the Equity and Inclusive Education Strategy (2009) are summarized to provide background relative to the investigation of the aspects of validity. Finally, formal research questions of the study are presented.

At this point, the definitions of and relationships between the terms of measurement and mathematics are presented, aiming to provide a better understanding of the conceptual structure and significance of the research leading to research questions. Measurement has its roots in the earliest times since humans needed to measure time (Sloley, 1931), distance, and other quantities. Measurement means the process of systematic assignment of numbers on variables to represent characteristics of persons, objects, or events (Vanderberg & Lance, 2000). Historically, evaluation of measurement in psychological and educational fields has been rooted in classical test theory, which defines true scores as the difference between observed and error scores (Crocker & Algina, 1986; Lord & Novick, 1968 as cited in Vanderberg & Lance, 2000). Unlike physical attributes such as weight or height, psychological attributes (i.e., constructs) of persons cannot be measured directly since they are hypothetical concepts (Crocker & Algina, 1986). For that reason, it is very common to encounter the terms of assessment and measurement interchangeably in the literature of psychological constructs. Because the existence of psychological or academic constructs can never be confirmed absolutely but can be inferred from observations of examinee behaviors (Crocker & Algina, 1986), development and validation of measurement instruments to assess these unobserved (latent) constructs is vital for educational studies. Educational measurement and test score validation guide educational researchers, teachers, policymakers, and other stakeholders to improve students' academic learning and achieve higher educational standards.

Although being one of the oldest concepts of humanity, describing mathematics as a philosophical and scientific construct is challenging. Mathematical constructs are not simply composed of pure mathematical knowledge and mathematical theories. Mathematical skills and their applications that build on mathematical constructs such as problem solving, reasoning, and computational strategies are also interwoven. The Programme for International Student Assessment (PISA) provides a well-respected and internationally accepted description of mathematical literacy. This research is structured based on the definition of PISA's mathematical literacy as a general guideline, which is stated as:

> "the capacity to identify, understand and engage in mathematics, and to make well-founded judgments about the role that mathematics plays in an individual's current and future private life, occupational life, social life with peers and relatives, and life as a constructive, concerned and reflective citizen."
>
> (OECD, 2002).

As seen from the above description, students are globally desired to be mathematics-literate via learning mathematical knowledge and engaging mathematical applications for the sake of being constructive citizens in their individual, occupational, and social lives. In this era, as being one of the most important components of the STEM disciplines, students' academic success in mathematics contributes to humanity and society (National Research Council, 2011), today's economy, and empowers remarkably the economy of the future (Schmidt, 2011). Therefore, measurement of mathematics achievement of students enrolled in K-12 education is a significant activity for a nation's educational, social, and economic development.

What is measured as common constructs of mathematics achievement might change based on the curriculum and/or state standards. In Canada, for example, EQAO accepts the mathematical literacy definition by the OECD and focuses on Ontario's curriculum expectations as the primary

resource while constructing a mathematical framework for student assessment. Education Quality and Accountability Office (EQAO) assess mathematical constructs such as Number Sense and Algebra, Linear Relations, Analytic Geometry, and Measurement and Geometry as mathematical concepts defined in the Ontario Curriculum. In addition to these concepts, the EQAO provincial assessment system, Ontario's mathematical curriculum, and the equity and inclusive education strategy (2009) are summarized in the policy and literature review section of this study to provide a relative background to the investigation of an evidentiary aspect of ninth grade mathematics test score validity.

1.2.Research Questions

Research questions of the study are presented in this section, after summarizing the study purpose and significance and providing descriptions of student subpopulations to elaborate the research focus. Those questions were designed to obtain empirical evidence and draw conclusions about evidentiary and differential validity of the EQAO's mathematics tests for ninth graders. The methodology used to answer these research inquiries are described in more detail in the methods chapter.

Students' assessments administered by EQAO serve as a tool to understand curriculum expectations and state education standards, evaluate students' school performance, and provide solutions and strategies for students', schools', and educators' success. Those measurement instruments must also consider equity and diversity of the student population in addition to improving students' learning. In this respect, gathering validity evidence for those tests is crucial because that concept refers to appropriate test score interpretation and use for the benefit of improving educational quality of the province. When the equity strategy of the province and curriculum expectations were reviewed, three goals were identified from materials provided by the

Ministry of Education, Ontario: 1- reaching the highest level of student success, 2- reducing the achievement gap between students, 3- enhancing overall public confidence by means of improving quality of education (Ontario's Equity and Inclusive Education Strategy, 2009). The importance of an inclusive and equal education strategy is not only mentioned by Ontario's education policies as a primary focus of education but also a global requirement defined by UNESCO (2008) to obtain a well-functioning and high standard education worldwide. This universal goal of equal education strategies, as stated in Ontario's educational policies, aims to reach the highest level of educational success with less achievement gaps between individuals, student groups, and school districts.

To understand the strengths and weaknesses in the existing structure of the education system and enhance students' academic success, assessment tools must be constructed, used, and interpreted appropriately. Validation of measurement instruments operates as one of the most crucial components for quality improvement of education while simultaneously contributing to the enhancement of society. Interpreting test score validity is not the only necessary task to achieve this goal; considering the diverse demographic structure of the student population, test equality is as important as the appropriate test score interpretation, which can also be investigated as a segment of validity. For that reason, this research investigates statistical evidence obtained from EQAO's students assessment data for test score meaning and interpretation. Empirical data obtained from ninth grade mathematical test administrations is used to conclude the appropriateness of test score use and test equity among student subgroups.

For this specific research, gender and achievement level groups were chosen as the subpopulations used for the evaluation of test equity and score validity. In addition to a less ambiguous student group like gender, definitions of success level groups need to be provided by referring to the achievement chart of the 9th grade mathematics curriculum of Ontario.

There are four levels defining students' achievement in the 9th grade curriculum. Those are level 1 indicating a limited level of achievement; level 2 indicating some degree of achievement; level 3 for a considerable achievement; and level 4 a high degree or thorough achievement. All these levels will also be mentioned later in the education policy review section of this dissertation manuscript. Decisions regarding these levels are made based on evaluations of students' efforts throughout the school year and are commented to the students by means of "Provincial Report Card", for grades 9 to 12.

In the light of introductory significance and definition of student subpopulation groups, which are to be analyzed to gather validity evidence, research questions of this study are presented as:

1- Does the EQAO Mathematics achievement test for 9^{th} graders measure the same constructs such as number sense and algebra, linear relations, analytic geometry, and measurement geometry across achievement groups for the 2015 winter academic test administrations?

a- Does the measurement model exhibit configural invariance (i.e., equal pattern of loadings) across groups, that is, all achievement groups relate the same subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance across achievement groups, that is, strength of the relationship between items and the relevant constructs (i.e., magnitude of loadings) are the same for examinees of each achievement group?

c- Does the measurement model exhibit strong/scalar invariance across achievement groups, that is, means of items are equal across groups?

d- Does the measurement model exhibit strict invariance across achievement groups, that is, factor residual variances are equal across groups?

2- Does the EQAO Mathematics achievement test for 9th graders measure the same constructs such as number sense and algebra, linear relations, analytic geometry, and measurement geometry between gender groups for the 2015 winter academic test administrations?

a- Does the measurement model exhibit configural invariance across groups, that is, both male and female students relate the same subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance between groups, that is, strength of the relationship between items and the relevant constructs (i.e., magnitude of loadings) are the same for examinees of each group?

c- Does the measurement model exhibit strong/scalar invariance between gender groups, that is, means of items are equal across groups?

d- Does the measurement model exhibit strict invariance across gender groups, that is, factor residual variances are equal across groups?

3- Does the EQAO Mathematics achievement test for 9th graders measure the same constructs, such as number sense and algebra, linear relations, and measurement geometry for achievement groups for the 2015 winter applied test administrations?

a- Does the measurement model exhibit configural invariance across groups, that is, all achievement groups relate the same subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance across achievement groups, that is, strength of the relationship between items and the relevant constructs are the same for examinees of each achievement group?

c- Does the measurement model exhibit strong/scalar invariance across achievement groups, that is, item means are equal across groups?

d- Does the measurement model exhibit strict invariance across achievement groups, that is, factor residual variances are equal across groups?

4- Does the EQAO Mathematics achievement test for 9th graders measure the same constructs, such as number sense and algebra, linear relations, and measurement geometry for gender groups for the 2015 winter applied test administrations?

a- Does the measurement model exhibit configural invariance across groups, that is, females and males relate the same subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance across achievement groups, that is, strength of the relationship between items and the relevant constructs are the same for examinees of gender groups?

c- Does the measurement model exhibit strong/scalar invariance between genders, that is, item means are equal across groups?

d- Does the measurement model exhibit strict invariance across gender groups, that is, factor residual variances are equal across groups?

Similar research questions will be investigated for spring semester test administrations:

5- Does the EQAO Mathematics achievement test for 9th graders measure the same constructs such as number sense and algebra, linear relations, analytic geometry, and measurement geometry across achievement groups for the 2015 spring academic test administrations?

a- Does the measurement model exhibit configural invariance across groups, that is, all achievement groups relate the same subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance across achievement groups, that is, strength of the relationship between items and the relevant constructs (i.e. magnitude of loadings) are the same for examinees of each achievement group?

c- Does the measurement model exhibit strong/scalar invariance across achievement groups, that is, means of items are equal across groups?

d- Does the measurement model exhibit strict invariance across achievement groups, that is, factor residual variances are equal across groups?

6- Does the EQAO Mathematics achievement test for 9^{th} graders measure the same constructs such as number sense and algebra, linear relations, analytic geometry, and measurement geometry across gender groups for the 2015 spring academic test administrations?

a- Does the measurement model exhibit configural invariance across groups, that is, all gender groups relate the same subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance across genders, that is, strength of the relationship between items and the relevant constructs (i.e., magnitude of loadings) are the same for examinees of each gender group?

c- Does the measurement model exhibit strong/scalar invariance between gender groups, that is, means of items are equal across groups?

d- Does the measurement model exhibit strict invariance between gender groups, that is, factor residual variances are equal across groups?

7- Does the EQAO Mathematics achievement test for 9th graders measure the same constructs, such as number sense and algebra, linear relations, and measurement geometry for achievement groups for the 2015 spring applied test administrations?

a- Does the measurement model exhibit configural invariance across groups, that is, all achievement groups relate the same subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance across achievement levels, that is, strength of the relationship between items and the relevant constructs are the same for examinees of each achievement group?

c- Does the measurement model exhibit strong/scalar invariance across achievement groups, that is, item means are equal across groups?

d- Does the measurement model exhibit strict invariance across achievement groups, that is, factor residual variances are equal across groups?

8- Does the EQAO Mathematics achievement test for 9th graders measure the same constructs, such as number sense and algebra, linear relations, and measurement geometry for gender groups for the 2015 spring applied test administrations?

a- Does the measurement model exhibit configural invariance across groups, that is, females and males similarly relate the subsets of items with the same constructs?

b- Does the measurement model exhibit weak/metric invariance across gender groups, that is, strength of the relationship between items and the relevant constructs or factor loadings are the same for male and female examinees?

c- Does the measurement model exhibit strong/scalar invariance across achievement groups, that is, item means are equal across groups?

d- Does the measurement model exhibit strict invariance across achievement groups, that is, factor residual variances are equal across groups?

In addition to the measurement invariance analyses for evidentiary validation, by means of differential function analysis below-mentioned research questions are going to be investigated to assess possible item bias for different population groups.

9- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any achievement group for the winter 2015 applied test administration?

10- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any gender group in the winter 2015 applied test administration?

11- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any achievement group for the winter 2015 academic test administration?

12- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any gender group in the winter 2015 academic test administration?

13- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any achievement group for the spring 2015 applied test administration?

14- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any gender group in the spring 2015 applied test administration?

15- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any achievement group for the spring 2015 academic test administration?

16- Do any of the mathematics items in the test have violations of equivalence according to Raju's differential item functioning method in favor of any gender group in the spring 2015 academic test administration?

Chapter 2: Policy and Literature Review

In this chapter, mathematical constructs in EQAO student assessments are described by means of reviewing the EQAO test framework, mathematics curriculum of the state, and provincial equity and achievement policies. Following the curriculum and policy review, different aspects of validity in the existing literature and examples of validation studies are presented.

2.1.EQAO Provincial Assessments

Education Quality and Accountability Office is an arm-length government agency of the province of Ontario, Canada, assessing the achievement of students in reading, writing, and mathematics based on the expectations in the Ontario curriculum. This agency also reports students' mathematical learning test results to parents, educators, and government. The Ministry of Education, district school boards, and schools use these results to improve learning, teaching, and student achievement to determine the strengths of individual students and identify the learning objectives needing improvement in mathematics education for the ninth graders. Based on these assessment reports, areas for educational improvement are specified, and instructional strategies for targeted improvements are provided for staff and stakeholder development. An Individual Student Report is also provided to each student who takes an EQAO assessment. All these test results serve as a resource to provide guidance to improve students' learning and revise or design new teaching strategies (EQAO, 2009).

Four provincial assessments are conducted each year by EQAO, these are : 1) the Assessment of Reading, Writing and Mathematics, Primary Division (1^{st} - 3^{rd} grades), 2) the Assessment of Reading, Writing and Mathematics, Junior Division (4^{th} - 6^{th} grades), 3) the Grade 9 Assessment of Mathematics, 4) the Ontario Secondary School Literacy Test.

In Ontario, there are two main kinds of educational assessment, large-scale educational testing, and classroom assessment, which are essential elements of the educational system. Large-scale testing measures students' academic success across the province at certain times in students' schooling, along with critical content fields and cognitive objectives (EQAO, 2009) as summarized below.

EQAO's Large Scale Assessments:

Purpose: To provide comparable year-to-year data on student achievement for public information.

Provide data which is reliable, objective, and high-quality that can direct school boards' improvement and target setting.

Expectations from the prescribed curriculum are measured by EQAO large-scale-assessments by tasks and items from the domain of the assessed curriculum.

The same items (in a year) or psychometrically comparable items (from year to year) are administered to all students.

Administration, scoring, and reporting are conducted in a consistent and standardized manner to ensure the comparability across the province.

To ensure objectivity and consistency, all scorers are trained and monitored and use the same scoring guidance.

The purpose of the ninth-grade mathematical assessment by EQAO and reporting of the students' scores are briefly presented in this section. As mentioned before, the main purpose of the Grade 9 Mathematics Assessment of Mathematics is to measure students' knowledge and skills stated in the Ontario curriculum which are expected to be learned by the end of ninth grade of formal education. This assessment also provides information about students' achievement status based on

these expectations. The results of this assessment are reported as individual student, school, school board reports, and province report.

Individual student reports display students' success with each assessment item. School reports include school-level performance results and board reports provide overall board-level performance results and comparisons to provincial success results. Those comparisons draw attention to mathematical strengths and areas of improvements. If the number of students reported on for a school or board is too small, EQAO does not provide these results publicly to prevent the identification of individual students. Provincial reports cover the overall results of the province and results of school boards. In these reports, students' demographics, and participation information; results by sub-groups such as gender, English language learners, and special needs students are also provided. In addition to these results, instructional strategies for success and school success stories as case studies are presented.

2.2.Benefits of EQAO Assessment

EQAO claims that by means of the 9th Grade Assessment; reliable, valid, and year-to-year data on student achievement is provided to the Ontario school system. Other assessment information such as demographics, attendance, and pass rates can be confidentially used along with these data to determine the success of improvement strategies of schools and boards such as staff development or new learning resources.

In addition to the specific reporting, there are some benefits of EQAO 9th Grade Mathematics Assessment such as: to improve planning and target settings in schools and boards by use of the data; to support an outstanding implementation of the curriculum; to provide a better understanding of assessment practices and curriculum level achievement among educators across the province; and to advance the perception of assessment practices among the society.

2.3.The Link of Assessment to the Ontario Curriculum

EQAO's Grade 9 Assessment is a standards-referenced, large-scale assessment which is based on the Ontario Curriculum expectations and standards (level of achievement) for student performance. The curriculum expectations describe the knowledge and skills that students are expected to achieve, demonstrate, and apply in their classwork, test performance, and in other various activities where their achievement is assessed. The EQAO Grade 9 assessment recognizes the two different mathematics courses: Principles of Mathematics (Academic Mathematics) and Foundations of Mathematics (Applied Mathematics). For each type of course, there are different forms or versions of test frameworks assessing various learning objectives summarized in the EQAO framework (2009) for each type of course. The academic course mainly focuses on the study of mathematical theories and abstract problems. On the other hand, practical applications, and concrete examples are underlined through the study for the applied course (The Ontario Curriculum, Grades 9 and 10: Mathematics, 2005). Table 1 summarizes the strands of both courses and lists curriculum expectations under each strand in a similar manner to that stated in the Ontario curriculum.

Table 1. Strands and Curriculum Expectations in the Grade 9 Mathematics Courses

Principles of Mathematics (Academic)	Foundations of Mathematics (Applied)
Number Sense and Algebra	*Number Sense and Algebra*
• Operating with Exponents • Manipulating Expressions and Solving Equations	• Solving Problems Involving Proportional Reasoning • Simplifying Expressions and Solving Equations
Linear Relations	*Linear Relations*

Table 1. (cont'd)

• Using Data Management to Investigate Relationships • Understanding Characteristics of Linear Relations • Connecting Various Representations of Linear Relations	• Using Data Management to Investigate Relationships • Determining Characteristics of Linear Relations • Investigating Constant Rate of Change • Connecting Various Representations of Linear Relations and Solving Problems Using the Representations
Measurement and Geometry	*Measurement and Geometry*
• Investigating the Optimal Values of Measurements • Solving Problems Involving Perimeter, Area, Surface Area, and Volume • Investigating and Applying Geometric Relationships	• Investigating the Optimal Values of Measurements of Rectangles • Solving Problems Involving Perimeter, Area, and Volume • Investigating and Applying Geometric Relationships
Analytic Geometry	
• Investigating the Relationship Between the Equation of a Relation and the Shape of Its Graph • Investigating the Properties of Slope • Using the Properties of Linear Relations to Solve Problems	

Adapted from *The Ontario Curriculum, Grades 9 and 10: Mathematics* (2005). (Notes. Strand names = italic, expectations = bullets).

The link between EQAO items and curriculum expectations is ensured by an item-writing committee for each assessment. EQAO recruits 10-20 item developers who are experienced educators in the field of mathematics. They meet twice a year to write and revise test items and refers them to Ontario Curriculum expectations. The item-writing committee creates a blueprint for all test administrations and matches test specifications to the learning standards summarized in Table 1. A more detailed description of item development and association between test blueprint and Ontario's curriculum expectations is provided in EQAO's technical report (EQAO, 2017), which is summarized in brief in this section.

2.4.Academic Mathematics (Principles) and Applied Mathematics (Foundations) Process Descriptors

These mathematical process descriptors from the Ontario Curriculum are expected to be integrated into student learning related to all strands. For the duration of the mathematics course, students are believed to develop their skills and abilities in the following areas:

Problem Solving

Develop, select, apply, and compare a variety of problem-solving strategies while they solve problems and conduct investigations to enhance their mathematical understanding.

Reasoning and Proving

Develop and apply reasoning skills such as the realization of relationships, generalization through introductory reasoning, use of counterexamples, to make mathematical inferences, assess inferences, and rationalize conclusions, and plan and construct organized mathematical arguments.

Reflecting

Demonstrate that they are reflecting on and monitoring their thinking to guide for clarification of their understanding while they achieve an investigation or solve a problem. For instance, assessing the efficacy of strategies and procedures used, recommending alternative approaches, judging the rationality of results, and verifying solutions.

Selecting Tools and Computational Strategies

Select and use a diversity of actual (concrete), visual, and electronic learning instruments and suitable computational approaches to explore mathematical ideas and to solve problems.

Connecting

Make connections among mathematical concepts and processes and relate mathematical concepts to situations and phenomena brought from other contexts such as other curriculum fields, daily life, present events, art, culture, and sports.

Representing

Establish a diversity of representations of mathematical ideas such as numeric, geometric, algebraic, graphical, pictorial representations, on screen dynamic representations. Connect and contrast them and apply suitable representations to solve problems.

Communicating

Communicate mathematical thinking verbally and visually. Writing and practicing mathematical vocabulary and diversity of suitable representations and noticing mathematical conventions in mathematical writing tasks.

2.5.Alignment between Definition of Mathematics and Current Research

Current research in mathematics teaching and learning identifies that children learn more mathematics when mathematics education is established on their ways of thinking and engages them in problem solving (Yackel, 1997; Yackel & Cobb, 1996; Zack & Gravel, 2001 as cited in EQAO, 2009).

Students also advance from teacher assistance to see the connections among diverse mathematical ideas (Boaler, 2002 as cited in EQAO, 2009). Therefore, mathematical concepts are not just transferred but are the outcome of questioning, probing, making mistakes, reflecting, and reworking. This is an active process where students play an active role in aiming to make sense of their experiences. These procedures of constructing new knowledge can appear more conveniently and efficiently provided that the students are working in a rich learning setting.

The grade nine assessment provides students opportunities to display an extensive scope of mathematical processes in the content strands. This helps to raise the focus on promoting understanding by means of meaningful mathematical activity during the lesson.

2.6.Understanding Ontario's Student Achievement Levels

EQAO uses the same description of the Ontario Ministry of National Education levels of achievement that are used for the achievement levels used on its assessments. According to these descriptions.

> "Below Level 1" specifies insufficient achievement level which does not pass.
>
> "Level 1" refers a passable level of achievement. In this level achievement is considered as *below* the provincial standard.

> "Level 2" symbolizes a moderate level of achievement that is still *below but approaching* the standard.
>
> "Level 3" identifies a high level of achievement which is *at* the provincial standard.
>
> "Level 4" depicts a very high to outstanding level of achievement that is *above* the provincial standard.
>
> (Ministry of Education, Ontario, the Ontario Curriculum Grades 9 to 12, 2005).

The characteristics provided for Level 3 in the Ontario Curriculum achievement charts agree with the provincial standard for the achievement of the curriculum expectations. Parents of students who achieve Level 3 can be assured that their children perform at the provincial standard and will be ready for work in the next grade. Also, students who succeeded at Level 4 have achieved expectations beyond that described level for a specific course. It means that the student has achieved all or almost all the expectations for the mentioned course, and he or she displays the ability to use the knowledge and skills specified for that course in more experienced or sophisticated ways than a student achieving the education standards at Level 3.

After scoring all the items in a student's performance, the data from the operational items are used to decide the student's level of performance. The Individual Student Reports display both the level and the extent within the level at which the student performed. This facilitates both parents and teachers to plan for development and progress.

2.7.Validity Aspects in Literature

Even though construct validity is seen as the consensus for all types of validity in the field of educational measurement (American Psychological Association, 1999), different aspects and various terminology for test score validation exist in the validity literature.

There are two distinct classifications of validity that have been labeled as evidentiary and consequential aspects (Messick, 1989; as cited in Reckase, 1998). Messick (1989) defines validity as "an inductive summary of both the existing evidence for and potential consequences of score interpretation and use" and claims that validation substantially serves to guide both current use of the test and prevailing research to gain an understanding of the meaning of test scores. As indicated by this definition and classification, the distinctions between these two aspects are summarized in the following table.

