

**MEASUREMENT INVARIANCE OF A SUMMATIVE ACHIEVEMENT ASSESSMENT
OVER TIME: IS STATUS REALLY READY FOR GROWTH?**

By

Steven Guy Viger

A DISSERTATION

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

Measurement and Quantitative Methods – Doctor of Philosophy

2014

ABSTRACT

MEASUREMENT INVARIANCE OF A SUMMATIVE ACHIEVEMENT ASSESSMENT OVER TIME: IS STATUS REALLY READY FOR GROWTH?

By

Steven Guy Viger

The current study investigates the phenomenon of measurement invariance by examining the construct stability of a summative mathematics achievement instrument over time gleaned from an existing data set. In doing so, not only is the general question of measurement invariance of the particular instrument addressed, but also in the context of growth studies. The onus of the study as well as the results are presented in light of the current political context of large scale K-12 assessment and the shifting of emphasis from status to growth. As the reader will discover, great pressure is placed on results not necessarily intended to serve as the metric required by policy. The results and implications are framed in both measurement and practical contexts.

The final product of this tremendous journey is dedicated to my incredible family, those still present and those who have left us. Without your support, motivation, love and patience, the finish line might have continued to elude me.

To Richard S. Viger and Richard W. Viger...I finished the game; thank you for pushing me across the goal line.

ACKNOWLEDGMENTS

A heartfelt thank you goes out to the faculty and staff of the Measurement and Quantitative Methods program; especially Drs. Mark Reckase and Kimberly Maier. The wisdom you've shared, the patience, the understanding and the dedication to producing quality students and brilliant researchers forced me to continuously raise my own bar to meet your high standards. Your support was sometimes of the tough love variety, but was never absent. As a result of this I pushed myself further than I thought possible and I now carry the same high standards to my work and life.

Thank you to the Michigan Department of Education for providing access to data and for support in many other areas. They are an organization dedicated to fulfilling policy with the upmost integrity. I've never worked with a better group of professionals.

To my wife, Andrea Jackson, you have always been my biggest fan, loudest cheerleader and most stubborn supporter. From the moment we met I shared with you my goals and not once did you show the slightest bit of doubt in my ability to follow through. It is because of you I was able to press on and keep my eyes on the prize.

I'm also proud to acknowledge and pay tribute to the critical role my children Caitlyn, Casey and Chase Viger played in the completion of this major goal. Even when you didn't understand what I was doing you never complained. Whenever I needed a hug or to get my mind off of my studies you provided the ultimate escape. Perhaps most importantly, I couldn't think of a greater motivation to finish my goals then to provide you with a working example of what hard

work and dedication to your dreams will lead to. Never wanting to fail your children is the greatest motivation any parent can have.

To the rest of my family...you never pushed, never rushed, never asked too many questions or inquired as to 'when' I'd be done. You just always assumed I would be done at some point. I'm so glad that time is now and that I could share my success with you all. You helped make me the man and the scholar I have become. I'm eternally grateful for my loving family. What a great example of the right way to do things and love each other.

Finally, I'd like to acknowledge the support of my friends. Some of you have motivated me just by being who you are while others have directly pushed and engaged me to finish my goals. Regardless of your style, you knew what made me move and used that to urge me along. Sometimes that meant a 'normal' conversation, sometimes a hug, or even going out of your way to give me private access to a place where I can get my work done in peace. All of my friends fit the bill but I need to specifically acknowledge Ryan Newton, Colleen Kelly, and Rosalie Kern for their tremendous support through thick and thin.

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
Chapter 1: Introduction	1
Research Questions	3
Chapter 2: Literature and Policy Review	5
A Paradigm for the Discussion of Validity Evidence.....	5
Policy Changes as an Influence on the Validity Argument	6
Measurement Invariance	10
Factor Analytic Strategies to Determining Measurement Invariance	11
The Standard Invariance Model Testing Sequence	17
Configural Invariance	18
Metric Invariance	19
Scalar Invariance	19
Strict Invariance	20
Outcome Measures: Fit Indices	20
Absolute Fit	22
Descriptive Fit	24
Comparative Fit Indices	25
Akaike Information Criterion (AIC)	25
Bayesian Information Criterion (BIC)	25
The Sample-Size Adjusted BIC (SABIC)	26
Construct Stability, Measurement Invariance and Validity Evidence	26
Content Based Evidence	28
Factorial Evidence	29
Chapter 3: Method	33
Data	33
Instrument	35
Analysis	38
Chapter 4: Results	45
Descriptive Statistics/Previous Achievement	45
Confirmatory Factor Analysis of One and Four Factor Models by Sample	46
Measurement Invariance Tests of One and Four Factor Models	48
Chapter 5: Discussion	50
Related to the original Fall 2009 administration and applied across groups:	
Does the Rasch model fit the data	50

Do the data fit the linear confirmatory model implied by the blueprint (content strands) for the test?.....	51
Does one of the models fit significantly better than the other model?	51
Do the aforementioned models exhibit measurement invariance across groups/study conditions?	52
Implications	52
Chapter 6: Limitations and Future Research	58
APPENDICES	60
Appendix A – Scale Score Distributions (Pre and Post) for each of the Study Groups ...	61
Appendix B – Item Characteristic Curves by Group	74
Appendix C – IRT Calibration Values by Group.....	77
REFERENCES	80

LIST OF TABLES

Table 1 – Sample Demographic Characteristics	35
Table 2 – Assessment breakdown by strands (numerals are item numbers)	36
Table 3 – Model 1, Configural Invariance	41
Table 4 – Model 1, Metric Invariance	41
Table 5 – Model 2, Configural Invariance	42
Table 6 – Model 2, Metric Invariance	42
Table 7 – Michigan MEAP Transition Table	44
Table 8 – Previous Performance (Mean Scale Score).....	46
Table 9 – Previous Performance (Percent Proficient).....	46
Table 10 – Group level Model Fit (Single Factor Model)	47
Table 11 – Group level Model Fit (Blueprint Based/Four Factor Model)	47
Table 12 – Measurement Invariance Study Results.....	49
Table 13 – Outcome Variable Group Differences	53
Table 14 – IRT Calibration Values by Group	78

LIST OF FIGURES

Figure 1 – Performance Level Change by Study Condition	54
Figure 2 – Multigroup IRT Test Characteristic Curves	56
Figure 3 – Test Information Functions from Multiple Group IRT Run	57
Figure 4 – Post-test Scale Score Distribution (Grade 8)	62
Figure 5 – Post-test Performance Level Frequencies (Grade 8)	63
Figure 6 – Pre-test Scale Score Distribution (Grade 8)	64
Figure 7 – Pre-test Performance Level Frequencies (Grade 8)	65
Figure 8 – Post-test Scale Score Distribution (Grade 9)	66
Figure 9 – Post-test Performance Level Frequencies (Grade 9)	67
Figure 10 – Pre-test Scale Score Distribution (Grade 9)	68
Figure 11 – Pre-test Performance Level Frequencies (Grade 9)	69
Figure 12 – Post-test Scale Score Distribution (Grade 10)	70
Figure 13 – Post-test Performance Level Frequencies (Grade 10)	71
Figure 14 – Pre-test Scale Score Distribution (Grade 10)	72
Figure 15 – Pre-test Performance Level Frequencies (Grade 10)	73
Figure 16 – Item Characteristic Curves by Group	75

Chapter 1: Introduction

In standards based assessments, such as a State Educational Agency's (SEA) summative K-12 achievement test, items are indicators of content standards that as a whole, are used as a mechanism to place students on a construct (or constructs) underlying continuum. Further, when scores are reported they are often transformed to the desired reporting scale. While these metrics vary widely, all are able to maintain students' relative standing to others as well as to criterion referenced cut-scores.

The author contends that it is extremely difficult to truly assess student growth as the students are changing cognitively, physically, and emotionally as a function of their development in ways that are not often measured, are difficult to measure, or in the least we have no direct data to link to. At the same time, the instruments are often changing as the students proceed with their schooling due to a constantly changing set of content standards and performance standards. Consequently, the interaction of persons and items that forms the foundation of modern scaling is modeled in the presence of often shifting sets of items and persons with naturally changing expectations or criteria.

Taking more of a purist approach suggests looking at what can and cannot be under our control as we seek to evaluate construct stability. It seems reasonable that as we are not able to control the changes of the people over time we can see how they change over time by holding constant the instrument with its intended underlying construct(s). To link back to the concept of internal structure, by allowing time to pass and hence for the students to develop by holding the actual instrument constant, also affords the opportunity to see how what is intended to be measured may or may not change over time. The results of such a factorial/structure evidence analysis can certainly inform whether or not changes in scores over time on a constant

instrument are the function of maturation and/or a change in underlying construct (possibly represented by changes in the pattern of inter-item correlations on the instrument). Factor analysis is one such analytical vehicle to use that will provide useful information in making such decisions.

In the current study, there is an underlying scale purported to measure a specific intended construct. The scale is made up of multiple items (or subscales, or multi-item parcels). In factor analytic terms, the items serve as indicators of the trait or factor in a common factor model. The author makes use of this scale in samples from distinct populations: the original sample and the three samples in which later data were collected. For any such use of scale scores, there is a critical assumption that the scale is measuring the same trait in all of the groups. If that assumption holds, then comparisons and analyses of those scores are acceptable and yield meaningful interpretations. But if that assumption is not true, then such comparisons and analyses do not yield meaningful results. When constructs shift across grades, such as when mathematics assessments move from testing arithmetic skills in third grade to testing pre-algebra and geometry skills in later grades, the growth model results may lead to imprecise longitudinal interpretations (Reckase 2004; Martineau 2006).

To this end, this study leverages confirmatory factor analysis techniques as well as the literature around the concept of measurement invariance to examine the factorial stability of a mathematics achievement test given to a sample of students one instructional year following the intended time of testing, two years after the intended time of testing and three years after the intended time of testing to determine to what degree measurement invariance over this cross-section of students based on grade remains the same. To the extent that the structure holds over time, this supports the ability to glean similarly interpreted growth data by use of a parallel form

of an assessment in a pre-test/post-test paradigm for growth. To the extent that the structure does not hold over time, this suggests an unintended relationship of other variables to the construct being measured if the paradigm is a simple gains (pre-test/post-test) type of approach. That is to say, that the development and everything occurring in the passage of time that differentiates the cross-sections of students, is also related to the achievement which would invalidate the instrument for the intended use in a gain score approach as it no longer measures the same construct and is not able to be scaled together in a meaningful way. Or at least, as scaled (likely horizontally), the intended inferences would not be supported.

The current study will address the following research questions within a measurement invariance paradigm driven by factor analytic strategies.

Research Questions

1. Related to the original Fall 2009 administration and applied across groups:
 - a. Does the Rasch model fit the data?
 - b. Do the data fit the linear confirmatory model implied by the blueprint (content strands) for the test?
 - c. Does one of the models in a. and b. fit significantly better than the other model?
2. Are the measurement models posited in Question 1 invariant to additional years of instruction?
 - a. Do the measurement models (unidimensional and multidimensional) exhibit configural invariance across groups such that both groups associate the same subsets of items with the same constructs?

- b. Do the measurement models (unidimensional and multidimensional) exhibit metric invariance across groups, indicating that overall, the strength of the relationships between items and their underlying constructs are the same for both groups?
- c. Do the measurement models (unidimensional and multidimensional) hold to the property of strict invariance across groups, suggesting that factor patterns, loadings, intercepts and residual variances are equal across groups?
- d. Does the comparative fit of both the unidimensional and multidimensional change across groups?

The next chapter will present both a review of the literature that speaks more to the motivation for the current study as well as delving into literature around the particular method and measurement paradigm explored. Taken as a whole, the notion of invariance or measurement stability over time speaks directly to the validity of inferences one can support. As such, the way in which a study such as this fits into validity arguments will be discussed.

Chapter 2: Literature and Policy Review

In this chapter, a brief introduction to validity evidence and the support of inferences will be provided to serve as a framework from which the criticality of this study, and others like it, can be deduced when considered in tandem with broad sweeping K-12 assessment policies. Put differently, the concept of validity is presented first to present the context of the evaluation and is followed by a review of policy and statutory changes brought into place which were created independently of the research literature and in some cases independent of AERA/NCME/APA standards. Once that context has been presented, a review of the literature pertinent to the study of measurement invariance as well as the confirmatory factor analysis strategy used to investigate the measurement invariance phenomenon will be provided.

A Paradigm for the Discussion of Validity Evidence

Generally speaking, validity refers to “the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests.” (Messick, 1989, 1994, 1995). In practical terms, validity can be used to describe how well one can legitimately trust the results of a test as interpreted for a specific purpose. In the world of operational psychometrics, it is the “specific purpose” portion of that definition that tends to vary from location to location, assessment to assessment...proposed use to proposed use. It logically follows that validity is a property of inferences, not instruments. As a result, validity must be established for each intended interpretation. It is because of this philosophical framework, which the author endorses, that it also becomes problematic to think of an instrument as valid or not. Validity is not a property of an instrument, it is a property of the inference one makes from the data produced by the instrument (Kane, 2006). As a result, each intended use must be supported by an accumulation of evidence suggesting that the instrument, and scaling/scoring/reporting

mechanism, is valid for that intended use and subsequent inferences made as the result. Shepard (1997) argues that intended effects and likely side effects are clearly within the responsibility of the test developer. Furthermore, persistent unanticipated effects are also the responsibility of the test developer. Moss (1998) suggests greater responsibility for the test developer and argues that considerations of test consequences should encompass the anticipated uses of test scores. In other words, test developers are obligated to attempt to maximize positive consequences and minimize negative consequences. Further, test developers should consider the consequences of testing in general rather than the immediate consequences of using scores from a specific test. For example, Moss argues that testing is reactive with test takers and test users. The administration of a test in a school changes the school, whether information from scores are intentionally used or ignored. How they are used, is likely driven more by policy than by proper measurement and psychometric considerations. As already alluded to, it is the responsibility of the test developer to be proactive in considering the immediate intended uses but also perhaps the more forward thinking unintended uses and/or consequences.

Policy Changes as an Influence on the Validity Argument

A November, 2005 announcement by the United States Department of Education (USED, 2005) encouraged states to propose pilot programs for growth-based accountability models for use in the 2005–2006 and 2006–2007 school years. Seven requirements for the pilot programs were given, with the first three viewed largely as alignment elements and the last four considered foundational elements. The alignment elements were as follows:

1. The accountability model must ensure that all students are proficient by 2013–14 and set annual goals to ensure that the achievement gap is closing for all groups of students.
2. The accountability model must not set expectations for annual achievement based upon

student background and school characteristics.

3. The accountability model must hold schools accountable for student achievement in reading/language arts and mathematics.

The foundational elements covered:

4. The accountability model must ensure that all students in the tested grades are included in the assessment and accountability system. Schools and districts must be held accountable for the performance of student subgroups. The accountability model includes all schools and districts.
5. The state's assessment system, the basis for the accountability model, must receive approval through the NCLB (No Child Left Behind) peer review process for the 2005–2006 school year. In addition, the full NCLB assessment system in grades 3–8 and in high school in reading/language arts and mathematics must have been in place for two testing cycles.
6. The accountability model and related state data system must track student progress.
7. The accountability model must include student participation rates in the state assessment system and student achievement on an additional academic indicator.

Following review, the USED proposal review team published a document summarizing cross-cutting issues that influenced their decisions to approve or deny states' proposals (USED, 2006). In particular, the guidance document indicated that states shall: (a) incorporate available years of existing achievement data, instead of relying on only two years of data; (b) align growth timeframes with school grade configuration and district enrollment; (c) make growth projections for all students, not just those below proficient; (d) hold schools accountable for the same subgroups as under the status model; (e) not use wide confidence intervals; (f) not reset growth

targets each year and (g) not average scores between proficient and non-proficient students. Although these issues were noted as influential in the peer review group's decisions, not all proposals approved through the growth model pilot peer review process met all of these conditions (CCSSO Accountability Systems and Reporting Working Group, 2009).

Figuring out a paradigm to address all of those issues is challenging at best and each state faces their own unique data idiosyncrasies. Approaches proposed varied from simple pre-test/post-test designs to elaborate vertical scaling designs, student growth percentiles and other projection methods and regression-based approaches as well as standards-based transition tables (Castellanos and Ho, 2012). All of these have strengths and weaknesses some of which take a toll monetarily and on the human resource side in that some require testing above and beyond the current status measures.

The findings of the CCSSO working group also suggest that oversight over these practices is not necessarily as tight as it could be and if this somewhat loose approach trickles down to state agencies, who are naturally reactionary to such policy changes and initiatives, then there might be many assumptions being made that could potentially go unchecked and lead to distortions in interpretations of the results down the road.

There is also a parallel push to make assessments more instructionally useful and relevant so that the content specifications that feed both the curriculum and assessment paths can be in-sync. The consequences of test score use take on increasing importance in the current era in which educators are attempting to leverage the information in test scores to improve student learning (Perie, Marion & Gong, 2007). Since then, states have come to realize that the proficiency requirements of 2013-2014 are not likely to be realized and have submitted waiver applications to absolve themselves of those requirements but with the caveat that while the proficiency

markers might be able to be reset or relaxed, an enhanced focus on growth must take place that gives credit to all students, not just those bordering the proficiency marker.

As a result, it should not come as a surprise that growth modeling is a huge topic right now in the psychometric and educational measurement literature and is occupying much of the reworking of ESEA and waiver applications. Compared with the original uses and purposes of test scores laid out in the NCLB of 2001, the amount of utility and information attempting to be gleaned from a measure intended to measure proficiency is enormous.

While some may argue that anticipating the intended uses is not the sole responsibility of the test developer (Reckase, 1998), the CCSSO working group pointed out the obvious that the burden does trickle down to the State agencies who are the responsible party for the content of the assessment. This is important because common growth definitions include the necessity of at least two substantively and statistically comparable measures of status to deduce anything meaningful from the change or difference in measures over time. Therefore, the idea of assuming a stable construct, without actually confirming that assumption, has the potential for serious implications if problems are present that invalidate the assumptions around the assessments and the use of the scores.

The use of transition tables based on horizontally scaled, yet with underlying vertically articulated performance standards, offer an alternative way to assess whether or not students are on track towards standards based proficiency. Unfortunately, since the construct is assumed to not be stable over time, the underlying measures are not useful in determining why a student is no longer (or is now) proficient. The former case is likely to be the target of intervention or instruction while the latter is useful in helping in determining ‘what works’.

A stable construct measured from indicators from a common domain can naturally lead to a range of outcomes that are comparable. Therefore, when changes over time are noted, it is actually feasible to drill into the measure a bit to determine where the differences occurred. However, before going down that path it is important that the proper housekeeping has occurred with respect to the measurements. First and foremost, we need a way of ensuring that we are measuring the same thing, or a stable construct, over time before we make inferences over time. This becomes an issue of measurement invariance over time.

Measurement Invariance

Mellenburgh (1989), Meredith (1993), and Meredith and Millsap (1992) provided a statistical definition of measurement invariance (MI) in which an observed score is said to be measurement invariant if a person's probability of an observed score does not depend on his/her group membership, conditional on the true score. That is, respondents from different groups, but with the same true score, will have the same observed score. Or, given a person's true score, knowing a person's group membership does not alter the person's probability of getting a specific observed score. (Wu, Li, and Zumbo, 2006). As such, measurement invariance is a rather blanket term that is used to refer to several different phenomenon. From a mechanical standpoint, measurement invariance can refer to the invariance of factor loadings, intercepts, or errors (Meredith, 1993). Unfortunately, in most large scale assessments the concept of invariance is much more of an assumption than it is a quality of measurement to be empirically investigated.

As comparability and bridging studies emerge due to the shifting of assessment modalities from paper/pencil to a digital environment and status measures and interim assessments (purposed as learning tools and feedback mechanisms) are used in the development of proxy measures of growth, the assumptions of invariance over subgroups becomes an

important issue that can become an assumption rather than empirical evidence. As mentioned, a popular methodology is to use factor analytic methods to determine whether or not structures hold across groups. Confirmatory factor analysis is usually preferred because clearly the purpose of an instrument is to measure something and we really should know what we wish to measure before the administration. How we define that something is not always clear nor is the construct and hence is not easily articulated. Psychological batteries are often developed using factor analysis paradigms applied over multiple responses and potential indicators of an underlying construct. The factor analytic strategies can be used for the purpose of variable reduction and clarification to get at the most salient indicators of the hypothesized constructs.

Factor Analytic Strategies to Determining Measurement Invariance

A common paradigm to investigate measurement invariance is in the context of factor analysis with applications of both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) often seen in the literature. Factor analysis can inform score validity as well as to help understand the theoretical nature of constructs (Thompson, 2009). A major use is then in the development of the operational construct and the operational representativeness of the theoretical constructs (Gorsuch, 1983). Taking a more exploratory approach lends itself well to helping to understand the theoretical nature of the construct in relation to the data. Similarly, Muthen and Muthen (2009) suggest that EFA can be used to explore the dimensionality of a measurement instrument by determining the smallest number of interpretable factors needed to explain the correlation matrix among a set of observed, or measured, variables.

As was previously alluded to, there are two discrete classes of factor analysis techniques, exploratory and confirmatory approaches (Thompson, 2009). EFA approaches are from the family of analytic techniques attributed to Spearman (1904) and are typically applied when the

researcher has little or no *a priori* expectations regarding either the number or nature of the latent variables or factors underlying a measurement instrument. Although it is typical when there is a strong hypothesis concerning the structure of the measurement to use a CFA approach, there is no requirement to declare that model in EFA as the analysis does not require, nor allow for, these expectations to enter into the calculations. This is not entirely true though as some programs require that you specify the number of factors and by specifying a given rotation method one is allowing (or not) the underlying factors to be correlated. However, if permitted, EFA programs will often extract as many factors as there are indicators. It is then up to the analyst to perform rotations and seek guidance in the interpreting the factors before deciding what the model actually demonstrated.

Of course, a major distinction is that with the CFA approach the researcher has already declared the number of factors as well as the relationships between the observed indicators and the underlying factors. Commonly referenced to Jöreskog (1971), CFA models require that the researcher provide specific direction with regards to the number of latent variables/factors, the relationships of the measured variables (i.e. items) to those latent variables, and the degree to which the latent variables are correlated. Put simply, researchers without a theory regarding the underlying structure of an instrument cannot use CFA techniques as they have nothing to confirm (the big ‘C’ in confirmatory factor analysis). However, it doesn’t preclude researchers with theories from resorting to exploratory techniques should the theories not pan out as intended. Such a practice could lead to capitalizing on chance in that all possible combinations could potentially be worked out until the best fitting model is brought to light. These concerns are valid in that the rotations come up with infinite solutions that account for the same variance/covariance matrix yet can differ greatly in interpretation. As a result, one’s ability to

interpret the construct depends not only on a strong understanding of the content (or access to someone who has that!) but also on their point of view on the space occupied by the factor solutions. With so many possible loading patterns, many possibilities will be seen as propitious but nothing speaks to whether or not the model is ‘correct’ as there is no comparison in EFA. CFA has that power in that not only is there model fit indices based on absolute criteria but also those that are based on comparative fit.

Factor interpretation is a difficult and somewhat subjective endeavor. In the context of a principal component analyses or orthogonally rotated factor analysis solutions, the goal is to determine a set of factors where the loadings are strong for some indicators and near zero for the rest, explicitly disallowing the presence of cross loadings. The strict requirement of zero cross-loadings in CFA has come under scrutiny because this requirement often does not fit the data well and has leads to a tendency to rely on the extensive use of model modification indices to find a well-fitting model (Asparhouv & Muthen, 2012). Browne (2001) suggests that in such cases, searching for a well-fitting measurement model may be better carried out by EFA in that all of the possible models are simultaneously tested rather than those specified by the researcher and subsequently tweaked using univariate modification indices which give projections of model fit if variables are removed from the analysis. While it is true that an inherent weakness of exploratory approaches is that they tend to capitalize on chance in creating factors based sometimes on the weakest of correlations that are considered large due to sample size, really the danger lies in putting too much faith in factors that, regardless of rotation methods, often leaves factors difficult to interpret substantively (Thompson, 2009).

In the CFA framework proposed models may not be supported by much empirical evidence that would back-up their selection in the first place; that is their interpretation related

back to the original theory may be equally difficult to interpret in relations to the theoretical model and empirical research questions. Furthermore, there is no direct index of which is the “correct model” and in many instances one could come up with a model of good fit, perhaps better than already published research, which is of little interpretive value and more importantly fails to serve as content evidence within the validity argument for use of the scores as status and/or growth measures.

With this said, carefully designed assessments often adhere to a strict assessment blueprint which is driven heavily by grade level content standards or curriculum standards. As such, there is a pre-determined structure to the measurement instrument that is implied by the test design and table of specifications. When such structures are imposed on the data, confirmatory factor analysis (CFA) is the form of the factor analytical model that is most appropriate. In invoking this strategy, the covariation among manifest indicators is examined in order to confirm the hypothesized underlying latent constructs, as specified in advance by the researcher and supported substantively by the literature. CFA is a theory driven technique in which the researcher specifies (1) the number of factors and their inter-correlation, (2) which items load on which factor and (3) whether errors are correlated. Statistical tests can then be conducted to determine whether the data confirm the theoretical model; thus the model is thought of as confirmatory (Bollen, 1989). A powerful aspect of CFA, leveraged in the context of this study, is that a researcher is able to simultaneously conduct multiple group analyses across time or samples, in order to evaluate measurement invariance/equivalence across those groups.

Issues of MI are relevant in longitudinal research and growth studies. When a scale is administered over repeated occasions to the same sample of people or cross sectionally, the question of MI involves the issue of whether the scale is measuring the same construct at

different occasions. In traditional *validity* studies, it is common practice to assume that as students mature cognitively, their scores on a given instrument should increase as a function of age. That difference in mean scores is definitely of interest, but even the finding that scores do significantly increase as a function of age (grade level in the present study) still does not mean that the construct is the same and is reflected the same in that underlying scale score. Any discussion of growth based on a repeated measure or parallel forms paradigm, in the absence of a vertical scale, should be done with caution as to not infer too much. Whatever growth, or lack of, does occur is certainly attributable to multiple factors. Many of them lack good measurement or data. So, approaching the validity argument from the aspect of structural or factorial validity is another way to get at the appropriateness of the inference before it is made.