Table 2. Aspects of Validity

	Test interpretation	*Test use*
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Value implications	Social consequences

Messick (1989, p.20, Table 2.1).

As seen from the table, construct validity is highlighted for test interpretation and test use when considering the evidential basis of validity. Both phases of validity are interconnected and overlying; for example, social consequences might be a form of evidence (Messick, 1989). One specific example from this research is: test fairness for different examinee groups is an evidence for construct validity, concurrently it is a social consequence especially in the circumstance of its lack. Reckase (1998) investigates also consequential approach of validity in detail, although that aspect would not be discussed in the scope of this research.

Evidential validity might be a highly related phenomenon to test fairness or test equity considering the wish for acquiring evidence of the test results use and score interpretation which should be equal for every test taker. While describing principles for test fairness, Kunnan (2010) states that a test should have equivalent construct validity in terms of its test score interpretation for all

examinees; specifically, it should not be biased against any examinee groups. Therefore, validity interpretation is commonly built upon the use of validity coefficients and comparative inferences across different sub-populations (Wainer et al., 1993). Evaluating the validity via test invariance across groups can be also referred as differential validity (Young & Kobrin, 2001). Since composition of test takers has become more and more diverse, differential validity achieved more significance in validation studies. Since this research specifically evaluates equivalence of test scores across different student populations, differential validity, test fairness and test equity are used interchangeably in this dissertation as key terms to represent evidentiary validation of test scores.

Based on the structure of Messick's validity approach, Kane (1992) has presented an argument-based approach to validity. The argument-based approach can be practical and/or interpretive. The former addresses issues in practical test score use, and the latter involves interpretation of test score meanings and implications. Practical arguments can be questionable and not proven being not strictly evaluated using traditional mathematics and logic. On the contrary, interpretive arguments are highly plausible seeking all available evidence from test scores to use of test score inferences and decisions. Therefore, interpretive arguments are followed by a discussion which has a significant influence on validation.

Validity inferences can be various, including theory-based or technical inferences. Most interpretations are theory based, such as construct -component- representation, and explanatory test score interpretation. Model fit to data or assumptions can be examples of technical inferences for item response models or structural equation models.

To assess validity and test fairness in this research; a factorial analysis approach for measurement invariance is used to evaluate the validity of inferences customarily made based on test score such

as students' success level to decide whether they are ready to work in the next grade level. Since a valid test needs to be fair across different groups of students, both gender and achievement level variables are taken into consideration for inferences of factorial validity which is evaluated by measurement invariance analyses. Results of the statistics can be significant evidence for component representation across achievement and gender groups. Differential item functioning analyses can serve as convincing evidence for the fit of the sub-populations data to an IRT model, and for testing possibilities of group bias.

2.8.Examples of Validity Research

In light of the research purpose and statistical approach of this study, relevant literature was reviewed to provide research examples of evidentiary validity arguments of large-scale mathematics tests in the line of factorial, correlational, and DIF based techniques.

Kupermintz et al. (1995) examined the validity and usefulness of the 8^{th} and 10^{th} grade National Educational Longitudinal Study of 1988 (NELS: 88) math tests using full-information item factor analysis and concluded that math knowledge and math reasoning were identified as two different factors for both grades. Based on regression analyses, they concluded student attitudes and program experiences related to knowledge, while gender, SES, and ethnicity differences related to reasoning. Both dimensions were found as relevant with teacher's emphasis on higher order thinking, students' home computer use, and early experience of advanced math courses. The authors also recommended that multidimensional achievement scores should be used by national educational research instead of total scores.

Kupermintz et al. (1997) also conducted a follow up study on the 12^{th} grade data with the same analytical approach that resulted in similar findings of 8^{th} and 10^{th} grade data. They also draw attention that even limited sample of student achievement data could warrant a multidimensional

approach to score construction, interpretation, validation, and use.

Gierl et al. (2005) examined the content and cognitive dimensions of the SAT exam. They found that arithmetic, geometry, and miscellaneous are distinct content areas while algebra is not. Confirmatory analyses were conducted, and four cognitive skill categories as basic math, advanced math, managing complexity, and modeling and insight were concluded as distinct skill clusters across the content areas. In addition, the authors provided a summary of previous studies for dimensionality assessment of SAT (Cook et al. 1990, Diones et al. 1996, Dorans et al. 1987).

In another study, Gierl and Khaliq (2001) evaluated data from the 1996 and 1997 administrations of a 9th grade mathematics achievement test from the province of Alberta, Canada. Authors applied DIF analyses to detect content-related gender differences and concluded that males outperformed females on items require substantial spatial processing, and females performed better on items requiring memorization, but the differences were small.

Walker and Beretvas (2001) used DIF analysis to evaluate construct validity for open-ended math items of the Washington Assessment of Student Learning (WASL) 1998 administration. The study examined DIF between proficient and non-proficient 4th and 7th grade examinees on their communication of mathematics. Although mathematical communication is accepted as a factor of math ability, occurrences of DIF in this study contributed different score interpretations between examinee groups. Based on their findings, the authors suggested that two different test scores should be reported being general mathematics ability and mathematics communication.

Differential validity studies were widely conducted for SAT scores for both verbal and mathematical areas leading to varying conclusions. For example, Wainer and Steinberg (1992) reported female test takers were performing significantly lower than males in mathematical SAT

test, similarly to Clark & Grandy (1984), Crawford et al. (1986) Hogrebe et al. (1983).

On the other hand, some other empirical studies revealed that females were better performing in mathematics in SAT assessments (Larson & Scontrino, 1976; Noble et al., 1996; Ramist et al., 1994; Rowan, 1978; Saka, 1991; Wilson,1983).

2.9.Research Significance

After reviewing EQAO provincial assessment, curriculum expectations, and equity strategy of the state, both invariance and test equity could be claimed as crucial concepts for the assessment of mathematical achievements of the 9^{th} grade students.

The Ministry of Education explains their aims to create the best publicly funded education system in the world via 1) reaching a high level of student achievement, 2) reducing the gaps in student achievement, 3) enhancing public confidence in publicly funded education, as stated in Ontario's Equity and Inclusive Education Strategy (2009).

To accomplish these priorities, equitable and inclusive education is essential, which is also perceived as vital internationally to provide a high-quality education for all learners (UNESCO, 2008).

The Equity and Inclusive Education Strategy (2009) states that Ontario's student diversity can be one of its strongest assets, and the full range of diversity must be valued and respected. Strategy also underlines that equitable and inclusive education is also crucial to construct a cohesive society and strong economy that will ensure Ontario's future prosperity.

Ontario's government is also dedicated to reducing achievement gaps across different demographic groups of students and increasing the level of student achievement for all. Some of the student groups such as recent immigrants, children from low-income families, Aboriginal students, male

students, and students with special education needs are defined as being at risk of lower achievement (Ontario's Equity and Inclusive Education Strategy, 2009). To increase outcomes for students at risk, all stakeholders must make efforts to define and eliminate barriers and must actively pursue to create the learning environments and necessary conditions required for student success. Ensuring that all students are engaged and respected in a diverse Ontario as such, students see themselves appreciated in their learning environment. Therefore, it is assured that equity and excellence are working together (Ontario's Equity and Inclusive Education Strategy, 2009).

Since lower achievers and male students might be at risk in terms of learning success as stated in the equity policy of the province, providing empirical evidence for test fairness via measurement invariance for disadvantaged groups is important. From the statistical perspective, measurement invariance indicates similarity of factor loadings, group means, and errors (Meredith, 1993). However, the concept of invariance is generally an assumption for most large-scale assessments rather than a measurement quality to be examined empirically (Viger, 2014). The present research is compelling in terms of providing empirical evidence for measurement invariance as quality of measurement for the 9th Grade Mathematical Assessment by EQAO for both gender and achievement level groups. Moreover, in the case of violation of measurement invariance, improving test quality and test fairness in the guidance of differential item statistics analysis is going to serve for a more objective and equitable assessment for students' mathematics achievement in Ontario.

Validity or test invariance can be evaluated widely by group comparison. Populations of examinees are studied for test equity on important demographics such as race, socioeconomic status, or gender. This aspect of validity has become significantly important given diversity of test takers was increased prominently within the past few decades (Young & Kobrin, 2001).

The evidentiary approach to differential validity is also an important topic in psychometrics as it relates to fair test score use and equity issues (Young & Kobrin, 2001). Sheppard (1992) provided an excellent conceptualization of invalidity as "something that distorts the meaning of test results for some groups", proofing group differences are sources of test bias. Due to the fact that fairness is a social concept, more than a mere technical investigation (Young, 2001), findings of the present research would have been considered as a consequential validation to some degree, without mentioning this aspect of validity argument specifically.

Chapter 3: Method

3.1.Sample

For this dissertation study, datasets from the 9th Grade Math Achievement from EQAO were used. There are 2 administrations of the mathematics assessments identified as “Winter” and “Spring” and each administration has 2 different program types: “Applied” and “Academic”. The data from 2015 have been investigated, and statistics were compared across test administrations. In this research, a total of 4 test administrations have been evaluated from two semesters for both applied and academic mathematics courses. Figure 1 explains the structure of the test administration datasets, which were used for data analyses in this dissertation.

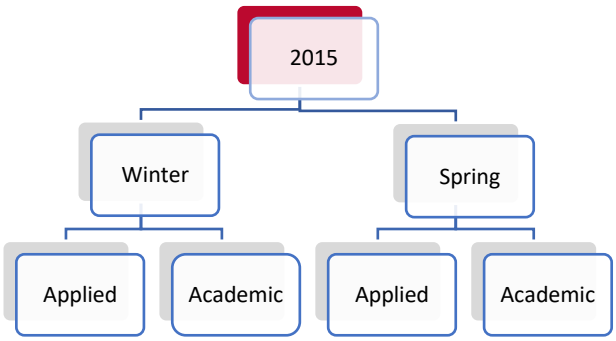

Figure 1. Datasets used in this Research.

In 2015 winter semester, 40,939 students were evaluated by the 9th Grade Academic Mathematics Assessment. In this test administration data, there is one value of -1 for gender, which does not indicate a meaningful gender demography representation; therefore, this subject has been removed from the data before statistical analyses. Also, there were only 83 (0.2%) students whose achievement level is below 1 and this low cell value created model convergence obstacles during

SEM modeling. As a result, level 1 and below level 1 group were combined into one group. After these moderate data management modifications, student group frequencies for gender and achievement levels, distributions of 40,938 students are presented in the table below.

Table 3. Gender and Achievement Level Frequencies of Winter 2015 Academic Math Assessment

	Male (%)	Female (%)	Row Total (%)
Below Level 1	54 (0)	29 (0)	83 (0)
Level 1	819 (2)	1103 (2.7)	1922 (4.7)
Level 2	2113 (5.2)	2662 (6.5)	4375 (11.7)
Level 3	14518 (35.5)	14802 (36.2)	29320 (71.7)
Level 4	2569 (6.3)	2269 (5.6)	4838 (11.9)
Column Total (%)	20073 (53)	20865 (47)	40939 (100)

There are 15,994 students in the 9th Grade Applied Mathematics Assessment data for the winter administration of the 2015 academic year. There is not any missing data on gender or achievement level variables. In this dataset, 8,995 (56%) of students are male, and 6,999 (44%) are female.

The majority of the students belongs to achievement level 3 (38%) and achievement level 2 (35%), followed by level 1(13%), level 4 (10%), and below level 1 (4%). Table 5 presents frequencies for each cell of this dataset across gender and achievement level groups.

Table 4. Gender and Achievement Level Frequencies of Winter 2015 Applied Math Assessment

	Male (%)	Female (%)	Row Total (%)
Below Level 1	318 (2)	270 (2)	588 (4)
Level 1	1124 (7)	963 (6)	2087 (13)
Level 2	3087 (19)	2550 (16)	5637 (35)
Level 3	3445 (21.6)	2574 (16)	6019 (38)
Level 4	1021 (6.4)	642 (4)	1663 (10)
Column Total (%)	8995 (56)	6999 (44)	15994 (100)

In 2015 spring academic administration, there were 45,403 ninth grade students, 48 percent of them were male, and 52 percent of them female. Among achievement groups, 82 students who were not in passable level (Below Level 1) were not included in the statistical analyses. Most of the students were at the state standard (72%), followed by students who are above the standard (13.5%), approaching the standard (10.8%), and below the standard (4%).

Table 5. Gender and Achievement Level Frequencies of Spring 2015 Academic Math Assessment

	Male (%)	Female (%)	Row Total (%)
Below Level 1	47 (0)	35 (0)	82 (0)
Level 1	856 (1.9)	939 (2.1)	1795(4)
Level 2	2210 (4.9)	2663 (5.9)	4873 (10.8)
Level 3	15928 (35)	16612 (37)	32540 (72)
Level 4	3048 (6.7)	3065 (6.8)	6113 (13.5)
Column Total (%)	22089 (48)	23314 (52)	45403 (100)

Lastly, there were 16,217 test takers, 56% were male and 44% were female, in spring 2015 applied mathematics assessment dataset. Students at the standard were 36.9% of the total population, 34% of them were approaching the standard, 11% of them were above, and 11% of them were below the learning standard defined by state curriculum. Table 6 displays frequencies for both gender and achievement level group variables for this specific test data.

Table 6. Gender and Achievement Level Frequencies of Spring 2015 Applied Math Assessment

	Male (%)	Female (%)	Row Total (%)
Below Level 1	447 (2.8)	261 (1.6)	588 (4.4)
Level 1	1031 (6.4)	862 (5.3)	2087 (11.7)
Level 2	2935 (18)	2568 (16)	5637 (34)
Level 3	3605 (22)	2739 (16.9)	6019 (36.9)
Level 4	1065 (6.7)	704 (4.3)	1663 (11)
Column Total (%)	9083 (56)	7134 (44)	16217 (100)

3.2.Instruments

In 2015, there were 2 different mathematical assessments administered by EQAO. Academic mathematics test administration consisting of 24 multiple choice and 7 open response items. In this research, multiple choice items have been evaluated in terms of evidentiary validity via differential statistics using group comparisons. Academic mathematics assessment measured 4 content areas as following: Number Sense and Algebra (5 items), Linear Relations (6 items), Analytic Geometry (7 items), Measurement and Geometry (6 items). Understanding, demonstration, problem solving, and investigation of linear relations, or geometry concepts, and applying data management techniques to all these content fields were the main cognitive skills

tested with these multiple-choice items. Like academic mathematics course test, applied course assessment was consisted of 24 multiple choice items; however, there were three content strands as Linear Relations (11 items), Number Sense and Algebra (7 items), and Measurement and Geometry (6 items). Data management techniques, connecting mathematical relationships, demonstration and application of rules, and description of relations were assessed in this specific administration similarly to academic course cognitive skills.

In the spring academic test, Linear Relations (6 items), Number Sense and Algebra (5 items), Measurement and Geometry (6 items), and Analytic Geometry (7 items) were the measured content fields. Through investigation of various tools, identifying/determining/explaining linear and non-linear relations were other cognitive tasks assessed in this test besides the above-mentioned ones. Likewise, identifying and explaining the values of geometric concepts has been among the tasks that were being measured. Similarly, three content areas were assessed on the applied mathematics test in the spring 2015 administration which are: Linear Relations (11 items), Number Sense and Algebra (7 items), and Measurement and Geometry (6 items). Understanding characteristics of linear relations, explanations of mathematical concepts, and problem solving were the examples for some measured skills in this test administration.

In addition to fundamental mathematical knowledge and skills mentioned here, all cognitive skills and procedures assessed in these four tests are explained in detail in the policy review chapter of this dissertation. Empirical evidence from these four tests was gathered using statistical methods is also discussed in the following section of this research writing.

3.3.Data Analyses

3.3.1.Reliability and Item Statistics

As a necessary condition for validity studies, reliability analyses for the four mathematical datasets were conducted beforehand. By means of "psych" and "ltm" packages in R version 3.4.1., reliability and item statistics were computed, and findings of the analyses were reported in the results chapter of this manuscript. Cronbach's alpha value for the internal consistency of test scores, and alpha values when a specific item was removed from the instrument were calculated to identify whether there were any specific items causing any decline in internal consistency of the instrument. As part of reliability analysis, item statistics such as difficulty and discrimination parameters were also reported. Evaluation of these statistics is necessary to decide whether there are any specific test items reducing test score reliability significantly or to decide whether there was a need for removing any specific item to improve the reliability of overall test scores.

3.3.2.Measurement Invariance

For evidentiary validity, group differences between gender and achievement level groups were evaluated. Multiple group confirmatory factor analysis was conducted in R using the "lavaan" package to evaluate measurement equivalence of the test scores. Measurement invariance analysis signifies that the meanings of specific latent variables, or constructs, are identical across different groups of test takers (Meredith & Teressi, 2006); that is, it tests the psychometric equivalence of a latent construct, and claims that this construct refers to the same meaning to different sample groups (Putnick & Bornstein, 2016). Hence, evaluating measurement invariance statistics would inform about test score meaning equivalence across gender and achievement groups and serve as a validity evidence for appropriate test score interpretation. For this specific purpose, structural

equation modeling approach to measurement invariance was used to evaluate score meaning equivalence. For statistical procedures of invariance analyses, Hirshchfeld and Brachel's (2014), and Timmons' (2010) guidelines were followed hand on hand. Establishing measurement invariance for a scale implies that observed mean differences across groups could have resulted from characteristic differences of the examinee groups rather than resulted by different relations between scores and underlying constructs; on the other hand, if there is no stable statistical relation found between underlying constructs and scale scores, mean score differences between groups might be caused by dissimilar relations between latent constructs and scores (Hirshchfeld & Brachel, 2014).

Structural equation modeling (SEM) computations investigate the relations between latent variables (underlying constructs) and manifest variables (observed responses), which are explicitly examined under the model. Hirshchfeld et al. (2014) provided Holzinger and Swineford's (1939) study as an example of conceptual understanding and to present a graphical representation of the measurement invariance procedure.

In their data, there were test scores of 300 students on nine different verbal ability tests, including three interrelated (or non-orthogonal) latent constructs as speed, textual, visual. In the figure below, underlying constructs are represented by ovals and observed scores are represented by rectangles, while the arrows display which item (observed scores) loads on which factor (latent construct). The reason the paths are represented by directed arrows is because of the assumption of measurement model, in which, the latent variables affect the individual items or observed scores.

When fitting this model to the data six parameters are being estimated: 1) a regression coefficient (for example, the loading of test "x1" on latent variable "textual"), 2) a regression intercept, 3) a regression residual variance, 4) the means of the latent variables, 5) the variances of the latent

variables, and 6) the covariances of the latent factors (Wu et al., 2007).

Through these fundamental estimates of measurement models, multiple group confirmatory factor analysis (measurement invariance) tests the equivalence of above-mentioned regression parameters across two or more groups of subjects.

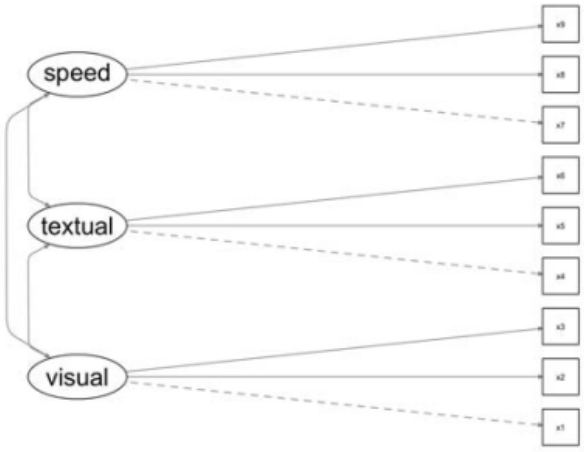

Figure 2. Measurement Model for the Holzinger and Swineford Data

(Hirschfeld & Brachel, 2014).

There are four levels of measurement invariance procedure within SEM, named as configural, weak (or metric), strong (or scalar), and strict invariance - that relates and estimates the afore-mentioned regression parameters. Configural invariance is supported when the number of latent variables and the pattern of loadings of latent variables on observed variables are the same across groups. For example, items x7, x8, and x9 are affected by the same latent variable "speed", items x4, x5, and x6 by "textual'. Weak invariance requires that the individual items and latent variables have similar factor loading values. This level of invariance is important to provide information about the relationship between items and latent variables across groups. Strong invariance requires

that in addition to item loadings, item intercept, threshold, or mean values are the same across groups based on defined SEM model. This measurement invariance level infers that there is no systematic response bias, which is a requirement to compare the latent variable mean differences across different groups (Chen, 2008). Timmons (2010) specified this stage of invariance as the equivalency of indicator means across groups. Lastly, strict invariance requires that not only the loadings and means but also the residual variances of indicators are equal across groups. However, strict invariance is not necessarily recommended because the criterion is too difficult to establish in practical test administrations (Timmons, 2010).

Once the measurement model is shown to be invariant across groups, researchers can test statistical hypotheses regarding means and relations of latent constructs. In the example above, if invariance requirements were established, a researcher could test the mean differences for textual ability across groups or whether these constructs are related to mathematics achievement across gender, grade level, or different factors, grouping students.

In this research, mathematical achievement of 9th grade high school students is the latent variable, and observed scores are obtained from 24 multiple choice mathematical items for each of the four test administrations. Stages of configural, weak, strong, and strict invariances were assessed using a confirmatory factor analytic model.

The below equation defines the steps of invariance assessment, which were also summarized in detail in Table 7. In this formula, tau (τ) indicates the intercepts or means for latent factors, and lambda (λ) describes the relationship between true and observed scores. Also, mean residuals for observed scores is zero (ϵ), which might be dropped out from the equation (Gregorich, 2006).

$$mean\ observed\ \gamma \cong \tau + (\lambda \times mean\ true\ \gamma) + \epsilon$$

Statistical evaluation of measurement invariance analyses was also summarized for each step in the table below. As seen in these factorial stages, configural invariance was evaluated to assess the similarity of loading patterns, weak invariance was used to compare the equivalence of factor loading parameter values, strong invariance was evaluated the equivalence of indicator means, and strict invariance was used to test the similarity of indicator residuals.

Table 7. Four Stage Invariance Procedure

Invariance Stage	*Definition*	*Criterion*	*Evaluation*	*Equation*
Configural	Equal patterns of factor loadings across groups	Goodness of fit statistics for overall factorial model	$\chi 2, CFI, TLI, RMSEA$	
Weak/metric	Factor loadings across groups are equivalent	CFI difference should be less than .01 compared to baseline model; RMSEA fit in one another's confidence intervals	$\Delta CFI, \Delta RMSEA$	$\lambda_{I1} = \lambda_{I2}$
Strong/scalar	Equal indicator means (or intercepts) across groups	CFI difference should be less than metric invariance with .01 cutoff point; RMSEA values within each other's CI	$\Delta CFI, \Delta RMSEA$	$\tau_{I1} = \tau_{I2}$
Strict	Indicator residual variances across groups are equal	CFI difference should be smaller than.01 compared to strong invariance; falling RMSEA values in each other's CI	$\Delta CFI, \Delta RMSEA$	$\theta_{I1} = \theta_{12}$

Each level of the four-stage measurement invariance procedure was assessed by comparison of comparative fit index (CFI), and root mean square error of approximation (RMSEA) statistics. The most-widely used criterion for the more constrained model is the CFI difference between the base and constrained models being less than .01 (Hirschfeld & Brachel, 2014; Timmons, 2010). For example, if the CFI change between configural model and weak model is less than .01, this would

indicate there is weak invariance. Likewise, the same comparison between weak and strong invariance models is made to assess strong invariance. Comparison of RMSEA values can be used as another criterion of the invariance assessment. If RMSEA values fall within each other's confidence intervals for the two levels of invariance, this indicates the invariance for the specific stage holds (Timmons, 2010). To calculate model fit statistics and evaluate four-stage measurement invariance steps, "lavaan" package in R programming software was used.