A central principle of MI is that measures across groups are considered to be on the same scale if relationships between the indicators and the trait are the same across groups. This statement can be translated into factor analytic terms: Given multiple items that make up a scale, if the loadings for those items on the single underlying factor are the same across groups, then measurement invariance is supported. When framed in these factor analytic terms, this property is called factorial invariance, and represents one approach to the study of MI. The various aspects of MI can be investigated using confirmatory factor analysis models. As will be seen, these models can be supplemented with a model for structured means so as to allow for the study of group differences in means on latent variables. That is, growth can be studied once invariance (or the degree of invariance) can be assessed.

In the context of using CFA or EFA to evaluate the measurement qualities of an instrument, the item level data is really the data with which we start. Therefore, it is important to

start by specifying the data model. The data model here (Equation 1) is one in which nonzero means on the measured and latent variables are assumed:

$$x = \tau_x + \Lambda_x \xi + \delta$$

Equation 1

where τ_x represents a vector of intercept terms for the x measured variables, Λ_x is the factor loading matrix, ξ is the vector of latent variables and δ is the vector of error of measurement terms for the x measured variables. Further, κ is defined as the vector of means on the latent variables. The following covariance model (equation 2) can be derived from equation 1 and expresses the population variances and covariances of the measured variables as a function of the parameters in Λ_x , Φ , and Θ_δ which are parameters in the matrix of factor loadings, the variance/covariance matrix of the latent variables, and the variance/covariance matrix of error terms respectively.

$$\Sigma_{xx} = \Lambda_x \Phi \Lambda'_x + \Theta_\delta$$

Equation 2

The mean structure model can also be derived (equation 3) which expresses the population means of the measured variables as a function of the vector of intercept terms, the factor loadings and the vector of means on the latent variables.

$$\mu_x = \tau_x + \Lambda_x \kappa$$

Equation 3

Overall, the full model for means and covariances has five parameter matrices: Λ_x , Φ , τ_x , κ and Θ_δ .

A model is specified by designating fixed, free, and constrained parameters in these 5 matrices.

This model can be fit to data (sample covariance matrix, S , and sample mean vector, \bar{x}) by

obtaining estimates of the parameters such that the resulting implied population covariance matrix and mean vector ($\hat{\Sigma}$ and $\hat{\mu}$ respectively) are as similar as possible to their sample counterparts. In fact, a critical aspect of model fit is the degree to which the implied and observed are one and the same. Equations 2 and 3 are easily generalizable to multiple groups. For instance, the covariance structure is generalized to equation 4.

$$\Sigma_{xx}^{(g)} = \Lambda_x^{(g)} \Phi^{(g)} \Lambda_x'^{(g)} + \theta_{\delta}^{(g)}$$

Equation 4

Similarly, the mean structure is generalized in equation 5.

$$\mu_x^{(g)} = \tau_x^{(g)} + \Lambda_x^{(g)} \kappa^{(g)}$$

Equation 5

Where, g represents the g th of G populations. The model is specified in terms of the parameter matrices for each group, possibly including equality constraints on selected parameters across groups. The model is fit to the multiple samples simultaneously and in a deliberate order as more and more constraints are placed on the multiple group models.

The Standard Invariance Model Testing Sequence

While numerous theoretical formulations for measurement invariance have already been posited in this text, this study concerns itself primarily with levels of invariance that tap the psychometric properties of the measures. Little (1997) refers to these degrees of measurement invariance as Category 1, which subsumes the common taxonomy of configural, metric and strict invariance most frequently found in the MI literature (e.g. Horn & McArdle, 1992; Meredith, 1993).

The notion of MI usually is raised with reference to a single scale and the question of whether it measures the same trait in different groups. This question can be studied using the

multi-sample CFA model, usually with structured means. In the simple MI case, there would be only one factor, and the indicators of that factor would be the scale items (or subscales, parcels, etc.). However, the matrix representation of the models shows that the concepts and procedures apply equally in the case of multiple implied factors (latent variables). The multi-sample CFA model with structured means can be used to investigate MI and to test for group differences in factor/latent variable means. This is achieved by testing a sequence of models, beginning with an unconstrained model and progressively introducing equality constraints on parameters. In doing so, it is preferred that the sequence be determined a priori. The following subsections discuss this progression in detail.

Configural Invariance

Configural (Horn, McArdle and Mason, 1983), Weak (Meredith, 1993) or pattern invariance (Millsap, 1997), is considered the lowest or weakest level of measurement invariance that can be obtained. This type of invariance refers to the pattern of salient (non-zero) and non-salient (zero or near zero) loadings which define the structure of a measurement instrument. Configural invariance is supported if the specified model with zero-loadings on non-target factors fits the data well in all groups. Put another way, configural invariance holds when the same items load on the same factors for both groups of interest (e.g. grade 8 vs. grade 10). In CFA literature, configural models are also used as the baseline model in comparisons of competing model. However, this is interesting in that it could be seen as rather subjective, or possibly over-restrictive, if 0 loadings were in fact specified in the model. This is certainly a topic related closely to rotation technique in EFA.

Metric Invariance

Metric (Thurstone, 1947), weak (Meredith, 1993), or factor pattern invariance (Millsap, 1995) is more restrictive than configural invariance. This level of invariance requires that the loadings in a CFA be constrained to be equivalent in each group while permitting the factor variances and covariances to vary across groups. In other words, $\Lambda^1 = \Lambda^2 = \Lambda^G$. Given that such statistics rarely demonstrate equality, the statement really means that the loadings in one group are proportionately equivalent to corresponding loadings in other groups (Bontempo & Hofer, 2007). In order to make such proportionally equivalent statements, the common-factor variances must be freely estimated in all but the first group (or whichever group is chosen as a reference). This is because loadings standardized to the common-factor variance each differ from the corresponding loading in another group by the same proportion. The proportion is the ratio of the variance in each group. The presence of metric invariance can support researcher's claims that there are similar interpretations of the factors across groups but most would recommend a more stringent level of invariance testing to support that the statements are equivalent.

Scalar Invariance

Scalar (SteenKamp & Baumgartner, 1998) or Strong (Meredith, 1993) invariance is more restrictive than metric/weak invariance because it constrains factor loadings as well as intercepts to be equal across groups. In other words, equality constraints across groups are applied to factor loading parameters and the intercept parameters; $\Lambda^1 = \Lambda^2 = \Lambda^G$ & $\tau^1 = \tau^2 = \tau^G$. By applying these constraints, we are saying that any observed mean differences at the item level are accounted for by the common-factor mean. If this assumption holds, the comparison of factor means across groups is reasonable.

Strict Invariance

The final sets of constraints are the most restrictive and hence they are associated with demonstrating strict invariance. By saying that groups hold to the principle of strict invariance we are specifying equality constraints on factor loadings, intercepts and errors across groups.

$\Lambda^1 = \Lambda^2 = \Lambda^G$, $\tau^1 = \tau^2 = \tau^G$ & $\Theta^1 = \Theta^2 = \Theta^G$. In this paradigm, all parameters except for the latent variable level are constrained to be equal. So, the latent variable, factor, means and covariances can be used in comparisons.

A model is said to be identified when, for a given research problem and data set, sufficient constraints are imposed such that there is a single set of parameter estimates yielded by the analysis (Thompson, 2004). Some mechanical processes must also take place to achieve model identification. Specifically, your latent constructs have to be assigned a scale of measurement. One way to accomplish this in a multiple group setting such as this is to specify the value of the factor loading of one item per factor to unity (1.0). These items are then referred to as marker, or reference, items. In the context of multiple group factor analysis paradigms, it is critical that the same item be fixed to unity for all groups examined. The procedures just mentioned should be considered as being part of the decision sequence in determining the level of invariance between groups.

Outcome Measures: Fit Indices

Fit refers to the ability of a model to reproduce the data (i.e., usually the variance-covariance matrix). A good-fitting model is one that is reasonably consistent with the data; a good-fitting measurement model is required before interpreting the causal paths of the structural model.

It should be noted that a good-fitting model is not necessarily a valid model. Models with arguably ridiculous results (e.g., paths that are clearly the wrong sign) and models with poor discriminant validity or Heywood cases can be “good-fitting” models. Therefore, parameter estimates must be carefully examined to determine if one has a reasonable model as well as a good-fitting model. It is important to realize that one might obtain a good-fitting model, yet it is still possible to improve the model and remove specification error. Of course, having a good-fitting model does not prove that the model is correctly specified. Finally, it should be noted that a model all of whose parameters are statistically significant can be from a poor fitting model. So, does this mean all hope is lost and this is a meaningless endeavor? Absolutely not, it just means we should be cautious with inferences and overstressing the reaches of the generalizability of results (as usual).

The appropriateness of one fit index compared to another is not a new ‘argument’ in the literature. Some researchers (e.g., Barrett, 2007) do not believe that fit indices add anything to the analysis, and only the chi square should be interpreted. The primary concern driving the χ^2 argument is that fit indices allow researchers to claim that a miss-specified model is not a bad model. Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne (2007) argue that cutoffs for a fit index can be misleading and subject to misuse in that they are generally rules of thumb not driven by empirical evidence. Therefore, the author contends that they are useful but much like the allusion to Messick’s validity paradigm, they are only useful when used as intended.

There is also the potential of “cherry picking” a fit index; computing several fit indices and picking the one index, or indices, that best confirms the research hypothesis rather than what is appropriate given the data and the intended inferences. Choosing not to use a commonly referenced index (like the TLI or the RMSEA) requires justification, especially if one wishes to

publish in high quality journals. Others, such as Kenny, Kaniskan, and McCoach (2011), have argued that fit indices should not even be computed for small degrees of freedom models. What is more important in those situations is to locate the source of specification error (Kenny & McCoach, 2011). Bollen and Long (1993) is a great reference that discusses in great detail many of the indices mentioned here. A crucial consideration discussed in choice of a fit index is the penalty it places for complexity. The penalty can be thought of as how much a χ^2 needs to change for the fit index not to change. Another crucial consideration is what the purpose is and your main research question. Answering the question of whether or not a model fits is different than answering which model fits better or which model fits better across groups. Here the purposes are twofold. During the establishment of configural variance, the author must first establish that the baselines are in fact decent fitting models. That would require an absolute measure of fit whereas the comparative fit investigations are best addressed with indices of relative fit.

Absolute Fit

The most common approach is to utilize the chi square distribution. For models with about 75 to 200 cases, the χ^2 test is a reasonable measure of fit. However, for models with roughly 400 or more cases, the chi square is almost always statistically significant. Chi square is also affected by the size of the correlations in the model: the larger the correlations, the poorer the fit. Sometimes chi square is more interpretable if it is transformed into a Z value using the following approximation:

$$Z_{\chi^2} = \sqrt{2(\chi^2)} - \sqrt{2(df) - 1}$$

A problem with this fit index is that there is no universally agreed upon standard as to what is a good and a bad fitting model. Using areas on the standard normal curve does not

remediate the sensitivity to sample sizes. The chi square test is too liberal (i.e., too many Type 1) errors when variables have non-normal distributions, especially distributions with kurtosis. Moreover, with small sample sizes, there are too many Type 1 errors. Of important note is that two very popular fit indices, TLI and RMSEA, are largely based on the χ^2/df concept.

The root mean square error of approximation (RMSEA) is currently the most popular measure of model fit and it now reported in virtually all papers that use CFA or SEM and some refer to the measure as the “Ramsey.” This absolute measure of fit is based on the non-centrality parameter. Its computational formula is: $\sqrt{\chi^2 - df} / \sqrt{df(N - 1)}$, where N the sample size and df the degrees of freedom of the model. If χ^2 is less than df, then the RMSEA is set to zero. The penalty for complexity is the χ^2 to df ratio. The measure is positively biased (i.e., tends to be too large) and the amount of the bias depends on smallness of sample size and df, primarily the latter.

MacCallum, Browne and Sugawara (1996) have used 0.01, 0.05, and 0.08 to indicate excellent, good, and mediocre fit respectively. However, others have suggested 0.10 as the cutoff for poor fitting models. These are definitions for the population. That is, a given model may have a population value of 0.05 (which would not be known), but in the sample it might be greater than 0.10. There is greater sampling error for small df and low N models, especially for the former. Thus, models with small df and low N can have artificially large values of the RMSEA. For instance, a chi square of 2.098 (a value not statistically significant), with a df of 1 and N of 70 yields an RMSEA of 0.126. For this reason, Kenny, Kaniskan, and McCoach (2011) argue to not even compute the RMSEA for low df models. A confidence interval can be computed for the RMSEA. Ideally the lower value of the 90% confidence interval includes or is very near zero (or no worse than 0.05) and the upper value is not very large, i.e., less than .08.

The width of the confidence interval is very informative about the precision in the estimate of the RMSEA. A value less than .08 is generally considered a good fit (Hu & Bentler, 1999).

Descriptive Fit

Incremental (sometimes called relative) fit indices are analogous to R^2 ; a value of zero indicates having the worst possible model and a value of one indicates having the best possible. In that respect, the model(s) of most interest are essentially put on a continuum ranging from the null or independence model (worst) to the ideal (a perfectly fitting model) with the theoretical frameworks typically falling in between.

The Bentler-Bonett Index (1980) or Normed Fit Index (NFI) is credited as being one of the very first measures of incremental fit proposed in the literature. The best model is defined as model with a χ^2 of zero and the worst model by the χ^2 of the null model. Formulaically, the index can be seen as:

$$\frac{\chi^2_{Null} - \chi^2_{Proposed Model}}{\chi^2_{Null}}$$

Traditionally, a value between .90 and .95 is considered marginal, above .95 is good, and below .90 is considered to be a poor fitting model. A major disadvantage of this measure is that it cannot be smaller if more parameters are added to the model. That is, there is a penalty of 0 for complexity; the more parameters added to the model, the larger the index. When comparing models with vastly different model specifications, such as this study, the NFI doesn't perform well and these differences would be quite misleading.

One remedy is to use the Tucker Lewis Index or Non-normed Fit Index (NNFI) which overcomes the non-penalty problem of the Bentler-Bonett index. The Tucker-Lewis index has such a penalty and leverages the historically preferred method of looking at the χ^2/df . The TLI is computed as follows:

$$\frac{\frac{\chi^2}{df} \text{Null Model} - \frac{\chi^2}{df} \text{Proposed Model}}{\frac{\chi^2}{df} \text{Null Model} - 1}$$

A weakness of the correction is that the index can rise above 1, however it is capped at 1 for practical purposes. Interpreted just as the Bentler-Bonett index, values closer to 1 indicate greater fit. An artifact is that for a given model, a lower χ^2/df (as long as it is not less than one) implies a better fitting model. The penalty for complexity is χ^2/df . That is, if that ratio doesn't change, the TLI does not change. Also worth noting, the TLI depends on the average size of the correlations in the data. If the average correlation between variables is not high, then the TLI will not be very high.

Comparative Fit Indices

Akaike Information Criterion (AIC)

The AIC is a comparative measure of fit and so it is meaningful only when two different models are estimated. Lower values indicate a better fit and so the model with the lowest AIC is the best fitting model. There are somewhat different formulas given for the AIC in the literature, but those differences are not really meaningful as it is the difference in AIC that really matters: $\chi^2 + k(k - 1) - 2df$, where k is the number of variables in the model and df is the degrees of freedom of the model. Note that $k(k - 1) - 2df$ equals the number of free parameters in the model. The AIC makes the researcher pay a penalty of two for every parameter that is estimated.

Bayesian Information Criterion (BIC)

Whereas the AIC has a penalty of 2 for every parameter estimated, the BIC increases the penalty as sample size increases: $\chi^2 + \ln(N)[k(k + 1)/2 - df]$, where $\ln(N)$ is the natural logarithm

of the number of cases in the sample. (If means are included in the model, then replace $k(k + 1)/2$ with $k(k + 3)/2$). As can be seen, the BIC places a VERY high value on parsimony.

The Sample-Size Adjusted BIC (SABIC)

The Sample-size adjusted BIC or SABIC like the BIC places a penalty for adding parameters based on sample, size but not as high a penalty as the BIC. The SABIC is not given in Amos, but is given in Mplus. Several recent simulation studies (Enders & Tofighi, 2008; Tofighi, & Enders, 2007) have suggested that the SABIC is a useful tool in comparing models. Its formula is: $\chi^2 + [(N + 2)/24][k(k + 1)/2 - df]$.

With all of the comparative fit indices, the goal is to obtain estimates as close to 0 as possible. Essentially, when comparing two competing models, values with lower comparative fit indices are a better way to explain the data. Especially when referencing those that penalize/adjust for model complexity and increasing sample size.

Construct Stability, Measurement Invariance and Validity Evidence

What should be evident thus far, is depending on the theoretical viewpoint being utilized to assess the comparability/interchangeability arguments, the methodology employed could focus on equating, test specification matches, alignment, measurement invariance in the spirit previously reviewed or most likely, a convoluted combination of all of the above. Regardless of how one wishes to frame the issue, it is always a validity question as it speaks directly to the intended inferences or uses of scores and assessments that produce those scores.

The heart of assessment is the construct and the generally conceived notion that one is measuring what they target to measure for the intended uses and purposes. Truly that is a massive set of assumptions that requires multiple inputs to assess. As such, there is guidance given that isn't meant to be exhaustive in any respect but helps to appropriately categorize many

of the analyses responsible psychometricians and test developers already undergo in the process of accumulating validity evidence in a large-scale assessment. Messick (1989, 1995) identifies five sources of evidence to support construct validity: content, response process, internal structure, relations to other variables, and consequences. These are not different types of validity; they are best thought of as categories of evidence that can be collected to support the construct validity of inferences made from assessment scores. Evidence should always be sought from several different sources to support any given interpretation, and strong evidence from one source does not negate the need to seek evidence from other sources.

While accumulating evidence, one should specifically consider two threats to validity: inadequate sampling of the content domain (construct underrepresentation) and factors exerting nonrandom influence on scores (bias, or construct-irrelevant variance). I will return to the issue of the inadequate sampling of the content domain in a later section; however the second factor is a major source of concern in the proposed study.

Nonrandom influence on scores is especially difficult to contain and identify. In terms of apportioning variance, a frame of reference or theoretical viewpoint must be stated clearly or else it becomes a rather difficult task as sources of bias may be considered relevant or irrelevant in certain uses and purposes. Depending on the intended use, to what extent does teaching a student actually contribute construct irrelevant variance? Is the delivery of curriculum actually a non-random influence on scores? Such questions may at first make you chuckle; when is teaching a bad thing? However, if you think about it, the concept might not be so odd when trying to conceptualize growth. If an instrument is intended to measure whether a student has a certain achievement level, as measured by a standardized test designed to tap content taught up until a certain point in time, then it could be argued that any activity after that designated point in time

could potentially contribute construct irrelevant variance. In that context, it may also call into question whether or not a repeated measures test-retest paradigm using a parallel form, or the same form, of the assessment used as a measure of growth would find supportive evidence to validate the instruments use, and the inferences made, in that context. Aside from policy considerations surrounding the use of measures for growth, there are also assumptions regarding instrument content that drive the appropriate measure and method of measuring and determining growth that may or may not be considered. This issue actually returns us to the concern of inadequate content sampling; an issue of test development and design.

Content Based Evidence

For the purpose of making decisions on a student's status measure, whether it is via a proficient/not-proficient or multi-level designation based on that score, the practice of creating a table of test specifications based on content standards and creating a content weighting scheme of what needs to be assessed is typically the beginning. Items are then written towards the table of specifications, reviewed by content experts who are able to judge the appropriateness of the items as well as the depth of knowledge (DOK) and to ensure there is no initial hint of potential bias towards one or more subgroups. Items are then field tested, the best items are selected based on a carefully balanced blend of psychometric qualities and alignment to the test blueprint and table of specifications, and those create the foundation for an operational assessment. Following the first operational assessment, it is common to delay for a short time the reporting of the results. This is to allow stake holders and appropriate personnel the ability to articulate achievement level descriptors (ALDs) which are the first step in bridging the performance on the assessment with the holistic expectation linked to the ALDs.

Following this, standard setting meetings take place in which cut scores are recommended. Standard setting takes on many forms but at the heart of most commonly used methods are those which leverage performance data on items and total scores. The recommendations are then taken to the appropriate approving bodies for the purpose of becoming policy. All of this is a very long and deliberate process that takes years to accomplish and if done properly results in a strong assessment for its intended uses and purposes.

Up till this point, it is important to note that the only intended use and purpose supported is the use as a status measure of achievement, in as much as the instrument still abides by the predefined blueprints and tables of specifications developed during the initial scaling and used for standard setting, where the score can be applied towards the criterion reference targets and a performance level derived. The language contained within the ALD becomes the operational definition of the intended inference for which validity evidence has been accumulated; and this is a status measure only.

Factorial Evidence

Test development processes support the accumulation of content related evidence towards the construct validity argument and suggest an internal structure (Messick 1995; Cook and Beckman, 2006). Reliability and factor analysis data are generally considered evidence of internal structure. That is to say, scores intended to measure a single construct should yield homogenous results, whereas scores intended to measure multiple constructs should demonstrate heterogenous responses in a pattern predicted by the constructs.

Just as constructs can be defined by blueprints and tables of specifications, they can also be implicitly assumed and defined by a choice of measurement model (e.g. Rasch, 2PL, 3PL, GPCM). These IRT-based models look precisely at that ever important person and item

interaction that the blueprint itself does not address, nor is it purposed to address. When data are scaled using one of the unidimensional IRT models, the assumption is that there is one underlying construct that is measured by the collection of operational (scored) items on the assessment. Interestingly, the specification of a Rasch model implies that the items themselves measure the construct equally well in that the discrimination parameters are assumed to be equal.

It is important to point out that internal consistency should be seen as a necessary but not sufficient condition for measuring homogeneity or unidimensionality in a sample of test items. Essentially, that conception of reliability assumes that unidimensionality exists in a sample of test items (Tavakol and Dennick, 2011; Green, Lissitz and Mulaik, 1977). And of course, the fitting of a unidimensional IRT model further makes that presumption. Furthermore, systematic variation in responses to specific items among subgroups who were expected to perform similarly (i.e. DIF) suggests a flaw in internal structure, whereas confirmation of predicted differences provides supporting evidence in this category. Dimensionality is a characteristic of the interaction of persons and items. In the context of the proposed study, if students on the first occasion of an assessment consistently answer a question one way and on subsequent administrations answer another way, regardless of other responses, this will weaken (or support, if this was expected) the validity of intended interpretations with respect to the desire to generalize analysis. In this context, a lack of DIF can be considered an associated supporting measure, but not necessary pre-condition, to measurement invariance. The current study does not propose to examine DIF directly but the author introduces it here as an analogue to the invariance issue. DIF is essentially lack of invariance at the item level which in the IRT paradigm plays out as different item parameters but in factor analytic terms leads to lack of equality in factor loadings or the presence of metric invariance.

With regards to the relations to other variables, correlation with scores from another instrument or outcome for which correlation would be expected, or lack of correlation where it would not, supports interpretation consistent with the underlying construct. This idea can be extended to other variables believed to account for variability where a lack of significant relationship is one way to accumulate evidence for construct validity of inferences via the correlation with scores or variables that theoretically should (or should not) be related to the score of interest. Both of these ideas, support for the internal structure and relationship to other variables, become crucial when the focus of assessment scores (i.e. the intended uses) switch to not only support status, but also growth. Of course one should never forget that the assessment scores should be instructionally relevant given that an obvious, yet unintended consequence, of assessment is that curriculum decisions can become somewhat guided by the potential content of a high-stakes assessment (Perie et al, 2007). One can only reason, rationally, that the practice of tying assessment scores to evaluation and accountability would tend to get the attention of those being evaluated and being held accountable. In growth modeling the correlation or relation of the construct to the variable of time (which is confounded with instruction- type, kind, quality, etc) is a potential threat to or enhancing piece of the validity argument; as has been discussed this is buried in the intended (whether implicit or explicit) uses of the assessment scores. That is to say, it's an issue of measurement invariance over time.

In a factor analytic paradigm, we need to at least be able to assume configural invariance such that the data collected at each point in time, decompose into the same number of factors, with the same items associated with each factor (Meredith, 1993). If that hypothesis holds, there is support for the assumption that participants belonging to different groups conceptualize the constructs in the same way (Riordan & Vandenberg, 1994). It is this specific issue, whether or

not participants in the different groups specific to this study (e.g. students exposed to differing levels of instruction and additional years of cognitive growth) conceptualize the mathematics achievement construct in the same way, within the confines of the measurement invariance paradigm.

Chapter 3: Method

This study utilizes original census data from the statewide administration of an 8th grade mathematics assessment as well as data collected as part of a special study purposed at determining off level testing behavior of students. The instrument under investigation is the base form for that assessment containing the collection of item responses for all students tested under standard conditions. It was administered initially in Fall 2009 to then 8th grade students as their summative assessment (technically assessing 7th grade content). The same instrument was administered to subsamples of then current 8th, 9th, and 10th grade students in the Spring of 2011. This creates four sets of student responses: the initial census sample tested “on-grade”, a sample one grade of instruction above the intended level, a sample two grades of instruction above the intended level and a sample three grades of instruction above the intended sample. For purposes of the current study, the intended population will be referred to as the control condition and the other groups as treatment groups 1-3 for 8th, 9th, and 10th grade samples respectively.

Data

During the Fall 2009 administration, 118,891 then eighth grade students sat for the assessment in question over a two week period of October. This sample is actually considered a census in that anyone considered to be tested under standard conditions actually received the same standard form of content with the only difference being unique sets of field test items embedded at the same location across the various forms. In as much as this is considered a census, there is no attempt made to compare to a larger group. Treatments 1-3 are convenience samples obtained by schools volunteering to take part in a study designed to investigate the viability of off-grade testing. They were not drawn to be representative of the state 8th grade testing population, but some degree of similarity was desired. **It is important to note, that the**

current study is a re-analysis of pre-existing data sets. The data sets were provided to the author as deidentified longitudinal student profiles already linking previous assessment performance, demographics and data from the study of off-grade testing. While this was desirable in that the data were completely anonymous to the researcher it also prevented many follow up questions that would have been helpful after the analysis was complete. Some of these issues are elaborated further in the discussion and limitations sections. of Table 1 presents the sample characteristics in terms of gender, ethnicity, special education designation, limited English proficiency classification and eligibility for free/reduced lunch (here used as a proxy for economically disadvantaged) for all of the samples.