3.3.3.Differential Item Functioning

At the item level, test fairness or equity of scores across different examinee groups is assessed by differential item functioning (DIF) analysis. This statistical method provides information about whether each item assesses examinees' ability and knowledge without functioning in favor of a specific student group, given the ability level of the examinees are the same.

One of the procedures of item bias research or differential item functioning is the computation of the area between two item characteristic curves (ICC) for two sample groups (Ironson & Subkoviak, 1979; Shepard, Camilli, & Averill, I981; Shepard, Camilli, & Williams, 1984; Rudner, Geston, & Knight, 1980a, 1980b; as cited in Raju, 1988). Raju (1988) presented formulas to compute the exact area between ICCs of two sample groups for one, two, and three parameter IRT models, and provided a discussion of the significance of area measures in respect to item bias. For this specific method, signed and unsigned areas (SA and UA) between two ICCs for two groups are defined. The SA refers to *difference* between curves and the UA refers to *distance* between them. Computations for SA and UA are presented in equations 5 and 6 (Raju, 1990) by use of integral calculations for the difference and distance between two item characteristics curves for each student group. When lower asymptotes, c parameters, are equal, the ICCs must intersect at one specific theta, or ability level. On the other hand, when $c1 \neq c2$, theta would be an arbitrary

point between -∞ and +∞ and two three-parameter ICCs do not intersect, which is not the case for this research.

To sum up, Raju's (1988; 1990) item response theory based DIF method is used to evaluate item fairness across gender groups for the present research. This method computes the areas between the item response functions for each group and assesses if there is a significant difference between the two item curves (Raju, 1988; Raju, 1990). Under the item response theory, difference and distance between item characteristic curves are calculated, based on the equivalence of item parameters for two different groups for the same test item (Cohen & Kim, 1993), hence, IRT based computations would be the major advantage of choosing this method. For larger samples and longer tests, false positive error rate is also decreased, especially for signed area estimates (Cohen & Kim, 1993).

Computations of differences and distances between two student group parameters were performed using "difR' package in R statistical software. This computation provided statistics for pairwise comparisons of subpopulations to evaluate whether each test item is functioning against any specific student group.

Chapter 4: Results

4.1.Reliability and Item Statistics

In this section, item and reliability statistics are presented for 24 multiple choice items of the 9th Grade Applied Mathematics Assessment data for winter 2015 administration.

The raw data is polytomous, scored from -4 to 4. Negative values represent incorrect responses and positive values represent correct responses. Based on the answer key, the data were re-coded dichotomously as 1 for correct responses and 0 for the incorrect responses. The reason for recoding is EQAO uses binary data for multiple choice items and the 3PL model for IRT parameter estimation. Recoded data was used for IRT model fit and DIF analyses, whereas raw data was used for the measurement invariance procedure. Table 8 and Table 9 present reliability and item analyses results of four test administrations for recoded (dichotomous) data.

The internal consistency reliability measure, Cronbach's alpha, was .7919 for academic mathematics test administration scores. Also, alpha values for the test scores were calculated if the specific test item was removed from the test administration (rem.a). All these values were smaller than .7919, indicating none of the items needed to be excluded from the test to improve the internal consistency of the scores. To evaluate discrimination of each test item, item-total correlation values (raw.r.), which is a biserial correlation, were calculated along with the discrimination values when each of the items was removed from the assessment (r.drop). Items with .20 and higher correlation could be considered having adequate discrimination. For the academic assessment, discrimination values were between .30 to .53, and there was not an improvement for the discrimination of the included items when the specific test item was removed from the assessment,

because all drop.r values were smaller than the raw.r. values. Proportion correct values were between .89 and .47, indicating item 1 was the least difficult item of the assessment and 89% of the students responded that correct, and item 8 was the most difficult item with 47% of students were responding to that item correctly. Proportion of missingness was 3% or less for all items, which was below the 10% cutoff value.

For the applied mathematics assessment scores, internal consistency was .7537, which could also be considered as good reliability (above .70). Dropping Item 24 yields an alpha value of .7542, which is greater than .7537, however, the increase of the value is only .0005 which is definitely considered as a negligible improvement. Also, discrimination of this specific item was .23 which is above .20 critical point, indicating this item should not be removed from the test. Similarly, there would not be any improvement of the discrimination values, ranging between .23 and .48, when any of the items were removed from the test. Item 14 was the most difficult item with 31% of correct response percentage, while item 1 was the easiest one with 82% of the students responded it correct. Missing proportions were also negligible, ranging between .01 and .02.

Table 8. Reliability and Item Statistics for Winter 2015 Assessments

	Academic Math						Applied Math					
	rem.a	raw.r	drop.r	i.pr	c.pr	m.pr	rem.a	raw.r	drop.r	i.pr	c.pr	m.pr
MC01	0.789	0.30	0.23	0.11	0.89	0	0.7493	0.31	0.23	0.18	0.82	0.01
MC02	0.7879	0.35	0.26	0.24	0.76	0.01	0.7441	0.42	0.31	0.46	0.54	0.01
MC03	0.7844	0.42	0.33	0.48	0.52	0.02	0.7448	0.39	0.29	0.56	0.44	0.01
MC04	0.7844	0.43	0.34	0.28	0.72	0.01	0.746	0.38	0.28	0.65	0.35	0.01
MC05	0.7838	0.44	0.34	0.32	0.68	0.01	0.7493	0.34	0.24	0.58	0.42	0.01
MC06	0.7856	0.4	0.31	0.26	0.74	0.01	0.7457	0.38	0.28	0.66	0.34	0.01

Table 8. (cont'd)

MC07	0.7914	0.3	0.2	0.3	0.7	0.01	0.751	0.31	0.2	0.33	0.67	0.02
MC08	0.7816	0.47	0.38	0.53	0.47	0.03	0.7471	0.37	0.26	0.54	0.46	0.02
MC09	0.7841	0.42	0.35	0.17	0.83	0.03	0.7423	0.44	0.34	0.35	0.65	0.02
MC10	0.7857	0.4	0.31	0.23	0.77	0.03	0.7393	0.48	0.39	0.32	0.68	0.02
MC11	0.7865	0.37	0.3	0.17	0.83	0.01	0.7426	0.44	0.34	0.4	0.6	0.01
MC12	0.7885	0.37	0.26	0.48	0.52	0	0.7526	0.28	0.18	0.33	0.67	0.01
MC13	0.786	0.39	0.31	0.26	0.74	0	0.7404	0.46	0.37	0.4	0.6	0.02
MC14	0.7837	0.44	0.35	0.3	0.7	0.01	0.7433	0.42	0.32	0.69	0.31	0.01
MC15	0.7795	0.52	0.45	0.18	0.82	0.01	0.7385	0.48	0.39	0.36	0.64	0.01
MC16	0.7784	0.53	0.46	0.25	0.75	0.01	0.7484	0.34	0.24	0.67	0.33	0.02
MC17	0.7844	0.43	0.34	0.36	0.64	0.01	0.7451	0.4	0.3	0.46	0.54	0.02
MC18	0.7855	0.39	0.31	0.19	0.81	0.01	0.7451	0.39	0.3	0.29	0.71	0.01
MC19	0.7854	0.42	0.32	0.38	0.62	0.01	0.7453	0.39	0.3	0.68	0.32	0.02
MC20	0.7818	0.47	0.38	0.34	0.66	0.01	0.7471	0.36	0.27	0.22	0.78	0.01
MC21	0.7859	0.41	0.31	0.43	0.57	0.01	0.7455	0.4	0.29	0.5	0.5	0.01
MC22	0.7808	0.49	0.4	0.44	0.56	0.01	0.7477	0.36	0.26	0.47	0.53	0.01
MC23	0.7829	0.45	0.37	0.27	0.73	0.01	0.7393	0.47	0.38	0.49	0.51	0.02
MC24	0.7861	0.39	0.3	0.22	0.78	0.01	0.7542	0.23	0.13	0.21	0.79	0.02
	alpha = .7919, n = 40939						alpha = .7537, n = 15994					

Notes. rem.a = alpha when item removed from scale, raw.r = discrimination of item; drop.r = discrimination when item dropped from scale; i.pr = incorrect proportion; c.pr = correct proportion; m.pr = missing proportion.

Internal consistency for academic mathematics assessment was .8213, indicating good reliability for the spring test administration. There was no improvement on the Cronbach's alpha value when any of the items were removed from the assessment. Like the reliability statistics, there was no

increase in item discriminations if any of the test items were dropped from the test. All items have a fair amount of discrimination above the critical value of .20, to be specific between .20 and .46. The most difficult item is item 2 with a correct response percentage of 36%, while the least difficult item is item 19 with 92% correct response rate. Item missingness are all negligible being equal or less than .01.

Cronbach's alpha for the applied mathematics test is .7543, which is slightly lower than a good reliability value of .80 for large scale assessments. There are not any test items causing an increase of alpha if removed from the scale. Item discriminations range between .23 and .48, which are above the critical value of .20. The least difficult item is item 9, with 82% of students responded correct, and the most difficult one is item 14 with 30% correct response rate. Missing response proportions for items are lower than .02 which is below critical value of .10. Based on the reliability and item statistics, there is no specific item needed for removal from the spring administration of applied mathematics test.

Table 9. Reliability and Item Statistics for Spring 2015 Assessments

	Academic Math						Applied Math					
	rem.a	raw.r	drop.r	i.pr	c.pr	m.pr	rem.a	raw.r	drop.r	i.pr	c.pr	m.pr
MC01	0.8148	0.44	0.37	0.25	0.75	0	0.7497	0.34	0.24	0.33	0.67	0.01
MC02	0.8178	0.39	0.3	0.64	0.36	0.01	0.7465	0.39	0.29	0.47	0.53	0.01
MC03	0.8199	0.32	0.24	0.22	0.78	0	0.7408	0.47	0.39	0.28	0.72	0.01
MC04	0.8152	0.44	0.36	0.28	0.72	0.01	0.744	0.42	0.32	0.56	0.44	0.01
MC05	0.8129	0.49	0.41	0.33	0.67	0.01	0.7525	0.3	0.2	0.37	0.63	0.01
MC06	0.8158	0.44	0.35	0.4	0.6	0	0.7452	0.41	0.31	0.62	0.38	0.01
MC07	0.8125	0.5	0.42	0.3	0.7	0.01	0.7412	0.46	0.37	0.57	0.43	0.02

Table 9. (cont'd)

MC08	0.8153	0.43	0.36	0.21	0.79	0.01	0.7473	0.38	0.27	0.56	0.44	0.02
MC09	0.8139	0.47	0.38	0.33	0.67	0.01	0.7487	0.34	0.26	0.18	0.82	0.02
MC10	0.8161	0.41	0.34	0.16	0.84	0.01	0.7401	0.48	0.39	0.3	0.7	0.02
MC11	0.8114	0.52	0.44	0.42	0.58	0.01	0.7439	0.43	0.33	0.36	0.64	0.01
MC12	0.8112	0.52	0.44	0.43	0.57	0.01	0.7532	0.29	0.19	0.34	0.66	0.01
MC13	0.8131	0.49	0.41	0.34	0.66	0.01	0.7395	0.48	0.39	0.51	0.49	0.02
MC14	0.8152	0.43	0.36	0.24	0.76	0.01	0.7452	0.4	0.3	0.7	0.3	0.01
MC15	0.8129	0.49	0.4	0.44	0.56	0	0.7405	0.47	0.38	0.52	0.48	0.01
MC16	0.8112	0.53	0.46	0.22	0.78	0.01	0.7459	0.4	0.3	0.39	0.61	0.01
MC17	0.8155	0.44	0.35	0.33	0.67	0.01	0.7423	0.45	0.36	0.28	0.72	0.02
MC18	0.8166	0.4	0.33	0.21	0.79	0.01	0.7457	0.39	0.3	0.3	0.7	0.02
MC19	0.8192	0.31	0.25	0.08	0.92	0	0.7495	0.35	0.24	0.62	0.38	0.01
MC20	0.8152	0.44	0.36	0.32	0.68	0.01	0.7512	0.32	0.22	0.49	0.51	0.02
MC21	0.8186	0.37	0.28	0.33	0.67	0.01	0.7530	0.28	0.18	0.69	0.31	0.02
MC22	0.8131	0.49	0.4	0.42	0.58	0.01	0.7427	0.44	0.34	0.51	0.49	0.02
MC23	0.8144	0.46	0.38	0.21	0.79	0.01	0.7467	0.38	0.29	0.29	0.71	0.02
MC24	0.8162	0.41	0.34	0.19	0.81	0.01	0.7547	0.23	0.14	0.2	0.8	0.02
				alpha = .8213, n = 45403						alpha = .7543, n = 16217		

Notes. rem.a = alpha when item removed from scale, raw.r = discrimination of item; drop.r = discrimination when item dropped from scale; i.pr = incorrect proportion; c.pr = correct proportion; m.pr = missing proportion.

4.2.Measurement Invariance

Four step measurement invariance procedure described in literature and method chapters was followed for evaluating test score equivalence across achievement level and gender groups.

Maximum likelihood estimation is the most common method of SEM parameter estimation, which can only be used if the multivariate normality assumption for the data holds (Kaplan, 2001). Because ML estimators are asymptotically efficient, unbiased, and consistent, it can be used for non-normal data as well unless there is excess multivariate kurtosis (Bollen & Bauldry, 2015). For that reason, this method was used for parameter estimation for the four datasets investigated in this study.

There were 24 items in the winter 2015 test administration for academic mathematics achievement. For measurement invariance analysis, confirmatory factorial approach was followed. Latent constructs of this test are Number Sense and Algebra (5 items), Linear Relations (6 items), Analytic Geometry (7 items), Measurement and Geometry (6 items).

Students in the lowest two overall outcome level were eliminated from the sample because of model convergence obstacles. Those were students who could not reach a passable level ($n_0 = 85$) and students below provincial standard with a passable level of achievement ($n_1 = 1922$). Since number of students removed from the sample was fairly small compared to the whole sample, this subject loss could not prevent the invariance procedure and not leads to a lack of convergence. To assess construct validity, students still below but approaching the standard, in group 2 ($n_2 = 4765$); students at the provincial standard, in group 3 ($n_3 = 29321$); and students above the state standard, in group 4 ($n_4 = 4848$) were compared using measurement invariance.

Configural invariance requires that the patterns of loadings should be the same across groups. Measure of fit statistics indicates patterns of loadings were equivalent (CFI= .977, TLI = .962, RMSEA = .011 [.010, .011]). Metric invariance requires parameters of factor loadings are equal across examinee groups. To assess metric invariance, the CFI values of configural and metric invariance are compared. If the difference is less than .01, conclusion of metric invariance would

be made (Cheung & Rensvold, 2002). There is metric invariance because the CFI difference is smaller than .01 for the weak invariance. Furthermore, the RMSEA values fell in one another's confidence intervals as a second criterion for the weak invariance (CFI metric = .967, TLI metric = .949, RMSEA = .011[.011, .012]).

To assess scalar invariance, comparisons of CFI and RMSEA values were made. The CFI values of metric and scalar invariance models were the same (CFI strong = .967, TLI metric = .949). The RMSEA values were also equivalent (RMSEA metric = = .011, (.011 - .012); RMSEA strong= .011, (.011 - .012)). For the step of strict invariance, SEM model failed to provide convergence statistics. However, strict factorial invariance across groups is stated to be too hard to establish; therefore, it is usually not recommended or required in practice (Timmons, 2010). To conclude, this is not interpreted as a lack of invariance.

Group comparison for gender was also made using measurement invariance procedure for academic test administration. As the first step assessment, configural invariance analysis was conducted for female and male students to detect whether the patterns of factor loadings are similar (CFI = .987, TLI = .946, RMSEA = .024). Second, metric invariance was assessed to see whether the factor loadings are equivalent across gender groups. For that reason, the difference between CFI values should be less than .01 and RMSEA values should fall in each other's confidence interval (configural CFI = .987, metric CFI = .986, RMSEA configural = .024 [.023, .025] , RMSEA metric = .023 [.022, .024]). Because the two of the conditions were satisfied, metric invariance would stand. Similarly, comparison of CFI values was made for metric and strong invariance as well as examining RMSEA confidence intervals. Strong invariance has the same CFI and RMSEA values with metric invariance, that is, this step of analysis holds measurement equivalence. Finally, CFI values and RMSEA values were assessed for strict invariance between

gender groups (strict CFI = .981, strict RMSEA = .025 [.024, .026]). Although this step is not usually required in test practices nor mostly achievable, because of the CFI criterion holds, that would be concluded strict invariance for female and male students also exist.

Three content areas were included in applied mathematics test administration: Linear Relations (11 items), Number Sense and Algebra (7 items), Measurement and Geometry (6 items). For this specific assessment, measurement invariance analysis could not provide a solution for the achievement groups. Model convergence does not necessarily mean a measurement bias or lack of invariance. When achievement groups were combined into two levels instead of the original four, a model solution has been established. Students who are below and approaching state learning standards were combined as low-achieving group, while students at and above the standards were defined as high-achieving level. After this arrangement of grouping variable, configural invariance resulted in high goodness of fit statistics (CFI = .924, TLI = .912, RMSEA = .014 [.013 - .015]). Metric and scalar invariance stages both existed although the latter step had slightly greater CFI value (CFI metric = .870, CFI scalar = .872), their RMSEA values were the same (RMSEA= .018 [.017, .019]). Although, strict invariance is usually not required for practical testing, that stage was also provided CFI value within the critical value, decreasing less than .01 (CFI strict = .865, RMSEA = .018 [.017, .019]).

When it was conducted for gender groups, configural invariance was resulted in good fit measures (CFI= .957, TLI= .949, RMSEA= .022 [.021, .023]). For the metric invariance, CFI and RMSEA values were compared to configural invariance step. Both criteria were sufficed, RMSEA values fall in each other's confidence intervals, and the CFI values were equivalent, indicating factor loadings for both groups are similar in the model (CFI = .957, TLI = .951, RMSEA = .022 [.021, .023]). Similarly, strong invariance holds for female and male students (CFI = .957, TLI = .951,

RMSEA = .022 [.021, .023]). Finally, it would be concluded that strict invariance existed for applied math assessment because CFI difference was within the cut off value .01 (CFI = .952, TLI = .948, RMSEA = .023 [.022, .023]).

In spring 2015 academic test, there were 24 items and four content categories: Linear relations (6 items), number sense and algebra (5 items), measurement and geometry (6 items), and analytic geometry (7 items). Students who did not meet the provincial standard were removed from the sample (n_0= 82); all other four groups were included in the model.

Configural invariance indicated achievement level groups 1,2,3, and 4 have similar loading patterns (CFI = .916, TLI = .900, RMSEA = .021, [.020 - .021]); their corresponding factor loadings are the same (CFI = .902, TLI = .891, RMSEA = .022, [.021 - .022]). Strong invariance concluded that the group means are equal (CFI = .902, TLI = .891, RMSEA = .021, [.020 - .021]), while strict invariance could not have converged at all, indicating inequivalent residuals. As mentioned earlier, this stage is not a necessary step for practical measurement procedures.

First step of invariance for gender groups indicates that factor loading patters were the same (CFI = .930, TLI = .921, RMSEA = .033, [.032 - .033]). Second step indicates that metric invariance also holds for gender groups indicating equal loading values (CFI = .923, TLI = .917, RMSEA = .033, [.033 - .034]). Third step, strong invariance, concludes that means for females and males are equal (CFI = .923, TLI = .917, RMSEA = .033, [.033 - .034]). Finally, residuals were found equivalent by testing strict invariance (CFI = .918, TLI = .915, RMSEA = .034, [.033 - .034]).

As being similar to the other test administrations, applied mathematics assessment had 24 multiple choice items in spring 2015 semester, representing 3 content areas: linear relations (11 items), number sense and algebra (7 items), and measurement & geometry (6 items). Before following the

four-step invariance analysis, group 1 and group 4 had to be removed from the sample due to convergence obstacles ($n_1 = 1795$, $n_4 = 6113$).

Similarity of loading patterns across groups 2 and 3 for achievement levels was tested via configural invariance. Goodness of fit measures indicate the loading patterns were the same for achievement groups (CFI = .913, TLI = .899, RMSEA = .021, [.021 - .022]). For this measurement model, CFI difference were slightly above the criterion, that's why metric invariance could be considered as present (CFI = .891, TLI = .876, RMSEA = .024, [.023 - .024]). Strong invariance indicates the equivalent means across groups. Because differences both CFI values RMSEAs are the same, this model satisfies strong invariance step (CFI = .891, TLI = .876, RMSEA = .024, [.023 - .024]). Lastly, strict invariance step was checked to see whether corresponding residuals are equivalent. Factor model indicates this stage does not hold due to CFI value difference was larger than .01 and RMSEA value was out of confidence interval of the previous stage (CFI = .629, TLI = .592, RMSEA = .043, [.043 - .044]). As stated before, last stage of measurement invariance procedure is not required for practical validity purposes.

Model comparison for gender groups concluded that this model has a good fit to test data, and males and females have similar factor loadings (CFI = .964, TLI = .954, RMSEA = .031, [.030 - .032]). Metric invariance was also held, indicating similar factor loading measures for the two groups (CFI = .964, TLI = .956, RMSEA = .031, [.030- .031]). Strong invariance procedure stated that indicator means are equivalent between gender groups (CFI = .964, TLI = .956, RMSEA = .031, [.030- .031]), besides, equal indicator residuals as presented by the strict invariance assessment (CFI = .962, TLI = .956, RMSEA = .030, [.030- .031]).

4.3.Differential Item Functioning

For item bias analysis, Raju's differential item functioning method was used. This technique compares two groups as one focal and one reference. This method measured the exact difference (signed) and distance (unsigned) between the areas of item characteristics curves for two groups of examinees. Comparison of Cohen's d (1969) effect size statistics across pairs of groups were calculated using the z-statistics for both signed and unsigned area differences.

At this section of the study, DIF measurements across gender and achievement groups are summarized. Four groups are included in achievement level DIF measurement: 1) insufficient or not passable level, 2) approaching the standard or below standard, 3) high achievement, i.e., at the provincial standard, and 4) outstanding achievement means above the state standard. In these comparisons, relatively lower achieving groups are assigned as reference, and higher achieving groups as focal. For gender comparisons, males are reference and female students are focal groups. Findings for signed area, unsigned area, and effect size measures for significance of DIF statistics of each four test administrations are also presented in the appendix later in this dissertation.