Table 1- Sample Demographic Characteristics

Group		Control		Treatment 1		Treatment 2		Treatment 3	
		Fall 2008		Spring 2009		Spring 2010		Spring 2011	
		8th Grade		8th Grade		9th Grade		10th Grade	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent
Total Sample		118891	100	1423	100	547	100	644	100
Gender	Female	58645	49.3	721	50.7	276	50.5	347	53.9
	Male	60246	50.7	702	49.3	271	49.5	297	46.1
Ethnicity	1	1125	0.9	6	0.4	7	1.3	4	0.6
	2	3025	2.5	0	0	4	0.7	6	0.9
	3	21643	18.2	344	24.2	144	26.3	150	23.3
	4	5132	4.3	113	7.9	22	4	28	4.3
	5	86431	72.7	901	63.3	366	66.9	455	70.7
	6	1395	1.2	21	1.5	4	0.7	1	0.2
	7	80	0.1	0	0	0	0	0	0
	8	60	0.1	0	0	0	0	0	0
	9	0	0	38	2.7	0	0	0	0
SES	Non-ED	66552	56	605	42.5	282	51.6	392	60.9
	ED	52339	44	818	57.5	265	48.4	252	39.1
LEP	Non-LEP	114962	96.7	1370	96.3	536	98	629	97.7
	LEP	3929	3.3	53	3.7	11	2	15	2.3
Spec.Ed.	Non-SE	106215	89.3	1326	93.2	529	96.7	615	95.5
	SE	12676	10.7	97	6.8	18	3.3	29	4.5

Instrument

The assessment contained 51 operational multiple choice items. Michigan is a fall testing state so the fall testing encompasses the previous year's content expectations only. The content standards being referred to are the Michigan Mathematics Grade Level Content Expectations (GLCEs) in effect during the testing related to this study (http://www.michigan.gov/documents/MathGLCE_140486_7.pdf).

The expectations are divided into strands with multiple domains within each. In practice, the skills and content addressed in these expectations are woven together into a coherent,

Mathematics curriculum. The domains in each mathematics strand are broader, more conceptual groupings. In several of the strands, the “domains” are similar to the “standards” in Principles and Standards for School Mathematics from the National Council of Teachers of Mathematics. For this particular assessment, five strands are possible for assessment: Numbers and Operations (N), Algebra (A), Measurement (M), Data and Probability (D), and Geometry (G). The particular grade and content area in this particular year did not include Measurement (M) items on the test blueprint. Therefore, only items belong to the A, D, G or N strands appeared on the assessment with the following item counts (out of 51) respectively: 22, 6, 9 and 14. What is clear is that the Numbers and Operations and Algebra items are most heavily weighted in the blueprint. These strands are mutually exclusive categories. Scores are reported at the strand level but not as a scale score; a raw score as well the total possible are given. Scale scores are based on item response theory scaling (specifically under the Rasch model) and represent a unidimensional scaling of all of the operational items together to form the underlying θ . The scale score is a linear transformation of that θ value. Table 2 lists the breakdown of items by strand.

Table 2 – Assessment breakdown by strands (numerals are item numbers)

Strand A:	1, 2, 3, 4, 11, 13, 14, 15, 16, 17, 18, 24, 27, 28, 29, 33, 34, 44, 46, 47, 48, 49
Strand D:	19, 20, 21, 22, 50, 51
Strand G:	12, 35, 36, 37, 38, 39, 40, 41, 42
Strand N:	5, 6, 7, 8, 9, 10, 23, 25, 26, 30, 31, 32, 43, 45

The blueprints underwent formal alignment procedures using the methodology developed by Norman Webb (http://www.michigan.gov/documents/Alignment_Analysis_of_Grades_3-8_Mathematics_Standards_and_the_MEAP_165665_7.pdf). While other models exist, this is a popular approach that state K-12 testing programs have used to satisfy the requirements for demonstrating alignment with the state standards in the given content area. Furthermore, careful item construction procedures are followed such that items are commissioned to individuals with

proper credentials to serve as content experts. The items are then vetted for initial review with state department level content leads at which point they are either accepted, denied or denied with revision requests. Those items that survive move on to an initial review by referent groups, commissioned by both the MDE and the development contractor, to determine: 1) is the item reflective of the intended content standard? 2) is the item written to the appropriate level of cognitive complexity (DOK as defined in Webb procedures in this case)? 3) Does the item contain any language/text/symbols/images that would unfairly advantage or disadvantage any subgroup of the intended population? 4) Is the content of the item appropriate for the grade level? Items, at this point, can be cleared for field testing immediately, cleared for field testing following revision, or marked as do not use (DNU). Items surviving are then field tested where a large array of statistics are calculated for the item and the results are then taken back to appropriate referent groups for further review to be used in combination with expert judgment to deem if an item should then be marked for further revision and additional field testing, ready for operational or do not use.

As alluded to in the review of the literature and policies, it typically takes 12-18 months and a minimum of \$2,000 to produce an item used in operational assessment. It is crucial for the reader to understand that because what it does is provide a very multifaceted, and very real, perspective that those in high-stakes assessment must take. That is, there is an enormous amount of faith put into the test blueprint and specifications understandably given the time, money and effort put into so many activities surrounding them. They hold the key to the validity arguments; if the blueprint is flawed or the construct is flawed the measurement cannot be valid for the intended inferences.

Analysis

As mentioned previously, a big assumption of CFA is that one has a model they want to confirm. It is not uncommon to start with an assessment blueprint as a confirmation approach (Thompson, 2004). However, those really define the intended content of an item as it relates to a curriculum or content standard. The interaction of a person with an item involves much more than an intended content standard. Factor analysis data deal with the interaction of the person with the measurement device so it may not necessarily be fruitful to consider as a solid baseline, a model built entirely on item specifications. Therefore, I propose to use an additional baseline model which is the Rasch model for dichotomous responses (Rasch, 1960; Wright & Stone, 1979). This was the model used to scale the original data and provide the scale scores in question. The model consists of a single latent construct measured by the collection (51) of operational items in the assessment, each with a unique error component.

The second model is based on the assessment blueprint and references the content strands of Table 2. There are 4 latent variables/factors each being measured by the observed variables referenced in that table. Each of those measured variables has a unique error component associated with it. The 4 latent factors are assumed to be correlated.

The instrument used in this study does not utilize a partial credit model and as a result all of items are dichotomously scored. Because of this, multiple-group CFA measurement models with binary indicators require a different parameterization which requires modifications to the aforementioned procedures (Jöreskog & Moustaki, 2001; Millsap, R. & Yun-Tein, 2004; Muthén, B. & Asparouhov, T., 2002). Essentially, each item on the measure is connected to its respective construct through a latent continuous response variable. This variable is cut by $m-1$ threshold parameters, where m represents the number of item score categories. Analyses are

then based on a matrix of tetrachoric correlations. The latent response variables require additional scaling factors in order to assess group differences in the common factor mean and variance.

To identify the model the following steps must be taken: (1) The intercept parameters for all latent response variables must be fixed to 0 in the first group; (2) Uniqueness variances need to be fixed to unity in the first group. As with any standard multiple-group confirmatory factor analysis, additional constraints are necessary in order to place the common-factor mean and variance on the same metric across groups.

The two most commonly referenced approaches for achieving this end are presented by Millsapp and Tein (2004) as well as Muthen and Aparouhov (2002). The Millsapp and Tein approach requires that the first $m-1$ thresholds be constrained across all groups and a second threshold or uniqueness (in the case of binary items, there would be no 2nd threshold) be constrained for one reference item in each group. Similarly, the Muthen and Asparouhov approach requires that thresholds and loadings are constrained in a reduced model and that tests of selected items are conducted against a full model where thresholds and loadings for these items are freed while maintaining model identification through fixing the specific-variance to unity for the selected items. However, it is important to note that in the documentation of these paradigms presented in the Mplus user's guide (Muthen & Muthen, 1998-2012), an important discussion occurs where critical differences are presented that discuss how the progression takes place in a different sequence when continuous variables are considered versus categorical (dichotomous in this case) variables. This is for a couple of reasons but namely the unique and somewhat simplified statistical properties of binary variables in addition to the idea, which is true here, that many of the measurement models

presented based on these binary outcomes are themselves part of an IRT based solution that leads to item characteristic curves (referred to in Muthen and Muthen as item probability curves). As such, the constraining of thresholds and factor loadings takes place in tandem as these parameters represent the IRT parameters of difficulty/scale location parameters and discrimination/scaling factor parameters respectively. As a result, there are fewer steps presented than one would typically note in an invariance study using continuous variables; the invariance progression typically goes through four iterations where various parameters are constrained making the multi-group models more stringent as the progression occurs.

The steps involved depend on the particular paradigm chosen and the parameterization schema selected. It is recommended, that when underlying IRT models are assumed and when most, if not all observed variables are categorical the weighted least squares indicator using the Θ parameterization (versus the Δ parameterization; Mplus default) is preferred. The key issue is that for categorical outcomes, the measurement parameters of interest are truly the factor loadings and the threshold parameters. When the Δ parameterization is considered, scale factors are also considered and the Θ parameterization adds in the ability to examine residual variances in addition to the other parameters. Variances for continuous latent response variables (factors) are estimated but residual variances for the observed categorical indicators are not estimated. The parameters that are estimated and fixed can be found in Tables 3-6 for both model 1 (Rasch) and model 2 (the blueprint based approach; table of specifications). They are presented in sequence for measurement invariance testing.

Table 3- Model 1, Configural Invariance

Parameter	Constraints	
<i>Control Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Constrained to be equal within group per Rasch requirements
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1
<hr/>		
<i>Treatment Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Constrained to be equal within group per Rasch requirements
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1

Table 4- Model 1, Metric Invariance

Parameter	Constraints	
<i>Control Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Held Equal Across Groups
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1
<hr/>		
<i>Treatment Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Held Equal Across Groups
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1

The procedures followed for Model 2 (Blueprint based) are similar to those presented in Tables 3 and 4. The differences come with the addition of multiple underlying factors and the correlation among factors to be considered. Similar to Model 1, the configural and metric invariance tests are presented in sequence.

Table 5 - Model 2, Configural Invariance

Parameter		Constraints
<i>Control Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Constrained to be equal within groups and factors, per Rasch requirements
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1
<i>Treatment Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Constrained to be equal within groups and factors, per Rasch requirements
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1

Table 6 - Model 2, Metric Invariance

Parameter		Constraints
<i>Control Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Held equal across groups, within factors
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1
<i>Treatment Group</i>		
Loadings	$\lambda(1) - \lambda(51)$	Held equal across groups, within factors
Thresholds	$\tau(1) - \tau(51)$	Free
Residuals	$\theta(1) - \theta(51)$	Fixed to 1
Factor means	(α)	Fixed to 0 for factor 1
Factor variances	(ψ)	Fixed to 1 for factor 1

In the literature review section regarding fit indices, multiple types with differing purposes, were presented. What is clear in the literature is that some indices are more appropriate than others and there is a great degree of correlation evident between the indexes. The rationale for one being better than the other, given it is an appropriate situation, is based primarily on theoretical arguments and partially on the pride of the author. Additionally, an important

consideration must necessarily be whether or not the analysis software and estimation paradigm are able to produce the desired metrics. Typically, the program (which is true in the case of MPlus) will disallow the calculation of an inappropriate index. The TLI, CFI and RMSE are used in the current study for all of the named rationales. Those indexes are useful for determining model fit which is only part of the research questions. As pointed out in literature reviews on the topic of GFIs (e.g. Cheung and Rensvold, 2002), most researchers take a market basket approach to the use of the indices as none are ‘known to be true and accurate’ and many have limitations depending on the structure and nature of the data.

For the purpose of determining if students have in fact shown substantively significant growth, two approaches were taken. The first makes several assumptions with the most critical being measurement invariance of the instrument over time. That is, the students test performance, whom had already been measured with a parallel form of the assessment during the correct time period, was scored during the experimental period using the raw to scale score conversion table as was used during the initial scaling of the instrument. So, there is the definite assumption in place that the IRT parameters are invariant over time. Nevertheless, their initial administration of the assessment as well as the follow up conducted during the treatment administration allowed two estimates of a student’s scale score as well as an estimate of the students’ performance level and sub-performance level. Specifically, I applied the agency’s transition table (see Table 7) to the pre and post measurements to determine if the students in the various treatments had at least achieved the amount of growth expected given instructions and increased rigor of performance and content expectations. Although the use of the IRT raw to scale score tables and the transition tables are outside of the scope of the intended use of these

measures, the author believes it will provide a useful context within which to discuss the various results.

Table 7 – Michigan MEAP Transition Table

Grade X MEAP Achievement		Grade X + 1 MEAP Achievement								
		Not Proficient			Partially Proficient		Proficient			Advanced
		Low	Mid	High	Low	High	Low	Mid	High	Mid
Not Proficient	Low	M	I	I	SI	SI	SI	SI	SI	SI
	Mid	D	M	I	I	SI	SI	SI	SI	SI
	High	D	D	M	I	I	SI	SI	SI	SI
Partially Proficient	Low	SD	D	D	M	I	I	SI	SI	SI
	High	SD	SD	D	D	M	I	I	SI	SI
Proficient	Low	SD	SD	SD	D	D	M	I	I	SI
	Mid	SD	SD	SD	SD	D	D	M	I	I
	High	SD	SD	SD	SD	SD	D	D	M	I
Advanced	Mid	SD	SD	SD	SD	SD	SD	D	D	M

NOTE: SI = Significant Improvement, I = Improvement, M = Maintain, D = Decline, SD = Significant Decline

As discussed in the introduction, the current study is not meant to create a new methodology nor is the author introducing a methodology that has never been used. The purpose of this study is to back up a little bit and take a look at how well some basic assumptions hold up when using a measure as a status index before we go forward into assuming that measure also is appropriate to be used as a measure of growth. I believe the extent to which the IRT model holds over time, in some ways, addresses the extent to which it may be appropriate to generalize these measures to a pre-test/post-test type of situation. Similarly, the degree to which the context or blue print based model holds over time is informative in that it speaks to the extent these subscales hold their meaning over time.

Chapter 4: Results

In this section, findings of the measurement invariance analyses performed on the MEAP Grade 8 Mathematics assessments are described in detail. The overall goals of this study were to: (1) evaluate the fit of both a single factor (Rash model) and four factor (blueprint based test design) in the original administration data and (2) to determine the extent to which those models are invariant to additional years of instruction. That is Measurement Invariance paradigms will be evaluated on the focus groups of the study. To this end, the results are sequenced as follows: (1) descriptive statistics of past performance for the groups of data referenced, (2) confirmatory factor analysis results of proposed models and (3) measurement invariance tests for the groups in question.

Descriptive Statistics/Previous Achievement

As all of the study participants had previously been administered a parallel and equated form of this assessment in October of their 8th grade year, previous performance was available and is also presented in tabular form for comparison purposes. In Tables 8 and 9, two different ways to express prior performance are provided. In Table 8, Mean scale scores and standard deviations are presented to give an indication of central tendency and variability differences among the groups and compared to the original census administration of the particular form of interest. The distribution of student scale scores (pre and post) are given in Appendix A. As can be seen, the study conditions were more variable and were centered on different means. Follow up paired *t*-tests on mean differences (invoking the Scheffe' procedure to control for inflated family wise type I error) revealed that all treatments were significantly different from the assumed census population value. Furthermore, they were significantly different from each other.

Table 8 – Previous Performance (Mean Scale Score)

	N	Mean Scaled Score	Scale Score SD
Census	118851	818.20	27.70
Treatment 1	1423	814.96	29.39
Treatment 2	547	824.91	33.72
Treatment 3	644	821.04	28.75

Additionally, ordinal performance levels (1 = Advanced, 2= Proficient, 3= Partially Proficient and 4 = Proficient) are also presented along with the percent proficient (sum of percent in performance level 1 and 2; used for accountability and reporting purposes). This metric is interesting in how it differs from Table 8 in that here, only treatment 2 showed significant proportional differences with the rest of the groups; there were a significantly lower proportion of students who were proficient in this group compared to the census and the others. These results highlight somewhat of a paradox in communicating these results. While we do have significant mean differences, the categorical placement of these students in terms of their pass/fail status was remarkably similar. This is the influence of the criterion referenced cut-score placement.

Table 9 – Previous Performance (Percent Proficient)

	N	% PL1	% PL2	% PL3	% PL4	% Proficient
Census	118851	42.7	31.8	18.5	7.0	74.5
Treatment 1	1423	34.4	39.6	21.3	4.8	74.0
Treatment 2	547	45.9	23.4	21.8	9.0	69.3
Treatment 3	644	34.6	40.4	20.8	4.2	75.0

Confirmatory Factor Analysis of One and Four Factor Models by Sample

Prior to submitting the data to multigroup invariance testing, the fit of the two measurement models (single factor and four factor) for each of the treatment and the control sample were evaluated in a CFA framework. The extent to which each of the models fit was examined using Mplus v. 7.11 (L. Muthén & B. O. Muthén, 2013). WLSMV estimation including a probit link and the THETA parameterization was used to estimate all models (L.

Muthén & B. Muthén, 2013). WLSMV provides weighted least squares parameter estimates using a diagonal weighted matrix with standard errors and mean- and- variance adjusted chi-squared test statistic that use a full weight matrix (B. Muthén, du Toit, & Spisic, 1997). Model fit was evaluated with relative fit indices CFI, TLI, and RMSEA. For the CFI and TLI indices values above .95 indicate a good fit. For the RMSEA, a value less than .06 is considered to indicate good fit.

Table 10-Group level Model Fit (Single Factor Model)

Group	N	Chisquare	DF	# free parameters	TLI	CFI	RMSE
Census	118851	121565.1	1224	102	0.951	0.953	0.029
Grade 8 Study	1423	2210.484	1224	102	0.966	0.967	0.024
Grade 9 Study	547	1558.031	1224	102	0.975	0.976	0.022
Grade 10 Study	644	1679.457	1224	102	0.956	0.957	0.024

For the single factor model (Table 10), all chi-square tests were significant ($p < .0001$), however, due to the binary nature of the variables and the WLSMV estimation utilized, the chi-squares are not trustworthy as global goodness of fit indices. CFI, TLI and RMSEA indicate adequate fit of the single factor model in all of the groups.

Table 11-Group level Model Fit (Blueprint Based/Four Factor Model)

Group	N	Chisquare	DF	# free parameters	TLI	CFI	RMSE
Census	118851	112980.015	1218	108	0.954	0.956	0.028
Grade 8 Study	1423	2190.486	1218	108	0.966	0.968	0.024
Grade 9 Study	547	1547.218	1218	108	0.976	0.977	0.022
Grade 10 Study	644	1657.152	1218	108	0.957	0.959	0.024

The results in Table 11 for the four-factor model also suggest adequate fit across all study groups. Of note, is the small increase (which is dependent on index) in fit gained by the additional constraints placed on the parameters. Although both models had adequate fit, the comparative gain was small.

Measurement Invariance Tests of One and Four Factor Models

In the discussion of invariance test procedures in previous sections there were multiple levels of invariance to be explored. Specifically (in order from least to most restrictive) tests for the degree to which configural, scalar, metric and strict parameterizations of measurement invariance assumptions hold across groups were outlined in Tables 3-6. When using maximum likelihood estimation, all of these tests of invariance are possible. However, one of the challenges with using binary data in a confirmatory factor analysis paradigm is that there is a reliance on alternative estimation procedures, such as the weighted least squares procedure invoked in MPlus. This technique requires that scale factors or residual variances be allowed to vary across groups; the metric invariance test constrains these to be equal so there is an obvious disconnect.

As a result, only the configural and scalar approaches are feasible within MPlus (Muthen and Muthen, 2013). For binary variables using weighted least squares estimation and the Θ parameterization, the configural setting has factor loadings and thresholds free across groups, residual variances fixed at one in all groups, and factor means fixed at zero in all groups. The metric of a factor is set by freeing all factor loadings and fixing the factor variance to one, the factor variance is fixed at one in all groups. The scalar setting has factor loadings and thresholds constrained to be equal across groups, residual variances fixed at one in one group and free in the other groups, and factor means fixed at zero in one group and free in the other groups. Again, the metric of a factor is set by freeing all factor loadings within a group and fixing the factor variance to one. Furthermore, the factor variance is fixed at one in one group and is free in the other groups. Table 12 presents the results of the invariance studies for three different models.

Models 1 and 2, single factor and blueprint based respectively were introduced previously in the paper as the models of interest.

Table 12- Measurement Invariance Study Results

Paradigm	Model	Chisquare	DF	P-value	TLI	CFI	RMSE
Configural	Single Factor/Rasch	5331.53	3672	p<.001	0.967	0.969	0.023
Scalar	Single Factor/Rasch	5677.13	3770	p<.001	0.964	0.964	0.024
Configural	4 Factor (blueprint)	5279.638	3654	p<.001	0.968	0.969	0.023
Scalar	4 Factor (blueprint)	5598.375	3740	p<.001	0.964	0.965	0.024

As can be seen from Table 12, the results of the invariance testing revealed adequate fit across all of the models and paradigms put forth for analysis. The fit indices are all within acceptable limits and are similar to those found when the same models were posited within groups before applying between group equality constraints. The results indicate that configural invariance is supported between groups. Therefore, the specified model with zero-loadings on non-target factors fits the data well in all groups; the same items load on the same factors for all of the groups considered.

The findings of scalar invariance holding across the groups subsumes the assumptions of configural invariance, the same pattern of loadings apply, but also goes a step further. Scalar invariance suggests that the factor loadings for the factors can be considered identical across groups. In multi-group confirmatory factor analysis terms, this suggests that the inter-item tetrachoric correlation matrices for all groups are statistically equivalent.

Chapter 5: Discussion

The current study was a journey with the overall goal to evaluate the degree to which measurement invariance held across time for a high-stakes mathematics achievement instrument. Put differently, the degree to which the measurement device would produce the same set of composite latent measurements over time, with an assumed increase in instructional time for each of the study conditions to determine to what degree the instrument might be sensitive to instruction. This is important due to the increased emphasis being placed on growth measures in K-12 accountability measures. Policy makers will need to determine if they wish to measure growth via a measure that is static and not as sensitive (i.e. invariant) over a typical time trajectory used in such high stakes decisions. In order to reach the conclusions brought forth so far and in the paragraphs to follow, the author proceeded to: (a) fit the model separately in each group; (b) fit the model in all groups allowing all parameters to be free (c) fit the model in all groups holding factor loadings equal to test the invariance of the factor loadings and (d) fit the model in all groups holding factor loadings and intercepts equal to test the invariance of the intercepts. By following such a prescribed sequence I was able to address all of my research questions which I will now step through sequentially.

Related to the original Fall 2009 administration and applied across groups:

Does the Rasch model fit the data?

According to Table 10, the single factor (parameterized to replicate the Rasch model) measurement model appeared to fit the data well in each of the groups. With the sample sizes in this study it wasn't surprising to discover that the chi-square tests were all significant so alternative model fit indexes were referenced as recommended appropriately in the literature.

Perhaps what was surprising in this case is that the ‘best fit’ wasn’t for the census population but rather the study condition group data appeared to fit the model a bit better than the larger group. Without first establishing that this model fit well in the groups and in the pooled groups it is untenable to look at differences in groups let alone proceed to test the same measurement configuration. Of course, this study looked at two models specifically and the invariance was of interest across both. Therefore, the next research question addressed was:

Do the data fit the linear confirmatory model implied by the blueprint (content strands) for the test?

As was the case with the single factor model, Table 11 indicates that all of the groups showed adequate fit to the blueprint based model. Again, the *poorest* fit seemed to be in the larger census population. The differences are negligible and there is not a valid test to determine if the differences in absolute fit are different across the four non-nested groups but they did share a common model parameterization.

Does one of the models fit significantly better than the other model?

The model fit indices in each of the groups and the pooled groups both showed remarkable similarity across the two models. In fact, in terms of parsimony it seems that little is gained by adding in the additional parameters needed to represent the blueprint based test design. Therefore, it seems that while they serve as useable reporting categories in terms of grouping items together by content specifications they do little to improve the fidelity of the measurement as a whole. Given that the single factor model will provide a more reliable underlying construct,

it would make more sense to go with the single factor representation as a matter of model parsimony.

Do the aforementioned models exhibit measurement invariance across groups/study conditions?

In this study, due to the dichotomous nature of the indicator variables and the estimation method employed by MPlus. It was not feasible to proceed to tests for strict and metric invariance as the IRT nature of the underlying models also do not make it possible to constrain it in the same way. The study found support for configural and scalar invariance of both models across the groups (see Table 12). The additional constraint of factor loading equality produced a significant difference test with the configural model. Taken together, it appears as if this measure is invariant across these groups in both model configurations. Therefore, the suggestion would be that one can treat these as parallel measures and mean differences and comparisons on latent variables will be permitted.

Implications

The developments of parallel assessments that take into account curriculum are expensive and it is likely that many agencies will attempt to use a more cost effective test-retest strategy for computing growth. Such a simple approach holds a lot of assumptions with the most important being that the latent trait(s) or scores that come out of the measurement need to hold a high degree of generalizability across time and at the least, there should be an indication that the structure of the assessment, as intended, is logical for all of the groups being assessed. The degree to which the configural and scalar invariance assumptions hold, either inhibits or enables the types of generalizations one would wish to make in a gain-score paradigm.

The results of this study seemed to indicate that is feasible to at least use this particular instrument as a pre-test and post-test measure and that differences noted are true differences in student ability rather than an artifact of the underlying inter-item correlations manifesting themselves in different structures over time. In fact, what has been found with the current study is that the tetrachoric correlation matrix, in which the correlation between the latent factors underlying the items as expressed by the binary indicators are expressed, is consistent over time and appears not to be sensitive to instruction or continued learning. In essence, the conclusion is that the instruments are invariant over time and therefore, the latent factors can be compared across groups.