For the Applied Mathematics winter administration for 9^{th} grade students, items 9,15 have small DIF against the lowest achievement group, items 10,18 function against students who are approaching the standard with small effect size ($0.21 \leq |d| \leq 0.44$) affirmed by signed area, while items 2,6,10,18 function in favor of group 2 , and items 7,15,16,17,20,22 function in favor of group 1 with negligible DIF by effect size statistics for unsigned area ($0.21 \leq |d| \leq 0.47$). Only item 20 has moderate DIF against group 2 based on signed area measures ($|d| = 0.65$).

Items 1,9,13,19 have small DIF against Group 1, while items 8,15 function against Group 3 ($0.21 \leq |d| \leq 0.31$). Items 10,17,18,22 display moderate DIF against Group 3 ($0.55 \leq |d| \leq 0.61$).

Unsigned area measures imply that items 1,8,18 function in favor of group 3, while items 9,17,19 function in favor of group 1, with negligible DIF ($0.21 \leq |d| \leq 0.38$). In addition, items 6,22 function in favor of group 3 with moderate DIF ($0.55 \leq |d| \leq 0.61$).

Signed area measures conclude that items 11,17 display small DIF in favor of group 4 while items 10,13,15,20 display negligible DIF against group 4 ($0.21 \leq |d| \leq 0.30$). Item 5 presents medium DIF against group 1 ($|d| = 0.50$). Unsigned area results show that items 17, 18 display function in favor of group 4, meantime, items 2,5,6,9,10,11,13,15,16,20 display small DIF against group 4 ($0.24 \leq |d| \leq 0.47$).

Items 2,12,13,20 function in favor of group 3, while items 14,16,21 function against group 3 with negligible DIF ($0.20 \leq |d| \leq 0.38$) based on signed area measures. In addition, items displaying moderate DIF are 3,9 against Group 2; items 8,10,15,17,18 against group 3 with moderate DIF ($0.52 \leq |d| \leq 0.68$), and item 22 in favor of group 2 with large DIF ($|d| = 0.99$). Unsigned area indicates that items 12,13,14,15,16,20,21 display negligible DIF against group 2; items 2,17,23 against group 3 ($0.21 \leq |d| \leq 0.49$).

Signed area measures identify items 6,17,18,23 as presenting negligible DIF in favor of group 4, and items 5,9,14,15 against group 4 ($0.20 \leq |d| \leq 0.48$). In addition, items 16 and 24 display medium DIF against group 4 ($|d| = 0.54; 0.59$). Items 11,15,17,18 have negligible DIF favoring group 4, items 2,3,5,9,10,13,14,21,23 function in favor of group 2, according to unsigned area effect sizes ($0.23 \leq |d| \leq 0.40$). Also, item 16 function in favor of group 4, while 6,12,24 function in favor of group 2 with moderate effect size ($0.54 \leq |d| \leq 0.73$).

Signed area measures reveal that items 11,17,18 and items 3,9,12,13,14,20,24 have negligible DIF, in favor of group 4 and group 3 respectively ($0.22 \leq |d| \leq 0.38$). Item 16 display large DIF against

group 4 (|d|= 0.92). Moreover, item 11,18 display negligible DIF in favor of group 4, while items 3,9,12,13,14,16,20,24 function against group 4 ($0.21 \leq |d| \leq 0.49$). Item 6 display large DIF against group 4 (|d|= 0.74).

For gender variable, it would be reasonable to consider items with large effect size as discriminating between female and male students due to significant effect size measure. Item 12 (SA); and 8,17 (SA) function in favor of male students; meanwhile items 8, 11 (SA), and 17 (both SA and US) function in favor of female students ($1.05 \leq |d| \leq 1.98$).

Tables in appendix present signed and unsigned area statistics and corresponding effect sizes for winter academic assessment scores. Z score difference between signed areas range between -6.09 and 0.019, and associated effect sizes change between -0.70 and 0.01. Items 1,6,10,16,20,22,24 perform different for achievement levels 1 and 2 with small effect size ($0.21 \leq |d| \leq 0.45$); also, item 19 functions in favor of students approaching province standard (level 2) with moderate effect size (|d| = 0.70, Z(SA) = -3.15). Unsigned z-values imply that items 1,4,11,20,21,22,23 show small or negligible DIF ($0.20 \leq |d| \leq 0.32$). Moderately significant DIF items are item 6 which functions in favor of 'approaching the standard' group (|d|= 0.56), and item 19 in favor of insufficient achievement level (|d|= 0.71). Items 15, 16, and 18 have large effect size values (|d| = 1.63, |d| = 1.08, |d| = 1.32) favoring relatively higher achievers.

When students at the standard level and insufficient level (group 3 and group 1) have been compared, items 1,3,12, and 19 ($0.23 \leq |d| \leq 0.37$) presented negligible DIF, items 4,13,17,20,21,22,23 indicated moderate DIF ($0.50 \leq |d| \leq 0.93$) , and items 6,15,16 had large DIF ($1.82 \leq |d| \leq 3.61$) based on signed area difference statistics. All these items were favoring students at the province standard. Unsigned area measures conclude that items 12, and 24 favor respectively

higher achievers (|d| = 0.38, |d| = 0.26), while items 1,3, and 19 have small or negligible DIF (0.21 ≤ |d| ≤ 0.34) favoring lower achievers.

Raju's signed area measures indicated three small DIF items as 12,13, and 22 (0.21 ≤ |d| ≤ 0.48), and three medium DIF items: 4,6,19 (0.50 ≤ |d| ≤ 0.72). These items were in favor of outstanding achievers, except item 22 was functioning in favor of underachievers. Unsigned area measures show that items 1,3 had (|d| = 0.22, |d| = 0.35) small DIF against outstanding achievers, while items 12,13 and 24 indicate small DIF (0.27 ≤ |d| ≤ 0.46) in favor of outstanding achievers. Items 4,6,21,23 display moderate DIF against outstanding achievers (0.61 ≤ |d| ≤ 0.74), item 19 functioning in favor of outstanding achievers (|d| ≤ 0.50).

Signed area effect size values affirm that items 4,20,21,22,23 display negligible DIF against students below standard according to effect size measures for signed area (0.21 ≤ |d| ≤ 0.41). Item 5 and 19 had bias (|d| = 0.32, |d| = 0.33) against higher achieving students. In addition to the small DIF presence, items 2,6,9,11,15,16,18 and 24 presented large DIF (0.93 ≤ |d| ≤ 4.88) against students approaching the standard. When unsigned area was evaluated, items 4,20,21,22,23 have small DIF in favor of students approaching the provincial standard. On the other hand, items 5,9,10,11,19 have medium DIF against lower achieving students (0.53 ≤ |d| ≤ 0.68), item 6 functions in favor of lower achieving students in a moderate level (|d|= 0.54).

Item 2 presents negligible DIF in favor of group 4, while item 11 displays negligible DIF in favor of group 2, according to signed area effect size findings (0.21 ≤ |d| ≤ 0.42). Unsigned area measure states that items 10 and 24 function in favor of group 4, while items 1,2,23 function against group 4 with negligible effect size (0.20 ≤ |d|≤ 0.42). Items 5,11,19 exhibit moderate DIF against group 2, along with items 21,22 functioning against group 4 (0.50 ≤ |d|≤ 0.68). In addition, item 6 display large DIF against group 4 (|d|= 0.85).

Signed area measures conclude that items 4,8,12,19 function in favor of outstanding achievers, while items 5,11,14,21,22,23 function in favor of high achievement group ($0.20 \leq |d| \leq 0.41$). Unsigned area measures indicate items 8,12,19 as in favor of outstanding achievers, items 3,4,5,6,11,14,21,22,23 being in favor of high achievers ($0.20 \leq |d| \leq 0.46$). To conclude, all of those DIF measures across items were negligible.

Signed area measures summarizing items 3,8,10,14,16,17,20,21,22 function in favor of female students with large effect size ($1.22 \leq |d| \leq 3.83$). Similarly, items 8,10,14,17,20,22 function in favor of females, and items 3,16,21 display DIF against female students, according to unsigned area effect sizes ($1.18 \leq |d| \leq 3.83$).

Similarly, Raju's DIF method was used with effect size measures to evaluate bias for spring applied test administration. Items 1,2,12,24 have small DIF against group 1 with small effect sizes, while item 18 and 23 function in favor of group 1 with moderate DIF ($.25 \leq |d| \leq .41$). Unsigned area concludes that items 2,10,17,22 display small DIF, items 3 and 18 display moderate DIF, and item 23 has large DIF in favor of group 2 ($.22 \leq |d| \leq .92$).

Signed area measures detect items 12,20 and 22 in favor of group 1, items 8,18, and 23 in favor of group 3 with negligible effect sizes ($.22 \leq |d| \leq .48$); unsigned area indicates items 8 and 10 as functioning for the advantage of group 1, while items 4,7,16,19,20 in benefit of group 2 ($.20 \leq |d| \leq .46$). Item 12 has moderate DIF in favor of group 1, items 18, and 22 in favor of group 2 ($.55 \leq |d| \leq .77$). Also, item 3 performs in favor of group 1 with large DIF ($|d| = 1.17$).

Signed area measures suggest that items 12 and 24 have small DIF against group 1, while items 2,8,13, and 18 function in favor of group 1 ($.22 \leq |d| \leq .30$). Besides, item 19 has moderate DIF against group 4 ($|d| = .65$). Unsigned area implies items 2,3,4,8,10,12,13,15,24 function in favor

of group 4, while item 24 functions against it ($.20 \leq |d| \leq .40$). Items 16,18,19,22 have medium to large DIF against group 4 ($.65 \leq |d| \leq .84$).

Based on signed area evaluation, items 2 and 24 function in favor of group 3, while items 5,11,14 function in favor of group 2 with small effect sizes ($.21 \leq |d| \leq .47$). Item 12 has moderate ($|d|= .63$), and 22 has large DIF ($|d| = .83$) against group 2. Moreover, items 14 and 24 have small DIF against group 2, while items 11 and 18 display moderate bias in favor of group 2. Also, items 3,6,8,15 have large bias against group 2, according to unsigned effect sizes ($.22 \leq |d| \leq .39$). Items 2 and 16 function in favor of group 3, item 14 functions against group 3 with negligible DIF. Item 12 had moderate DIF, and items 3,6,8,15 had large DIF in favor of group 3; while items 2,6 function against group 3 with negligible bias, items 10,11,18,22,23 with moderate to large DIF ($.22 \leq |d| \leq 1.08$).

Signed area measures imply that items 5,15,24 display small, and item 12 displays medium DIF in favor of group 4 ($.20 \leq |d| \leq .66$). According to unsigned area effect sizes, items 6,15,24 function in favor of group 4, while items 2,5,17 function against group 4 with negligible DIF ($.21 \leq |d| \leq .45$). Items 8,12 function in favor of group 4, while 18,21 against group 4 with large DIF ($.87 \leq |d| \leq 1.09$).

Signed areas indicate that items 4,6,10,15,18 have small DIF against group 4, while item 5 displays moderate, and item 8 displays large effect sizes in terms of DIF existence ($.21 \leq |d| \leq .83$). Items 2,13,22 function in favor of group 4 with moderate, and item 19 with large DIF ($.26 \leq |d| \leq 1.0$). Unsigned measures show that items 6,10,15,18,21 have negligible, and items 2,4,5,19,22 display moderate DIF in favor of group 3 ($.22 \leq |d| \leq .79$). Items 12,13 have small and item 8 has large DIF against group 3 ($.20 \leq |d| \leq .75$).

Items 14,16,18,21 function in favor of male students, and items 3,9,24 in favor of females with small DIF ($.20 \leq |d| \leq .48$). Items 4,5,10,17,20,23 have moderate DIF ($.52 \leq |d| \leq .69$), signed areas suggested. Items 2,4,10,11,15,17,19,20,22,23 function in favor of males, and items 6,8,12 display medium to large DIF against females ($.54 \leq |d| \leq 1.9$). Unsigned area measures show that item 14 has small DIF, items 3,4,10,13,17,23 with moderate DIF, and item 8 with large DIF against females ($.20 \leq |d| \leq .90$). Items 1,7,9,16,18,20,21,24 function against females with negligible DIF, and items 5,6 display medium bias ($|d| = .63$, $|d| = .64$). Item 8 has large DIF in favor of females, while items 2,11,15,19,22 function against female students ($.90 \leq |d| \leq 1.78$).

Spring academic test items were evaluated in terms of achievement group bias. Items 3,10,15,19,23 function in favor of group 1, and item 5 functions in favor of group 2 with negligible DIF ($.23 \leq |d| \leq .42$).

Item 24 functions in favor of group 1, while items 7 and 18 function against group 1 with moderate effect sizes ($.57 \leq |d| \leq .74$). Lastly, items 8,14,16,20,22 have large DIF against group 1 according to signed area measures ($.92 \leq |d| \leq 1.54$). Unsigned area statistics showed that items 5,15,18,19 and 23 have small DIF ($.24 \leq |d| \leq .43$), items 7,and 10 with medium DIF ($.66 \leq |d| \leq .78$), and items 3,8,14,16,20,22, and 24 with large DIF ($.91 \leq |d| \leq 1.54$).

Items 1,2,4,6,7,9,11,13 display negligible DIF ($.22 \leq |d| \leq .38$), item 15 moderate DIF ($|d| = .70$), and items 3,8,12,14,16,17,18,20,and 22 display large DIF based on signed areas ($.80 \leq |d| \leq 2.52$). Besides, unsigned area measures show that items 1,2,4,6,7,9,10,11,13 perform with small ($.22 \leq |d| \leq .39$), items 18 and 24 with medium ($.59 \leq |d| \leq .73$), and items 3,8,12,14,15,16,17,20 and 22 with large DIF ($.85 \leq |d| \leq 2.64$) similar to the unsigned measures.

Signed area effect sizes show items 3,7,11,12,14,15,16,22,24 have negligible ($.23 \leq |d| \leq .45$), and

item 4 has medium DIF (|d| = .50); while unsigned area presents items 3,5,7,11,12,16,19,22,24 with small (.26 ≤ |d| ≤ .42), and items 4,9,14, and 15 with moderate DIF (.50 ≤ |d| ≤ .78) in favor of group 3.

Items from 4 to 7, 12 to 15, 18 to 20, and item 22 display small DIF (.20 ≤ |d| ≤ .48), item 8 and 10 display medium (.54 ≤ |d| ≤ .61), and item 3,23,24 have large DIF (1.36 ≤ |d| ≤ 3.19), according to signed area effect sizes against group 2. Unsigned area shows items 4 through 6, and items 10,12,13,15,19 function with small, items 7,23 with moderate, and 3,8,14,16,20,21,22,24 with large DIF effect sizes (.82 ≤ |d| ≤ 2.14).

Items with negligible DIF are 3,5,7,16,21,22 functioning in favor of group 2, and items 4,12 and 24 functioning in favor of group 4 (.20 ≤ |d| ≤ .45), according to signed area statistics. Besides, item 14 has moderate DIF in favor of group 2 (|d| = .65). Unsigned area states items 3,4,15,22 have small DIF against group 4, items 5,6,12,16,24 have small DIF in favor of group 4 (.22 ≤ |d| ≤ .48), and item 14 displaying moderate DIF in favor of group 4 (|d| = .65).

When signed areas were reviewed, there were 7 items found with negligible amount of DIF; items 5,7,11,16,22 function in favor of group 3, while items 4 and 24 function against group 3 (.20 ≤ |d| ≤ .39), items with medium DIF are 9,14,15 in favor of group 3 (.52 ≤ |d| ≤ .56). Unsigned areas detected the same items with the signed area measures functioning in favor of and against group 3 with small DIF (.20 ≤ |d| ≤ .42), and moderate DIF as well (.55 ≤ |d| ≤ .67).

Gender comparisons to detect DIF revealed that items 5,7,11,14,16,22 function in favor of male students, and items 4 and 24 function in favor of female students with negligible DIF (.20≤ |d| ≤ .46). Items 9 and 15 display moderate DIF against female students (|d| = .51, |d|= .53) based on signed area measures. Unsigned area effect statistics pinpointed exact similar items with small DIF

(.20≤ |d| ≤ .48) and medium biases (|d| = .54, |d|= .58) with the same DIF direction.

The below tables summarize DIF statistics results for four of the mathematics assessments. Table 10 shows detected DIF items for achievement level groups. Although all DIF items were reported in this chapter, only large and moderate DIF items are presented in the below tables.

Table 10. Items with Significant DIF for Achievement Groups

	Winter Applied	Winter Academic	Spring Applied	Spring Academic
Number Sense and Algebra	9(F),12,18,2,15, 5,8,15	4(RF),6(F),21	3(RF),8(RF), 18(RF), 22(RF)	3(RF),9(RF), 14(RF)
Linear Relations	3(F),7,10,16(RF), 22(RF),11,	11,18(F),23	19(FR),23(RF)	8(RF),20(RF)
Measurement and Geometry	20,17,6(RF),24	9,19(F),20(RF)		7(F),22(RF)
Analytic Geometry		13(RF), 22(RF), 2(F),15(F),16(F), 24(F)		15(RF),24(F)

Notes. F = Advantageous for focal group. RF = Functions in favor of reference or focal groups for different area measures.

Table 11 presents DIF occurrences for gender groups. Male students at gender comparisons, and respectively lower achieving students at performance comparisons were assigned as reference groups. When the reference group was not advantageous for any specific DIF item, this item was labeled as F representing “Focal group”. If one item displays DIF without a specific pattern and functions against both reference and focal groups at pairwise comparisons, RF was noted in parenthesis, meaning that the item displays DIF for reference and focal groups at different times of comparisons.

Table 11. Items with Significant DIF for Gender Groups

	Winter Applied	Winter Academic	Spring Applied	Spring Academic
Number Sense and Algebra	2(RF),8(RF),12, 15	10,21(RF)	2(RF),22(RF) 8(F),12(F),18(F)	-
Linear Relations	7(RF),11(RF), 16,21,22	8,14	11(RF)19(RF)	9
Measurement and Geometry	4(RF),17,20	17,20	15(RF)	-
Analytic Geometry		22, 3(RF),16(RF), 24(RF)		15

Notes. F = Advantageous for focal group. RF = Functions in favor of reference or focal groups for different area measures.

Chapter 5: Conclusions and Discussion

In this research study, the evidentiary validity of EQAO's mathematical assessments for four different test administrations was evaluated. These tests are winter academic mathematics assessment, winter applied mathematics assessment, spring academic mathematics assessment, and spring applied mathematics assessment administered in 2015. As one of the common methods to gather validity evidence, a comparison of group differences was made using measurement invariance and differential item functioning statistics.

Prior to these statistical analyses, reliability, and item analyses of the four datasets were computed. Internal consistency, item difficulty, item discrimination, item missing proportions were interpreted to decide whether there was a necessary item to be removed from the statistics to improve reliability or item statistics measures. In conclusion, four test administration have fair amount of internal consistency as presented in the tables in the results chapter of the study. Another conclusion stating that there was no need for item deletion from any of the four datasets to improve test reliability, based on internal consistency and discrimination values were not improved when any of the test items were removed.

For differential validity conclusions, four stage measurement invariance procedure were followed to gather evidence from four datasets of academic and applied test administrations in winter and spring semesters. For winter 2015 academic test scores, configural, weak, and strong equivalence steps of measurement invariance were satisfied for achievement groups of students based on goodness of fit statistics evaluation and model comparisons using CFI differences. Strict invariance could not be established for this dataset; however, it was not a necessary or

recommended step for most practical test purposes. To conclude, the confirmatory factor structure of the winter academic test measured similar constructs for achievement level groups, these are “Number Sense and Algebra”, “Linear Relations”, “Analytic Geometry”, and “Measurement and Geometry”. Configural invariance indicates similar factor loading patterns, weak invariance indicates equivalent factor loadings, and strong invariance stage indicates equivalent indicator means for achievement groups. These results suggest that the formerly mentioned four mathematical constructs were understood and responded similarly by nine grade students regardless of their differences in achievement levels. Test score meaning for these four latent constructs can be similarly interpreted for all achievement groups serving as a differential validity evidence. Similarly, measurement invariance held for gender groups for this specific test administration. Although being not a required stage, strict invariance also existed for gender groups for this dataset, indicating equivalent indicator residuals. To sum up, four mathematical constructs assessed by winter academic test were similar across both achievement and gender groups of the test population.

For winter applied test administration, three latent constructs such as “Linear Relations”, “Number Sense and Algebra”, and “Measurement and Geometry” were evaluated with multiple choice test items. However, the measurement invariance model could not provide a statistical solution for the factorial model when four of the achievement groups were included in the analysis. The failure of this model to produce parameter estimates is one of the limitations of the present research study. Although there is a lack of validity evidence for the interpretation of the test results for these groups, the researcher does not claim that these four mathematical constructs are not appropriately measured by this specific test administration. In fact, this obstacle had been overcome when the four achievement groups were combined and reduced into two groups as lower and higher

achieving students. Measurement invariance analysis provided parameter estimates with the same structural model when it was re-analyzed with two achievement groups instead four different groups. This model simplification provided a solution for lack of parameter estimation and enabled keeping the factorial structure of the mathematical constructs as the same. In addition, that might indicate that these three constructs were closely related and using a structural equation model such as a combination of these constructs into one factor, removal of problematic items after a content validity analysis, or removal of one of the specific latent constructs might also provide a solution for this test administration dataset. These options would be revisited as future research with different factorial models as well in order to provide group comparison statistics for the original four level achievements, although that might cause a modification on the factorial structure of the latent constructs. Measurement invariance analyses for gender groups indicate configural, weak, strong, and strict invariances held, meaning these test scores of three mathematical constructs were understood and interpreted similarly by male and female students.

Similarly, the spring academic assessment measured four constructs as “Linear Relations”, “Number Sense and Algebra”, “Analytic Geometry”, and “Measurement and Geometry”. Based on goodness of fit statistics, configural, weak, and strong invariance steps of measurement invariance held for achievement level groups. The latest stage of the invariance analysis was not satisfied, indicating a lack of strict invariance; however, this stage is not required for practical applications because it is too hard to obtain. Gender group comparisons were also made using four stage measurement analysis procedures. Goodness of fit statistics and model comparisons via CFI differences refers that all four levels of measurement invariance held for male and female students. In conclusion, the spring academic mathematics test measures the same constructs as “Linear Relations”, “Number Sense and Algebra”, “Analytic Geometry”, and “Measurement and

Geometry" for both achievement and gender groups.

In the spring applied test, students' learning for three mathematical constructs as "Linear Relations", "Number Sense and Algebra", and "Measurement and Geometry" were evaluated. Identical to previous statistical analyses, strict invariance stage did not hold for this test administration data for achievement level groups of students. All required levels of measurement equivalence, which are configural, weak, and strong invariance procedures were supported. For gender groups, four levels of measurement invariance procedure held based on goodness of fit statistics inferences. In conclusion, test score interpretation for gender and achievement groups of these three mathematical constructs were found equivalent. Factor structure of the applied test administration concludes that loading patterns, loading measures, and indicator means for gender and achievement groups of students are identical, which is considered as empirical evidence for equivalent construct meaning and interpretation of this specific test administration data.

Findings of measurement invariance evaluation concludes that both applied and academic mathematic assessment scores have evidentiary validity in terms of four mathematical constructs measured by the EQAO test administrations. Although there are some model convergence obstacles for winter applied dataset across achievement level groups, after combining four groups into two for this dataset, all performance groups and gender groups yielded meaningful statistics for validation. For most of the analyses, strict invariance stage held even though that was not required for practical measurement cases. Therefore, test score interpretations for "Linear Relations", "Number Sense and Algebra", "Measurement and Geometry", and "Analytical Geometry" can be concluded as equal across achievement and gender groups. Both academic and applied tests measure those four mathematical abilities in an unambiguous way and those constructs represent similar or equivalent meanings across different examinee groups. Based on

these results and inferences, any future group score differences can be claimed as a real ability difference across population groups but not a test fairness issue caused by the measurement instrument. That would be considered as a strong validity evidence in terms of differential test equity.

For the convergence obstacles occurred for administrations, a content review might be helpful to identify problematic and less functioning test items. When these items are revisited and revised, there might be an improvement of model convergence statistics in the future measurement invariance procedures. Moreover, the removal of highly problematic items based on item statistics and content analysis might provide partial invariance evidence for further model construction steps, as well. However, this solution was not able to be applied in this current study because the researcher was not able to review actual test items.

One conclusion to be derived from item level test fairness evaluation is that all of four measured constructs mostly function in favor of higher achieving groups of students, despite of a few items resulting in conflicting DIF measures among the four test administrations. This conclusion can be claimed as an expected conclusion because it is hard to claim that there are not real group differences across achievement level groups. Since achievement level is not a mere demographical characteristic of students not effecting their mathematic performance, readers can discuss that test scores display true mean success differences across achievement level groups rather than being biased measurement instruments.

In terms of gender comparisons, applied tests produced more equal results than academic test administrations. In academic tests, female students were better performing in the field of number sense and algebra (N), and male students were gaining more success on analytic geometry (G) construct of assessment. Existing literature discusses widely that algebra skills require more

memorization and comprehension, that is related to verbal academic knowledge to an extended degree, while analytic geometry requires creative and analytical thinking skills as well as problem solving abilities and inquiry. These skills are considered as higher order level cognitive skills and require higher affect and motivational interests. If actual test items had been reviewed, suggestions to improve females' geometry skills and males' verbal and arithmetic skills could have been provided. However, without having these content analyses, any in-class and extracurricular suggestions could not have been discussed in this research. That is why this study was helpful to detect the achievement gaps between gender groups to some extent, but there would be no further suggestions or conclusions based on research findings within the scope of the research purposes for both algebra and geometry academic skills. That would be a future research inquiry for recommendations about curriculum or learning and teaching methods if the original test items would have been reviewed and analyzed in terms of curriculum standards.

One limitation for DIF analyses was Raju's method compares two groups at one time. This is not a drawback for gender group parameters, although being a disadvantage for achievement group statistics because there are four achievement level groups. Pairwise comparisons were made in this dissertation research, a possible concern regarding some of the DIF items is that they might be significant by chance due to multiple and non-independent group comparisons. For that reason, another DIF method such as Lord's chi-square comparison can be used to enable comparing these four groups at the same time and the results of these two methods can be compared as a future research. Cohen & Kim (1993) conducted a simulation study to evaluate the efficiency of Lord's chi-square method and Raju's signed and unsigned area statistics, they concluded that chi-square results would be more reliable over signed area estimates for small sample sizes, short tests and large DIF percentages for overall tests. However, for large sample sizes, longer tests, and when

DIF percent is not too large (less than 10%), either area measures or chi-square statistics would be equally efficient. This study might guide comparison analyses as a future study for the test equivalence and evidentiary validity assessments of the EQAO ninth grade mathematical constructs.

APPENDIX

Table A1. Winter 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 2 (Reference = Group 1, Focal = Group 2)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.282	-0.06 | [-0.48, 0.36]	-0.271	-0.06 | [-0.48, 0.36]
MC02r	0.585	0.12 | [-0.29, 0.54]	-0.993	-0.21 | [-0.63, 0.21]
MC03r	0.695	0.15 | [-0.27, 0.57]	-0.719	-0.15 | [-0.57, 0.26]
MC04r	-0.655	-0.14 | [-0.56, 0.28]	0.663	0.14 | [-0.28, 0.56]
MC05r	0.087	0.02 | [-0.40, 0.44]	0.486	0.10 | [-0.31, 0.52]
MC06r	-0.041	-0.01 | [-0.43, 0.41]	-1.583	-0.34 | [-0.76, 0.08]
MC07r	-0.570	-0.12 | [-0.54, 0.30]	2.306	0.49 | [0.07, 0.91]
MC08r	0.668	0.14 | [-0.28, 0.56]	-0.607	-0.13 | [-0.55, 0.29]
MC09r	-0.962	-0.21 | [-0.62, 0.21]	0.826	0.18 | [-0.24, 0.59]
MC10r	1.911	0.41 | [-0.01, 0.83]	-1.798	-0.38 | [-0.80, 0.03]
MC11r	-0.668	-0.14 | [-0.56, 0.28]	0.795	0.17 | [-0.25, 0.59]
MC12r	-0.279	-0.06 | [-0.48, 0.36]	0.239	0.05 | [-0.37, 0.47]
MC13r	0.463	0.10 | [-0.32, 0.52]	-0.881	-0.19 | [-0.61, 0.23]
MC14r	-0.004	0.00 | [-0.42, 0.42]	0.004	0.00 | [-0.42, 0.42]
MC15r	-2.065	-0.44 | [-0.86, -0.02]	1.308	0.28 | [-0.14, 0.70]
MC16r	-0.546	-0.12 | [-0.53, 0.30]	0.978	0.21 | [-0.21, 0.63]
MC17r	0.553	0.12 | [-0.30, 0.54]	1.591	0.34 | [-0.08, 0.76]
MC18r	1.690	0.36 | [-0.06, 0.78]	-1.043	-0.22 | [-0.64, 0.20]
MC19r	0.396	0.08 | [-0.33, 0.50]	-0.405	-0.09 | [-0.50, 0.33]
MC20r	3.024	0.65 | [0.23, 1.06]	2.204	0.47 | [0.05, 0.89]
MC21r	0.256	0.05 | [-0.36, 0.47]	-0.235	-0.05 | [-0.47, 0.37]
MC22r	0.572	0.12 | [-0.30, 0.54]	1.217	0.26 | [-0.16, 0.68]
MC23r	-0.123	-0.03 | [-0.44, 0.39]	0.118	0.03 | [-0.39, 0.44]
MC24r	0.216	0.05 | [-0.37, 0.46]	0.313	0.07 | [-0.35, 0.48]

Table A2. Winter 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 3 (Reference = Group 1, Focal = Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-1.046	-0.22 | [-0.63, 0.19]	-0.989	-0.21 | [-0.62, 0.20]
MC02r	-0.577	-0.12 | [-0.53, 0.29]	3.595	0.76 | [0.34, 1.17]
MC03r	0.499	0.11 | [-0.31, 0.52]	-0.461	-0.10 | [-0.51, 0.32]
MC04r	-0.485	-0.10 | [-0.52, 0.31]	0.562	0.12 | [-0.29, 0.53]
MC05r	0.687	0.14 | [-0.27, 0.56]	-0.512	-0.11 | [-0.52, 0.31]
MC06r	-0.241	-0.05 | [-0.46, 0.36]	-2.604	-0.55 | [-0.96, -0.14]
MC07r	0.425	0.09 | [-0.32, 0.50]	0.538	0.11 | [-0.30, 0.53]
MC08r	1.050	0.22 | [-0.19, 0.63]	-1.039	-0.22 | [-0.63, 0.19]
MC09r	-1.493	-0.31 | [-0.73, 0.10]	1.459	0.31 | [-0.11, 0.72]
MC10r	2.741	0.58 | [0.16, 0.99]	2.737	0.58 | [0.16, 0.99]
MC11r	-0.422	-0.09 | [-0.50, 0.32]	0.801	0.17 | [-0.24, 0.58]
MC12r	-0.580	-0.12 | [-0.54, 0.29]	0.201	0.04 | [-0.37, 0.46]
MC13r	-1.020	-0.21 | [-0.63, 0.20]	-0.840	-0.18 | [-0.59, 0.24]
MC14r	-0.004	0.00 | [-0.41, 0.41]	0.004	0.00 | [-0.41, 0.41]
MC15r	1.488	0.31 | [-0.10, 0.73]	2.726	0.57 | [0.16, 0.99]
MC16r	0.049	0.01 | [-0.40, 0.42]	5.0667	1.07 | [0.65, 1.48]
MC17r	2.659	0.56 | [0.15, 0.97]	1.507	0.32 | [-0.10, 0.73]
MC18r	2.587	0.55 | [0.13, 0.96]	-1.784	-0.38 | [-0.79, 0.04]
MC19r	-1.291	-0.27 | [-0.69, 0.14]	1.197	0.25 | [-0.16, 0.67]
MC20r	-0.754	-0.16 | [-0.57, 0.25]	-0.784	-0.17 | [-0.58, 0.25]
MC21r	0.338	0.07 | [-0.34, 0.48]	-0.308	-0.06 | [-0.48, 0.35]
MC22r	2.887	0.61 | [0.20, 1.02]	-2.887	-0.61 | [-1.02, -0.20]
MC23r	-0.155	-0.03 | [-0.45, 0.38]	0.162	0.03 | [-0.38, 0.45]
MC24r	-0.169	-0.04 | [-0.45, 0.38]	0.236	0.05 | [-0.36, 0.46]

Table A3. Winter 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 4 (Reference = Group 1, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.202	-0.05 | [-0.55, 0.45]	-0.200	-0.05 | [-0.55, 0.45]
MC02r	0.715	0.18 | [-0.32, 0.68]	0.946	0.24 | [-0.26, 0.74]
MC03r	0.743	0.19 | [-0.31, 0.69]	-0.727	-0.19 | [-0.69, 0.32]
MC04r	-0.485	-0.12 | [-0.62, 0.38]	-0.485	-0.12 | [-0.62, 0.38]
MC05r	1.959	0.50 | [0.00, 1.00]	1.740	0.44 | [-0.06, 0.95]
MC06r	-0.448	-0.11 | [-0.62, 0.39]	1.834	0.47 | [-0.03, 0.97]
MC07r	-0.245	-0.06 | [-0.56, 0.44]	-0.223	-0.06 | [-0.56, 0.44]
MC08r	0.329	0.08 | [-0.42, 0.59]	0.264	0.07 | [-0.43, 0.57]
MC09r	0.283	0.07 | [-0.43, 0.57]	1.487	0.38 | [-0.12, 0.88]
MC10r	1.067	0.27 | [-0.23, 0.77]	1.067	0.27 | [-0.23, 0.77]
MC11r	-0.985	-0.25 | [-0.75, 0.25]	0.973	0.25 | [-0.25, 0.75]
MC12r	-0.189	-0.05 | [-0.55, 0.45]	0.256	0.07 | [-0.44, 0.57]
MC13r	0.834	0.21 | [-0.29, 0.71]	1.543	0.39 | [-0.11, 0.90]
MC14r	-0.003	0.00 | [-0.50, 0.50]	0.004	0.00 | [-0.50, 0.50]
MC15r	0.938	0.24 | [-0.26, 0.74]	1.090	0.28 | [-0.22, 0.78]
MC16r	0.491	0.13 | [-0.38, 0.63]	1.357	0.35 | [-0.15, 0.85]
MC17r	-0.892	-0.23 | [-0.73, 0.27]	-1.585	-0.41 | [-0.91, 0.10]
MC18r	-0.656	-0.17 | [-0.67, 0.33]	-1.211	-0.31 | [-0.81, 0.19]
MC19r	0.134	0.03 | [-0.47, 0.54]	0.144	0.04 | [-0.46, 0.54]
MC20r	1.175	0.30 | [-0.20, 0.80]	1.170	0.30 | [-0.20, 0.80]
MC21r	0.514	0.13 | [-0.37, 0.63]	-0.507	-0.13 | [-0.63, 0.37]
MC22r	0.363	0.09 | [-0.41, 0.59]	0.331	0.08 | [-0.42, 0.59]
MC23r	-0.202	-0.05 | [-0.55, 0.45]	0.195	0.05 | [-0.45, 0.55]
MC24r	0.669	0.17 | [-0.33, 0.67]	0.535	0.14 | [-0.36, 0.64]

Table A4. Winter 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 3 (Reference = Group 2, Focal = Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	0.180	0.05 | [-0.45, 0.55]	0.214	0.05 | [-0.45, 0.56]
MC02r	-0.797	-0.20 | [-0.70, 0.30]	1.899	0.49 | [-0.02, 0.99]
MC03r	-2.059	-0.53 | [-1.03, -0.03]	2.706	0.69 | [0.19, 1.19]
MC04r	0.449	0.11 | [-0.39, 0.62]	-0.442	-0.11 | [-0.61, 0.39]
MC05r	0.674	0.17 | [-0.33, 0.67]	-0.524	-0.13 | [-0.63, 0.37]
MC06r	-0.586	-0.15 | [-0.65, 0.35]	-1.802	-0.46 | [-0.96, 0.04]
MC07r	0.444	0.11 | [-0.39, 0.61]	0.411	0.11 | [-0.40, 0.61]
MC08r	2.466	0.63 | [0.13, 1.13]	2.699	0.69 | [0.19, 1.19]
MC09r	-2.726	-0.70 | [-1.20, -0.20]	-2.228	-0.57 | [-1.07, -0.07]
MC10r	2.047	0.52 | [0.02, 1.02]	2.071	0.53 | [0.03, 1.03]
MC11r	3.129	0.80 | [0.30, 1.30]	3.059	0.78 | [0.28, 1.28]
MC12r	-1.460	-0.37 | [-0.87, 0.13]	-1.247	-0.32 | [-0.82, 0.18]
MC13r	-1.116	-0.29 | [-0.79, 0.22]	-1.021	-0.26 | [-0.76, 0.24]
MC14r	0.812	0.21 | [-0.29, 0.71]	-0.812	-0.21 | [-0.71, 0.29]
MC15r	2.665	0.68 | [0.18, 1.18]	-1.060	-0.27 | [-0.77, 0.23]
MC16r	1.473	0.38 | [-0.12, 0.88]	-1.398	-0.36 | [-0.86, 0.14]
MC17r	2.632	0.67 | [0.17, 1.17]	1.241	0.32 | [-0.18, 0.82]
MC18r	2.672	0.68 | [0.18, 1.18]	2.672	0.68 | [0.18, 1.18]
MC19r	-0.636	-0.16 | [-0.66, 0.34]	0.720	0.18 | [-0.32, 0.68]
MC20r	-0.830	-0.21 | [-0.71, 0.29]	-0.874	-0.22 | [-0.72, 0.28]
MC21r	1.744	0.45 | [-0.06, 0.95]	-1.250	-0.32 | [-0.82, 0.18]
MC22r	3.875	0.99 | [0.49, 1.49]	-2.522	-0.64 | [-1.15, -0.14]
MC23r	-0.703	-0.18 | [-0.68, 0.32]	0.902	0.23 | [-0.27, 0.73]
MC24r	-0.698	-0.18 | [-0.68, 0.32]	-0.627	-0.16 | [-0.66, 0.34]

Table A5. Winter 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 4 (Reference = Group 2, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.037	-0.01 | [-0.43, 0.42]	0.217	0.05 | [-0.38, 0.47]
MC02r	0.515	0.11 | [-0.31, 0.54]	1.454	0.31 | [-0.11, 0.74]
MC03r	0.276	0.06 | [-0.36, 0.48]	2.115	0.46 | [0.03, 0.88]
MC04r	0.445	0.10 | [-0.33, 0.52]	-0.762	-0.16 | [-0.59, 0.26]
MC05r	2.226	0.48 | [0.06, 0.91]	1.928	0.42 | [-0.01, 0.84]
MC06r	-1.267	-0.27 | [-0.70, 0.15]	3.360	0.73 | [0.30, 1.15]
MC07r	-0.240	-0.05 | [-0.48, 0.37]	-0.254	-0.05 | [-0.48, 0.37]
MC08r	0.289	0.06 | [-0.36, 0.49]	0.254	0.05 | [-0.37, 0.48]
MC09r	1.123	0.24 | [-0.18, 0.67]	1.280	0.28 | [-0.15, 0.70]
MC10r	0.835	0.18 | [-0.24, 0.60]	1.393	0.30 | [-0.12, 0.73]
MC11r	-0.896	-0.19 | [-0.62, 0.23]	-1.417	-0.31 | [-0.73, 0.12]
MC12r	0.701	0.15 | [-0.27, 0.58]	2.500	0.54 | [0.12, 0.97]
MC13r	0.342	0.07 | [-0.35, 0.50]	1.441	0.31 | [-0.11, 0.74]
MC14r	1.571	0.34 | [-0.08, 0.76]	1.422	0.31 | [-0.12, 0.73]
MC15r	1.373	0.30 | [-0.13, 0.72]	-1.555	-0.34 | [-0.76, 0.09]
MC16r	2.478	0.54 | [0.11, 0.96]	-2.478	-0.54 | [-0.96, -0.11]
MC17r	-1.037	-0.22 | [-0.65, 0.20]	-1.869	-0.40 | [-0.83, 0.02]
MC18r	-0.956	-0.21 | [-0.63, 0.22]	-1.054	-0.23 | [-0.65, 0.20]
MC19r	0.062	0.01 | [-0.41, 0.44]	0.163	0.04 | [-0.39, 0.46]
MC20r	0.888	0.19 | [-0.23, 0.62]	-0.888	-0.19 | [-0.62, 0.23]
MC21r	0.932	0.20 | [-0.22, 0.63]	0.915	0.20 | [-0.23, 0.62]
MC22r	0.355	0.08 | [-0.35, 0.50]	0.322	0.07 | [-0.35, 0.49]
MC23r	-1.075	-0.23 | [-0.66, 0.19]	1.079	0.23 | [-0.19, 0.66]
MC24r	2.722	0.59 | [0.16, 1.01]	2.827	0.61 | [0.19, 1.04]

Table A6. Winter 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 3 and 4 (Reference = Group 3, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.145	-0.03 | [-0.45, 0.39]	-0.143	-0.03 | [-0.45, 0.39]
MC02r	0.776	0.17 | [-0.25, 0.58]	0.761	0.16 | [-0.26, 0.58]
MC03r	1.485	0.32 | [-0.10, 0.74]	1.351	0.29 | [-0.13, 0.71]
MC04r	-0.014	0.00 | [-0.42, 0.42]	-2.038	-0.44 | [-0.85, -0.02]
MC05r	-0.261	-0.06 | [-0.47, 0.36]	0.550	0.12 | [-0.30, 0.54]
MC06r	-0.822	-0.18 | [-0.59, 0.24]	3.472	0.74 | [0.32, 1.16]
MC07r	-0.351	-0.07 | [-0.49, 0.34]	-0.392	-0.08 | [-0.50, 0.33]
MC08r	0.266	0.06 | [-0.36, 0.48]	0.243	0.05 | [-0.37, 0.47]
MC09r	1.548	0.33 | [-0.09, 0.75]	1.686	0.36 | [-0.06, 0.78]
MC10r	0.385	0.08 | [-0.34, 0.50]	-0.385	-0.08 | [-0.50, 0.34]
MC11r	-1.405	-0.30 | [-0.72, 0.12]	-1.653	-0.35 | [-0.77, 0.07]
MC12r	1.616	0.35 | [-0.07, 0.76]	1.684	0.36 | [-0.06, 0.78]
MC13r	1.244	0.27 | [-0.15, 0.68]	1.090	0.23 | [-0.19, 0.65]
MC14r	1.532	0.33 | [-0.09, 0.75]	0.988	0.21 | [-0.21, 0.63]
MC15r	0.499	0.11 | [-0.31, 0.53]	-0.630	-0.13 | [-0.55, 0.28]
MC16r	4.295	0.92 | [0.50, 1.34]	2.301	0.49 | [0.07, 0.91]
MC17r	-1.796	-0.38 | [-0.80, 0.04]	-2.174	-0.46 | [-0.88, -0.05]
MC18r	-1.136	-0.24 | [-0.66, 0.18]	-1.184	-0.25 | [-0.67, 0.17]
MC19r	0.179	0.04 | [-0.38, 0.46]	0.134	0.03 | [-0.39, 0.45]
MC20r	1.033	0.22 | [-0.20, 0.64]	0.995	0.21 | [-0.21, 0.63]
MC21r	0.669	0.14 | [-0.28, 0.56]	0.599	0.13 | [-0.29, 0.55]
MC22r	0.330	0.07 | [-0.35, 0.49]	0.317	0.07 | [-0.35, 0.49]
MC23r	-0.815	-0.17 | [-0.59, 0.24]	-0.897	-0.19 | [-0.61, 0.23]
MC24r	1.464	0.31 | [-0.11, 0.73]	1.413	0.30 | [-0.12, 0.72]

Table A7. Winter 2015 Applied Math Assessment DIF Effect Size Measures for Gender Groups (Reference = Males, Focal = Females)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.332	-0.06 | [-0.41, 0.29]	-0.627	-0.11 | [-0.46, 0.24]
MC02r	6.149	1.09 | [0.74, 1.44]	-5.209	-0.93 | [-1.28, -0.58]
MC03r	3.309	0.59 | [0.24, 0.94]	-2.641	-0.47 | [-0.82, -0.12]
MC04r	4.554	0.81 | [0.46, 1.16]	-3.917	-0.70 | [-1.05, -0.35]
MC05r	2.446	0.44 | [0.09, 0.78]	-1.902	-0.34 | [-0.69, 0.01]
MC06r	-1.366	-0.24 | [-0.59, 0.11]	-2.767	-0.49 | [-0.84, -0.14]
MC07r	5.393	0.96 | [0.61, 1.31]	-5.393	-0.96 | [-1.31, -0.61]
MC08r	7.792	1.39 | [1.04, 1.73]	-7.737	-1.38 | [-1.72, -1.03]
MC09r	1.409	0.25 | [-0.10, 0.60]	-2.144	-0.38 | [-0.73, -0.03]
MC10r	2.972	0.53 | [0.18, 0.88]	-2.972	-0.53 | [-0.88, -0.18]
MC11r	6.760	1.20 | [0.85, 1.55]	-6.723	-1.20 | [-1.54, -0.85]
MC12r	-5.879	-1.05 | [-1.39, -0.70]	-3.914	-0.70 | [-1.04, -0.35]
MC13r	-1.708	-0.30 | [-0.65, 0.04]	1.944	0.35 | [0.00, 0.69]
MC14r	0.263	0.05 | [-0.30, 0.40]	3.145	0.56 | [0.21, 0.91]
MC15r	2.620	0.47 | [0.12, 0.81]	-3.352	-0.60 | [-0.94, -0.25]
MC16r	3.767	0.67 | [0.32, 1.02]	4.833	0.86 | [0.51, 1.21]
MC17r	11.106	1.98 | [1.63, 2.32]	11.106	1.98 | [1.63, 2.32]
MC18r	3.766	0.67 | [0.32, 1.02]	3.766	0.67 | [0.32, 1.02]
MC19r	0.893	0.16 | [-0.19, 0.51]	0.894	0.16 | [-0.19, 0.51]
MC20r	3.615	0.64 | [0.29, 0.99]	3.613	0.64 | [0.29, 0.99]
MC21r	4.317	0.77 | [0.42, 1.12]	-3.382	-0.60 | [-0.95, -0.25]
MC22r	3.949	0.70 | [0.35, 1.05]	-2.995	-0.53 | [-0.88, -0.18]
MC23r	-2.529	-0.45 | [-0.80, -0.10]	2.462	0.44 | [0.09, 0.79]
MC24r	-2.184	-0.39 | [-0.74, -0.04]	-1.926	-0.34 | [-0.69, 0.01]