Table 13 – Outcome Variable Group Differences

	N	Mean Scaled Score	Scale Score SD	% Proficient
Grade 8	1423	814.96	29.39	65.5
Grade 9	547	824.91	33.72	76.6
Grade 10	644	821.04	28.75	79.3

Table 13 depicts the performance of the groups on the outcome measure. With the exception of the grade 9 compared to grade 10 students on percent proficient, all other differences are statistically significant. In addition, if the student scores were submitted to the performance level change matrix presented in Table 10, the results would be as presented in Figure 1. Perhaps the biggest surprise of that result set was the finding that after 3 additional years of instruction there was still nearly 10% of 10th grade students who's scores have decreased so much from when they originally sat for the assessment as incoming 8th grade students. The finding for the 9th grade students is similar. Is this a function of decay? Is the decay from lack of use of the skill set? It would have been interesting to explore these students' course taking activity following grade 8 to determine which, if any, took a very minimal approach to furthering themselves in the mathematical areas. While those results are troubling what is even more

shocking is that at the end of 8th grade, the students who were only 5 months removed from their initial assessment failed to show growth for the most part. More than 70% of these students exhibited significant decline, decline or merely maintained their standing along the continuum. One thing that can be confirmed, all of these 8th grade students were currently enrolled in pre-algebra or algebra at the time of this study. That is, all were on either a standard or advanced curriculum. No students were being instructed off grade-level.

Figure 1 - Performance Level Change by Study Condition

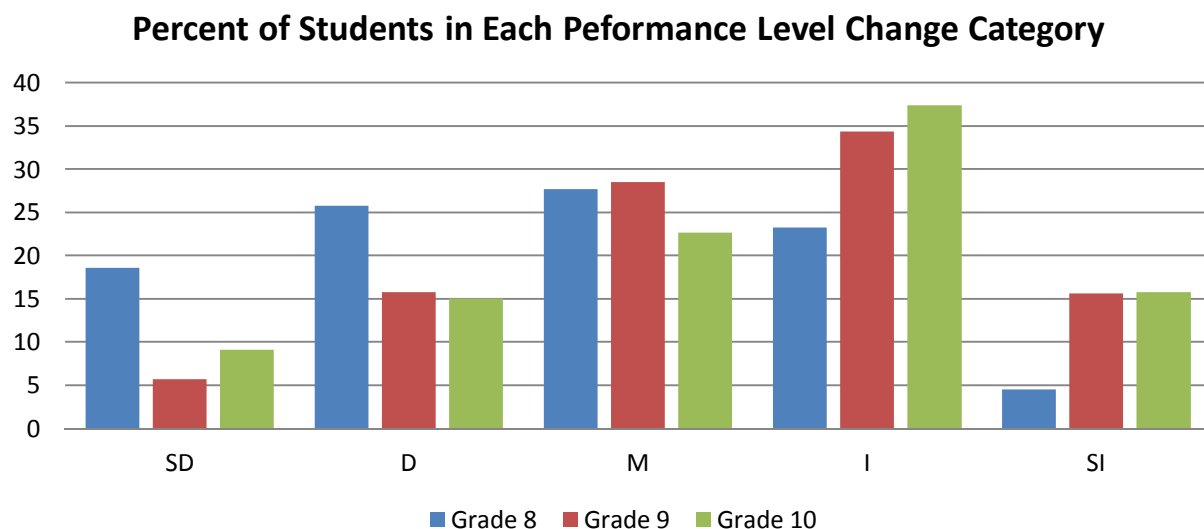
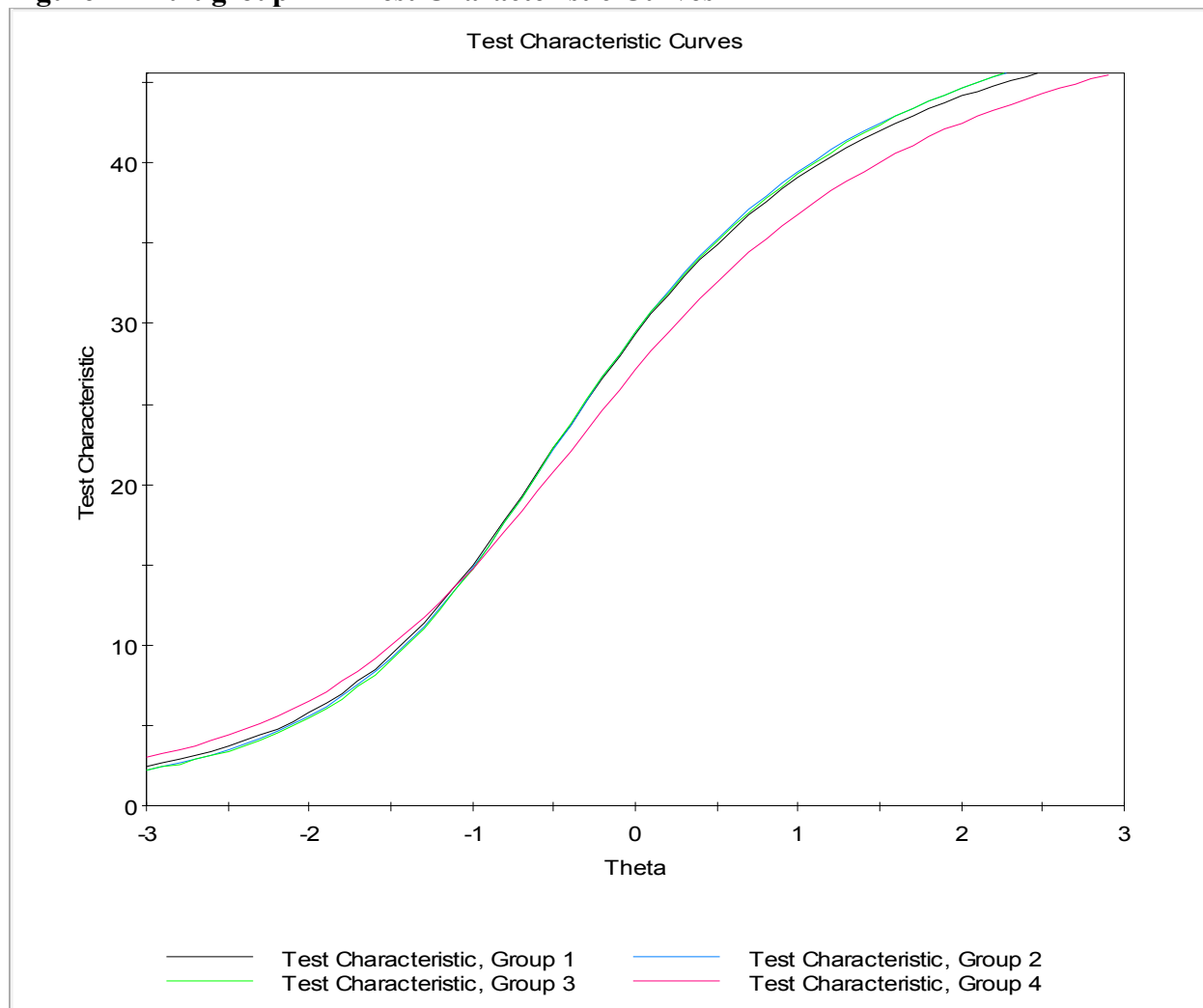


Figure 1 presents some pretty dire results, especially when combined with the finding of invariance across conditions. Supposedly, these are real differences; these declines represent a true decline in student standing on the same underlying construct measured at their initial testing session. However, although the MPlus results suggested invariance it seems that there are other sources of information that suggest otherwise.

To further investigate the invariance of the instrument and its sensitivity to the inherent differences between groups, instructional and otherwise, further analyses were conducted that provide further exploration into the phenomenon. An interesting artifact of the approach taken

with Mplus on the single factor model is that the model is actually a two-parameter logistic item response theory model. Therefore, the Mplus program, for the single factor model, suggested that there was invariance of the two 2PL IRT model across the groups. A slightly different approach to fitting the 2PL model to multiple groups was taken to determine if the same finding held. That is, is their invariance to the extent that we can be comfortable we're measuring the same thing? To accomplish this, a multiple group run of the two parameter model was carried out using the IRTPRO application version 2.1.1 (SSI, 2011). The original census population anchors the parameter estimation with the other groups getting placed on the same scale via the concurrent run. Figure 2 presents the test characteristic curves for the four groups. In the figure, group 1 is the 8th grade; group 2 is the 9th grade; group 3 is the 10th grade and group 4 the original census group respectively. As can be seen, while this chart suggests similar performance for the study conditions they also show that there is a rather dramatic difference in TCCs between the study conditions and the original census.

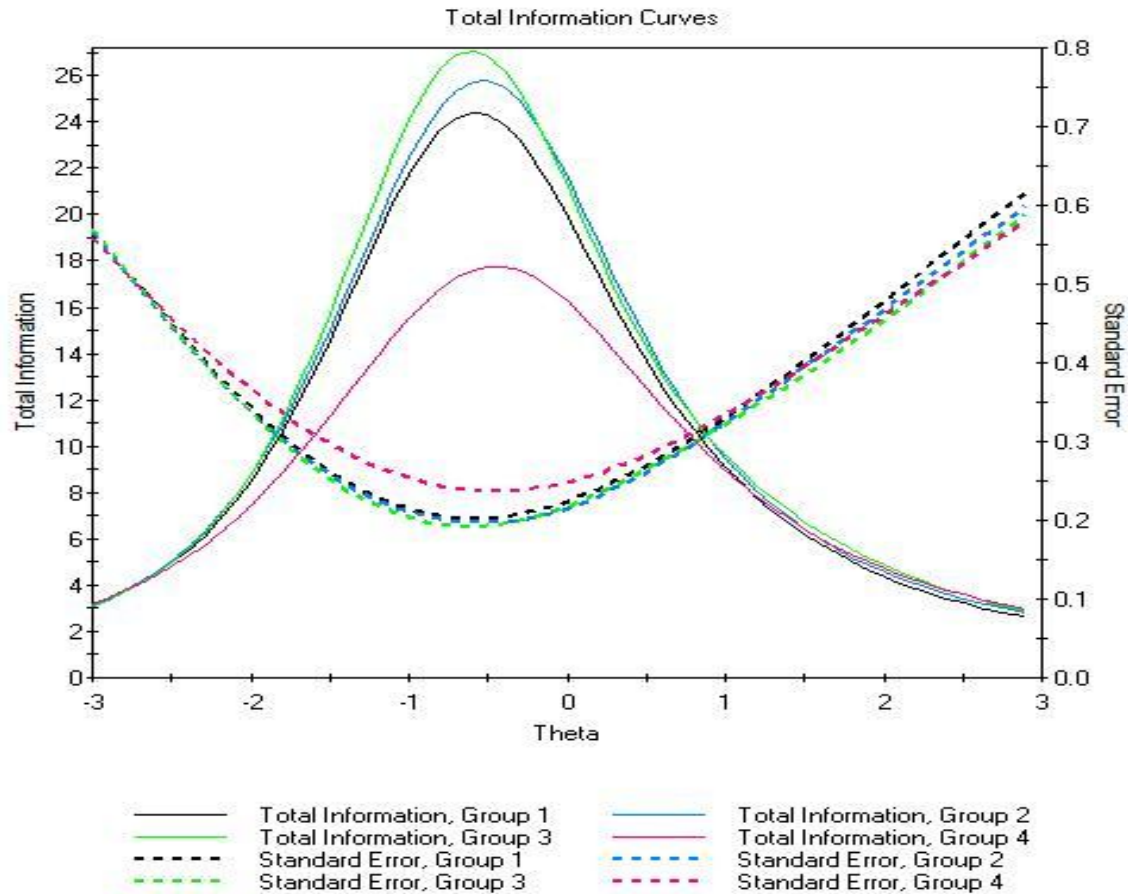
Figure 2- Multigroup IRT Test Characteristic Curves



Test information curves for the groups are also presented, along with their reciprocal standard error curves (see figure 3). These provide a slightly different picture in that they represent somewhat of a combination between the ability distribution of the assessment for each of the groups as well as their alignment to the item difficulty of the assessment. Figure 3 is based on the notion that the census scale is the appropriate parameterization. What is suggested is that the test information function for all of the groups is centered just below the origin of the theta scale (approximately -.5 on the logit scale). Scaling in IRTPro is accomplished by assuming a theta

distribution with a mean of 0 and standard deviation of 1. Therefore, the assessment is centered just below average value; the calibration goal.

Figure 3 – Test Information Functions from Multiple Group IRT Run



Appendix B and C presents the IRT calibration results for all 51 items for each of the groups in two ways. Appendix B are each of the item characteristic curves by group whereas Appendix C is a tabular presentation of item parameters by group (i.e. the data driving the charts in Appendix B). There are several areas where there are departures from a consensus value. Cell tables highlighted in yellow denote items departing from expectation. In this case, it seems reasonable to assume that the item difficulty, if anything, would decrease over time as opposed to increase.

Chapter 6: Limitations and Future Research

The biggest limitation in this study was the volunteer nature of the sample. The participants were purely voluntary and in fact most administrators revealed to the researcher that they allowed teachers to self-select their classes into the study and of course no students were required to participate and could back out at any time. Of course, with such situations motivation of the student being assessed is always discussed and particularly a possible lack of motivation for the student to perform well. In this case, all participants were aware that the study apparatus was not part of the state mandated battery of assessments and therefore the level of effort might not have been great. Additionally, this begs the question of how much extrinsic motivation the teachers of these students might have decided to not push or impart on the students. Of course, the lack of a random sample and random assignment limits the ability of the researcher to generalize to a great degree. Certainly it would not make sense for the author to assert that the measures are invariant across all grades represented in the study and across the universe of potential students/participants.