Table A8. Winter 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 2 (Reference = Group 1, Focal = Group = 2)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-1.161	-0.26 | [-0.69, 0.18]	-0.923	-0.20 | [-0.64, 0.23]
MC02r	0.185	0.04 | [-0.39, 0.48]	-0.227	-0.05 | [-0.49, 0.38]
MC03r	0.294	0.07 | [-0.37, 0.50]	-0.298	-0.07 | [-0.50, 0.37]
MC04r	0.496	0.11 | [-0.33, 0.55]	-1.209	-0.27 | [-0.70, 0.17]
MC05r	-0.127	-0.03 | [-0.46, 0.41]	0.129	0.03 | [-0.41, 0.46]
MC06r	-2.005	-0.45 | [-0.88, -0.01]	-2.537	-0.56 | [-1.00, -0.13]
MC07r	0.019	0.00 | [-0.43, 0.44]	-0.020	0.00 | [-0.44, 0.43]
MC08r	0.048	0.01 | [-0.42, 0.45]	-0.048	-0.01 | [-0.45, 0.42]
MC09r	0.490	0.11 | [-0.33, 0.54]	-0.385	-0.09 | [-0.52, 0.35]
MC10r	1.328	0.29 | [-0.14, 0.73]	0.760	0.17 | [-0.27, 0.60]
MC11r	-0.415	-0.09 | [-0.53, 0.34]	1.022	0.23 | [-0.21, 0.66]
MC12r	-0.846	-0.19 | [-0.62, 0.25]	0.818	0.18 | [-0.25, 0.62]
MC13r	0.181	0.04 | [-0.39, 0.48]	-0.199	-0.04 | [-0.48, 0.39]
MC14r	0.031	0.01 | [-0.43, 0.44]	-0.032	-0.01 | [-0.44, 0.43]
MC15r	-6.088	-1.35 | [-1.79, -0.92]	-7.320	-1.63 | [-2.06, -1.19]
MC16r	-1.996	-0.44 | [-0.88, -0.01]	-4.880	-1.08 | [-1.52, -0.65]
MC17r	0.063	0.01 | [-0.42, 0.45]	-0.064	-0.01 | [-0.45, 0.42]
MC18r	-5.925	-1.32 | [-1.75, -0.88]	-5.925	-1.32 | [-1.75, -0.88]
MC19r	-3.144	-0.70 | [-1.13, -0.26]	3.212	0.71 | [0.28, 1.15]
MC20r	0.953	0.21 | [-0.22, 0.65]	-1.429	-0.32 | [-0.75, 0.12]
MC21r	0.288	0.06 | [-0.37, 0.50]	-1.182	-0.26 | [-0.70, 0.17]
MC22r	1.279	0.28 | [-0.15, 0.72]	-1.414	-0.31 | [-0.75, 0.12]
MC23r	0.543	0.12 | [-0.31, 0.56]	-1.270	-0.28 | [-0.72, 0.15]
MC24r	1.342	0.30 | [-0.14, 0.73]	-0.867	-0.19 | [-0.63, 0.24]

Table A9. Winter 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 3 (Reference = Group 1, Focal = Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-1.046	-0.31 | [-0.60, -0.01]	1.693	0.25 | [-0.04, 0.55]
MC02r	-0.577	0.03 | [-0.27, 0.32]	-0.191	-0.03 | [-0.32, 0.27]
MC03r	0.499	-0.23 | [-0.52, 0.07]	1.407	0.21 | [-0.08, 0.51]
MC04r	-0.485	-0.93 | [-1.23, -0.64]	6.198	0.93 | [0.64, 1.23]
MC05r	0.687	-0.02 | [-0.31, 0.28]	0.120	0.02 | [-0.28, 0.31]
MC06r	-0.241	-1.82 | [-2.12, -1.53]	-12.978	-1.95 | [-2.25, -1.66]
MC07r	0.425	0.00 | [-0.29, 0.30]	-0.019	0.00 | [-0.30, 0.29]
MC08r	1.050	-0.05 | [-0.34, 0.25]	0.311	0.05 | [-0.25, 0.34]
MC09r	-1.493	0.06 | [-0.24, 0.35]	-0.504	-0.08 | [-0.37, 0.22]
MC10r	2.741	-0.03 | [-0.32, 0.27]	-3.182	-0.48 | [-0.77, -0.18]
MC11r	-0.423	-0.17 | [-0.46, 0.13]	0.685	0.10 | [-0.19, 0.40]
MC12r	-0.580	-0.37 | [-0.67, -0.08]	-2.496	-0.38 | [-0.67, -0.08]
MC13r	-1.020	-0.86 | [-1.16, -0.57]	5.672	0.85 | [0.56, 1.15]
MC14r	-0.004	0.00 | [-0.29, 0.30]	-0.029	0.00 | [-0.30, 0.29]
MC15r	1.488	-3.61 | [-3.90, -3.31]	-24.570	-3.70 | [-3.99, -3.40]
MC16r	0.049	-2.77 | [-3.06, -2.47]	-18.976	-2.85 | [-3.15, -2.56]
MC17r	2.659	-0.64 | [-0.94, -0.35]	-4.267	-0.64 | [-0.94, -0.35]
MC18r	2.587	-2.18 | [-2.48, -1.89]	14.504	2.18 | [1.89, 2.48]
MC19r	-1.291	-0.34 | [-0.64, -0.05]	2.281	0.34 | [0.05, 0.64]
MC20r	-0.754	-0.71 | [-1.01, -0.42]	4.730	0.71 | [0.42, 1.01]
MC21r	0.338	-0.50 | [-0.79, -0.20]	-3.312	-0.50 | [-0.79, -0.20]
MC22r	2.887	-0.86 | [-1.16, -0.57]	-5.941	-0.89 | [-1.19, -0.60]
MC23r	-0.155	-0.71 | [-1.00, -0.41]	4.703	0.71 | [0.41, 1.00]
MC24r	-0.169	0.09 | [-0.21, 0.38]	-1.742	-0.26 | [-0.56, 0.03]

Table A10. Winter 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 4 (Reference = Group 1, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	1.693	-0.19 | [-0.62, 0.25]	1.014	0.22 | [-0.21, 0.66]
MC02r	-0.191	0.03 | [-0.40, 0.47]	-0.188	-0.04 | [-0.47, 0.39]
MC03r	1.407	0.09 | [-0.34, 0.53]	1.568	0.35 | [-0.09, 0.78]
MC04r	6.198	-0.61 | [-1.04, -0.18]	2.771	0.61 | [0.18, 1.04]
MC05r	0.120	-0.02 | [-0.46, 0.41]	0.120	0.03 | [-0.41, 0.46]
MC06r	-12.978	-0.72 | [-1.15, -0.29]	3.267	0.72 | [0.29, 1.15]
MC07r	-0.019	0.00 | [-0.43, 0.44]	-0.018	0.00 | [-0.44, 0.43]
MC08r	0.311	-0.09 | [-0.53, 0.34]	0.425	0.09 | [-0.34, 0.53]
MC09r	-0.504	-0.02 | [-0.45, 0.41]	-0.608	-0.13 | [-0.57, 0.30]
MC10r	-3.182	-0.15 | [-0.58, 0.28]	-0.877	-0.19 | [-0.63, 0.24]
MC11r	0.685	0.04 | [-0.40, 0.47]	0.828	0.18 | [-0.25, 0.61]
MC12r	-2.496	-0.48 | [-0.91, -0.05]	-2.084	-0.46 | [-0.89, -0.03]
MC13r	5.672	-0.27 | [-0.70, 0.16]	-1.224	-0.27 | [-0.70, 0.16]
MC14r	-0.029	0.01 | [-0.43, 0.44]	-0.029	-0.01 | [-0.44, 0.43]
MC15r	-24.570	0.18 | [-0.25, 0.61]	0.831	0.18 | [-0.25, 0.62]
MC16r	-18.976	0.11 | [-0.32, 0.54]	0.511	0.11 | [-0.32, 0.54]
MC17r	-4.267	-0.08 | [-0.51, 0.36]	-0.336	-0.07 | [-0.51, 0.36]
MC18r	14.504	-0.12 | [-0.55, 0.31]	-0.545	-0.12 | [-0.55, 0.31]
MC19r	2.281	-0.50 | [-0.94, -0.07]	-2.284	-0.50 | [-0.94, -0.07]
MC20r	4.730	0.04 | [-0.39, 0.47]	0.199	0.04 | [-0.39, 0.48]
MC21r	-3.312	-0.04 | [-0.47, 0.39]	3.339	0.74 | [0.30, 1.17]
MC22r	-5.941	0.21 | [-0.22, 0.64]	1.798	0.40 | [-0.04, 0.83]
MC23r	4.703	0.16 | [-0.27, 0.59]	3.197	0.70 | [0.27, 1.14]
MC24r	-1.742	-0.11 | [-0.54, 0.32]	-1.575	-0.35 | [-0.78, 0.08]

Table A11. Winter 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 3 (Reference = Group 2, Focal Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	0.876	0.13 | [-0.16, 0.42]	1.105	0.16 | [-0.13, 0.45]
MC02r	-6.739	-0.99 | [-1.28, -0.70]	1.059	0.16 | [-0.13, 0.44]
MC03r	-0.341	-0.05 | [-0.34, 0.24]	0.334	0.05 | [-0.24, 0.34]
MC04r	-2.252	-0.33 | [-0.62, -0.04]	1.377	0.20 | [-0.09, 0.49]
MC05r	2.152	0.32 | [0.03, 0.61]	-4.198	-0.62 | [-0.91, -0.33]
MC06r	-11.045	-1.63 | [-1.91, -1.34]	3.668	0.54 | [0.25, 0.83]
MC07r	-1.191	-0.18 | [-0.46, 0.11]	0.469	0.07 | [-0.22, 0.36]
MC08r	-0.052	-0.01 | [-0.30, 0.28]	0.052	0.01 | [-0.28, 0.30]
MC09r	-15.192	-2.24 | [-2.52, -1.95]	-4.630	-0.68 | [-0.97, -0.39]
MC10r	-4.717	-0.69 | [-0.98, -0.41]	-4.542	-0.67 | [-0.96, -0.38]
MC11r	-6.340	-0.93 | [-1.22, -0.64]	-3.977	-0.59 | [-0.87, -0.30]
MC12r	0.735	0.11 | [-0.18, 0.40]	-0.788	-0.12 | [-0.40, 0.17]
MC13r	-0.226	-0.03 | [-0.32, 0.26]	0.202	0.03 | [-0.26, 0.32]
MC14r	-0.505	-0.07 | [-0.36, 0.21]	0.459	0.07 | [-0.22, 0.36]
MC15r	-33.150	-4.88 | [-5.17, -4.59]	-33.484	-4.93 | [-5.22, -4.64]
MC16r	-18.132	-2.67 | [-2.96, -2.38]	18.132	2.67 | [2.38, 2.96]
MC17r	-0.065	-0.01 | [-0.30, 0.28]	0.064	0.01 | [-0.28, 0.30]
MC18r	-19.950	-2.94 | [-3.22, -2.65]	19.950	2.94 | [2.65, 3.22]
MC19r	2.225	0.33 | [0.04, 0.62]	-3.608	-0.53 | [-0.82, -0.24]
MC20r	-2.469	-0.36 | [-0.65, -0.07]	1.995	0.29 | [0.01, 0.58]
MC21r	-2.796	-0.41 | [-0.70, -0.12]	2.306	0.34 | [0.05, 0.63]
MC22r	-2.091	-0.31 | [-0.60, -0.02]	1.955	0.29 | [0.00, 0.58]
MC23r	-2.259	-0.33 | [-0.62, -0.04]	1.586	0.23 | [-0.06, 0.52]
MC24r	-15.621	-2.30 | [-2.59, -2.01]	-7.723	-1.14 | [-1.43, -0.85]

Table A12. Winter 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 4 (Reference = Group 2, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	0.092	0.02 | [-0.38, 0.41]	1.006	0.20 | [-0.19, 0.60]
MC02r	-2.083	-0.42 | [-0.82, -0.02]	1.206	0.24 | [-0.15, 0.64]
MC03r	-0.262	-0.05 | [-0.45, 0.34]	0.352	0.07 | [-0.32, 0.47]
MC04r	-2.486	-0.50 | [-0.90, -0.11]	1.790	0.36 | [-0.03, 0.76]
MC05r	2.582	0.52 | [0.13, 0.92]	-2.640	-0.53 | [-0.93, -0.14]
MC06r	-2.821	-0.57 | [-0.97, -0.17]	4.229	0.85 | [0.46, 1.25]
MC07r	0.492	0.10 | [-0.30, 0.50]	0.688	0.14 | [-0.26, 0.53]
MC08r	-0.054	-0.01 | [-0.41, 0.39]	0.053	0.01 | [-0.39, 0.41]
MC09r	-0.603	-0.12 | [-0.52, 0.27]	-0.726	-0.15 | [-0.54, 0.25]
MC10r	-0.739	-0.15 | [-0.55, 0.25]	-0.974	-0.20 | [-0.59, 0.20]
MC11r	1.025	0.21 | [-0.19, 0.60]	-3.184	-0.64 | [-1.04, -0.25]
MC12r	0.483	0.10 | [-0.30, 0.49]	-0.799	-0.16 | [-0.56, 0.23]
MC13r	-0.314	-0.06 | [-0.46, 0.33]	0.202	0.04 | [-0.36, 0.44]
MC14r	-0.382	-0.08 | [-0.47, 0.32]	0.479	0.10 | [-0.30, 0.49]
MC15r	0.925	0.19 | [-0.21, 0.58]	0.953	0.19 | [-0.20, 0.59]
MC16r	0.521	0.11 | [-0.29, 0.50]	0.535	0.11 | [-0.29, 0.50]
MC17r	-0.078	-0.02 | [-0.41, 0.38]	0.066	0.01 | [-0.38, 0.41]
MC18r	-0.520	-0.10 | [-0.50, 0.29]	-0.515	-0.10 | [-0.50, 0.29]
MC19r	-0.577	-0.12 | [-0.51, 0.28]	-2.464	-0.50 | [-0.89, -0.10]
MC20r	0.185	0.04 | [-0.36, 0.43]	0.204	0.04 | [-0.35, 0.44]
MC21r	-0.407	-0.08 | [-0.48, 0.31]	3.359	0.68 | [0.28, 1.07]
MC22r	-0.337	-0.07 | [-0.46, 0.33]	2.614	0.53 | [0.13, 0.92]
MC23r	0.054	0.01 | [-0.39, 0.41]	2.055	0.42 | [0.02, 0.81]
MC24r	-0.923	-0.19 | [-0.58, 0.21]	-1.206	-0.24 | [-0.64, 0.15]

Table A13. Winter 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 3 and 4 (Reference = Group 3, Focal = Group 4)

	Z(SA)	d | [95% CI]	Z(UA)	d | [95% CI]
MC01r	-0.603	-0.09 | [-0.38, 0.20]	-0.603	-0.09 | [-0.38, 0.20]
MC02r	-1.054	-0.16 | [-0.44, 0.13]	1.054	0.16 | [-0.13, 0.44]
MC03r	1.154	0.17 | [-0.12, 0.46]	1.333	0.20 | [-0.09, 0.48]
MC04r	-1.333	-0.20 | [-0.48, 0.09]	1.333	0.20 | [-0.09, 0.48]
MC05r	2.095	0.31 | [0.02, 0.60]	2.344	0.34 | [0.06, 0.63]
MC06r	-1.075	-0.16 | [-0.45, 0.13]	1.897	0.28 | [-0.01, 0.57]
MC07r	0.575	0.08 | [-0.20, 0.37]	0.619	0.09 | [-0.20, 0.38]
MC08r	-1.848	-0.27 | [-0.56, 0.02]	-1.833	-0.27 | [-0.56, 0.02]
MC09r	-0.497	-0.07 | [-0.36, 0.22]	-0.497	-0.07 | [-0.36, 0.22]
MC10r	-0.677	-0.10 | [-0.39, 0.19]	-0.675	-0.10 | [-0.39, 0.19]
MC11r	2.260	0.33 | [0.04, 0.62]	2.366	0.35 | [0.06, 0.64]
MC12r	-1.548	-0.23 | [-0.52, 0.06]	-1.430	-0.21 | [-0.50, 0.08]
MC13r	-0.812	-0.12 | [-0.41, 0.17]	-0.803	-0.12 | [-0.41, 0.17]
MC14r	2.597	0.38 | [0.09, 0.67]	3.140	0.46 | [0.17, 0.75]
MC15r	1.279	0.19 | [-0.10, 0.48]	1.296	0.19 | [-0.10, 0.48]
MC16r	0.618	0.09 | [-0.20, 0.38]	0.623	0.09 | [-0.20, 0.38]
MC17r	-0.310	-0.05 | [-0.33, 0.24]	-0.306	-0.04 | [-0.33, 0.24]
MC18r	-0.473	-0.07 | [-0.36, 0.22]	-0.469	-0.07 | [-0.36, 0.22]
MC19r	-1.401	-0.21 | [-0.49, 0.08]	-1.381	-0.20 | [-0.49, 0.09]
MC20r	0.207	0.03 | [-0.26, 0.32]	0.209	0.03 | [-0.26, 0.32]
MC21r	2.023	0.30 | [0.01, 0.59]	3.098	0.46 | [0.17, 0.74]
MC22r	1.894	0.28 | [-0.01, 0.57]	2.306	0.34 | [0.05, 0.63]
MC23r	2.810	0.41 | [0.13, 0.70]	3.131	0.46 | [0.17, 0.75]
MC24r	-0.694	-0.10 | [-0.39, 0.19]	-0.694	-0.10 | [-0.39, 0.19]

Table A14. Winter 2015 Academic Math Assessment DIF Effect Size Measures for Gender Groups (Reference = Males, Focal = Males)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	1.224	0.17 | [-0.10, 0.45]	1.739	0.24 | [-0.03, 0.52]
MC02r	2.804	0.39 | [0.12, 0.67]	3.238	0.46 | [0.18, 0.73]
MC03r	8.689	1.22 | [0.95, 1.50]	-8.680	-1.22 | [-1.50, -0.94]
MC04r	-1.729	-0.24 | [-0.52, 0.03]	-1.691	-0.24 | [-0.51, 0.04]
MC05r	2.019	0.28 | [0.01, 0.56]	-3.599	-0.51 | [-0.78, -0.23]
MC06r	-2.301	-0.32 | [-0.60, -0.05]	-2.087	-0.29 | [-0.57, -0.02]
MC07r	-0.139	-0.02 | [-0.30, 0.26]	-1.502	-0.21 | [-0.49, 0.06]
MC08r	12.041	1.69 | [1.42, 1.97]	13.393	1.88 | [1.61, 2.16]
MC09r	-2.422	-0.34 | [-0.62, -0.06]	-2.238	-0.31 | [-0.59, -0.04]
MC10r	8.710	1.22 | [0.95, 1.50]	8.589	1.21 | [0.93, 1.48]
MC11r	3.933	0.55 | [0.28, 0.83]	-4.076	-0.57 | [-0.85, -0.30]
MC12r	3.135	0.44 | [0.17, 0.72]	2.677	0.38 | [0.10, 0.65]
MC13r	-1.233	-0.17 | [-0.45, 0.10]	4.535	0.64 | [0.36, 0.91]
MC14r	17.219	2.42 | [2.15, 2.70]	17.215	2.42 | [2.14, 2.70]
MC15r	4.301	0.60 | [0.33, 0.88]	-4.301	-0.60 | [-0.88, -0.33]
MC16r	10.409	1.46 | [1.19, 1.74]	-10.409	-1.46 | [-1.74, -1.19]
MC17r	8.411	1.18 | [0.91, 1.46]	8.370	1.18 | [0.90, 1.45]
MC18r	1.091	0.15 | [-0.12, 0.43]	-1.170	-0.16 | [-0.44, 0.11]
MC19r	-3.910	-0.55 | [-0.83, -0.27]	-3.910	-0.55 | [-0.83, -0.27]
MC20r	9.681	1.36 | [1.09, 1.64]	9.681	1.36 | [1.09, 1.64]
MC21r	27.250	3.83 | [3.56, 4.11]	-27.250	-3.83 | [-4.11, -3.56]
MC22r	19.843	2.79 | [2.51, 3.07]	19.726	2.77 | [2.50, 3.05]
MC23r	2.711	0.38 | [0.11, 0.66]	-3.088	-0.43 | [-0.71, -0.16]
MC24r	-4.925	-0.69 | [-0.97, -0.42]	-4.588	-0.65 | [-0.92, -0.37]

Table A15. Spring 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 2 (Reference = Group 1, Focal = Group 2)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-1.271	-0.27 | [-0.70, 0.15]	-0.722	-0.16 | [-0.58, 0.27]
MC02r	-1.171	-0.25 | [-0.68, 0.17]	1.364	0.29 | [-0.13, 0.72]
MC03r	-2.880	-0.62 | [-1.04, -0.20]	2.858	0.62 | [0.19, 1.04]
MC04r	0.354	0.08 | [-0.35, 0.50]	-0.434	-0.09 | [-0.52, 0.33]
MC05r	-0.420	-0.09 | [-0.51, 0.33]	0.545	0.12 | [-0.31, 0.54]
MC06r	NaN	|	NaN	|
MC07r	0.283	0.06 | [-0.36, 0.48]	-0.287	-0.06 | [-0.48, 0.36]
MC08r	-0.485	-0.10 | [-0.53, 0.32]	-0.496	-0.11 | [-0.53, 0.32]
MC09r	0.872	0.19 | [-0.23, 0.61]	0.830	0.18 | [-0.24, 0.60]
MC10r	-0.677	-0.15 | [-0.57, 0.28]	0.997	0.22 | [-0.21, 0.64]
MC11r	-0.436	-0.09 | [-0.52, 0.33]	0.603	0.13 | [-0.29, 0.55]
MC12r	-1.882	-0.41 | [-0.83, 0.02]	0.825	0.18 | [-0.24, 0.60]
MC13r	-0.041	-0.01 | [-0.43, 0.41]	0.041	0.01 | [-0.41, 0.43]
MC14r	0.101	0.02 | [-0.40, 0.44]	-0.105	-0.02 | [-0.45, 0.40]
MC15r	-0.460	-0.10 | [-0.52, 0.32]	0.454	0.10 | [-0.32, 0.52]
MC16r	-0.397	-0.09 | [-0.51, 0.34]	0.671	0.14 | [-0.28, 0.57]
MC17r	0.693	0.15 | [-0.27, 0.57]	1.008	0.22 | [-0.21, 0.64]
MC18r	2.631	0.57 | [0.14, 0.99]	3.411	0.74 | [0.31, 1.16]
MC19r	-0.330	-0.07 | [-0.49, 0.35]	0.327	0.07 | [-0.35, 0.49]
MC20r	-0.015	0.00 | [-0.43, 0.42]	0.015	0.00 | [-0.42, 0.43]
MC21r	-0.381	-0.08 | [-0.50, 0.34]	0.376	0.08 | [-0.34, 0.50]
MC22r	0.373	0.08 | [-0.34, 0.50]	1.044	0.23 | [-0.20, 0.65]
MC23r	3.409	0.74 | [0.31, 1.16]	4.279	0.92 | [0.50, 1.35]
MC24r	-1.264	-0.27 | [-0.70, 0.15]	-1.332	-0.29 | [-0.71, 0.14]

Table A16. Spring 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 3 (Reference = Group 1, Focal = Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.840	-0.18 | [-0.59, 0.24]	-0.706	-0.15 | [-0.56, 0.26]
MC02r	-0.831	-0.17 | [-0.59, 0.24]	0.784	0.16 | [-0.25, 0.58]
MC03r	0.404	0.08 | [-0.33, 0.50]	5.560	1.17 | [0.76, 1.58]
MC04r	-0.636	-0.13 | [-0.54, 0.28]	-1.547	-0.32 | [-0.74, 0.09]
MC05r	-0.643	-0.13 | [-0.55, 0.28]	0.335	0.07 | [-0.34, 0.48]
MC06r	NaN	|	NaN	|
MC07r	-0.918	-0.19 | [-0.60, 0.22]	-1.625	-0.34 | [-0.75, 0.07]
MC08r	-1.373	-0.29 | [-0.70, 0.12]	1.943	0.41 | [0.00, 0.82]
MC09r	-0.598	-0.13 | [-0.54, 0.29]	-0.613	-0.13 | [-0.54, 0.28]
MC10r	-1.127	-0.24 | [-0.65, 0.17]	0.990	0.21 | [-0.20, 0.62]
MC11r	-0.532	-0.11 | [-0.52, 0.30]	0.437	0.09 | [-0.32, 0.50]
MC12r	1.668	0.35 | [-0.06, 0.76]	2.631	0.55 | [0.14, 0.96]
MC13r	-0.278	-0.06 | [-0.47, 0.35]	0.919	0.19 | [-0.22, 0.60]
MC14r	-0.063	-0.01 | [-0.42, 0.40]	0.336	0.07 | [-0.34, 0.48]
MC15r	-0.532	-0.11 | [-0.52, 0.30]	0.583	0.12 | [-0.29, 0.53]
MC16r	-0.817	-0.17 | [-0.58, 0.24]	-1.138	-0.24 | [-0.65, 0.17]
MC17r	0.653	0.14 | [-0.27, 0.55]	0.635	0.13 | [-0.28, 0.54]
MC18r	-1.910	-0.40 | [-0.81, 0.01]	-3.442	-0.72 | [-1.13, -0.31]
MC19r	-0.238	-0.05 | [-0.46, 0.36]	-2.183	-0.46 | [-0.87, -0.05]
MC20r	1.058	0.22 | [-0.19, 0.63]	-0.935	-0.20 | [-0.61, 0.22]
MC21r	-0.138	-0.03 | [-0.44, 0.38]	0.135	0.03 | [-0.38, 0.44]
MC22r	1.328	0.28 | [-0.13, 0.69]	-3.667	-0.77 | [-1.18, -0.36]
MC23r	-2.293	-0.48 | [-0.89, -0.07]	-4.291	-0.90 | [-1.31, -0.49]
MC24r	0.717	0.15 | [-0.26, 0.56]	0.678	0.14 | [-0.27, 0.55]

Table A17. Spring 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 4 (Reference = Group 1, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.055	-0.01 | [-0.52, 0.49]	-0.054	-0.01 | [-0.52, 0.49]
MC02r	-1.037	-0.27 | [-0.77, 0.24]	0.921	0.24 | [-0.27, 0.74]
MC03r	0.746	0.19 | [-0.31, 0.70]	0.797	0.21 | [-0.30, 0.71]
MC04r	-0.439	-0.11 | [-0.62, 0.39]	0.958	0.25 | [-0.26, 0.75]
MC05r	-0.167	-0.04 | [-0.55, 0.46]	0.507	0.13 | [-0.37, 0.63]
MC06r	NaN	|	NaN	|
MC07r	-0.380	-0.10 | [-0.60, 0.41]	-0.360	-0.09 | [-0.60, 0.41]
MC08r	-0.991	-0.25 | [-0.76, 0.25]	1.315	0.34 | [-0.17, 0.84]
MC09r	-0.049	-0.01 | [-0.52, 0.49]	-0.049	-0.01 | [-0.52, 0.49]
MC10r	0.064	0.02 | [-0.49, 0.52]	0.927	0.24 | [-0.27, 0.74]
MC11r	-0.390	-0.10 | [-0.60, 0.40]	-0.390	-0.10 | [-0.60, 0.40]
MC12r	1.154	0.30 | [-0.21, 0.80]	1.918	0.49 | [-0.01, 1.00]
MC13r	-0.871	-0.22 | [-0.73, 0.28]	0.871	0.22 | [-0.28, 0.73]
MC14r	0.090	0.02 | [-0.48, 0.53]	0.345	0.09 | [-0.42, 0.59]
MC15r	-0.162	-0.04 | [-0.55, 0.46]	0.760	0.20 | [-0.31, 0.70]
MC16r	-0.685	-0.18 | [-0.68, 0.33]	-3.285	-0.84 | [-1.35, -0.34]
MC17r	0.677	0.17 | [-0.33, 0.68]	-0.688	-0.18 | [-0.68, 0.33]
MC18r	-0.884	-0.23 | [-0.73, 0.28]	-3.019	-0.78 | [-1.28, -0.27]
MC19r	-2.537	-0.65 | [-1.16, -0.15]	-2.543	-0.65 | [-1.16, -0.15]
MC20r	-0.119	-0.03 | [-0.53, 0.47]	-0.146	-0.04 | [-0.54, 0.47]
MC21r	-0.206	-0.05 | [-0.56, 0.45]	0.229	0.06 | [-0.44, 0.56]
MC22r	0.046	0.01 | [-0.49, 0.52]	-2.865	-0.74 | [-1.24, -0.23]
MC23r	-0.509	-0.13 | [-0.63, 0.37]	-0.598	-0.15 | [-0.66, 0.35]
MC24r	1.073	0.28 | [-0.23, 0.78]	-1.567	-0.40 | [-0.91, 0.10]

Table A18. Spring 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 3 (Reference = Group 2, Focal = Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.617	-0.12 | [-0.49, 0.26]	-0.617	-0.12 | [-0.49, 0.26]
MC02r	1.219	0.23 | [-0.14, 0.61]	-2.055	-0.39 | [-0.77, -0.02]
MC03r	14.071	2.70 | [2.32, 3.07]	11.160	2.14 | [1.76, 2.52]
MC04r	-0.646	-0.12 | [-0.50, 0.25]	0.575	0.11 | [-0.27, 0.49]
MC05r	-2.454	-0.47 | [-0.85, -0.09]	-3.547	-0.68 | [-1.06, -0.30]
MC06r	-13.503	-2.59 | [-2.96, -2.21]	10.987	2.11 | [1.73, 2.48]
MC07r	-0.295	-0.06 | [-0.43, 0.32]	0.271	0.05 | [-0.32, 0.43]
MC08r	-4.167	-0.80 | [-1.17, -0.42]	5.213	1.00 | [0.62, 1.38]
MC09r	-0.826	-0.16 | [-0.53, 0.22]	-0.884	-0.17 | [-0.55, 0.21]
MC10r	-3.553	-0.68 | [-1.06, -0.31]	-4.686	-0.90 | [-1.27, -0.52]
MC11r	-1.656	-0.32 | [-0.69, 0.06]	-2.587	-0.50 | [-0.87, -0.12]
MC12r	3.270	0.63 | [0.25, 1.00]	3.946	0.76 | [0.38, 1.13]
MC13r	0.040	0.01 | [-0.37, 0.38]	-0.038	-0.01 | [-0.38, 0.37]
MC14r	-1.453	-0.28 | [-0.65, 0.10]	1.322	0.25 | [-0.12, 0.63]
MC15r	-6.954	-1.33 | [-1.71, -0.96]	10.888	2.09 | [1.71, 2.46]
MC16r	-0.443	-0.08 | [-0.46, 0.29]	-1.143	-0.22 | [-0.59, 0.16]
MC17r	0.468	0.09 | [-0.29, 0.47]	0.465	0.09 | [-0.29, 0.46]
MC18r	-3.183	-0.61 | [-0.99, -0.23]	-4.076	-0.78 | [-1.16, -0.41]
MC19r	0.327	0.06 | [-0.31, 0.44]	-0.337	-0.06 | [-0.44, 0.31]
MC20r	0.015	0.00 | [-0.37, 0.38]	-0.015	0.00 | [-0.38, 0.37]
MC21r	0.432	0.08 | [-0.29, 0.46]	-0.443	-0.08 | [-0.46, 0.29]
MC22r	4.347	0.83 | [0.46, 1.21]	-5.609	-1.08 | [-1.45, -0.70]
MC23r	-4.466	-0.86 | [-1.23, -0.48]	-5.477	-1.05 | [-1.43, -0.67]
MC24r	1.117	0.21 | [-0.16, 0.59]	1.170	0.22 | [-0.15, 0.60]

Table A19. Spring 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 4 (Reference = Group 2, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.053	-0.01 | [-0.44, 0.41]	-0.050	-0.01 | [-0.44, 0.41]
MC02r	0.504	0.11 | [-0.32, 0.53]	-1.729	-0.37 | [-0.80, 0.05]
MC03r	0.859	0.19 | [-0.24, 0.61]	0.877	0.19 | [-0.23, 0.61]
MC04r	-0.566	-0.12 | [-0.55, 0.30]	0.737	0.16 | [-0.26, 0.58]
MC05r	1.515	0.33 | [-0.10, 0.75]	-2.096	-0.45 | [-0.88, -0.03]
MC06r	0.743	0.16 | [-0.26, 0.59]	1.360	0.29 | [-0.13, 0.72]
MC07r	-0.425	-0.09 | [-0.52, 0.33]	0.421	0.09 | [-0.33, 0.52]
MC08r	-2.208	-0.48 | [-0.90, -0.05]	4.002	0.87 | [0.44, 1.29]
MC09r	-0.053	-0.01 | [-0.44, 0.41]	-0.054	-0.01 | [-0.44, 0.41]
MC10r	0.643	0.14 | [-0.29, 0.56]	-0.654	-0.14 | [-0.57, 0.28]
MC11r	-0.271	-0.06 | [-0.48, 0.37]	-0.348	-0.08 | [-0.50, 0.35]
MC12r	3.029	0.66 | [0.23, 1.08]	4.930	1.07 | [0.64, 1.49]
MC13r	0.039	0.01 | [-0.42, 0.43]	-0.040	-0.01 | [-0.43, 0.42]
MC14r	-0.070	-0.02 | [-0.44, 0.41]	0.484	0.10 | [-0.32, 0.53]
MC15r	0.901	0.20 | [-0.23, 0.62]	1.198	0.26 | [-0.17, 0.68]
MC16r	0.203	0.04 | [-0.38, 0.47]	-0.801	-0.17 | [-0.60, 0.25]
MC17r	0.361	0.08 | [-0.35, 0.50]	-0.981	-0.21 | [-0.64, 0.21]
MC18r	-3.069	-0.66 | [-1.09, -0.24]	-5.038	-1.09 | [-1.52, -0.67]
MC19r	0.302	0.07 | [-0.36, 0.49]	-0.315	-0.07 | [-0.49, 0.36]
MC20r	0.014	0.00 | [-0.42, 0.43]	-0.015	0.00 | [-0.43, 0.42]
MC21r	0.350	0.08 | [-0.35, 0.50]	-0.336	-0.07 | [-0.50, 0.35]
MC22r	-0.525	-0.11 | [-0.54, 0.31]	-5.000	-1.08 | [-1.51, -0.66]
MC23r	-0.585	-0.13 | [-0.55, 0.30]	-0.658	-0.14 | [-0.57, 0.28]
MC24r	1.463	0.32 | [-0.11, 0.74]	1.348	0.29 | [-0.13, 0.72]

Table A20. Spring 2015 Applied Math Assessment DIF Effect Size Measures for Achievement Groups 3 and 4 (Reference = Group 3, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.048	-0.01 | [-0.42, 0.40]	-0.046	-0.01 | [-0.42, 0.40]
MC02r	-1.388	-0.29 | [-0.71, 0.12]	3.390	0.71 | [0.30, 1.13]
MC03r	0.730	0.15 | [-0.26, 0.57]	0.730	0.15 | [-0.26, 0.57]
MC04r	1.095	0.23 | [-0.18, 0.64]	2.377	0.50 | [0.09, 0.91]
MC05r	2.520	0.53 | [0.12, 0.94]	2.823	0.59 | [0.18, 1.01]
MC06r	1.472	0.31 | [-0.10, 0.72]	1.055	0.22 | [-0.19, 0.64]
MC07r	-0.358	-0.08 | [-0.49, 0.34]	-0.342	-0.07 | [-0.49, 0.34]
MC08r	3.917	0.83 | [0.41, 1.24]	-3.567	-0.75 | [-1.16, -0.34]
MC09r	-0.039	-0.01 | [-0.42, 0.40]	-0.038	-0.01 | [-0.42, 0.40]
MC10r	1.016	0.21 | [-0.20, 0.63]	1.063	0.22 | [-0.19, 0.64]
MC11r	-0.240	-0.05 | [-0.46, 0.36]	-0.231	-0.05 | [-0.46, 0.36]
MC12r	-0.608	-0.13 | [-0.54, 0.28]	-1.775	-0.37 | [-0.79, 0.04]
MC13r	-1.335	-0.28 | [-0.69, 0.13]	-0.948	-0.20 | [-0.61, 0.21]
MC14r	0.897	0.19 | [-0.22, 0.60]	0.812	0.17 | [-0.24, 0.58]
MC15r	1.122	0.24 | [-0.18, 0.65]	1.066	0.22 | [-0.19, 0.64]
MC16r	0.678	0.14 | [-0.27, 0.56]	0.890	0.19 | [-0.23, 0.60]
MC17r	-0.257	-0.05 | [-0.47, 0.36]	-0.333	-0.07 | [-0.48, 0.34]
MC18r	1.265	0.27 | [-0.15, 0.68]	1.789	0.38 | [-0.04, 0.79]
MC19r	-4.739	-1.00 | [-1.41, -0.59]	3.768	0.79 | [0.38, 1.21]
MC20r	-0.140	-0.03 | [-0.44, 0.38]	-0.122	-0.03 | [-0.44, 0.39]
MC21r	-1.236	-0.26 | [-0.67, 0.15]	1.712	0.36 | [-0.05, 0.77]
MC22r	-2.151	-0.45 | [-0.87, -0.04]	3.719	0.78 | [0.37, 1.20]
MC23r	-0.409	-0.09 | [-0.50, 0.33]	-0.400	-0.08 | [-0.50, 0.33]
MC24r	-0.632	-0.13 | [-0.55, 0.28]	-0.656	-0.14 | [-0.55, 0.27]

Table A21. Spring 2015 Applied Math Assessment DIF Effect Size Measures for Gender Groups (Reference = Males, Focal = Females)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	0.410	0.07 | [-0.27, 0.42]	-1.714	-0.30 | [-0.65, 0.04]
MC02r	6.377	1.13 | [0.78, 1.48]	-6.376	-1.13 | [-1.48, -0.78]
MC03r	-2.629	-0.47 | [-0.81, -0.12]	3.155	0.56 | [0.21, 0.91]
MC04r	3.670	0.65 | [0.30, 1.00]	3.811	0.68 | [0.33, 1.02]
MC05r	2.958	0.52 | [0.18, 0.87]	-3.620	-0.64 | [-0.99, -0.29]
MC06r	-3.037	-0.54 | [-0.89, -0.19]	-3.533	-0.63 | [-0.97, -0.28]
MC07r	2.237	0.40 | [0.05, 0.74]	-2.107	-0.37 | [-0.72, -0.03]
MC08r	-5.099	-0.90 | [-1.25, -0.56]	5.089	0.90 | [0.55, 1.25]
MC09r	-2.025	-0.36 | [-0.71, -0.01]	-2.018	-0.36 | [-0.71, -0.01]
MC10r	3.127	0.55 | [0.21, 0.90]	3.120	0.55 | [0.21, 0.90]
MC11r	7.852	1.39 | [1.04, 1.74]	-9.064	-1.61 | [-1.95, -1.26]
MC12r	-5.457	-0.97 | [-1.31, -0.62]	-5.428	-0.96 | [-1.31, -0.61]
MC13r	0.047	0.01 | [-0.34, 0.36]	2.947	0.52 | [0.17, 0.87]
MC14r	1.109	0.20 | [-0.15, 0.54]	1.144	0.20 | [-0.14, 0.55]
MC15r	6.874	1.22 | [0.87, 1.57]	-6.873	-1.22 | [-1.57, -0.87]
MC16r	2.284	0.40 | [0.06, 0.75]	-2.203	-0.39 | [-0.74, -0.04]
MC17r	3.325	0.59 | [0.24, 0.94]	3.319	0.59 | [0.24, 0.94]
MC18r	2.569	0.46 | [0.11, 0.80]	-2.607	-0.46 | [-0.81, -0.11]
MC19r	7.112	1.26 | [0.91, 1.61]	-6.744	-1.20 | [-1.54, -0.85]
MC20r	3.887	0.69 | [0.34, 1.04]	-2.541	-0.45 | [-0.80, -0.10]
MC21r	1.225	0.22 | [-0.13, 0.56]	-1.484	-0.26 | [-0.61, 0.08]
MC22r	10.732	1.90 | [1.55, 2.25]	-10.028	-1.78 | [-2.12, -1.43]
MC23r	3.796	0.67 | [0.33, 1.02]	3.227	0.57 | [0.22, 0.92]
MC24r	-2.717	-0.48 | [-0.83, -0.13]	-2.405	-0.43 | [-0.77, -0.08]

Table A22. Spring 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 2 (Reference = Group 1, Focal = Group 2)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	0.331	0.07 | [-0.36, 0.51]	0.407	-0.09 | [-0.52, 0.34]
MC02r	0.377	0.08 | [-0.35, 0.52]	-0.379	-0.08 | [-0.52, 0.35]
MC03r	1.899	0.42 | [-0.01, 0.85]	6.285	1.39 | [0.96, 1.82]
MC04r	-0.328	-0.07 | [-0.51, 0.36]	-0.832	-0.18 | [-0.62, 0.25]
MC05r	-1.905	-0.42 | [-0.86, 0.01]	1.923	0.43 | [-0.01, 0.86]
MC06r	0.072	0.02 | [-0.42, 0.45]	-0.468	-0.10 | [-0.54, 0.33]
MC07r	-3.365	-0.74 | [-1.18, -0.31]	3.539	0.78 | [0.35, 1.22]
MC08r	-4.178	-0.92 | [-1.36, -0.49]	4.807	1.06 | [0.63, 1.50]
MC09r	-0.039	-0.01 | [-0.44, 0.43]	0.038	0.01 | [-0.43, 0.44]
MC10r	1.259	0.28 | [-0.16, 0.71]	-2.982	-0.66 | [-1.09, -0.23]
MC11r	0.388	0.09 | [-0.35, 0.52]	-0.398	-0.09 | [-0.52, 0.35]
MC12r	0.768	0.17 | [-0.26, 0.60]	-1.262	-0.28 | [-0.71, 0.15]
MC13r	-0.587	-0.13 | [-0.56, 0.30]	0.586	0.13 | [-0.30, 0.56]
MC14r	-6.171	-1.37 | [-1.80, -0.93]	6.071	1.34 | [0.91, 1.78]
MC15r	1.017	0.23 | [-0.21, 0.66]	-1.104	-0.24 | [-0.68, 0.19]
MC16r	-6.656	-1.47 | [-1.91, -1.04]	5.769	1.28 | [0.84, 1.71]
MC17r	-0.564	-0.12 | [-0.56, 0.31]	0.541	0.12 | [-0.31, 0.55]
MC18r	-2.564	-0.57 | [-1.00, -0.13]	1.212	0.27 | [-0.17, 0.70]
MC19r	1.753	0.39 | [-0.05, 0.82]	1.486	0.33 | [-0.10, 0.76]
MC20r	-6.095	-1.35 | [-1.78, -0.92]	5.530	1.22 | [0.79, 1.66]
MC21r	-0.359	-0.08 | [-0.51, 0.35]	0.376	0.08 | [-0.35, 0.52]
MC22r	-6.963	-1.54 | [-1.97, -1.11]	6.940	1.54 | [1.10, 1.97]
MC23r	1.049	0.23 | [-0.20, 0.67]	-1.257	-0.28 | [-0.71, 0.16]
MC24r	2.785	0.62 | [0.18, 1.05]	-4.131	-0.91 | [-1.35, -0.48]

Table A23. Spring 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 3 (Reference = Group 1, Focal = Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-2.174	-0.32 | [-0.61, -0.03]	1.547	0.23 | [-0.06, 0.52]
MC02r	1.503	0.22 | [-0.07, 0.51]	-1.518	-0.22 | [-0.51, 0.06]
MC03r	-17.146	-2.52 | [-2.81, -2.23]	-17.220	-2.53 | [-2.82, -2.24]
MC04r	-1.894	-0.28 | [-0.57, 0.01]	1.732	0.25 | [-0.03, 0.54]
MC05r	-1.325	-0.19 | [-0.48, 0.09]	1.177	0.17 | [-0.12, 0.46]
MC06r	-1.696	-0.25 | [-0.54, 0.04]	1.542	0.23 | [-0.06, 0.51]
MC07r	-2.618	-0.38 | [-0.67, -0.10]	2.465	0.36 | [0.07, 0.65]
MC08r	-5.788	-0.85 | [-1.14, -0.56]	5.784	0.85 | [0.56, 1.14]
MC09r	-2.127	-0.31 | [-0.60, -0.02]	1.961	0.29 | [0.00, 0.58]
MC10r	-0.180	-0.03 | [-0.31, 0.26]	-2.683	-0.39 | [-0.68, -0.11]
MC11r	-2.557	-0.38 | [-0.66, -0.09]	-17.980	-2.64 | [-2.93, -2.35]
MC12r	-5.679	-0.83 | [-1.12, -0.55]	-5.953	-0.87 | [-1.16, -0.59]
MC13r	-1.552	-0.23 | [-0.52, 0.06]	1.493	0.22 | [-0.07, 0.51]
MC14r	-9.097	-1.34 | [-1.62, -1.05]	-9.097	-1.34 | [-1.62, -1.05]
MC15r	-4.789	-0.70 | [-0.99, -0.42]	-5.068	-0.74 | [-1.03, -0.46]
MC16r	-13.130	-1.93 | [-2.22, -1.64]	-13.132	-1.93 | [-2.22, -1.64]
MC17r	-5.471	-0.80 | [-1.09, -0.52]	-5.471	-0.80 | [-1.09, -0.52]
MC18r	-5.856	-0.86 | [-1.15, -0.57]	4.984	0.73 | [0.44, 1.02]
MC19r	-0.664	-0.10 | [-0.39, 0.19]	0.742	0.11 | [-0.18, 0.40]
MC20r	-8.785	-1.35 | [-1.78, -0.92]	-9.890	-1.45 | [-1.74, -1.17]
MC21r	-0.354	-0.08 | [-0.51, 0.35]	0.317	0.05 | [-0.24, 0.33]
MC22r	-7.635	-1.54 | [-1.97, -1.11]	-9.394	-1.38 | [-1.67, -1.09]
MC23r	0.703	0.23 | [-0.20, 0.67]	-0.988	-0.15 | [-0.43, 0.14]
MC24r	-0.311	-0.05 | [-0.33, 0.24]	-4.026	-0.59 | [-0.88, -0.30]

Table A24. Spring 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 1 and 4 (Reference = Group 1, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.140	-0.03 | [-0.45, 0.39]	-0.140	-0.03 | [-0.45, 0.39]
MC02r	0.560	0.12 | [-0.30, 0.53]	-0.789	-0.17 | [-0.58, 0.25]
MC03r	1.184	0.25 | [-0.16, 0.67]	1.248	0.26 | [-0.15, 0.68]
MC04r	-2.352	-0.50 | [-0.91, -0.08]	-2.352	-0.50 | [-0.91, -0.08]
MC05r	0.663	0.14 | [-0.28, 0.56]	1.935	0.41 | [-0.01, 0.83]
MC06r	0.200	0.04 | [-0.37, 0.46]	0.205	0.04 | [-0.37, 0.46]
MC07r	1.069	0.23 | [-0.19, 0.64]	1.984	0.42 | [0.01, 0.84]
MC08r	0.301	0.06 | [-0.35, 0.48]	0.304	0.06 | [-0.35, 0.48]
MC09r	0.404	0.09 | [-0.33, 0.50]	3.176	0.67 | [0.26, 1.09]
MC10r	0.736	0.16 | [-0.26, 0.57]	0.736	0.16 | [-0.26, 0.57]
MC11r	2.003	0.42 | [0.01, 0.84]	2.203	0.47 | [0.05, 0.88]
MC12r	-1.225	-0.26 | [-0.68, 0.16]	-1.212	-0.26 | [-0.67, 0.16]
MC13r	0.467	0.10 | [-0.32, 0.51]	0.503	0.11 | [-0.31, 0.52]
MC14r	2.099	0.45 | [0.03, 0.86]	2.980	0.63 | [0.22, 1.05]
MC15r	1.975	0.42 | [0.00, 0.83]	3.656	0.78 | [0.36, 1.19]
MC16r	1.326	0.28 | [-0.13, 0.70]	1.485	0.31 | [-0.10, 0.73]
MC17r	-0.511	-0.11 | [-0.52, 0.31]	-0.505	-0.11 | [-0.52, 0.31]
MC18r	-0.284	-0.06 | [-0.48, 0.36]	-0.282	-0.06 | [-0.48, 0.36]
MC19r	0.896	0.19 | [-0.23, 0.61]	1.118	0.24 | [-0.18, 0.65]
MC20r	0.501	0.11 | [-0.31, 0.52]	0.510	0.11 | [-0.31, 0.52]
MC21r	0.183	0.04 | [-0.38, 0.45]	0.589	0.12 | [-0.29, 0.54]
MC22r	1.500	0.32 | [-0.10, 0.73]	1.624	0.34 | [-0.07, 0.76]
MC23r	0.228	0.05 | [-0.37, 0.46]	0.227	0.05 | [-0.37, 0.46]
MC24r	-1.396	-0.30 | [-0.71, 0.12]	-1.643	-0.35 | [-0.76, 0.07]

Table A25. Spring 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 3 (Reference = Group 2, Focal = Group 3)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.580	-0.08 | [-0.37, 0.20]	0.500	0.07 | [-0.21, 0.35]
MC02r	-0.305	-0.04 | [-0.33, 0.24]	0.303	0.04 | [-0.24, 0.33]
MC03r	-20.069	-2.89 | [-3.17, -2.60]	-14.886	-2.14 | [-2.42, -1.86]
MC04r	-2.146	-0.31 | [-0.59, -0.03]	1.472	0.21 | [-0.07, 0.49]
MC05r	1.375	0.20 | [-0.08, 0.48]	-1.648	-0.24 | [-0.52, 0.04]
MC06r	-1.825	-0.26 | [-0.54, 0.02]	1.532	0.22 | [-0.06, 0.50]
MC07r	2.246	0.32 | [0.04, 0.60]	-3.925	-0.56 | [-0.85, -0.28]
MC08r	-3.753	-0.54 | [-0.82, -0.26]	-5.573	-0.80 | [-1.08, -0.52]
MC09r	0.038	0.01 | [-0.28, 0.29]	-0.038	-0.01 | [-0.29, 0.28]
MC10r	-4.214	-0.61 | [-0.89, -0.32]	1.897	0.27 | [-0.01, 0.55]
MC11r	-0.419	-0.06 | [-0.34, 0.22]	0.409	0.06 | [-0.22, 0.34]
MC12r	-2.167	-0.31 | [-0.59, -0.03]	2.089	0.30 | [0.02, 0.58]
MC13r	-3.336	-0.48 | [-0.76, -0.20]	2.932	0.42 | [0.14, 0.70]
MC14r	-2.371	-0.34 | [-0.62, -0.06]	-7.467	-1.07 | [-1.36, -0.79]
MC15r	-1.579	-0.23 | [-0.51, 0.05]	1.515	0.22 | [-0.06, 0.50]
MC16r	0.717	0.10 | [-0.18, 0.38]	-5.706	-0.82 | [-1.10, -0.54]
MC17r	0.482	0.07 | [-0.21, 0.35]	-0.526	-0.08 | [-0.36, 0.21]
MC18r	-1.712	-0.25 | [-0.53, 0.04]	-1.130	-0.16 | [-0.44, 0.12]
MC19r	-2.262	-0.33 | [-0.61, -0.04]	-2.018	-0.29 | [-0.57, -0.01]
MC20r	1.946	0.28 | [0.00, 0.56]	-6.134	-0.88 | [-1.16, -0.60]
MC21r	0.796	0.11 | [-0.17, 0.40]	-7.549	-1.09 | [-1.37, -0.80]
MC22r	3.364	0.48 | [0.20, 0.77]	-8.682	-1.25 | [-1.53, -0.97]
MC23r	-9.426	-1.36 | [-1.64, -1.07]	4.196	0.60 | [0.32, 0.89]
MC24r	-22.178	-3.19 | [-3.47, -2.91]	10.815	1.56 | [1.27, 1.84]

Table A26. Spring 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 2 and 4 (Reference = Group 2, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.146	-0.03 | [-0.41, 0.35]	-0.145	-0.03 | [-0.41, 0.35]
MC02r	-0.329	-0.06 | [-0.45, 0.32]	0.328	0.06 | [-0.32, 0.45]
MC03r	1.169	0.23 | [-0.15, 0.61]	1.145	0.22 | [-0.16, 0.61]
MC04r	-2.291	-0.45 | [-0.83, -0.06]	2.291	0.45 | [0.06, 0.83]
MC05r	2.147	0.42 | [0.04, 0.80]	-2.151	-0.42 | [-0.80, -0.04]
MC06r	0.199	0.04 | [-0.34, 0.42]	0.207	0.04 | [-0.34, 0.42]
MC07r	2.456	0.48 | [0.10, 0.86]	-2.456	-0.48 | [-0.86, -0.10]
MC08r	0.316	0.06 | [-0.32, 0.44]	0.314	0.06 | [-0.32, 0.44]
MC09r	0.039	0.01 | [-0.38, 0.39]	-0.038	-0.01 | [-0.39, 0.38]
MC10r	0.696	0.14 | [-0.25, 0.52]	0.801	0.16 | [-0.23, 0.54]
MC11r	-0.290	-0.06 | [-0.44, 0.33]	0.396	0.08 | [-0.31, 0.46]
MC12r	-1.284	-0.25 | [-0.63, 0.13]	-1.273	-0.25 | [-0.63, 0.13]
MC13r	0.504	0.10 | [-0.28, 0.48]	0.528	0.10 | [-0.28, 0.49]
MC14r	3.334	0.65 | [0.27, 1.03]	-3.334	-0.65 | [-1.03, -0.27]
MC15r	-0.335	-0.07 | [-0.45, 0.32]	2.205	0.43 | [0.05, 0.81]
MC16r	2.329	0.45 | [0.07, 0.84]	-2.329	-0.45 | [-0.84, -0.07]
MC17r	-0.291	-0.06 | [-0.44, 0.33]	-0.647	-0.13 | [-0.51, 0.26]
MC18r	-0.277	-0.05 | [-0.44, 0.33]	-0.297	-0.06 | [-0.44, 0.32]
MC19r	0.447	0.09 | [-0.30, 0.47]	-0.562	-0.11 | [-0.49, 0.27]
MC20r	0.544	0.11 | [-0.28, 0.49]	0.532	0.10 | [-0.28, 0.49]
MC21r	1.012	0.20 | [-0.19, 0.58]	0.976	0.19 | [-0.19, 0.57]
MC22r	1.907	0.37 | [-0.01, 0.76]	1.857	0.36 | [-0.02, 0.75]
MC23r	0.209	0.04 | [-0.34, 0.42]	0.210	0.04 | [-0.34, 0.42]
MC24r	-1.640	-0.32 | [-0.70, 0.06]	-1.640	-0.32 | [-0.70, 0.06]

Table A27. Spring 2015 Academic Math Assessment DIF Effect Size Measures for Achievement Groups 3 and 4 (Reference = Group 3, Focal = Group 4)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.136	-0.02 | [-0.30, 0.26]	-0.136	-0.02 | [-0.30, 0.26]
MC02r	-0.349	-0.05 | [-0.33, 0.23]	-0.349	-0.05 | [-0.33, 0.23]
MC03r	1.326	0.19 | [-0.09, 0.47]	1.326	0.19 | [-0.09, 0.47]
MC04r	-1.716	-0.24 | [-0.52, 0.03]	-1.716	-0.24 | [-0.52, 0.03]
MC05r	1.679	0.24 | [-0.04, 0.52]	1.679	0.24 | [-0.04, 0.52]
MC06r	0.209	0.03 | [-0.25, 0.31]	0.209	0.03 | [-0.25, 0.31]
MC07r	2.074	0.30 | [0.02, 0.58]	2.074	0.30 | [0.02, 0.58]
MC08r	0.320	0.05 | [-0.23, 0.33]	0.320	0.05 | [-0.23, 0.33]
MC09r	3.917	0.56 | [0.28, 0.84]	3.917	0.56 | [0.28, 0.84]
MC10r	0.742	0.11 | [-0.17, 0.39]	0.742	0.11 | [-0.17, 0.39]
MC11r	2.704	0.39 | [0.11, 0.67]	2.704	0.39 | [0.11, 0.67]
MC12r	-1.114	-0.16 | [-0.44, 0.12]	-1.114	-0.16 | [-0.44, 0.12]
MC13r	0.561	0.08 | [-0.20, 0.36]	0.561	0.08 | [-0.20, 0.36]
MC14r	3.628	0.52 | [0.24, 0.80]	3.628	0.52 | [0.24, 0.80]
MC15r	3.837	0.55 | [0.27, 0.83]	3.837	0.55 | [0.27, 0.83]
MC16r	2.252	0.32 | [0.04, 0.60]	2.252	0.32 | [0.04, 0.60]
MC17r	-0.482	-0.07 | [-0.35, 0.21]	-0.482	-0.07 | [-0.35, 0.21]
MC18r	-0.273	-0.04 | [-0.32, 0.24]	-0.273	-0.04 | [-0.32, 0.24]
MC19r	0.975	0.14 | [-0.14, 0.42]	0.975	0.14 | [-0.14, 0.42]
MC20r	0.532	0.08 | [-0.20, 0.36]	0.532	0.08 | [-0.20, 0.36]
MC21r	1.004	0.14 | [-0.14, 0.42]	1.004	0.14 | [-0.14, 0.42]
MC22r	1.743	0.25 | [-0.03, 0.53]	1.743	0.25 | [-0.03, 0.53]
MC23r	0.215	0.03 | [-0.25, 0.31]	0.215	0.03 | [-0.25, 0.31]
MC24r	-1.374	-0.20 | [-0.48, 0.08]	-1.374	-0.20 | [-0.48, 0.08]

Table A28. Spring 2015 Academic Math Assessment DIF Effect Size Measures for Gender Groups (Reference = Males, Focal = Females)

Item	Z(SA)	d (95% CI)	Z (UA)	d (95% CI)
MC01r	-0.137	-0.02 | [-0.29, 0.25]	-0.137	-0.02 | [-0.29, 0.25]
MC02r	-0.200	-0.03 | [-0.30, 0.24]	0.230	0.03 | [-0.24, 0.30]
MC03r	1.296	0.18 | [-0.09, 0.45]	1.313	0.18 | [-0.09, 0.45]
MC04r	-1.792	-0.25 | [-0.51, 0.02]	-1.775	-0.24 | [-0.51, 0.03]
MC05r	1.667	0.23 | [-0.04, 0.50]	1.688	0.23 | [-0.04, 0.50]
MC06r	0.209	0.03 | [-0.24, 0.30]	0.209	0.03 | [-0.24, 0.30]
MC07r	2.030	0.28 | [0.01, 0.55]	2.059	0.28 | [0.01, 0.55]
MC08r	0.316	0.04 | [-0.23, 0.31]	0.317	0.04 | [-0.23, 0.31]
MC09r	3.697	0.51 | [0.24, 0.78]	3.923	0.54 | [0.27, 0.81]
MC10r	0.714	0.10 | [-0.17, 0.37]	0.716	0.10 | [-0.17, 0.37]
MC11r	2.727	0.37 | [0.11, 0.64]	2.819	0.39 | [0.12, 0.65]
MC12r	-1.112	-0.15 | [-0.42, 0.12]	-1.105	-0.15 | [-0.42, 0.12]
MC13r	0.558	0.08 | [-0.19, 0.34]	0.560	0.08 | [-0.19, 0.35]
MC14r	3.392	0.46 | [0.20, 0.73]	3.474	0.48 | [0.21, 0.74]
MC15r	3.845	0.53 | [0.26, 0.80]	4.230	0.58 | [0.31, 0.85]
MC16r	2.106	0.29 | [0.02, 0.56]	2.123	0.29 | [0.02, 0.56]
MC17r	-0.484	-0.07 | [-0.33, 0.20]	-0.482	-0.07 | [-0.33, 0.20]
MC18r	-0.275	-0.04 | [-0.31, 0.23]	-0.275	-0.04 | [-0.31, 0.23]
MC19r	0.900	0.12 | [-0.15, 0.39]	0.902	0.12 | [-0.14, 0.39]
MC20r	0.528	0.07 | [-0.20, 0.34]	0.531	0.07 | [-0.20, 0.34]
MC21r	0.999	0.14 | [-0.13, 0.41]	1.014	0.14 | [-0.13, 0.41]
MC22r	1.766	0.24 | [-0.03, 0.51]	1.812	0.25 | [-0.02, 0.52]
MC23r	0.213	0.03 | [-0.24, 0.30]	0.213	0.03 | [-0.24, 0.30]
MC24r	-1.495	-0.20 | [-0.47, 0.06]	-1.492	-0.20 | [-0.47, 0.06]

REFERENCES

REFERENCES

Bollen, Kenneth A. & Bauldry, Shawn (2015). *Indicator: Methodology.* In Wright, James D. (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 750-4). Oxford, England: Elsevier Press.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *J. Pers. Soc. Psychol. 95*, 1005–1018. doi: 10.1037/a0013193

Clark, M. J., & Grandy, J. (1984). *Sex differences in the academic performance of SAT takers.* (College Board Report No. 84-8). New York: College Board.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

Cook, L.L., Eignor, D.R., Burton, E.B. (1990*). Aligning Score Scales for Achievement Tests in Multiple Content Areas.* Educational Testing Service: Princeton, NJ.

Crawford, P. L., Alferink, D. M., & Spencer, J. L. (1986). *Postdictions of college GPAs from ACT composite scores and high school GPAs: Comparisons by race and gender.* West Virginia State College (ERIC Document Reproduction Service No. ED 326 541).

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Toronto: Holt, Rinehart, and Winston, Inc.

Cureton, E. (1966). Corrected item-test correlations. *Psychometrika, 31*(1), 93-96.

De Ayala, R.J. (2009) *The theory and practice of Item Response Theory*. New York: The Guilford Press.

Diones, R., Bejar, I.I., Chaffin, R. (1996). The dimensionality of responses to SAT analogy items. *ETS Research Report Series,* 1-31.

Dorans, N. J., Lawrence, I. M. (1987). *The internal construct validity of the SAT.* Research Report. (Educational Testing Service, Princeton, NJ).

Education Quality and Accountability Office. (2009). *Framework Grade 9 Assessment of Mathematics*. Toronto, ON: EQAO.

Education Quality and Accountability Office. (2017). *EQAO's Technical Report for the 2014-2105 Assessments*. Toronto, ON: EQAO.

El Ghaziri, A., & Qannari, E.M. (2015). Measures of association between two datasets: application to sensory data. *Food Quality and Preference, 40* (A), 116-124.

Gierl, M., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying Content and Cognitive Skills That Produce Gender Differences in Mathematics: A Demonstration of the Multidimensionality-Based DIF Analysis Paradigm. *Journal of Educational Measurement, 40*(4), 281-306. Retrieved from http://www.jstor.org.proxy2.cl.msu.edu/stable/1435383

Gierl, M. J., Tan, X., Wang, C. (2005). *Identifying Content and Cognitive Dimensions on the SAT*. The College Board: New York.

Gierl, M., Khaliq, S. (2001). Identifying Sources of Differential Item and Bundle Functioning on Translated Achievement Tests: A Confirmatory Analysis. *Journal of Educational Measurement, 38*(2), 164-187. Retrieved from http://www.jstor.org.proxy2.cl.msu.edu/stable/1435261

Gregorich S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical care, 44*(11), 78 - 94. Retrieved from: https://doi.org/10.1097/01.mlr.0000245454.12228.8f

Hirschfeld, G., von Brachel, R. (2014) "Improving Multiple-Group confirmatory factor analysis in R – A tutorial in measurement invariance with continuous and ordinal indicators," *Practical Assessment, Research, and Evaluation*,*19*, Article 7. DOI: https://doi.org/10.7275/qazy-2946

Hogrebe, M. C., Ervin, L., Dwinell, P. L., & Newman, (1983). The moderating effects of gender and race in predicting the academic performance of college developmental students. *Educational and Psychological Measurement, 43,*523–530.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. https://doi.org/10.1037/0033-2909.112.3.527

Kaplan, D. (2001). *International Encyclopedia of the Social & Behavioral Sciences.* Retrieved from https://www.sciencedirect.com/topics/computer-science/multivariate-normality

Kunnan, A. (2010). Test fairness and Toulmin's argument structure. *Language Testing, 27*(2), 183-189.

Kupermintz, H., Ennis, M., Hamilton, L., Talbert, J., & Snow, R. (1995). Enhancing the Validity and Usefulness of Large-Scale Educational Assessments: I. NELS:88 Mathematics Achievement. *American Educational Research Journal, 32*(3), 525-554. Retrieved from http://www.jstor.org.proxy1.cl.msu.edu/stable/1163323

Kupermintz, H., & Snow, R. E. (1997). Enhancing the validity and usefulness of large-scale educational assessments: III. NELS:88 mathematics achievement to 12th grade. *American Educational Research Journal, 34*(1), 124–150. https://doi.org/10.2307/1163344

Larson, J. R., & Scontrino, M. P. (1976). The consistency of high school grade point average and of the verbal and mathematical portions of the Scholastic Aptitude Test of the College Entrance Examination Board, as predictors of college performance: An eight-year study. *Educational and Psychological Measurement, 36*, 439–443.

Little, R. J. A., & Rubin, D. B. (*2002*). *Statistical analysis with missing data* (2nd ed.). New York, NY: John Wiley & Sons.

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

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Meredith, W., Teresi, J. (2006). An essay on measurement and factorial invariance. *Medical Care, 44*(3), S69 –S77. doi: 10.1097/01.mlr.0000245438.73837.89

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

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5th ed.). Upper Saddle River, N.J.: Prentice-Hall.

National Research Council. (2011). *Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics.* Washington, DC: The National Academies Press. Retrieved from https://www.ltrr.arizona.edu/webhome/sheppard/TUSD/NRC2011.pdf

Noble, J., Crouse, J., & Schulz, M. (1996). *Differential prediction/impact on course placement for ethnic and gender groups* (Research Report No. 96-8). Iowa City, IA: American College Testing.

Ontario Ministry of Education. (2009). *Ontario's Equity and Inclusive Education Strategy:2009.* Retrieved from http://www.edu.gov.on.ca/eng/policyfunding/equity.pdf

Ontario Ministry of Education. (2005). *The Ontario curriculum grades 9 and 10: mathematics*. (ISBN 0-7794-7940-80). Toronto, ON: Queen's Printer for Ontario.

Organization of Economic Co-operation and Development. (2002). *Education at a glance 2002*. Retrieved from http://www.oecd.org/education/skills-beyond-school/educationataglance2002-home.htm

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report No. 93-1). New York: College Board.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–16. Retrieved from: https://doi.org/10.1111/j.1745-3992.1998.tb00827.x

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25–36.

Rowan, R. W. (1978). The predictive value of the ACT at Murray State University over a four-year college program. *Measurement and Evaluation in Guidance, 11*, 143–149.

Saka, T. T. (1991). High school GPA, SAT scores and college academic achievement for University of Hawaii freshmen. *Pacific Educational Research Journal, 7*, 19 –32.

Schmidt, W.H. (2011). *STEM reform: Which way to go?* Paper presented at the National Research Council Workshop on Successful STEM Education in K-12 Schools. Available at: http://www7.nationalacademies.org/bose/STEM_Schools_Workshop_Paper_Schmidt.pdf

Sloley, R. W. (1931). Primitive Methods of Measuring Time: With Special Reference to Egypt.

The Journal of Egyptian Archaeology, 17(3/4), 166-178. doi:10.2307/3854758

Timmons, A.C. (2010). *Establishing factorial invariance for multiple-group confirmatory factor analysis*. KUant Guides: Guide 22.1.

UNESCO (2008). *Overcoming inequality: why governance matters; EFA global monitoring report, 2009.* Paris: UNESCO. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000177683

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69. doi:10.1177/109442810031002

Viger, S.G. (2014). *Measurement invariance of a summative achievement assessment over time: is status really ready for growth?* (Doctoral dissertation). Available from ProQuest Dissertations & Theses Global database. (UMI No: 3617818).

Wainer, H., Saka, T., & Donoghue, J. R. (1993). Notes: The Validity of the SAT at the University of Hawaii: A Riddle Wrapped in an Enigma. *Educational Evaluation and Policy Analysis*, *15*(1), 91–98. https://doi.org/10.3102/01623737015001091

Wainer, H., & Steinberg, L. S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review, 62*(3), 323–336. https://doi.org/10.17763/haer.62.3.1p1555011301r133

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38*(2), 147-163.

Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year* (College Board Report No. 83– 2 and Educational Testing Service Research Report No. 83– 11). New York: College Board.

Yang, F.M., & Kao, S.T. (2014). Item response theory for measurement validity. *Shangai Archieves of Psychiatry 26*(3), 171-177. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4118016/

Young, J.W., & Kobrin, J.L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (College Board Research Report No. 2001-6). New York: The College Board.