Another limitation, and a future direction should I choose to pursue this research further, would be to look more closely at the modification indices and other univariate tests available to help in model refinement. In this study I posited two main models in a confirmatory environment. I never specified they were without question the correct models as that is not the driving force of this study. However, there are further refinements that could be made to determine the best set of indicators (items) that collectively leads to the most invariant model over time and instructional exposure. For instance, there are several items in Appendix A that show great deviations across groups in terms of their IRT parameters. It could be an artifact of the scaling technique, but those trace curves do suggest some large item level differences that are

likely cancelled out much like differential item functioning at times leads to DIF cancellation such that there might not be obvious bias at the test level yet it still exists at the item level. Purifying the measurement instrument is going to be key for true invariant measures over time. Of course, each content area should probably be expanded a bit by adding more items to the content strands and testing those out as intact tests. I believe the type and level of inferences that will be required of growth modeling in the future is going to well beyond the current norm and will push validity inferences to all time levels of thin. To me, that is why it is imperative we take a step backward before we rush forward and make too many assumptions. These are not arbitrary test scores. They relate to student standing, they relate to school standing, district standing, teacher standing and if you think about it to many extents the livelihood of the students going forward as well as those whose jobs depend on evaluations based largely on assessment data.

APPENDICES

Appendix A

Scale Score Distributions (Pre and Post) for each of the Study Groups

Figure 4 – Post-test Scale Score Distribution (Grade 8)

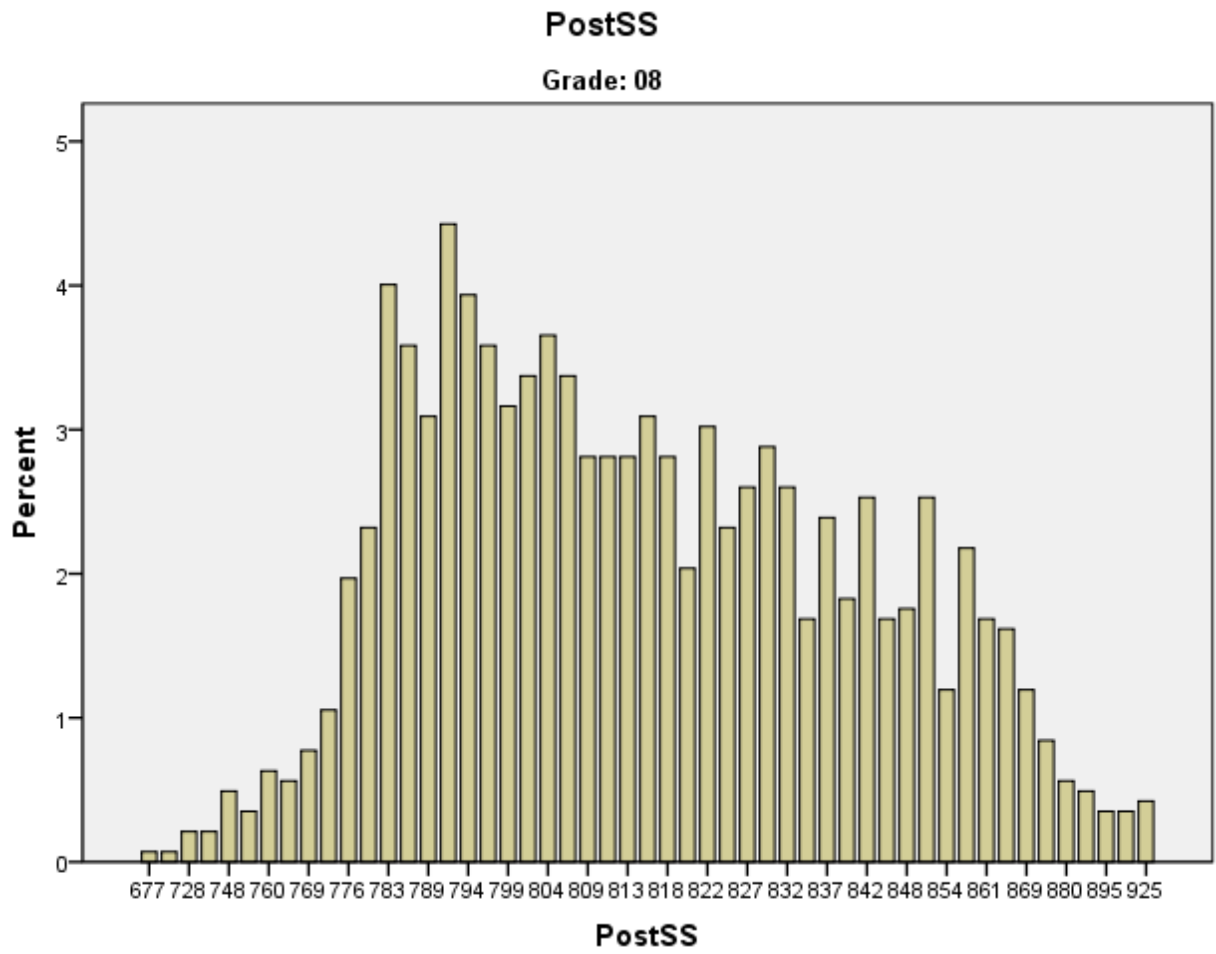


Figure 5 – Post-test Performance Level Frequencies (Grade 8)

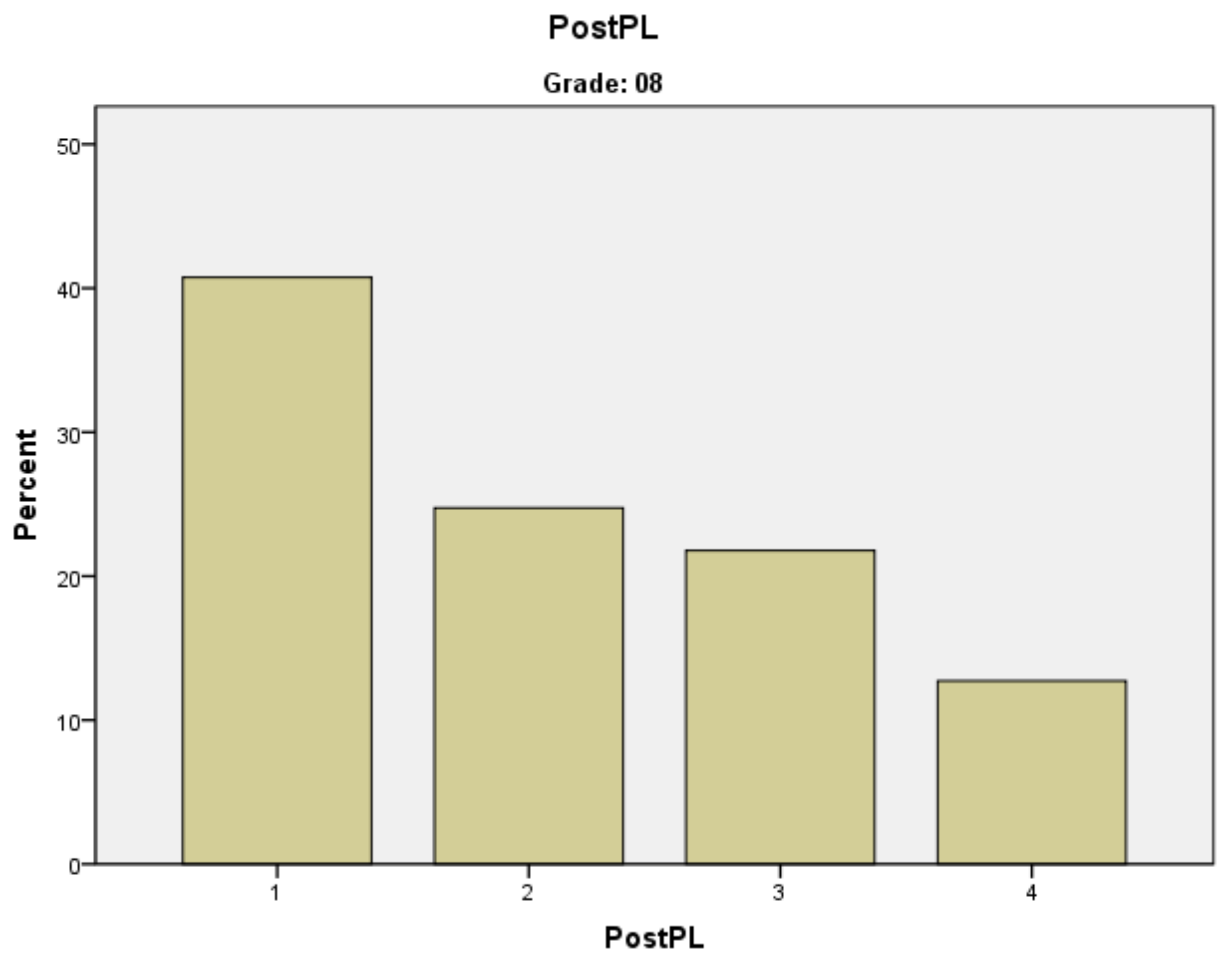


Figure 6 – Pre-test Scale Score Distribution (Grade 8)

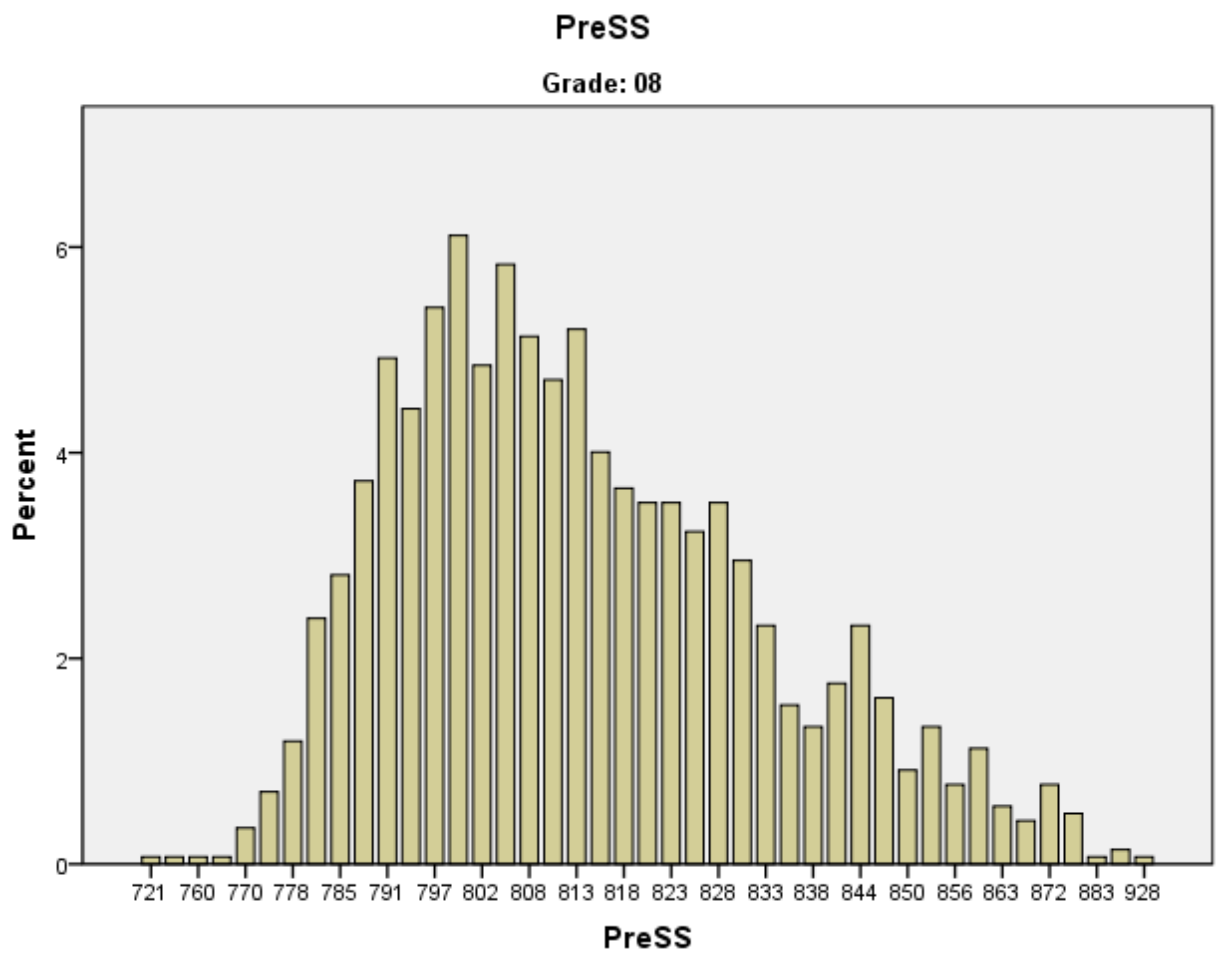


Figure 7 – Pre-test Performance Level Frequencies (Grade 8)

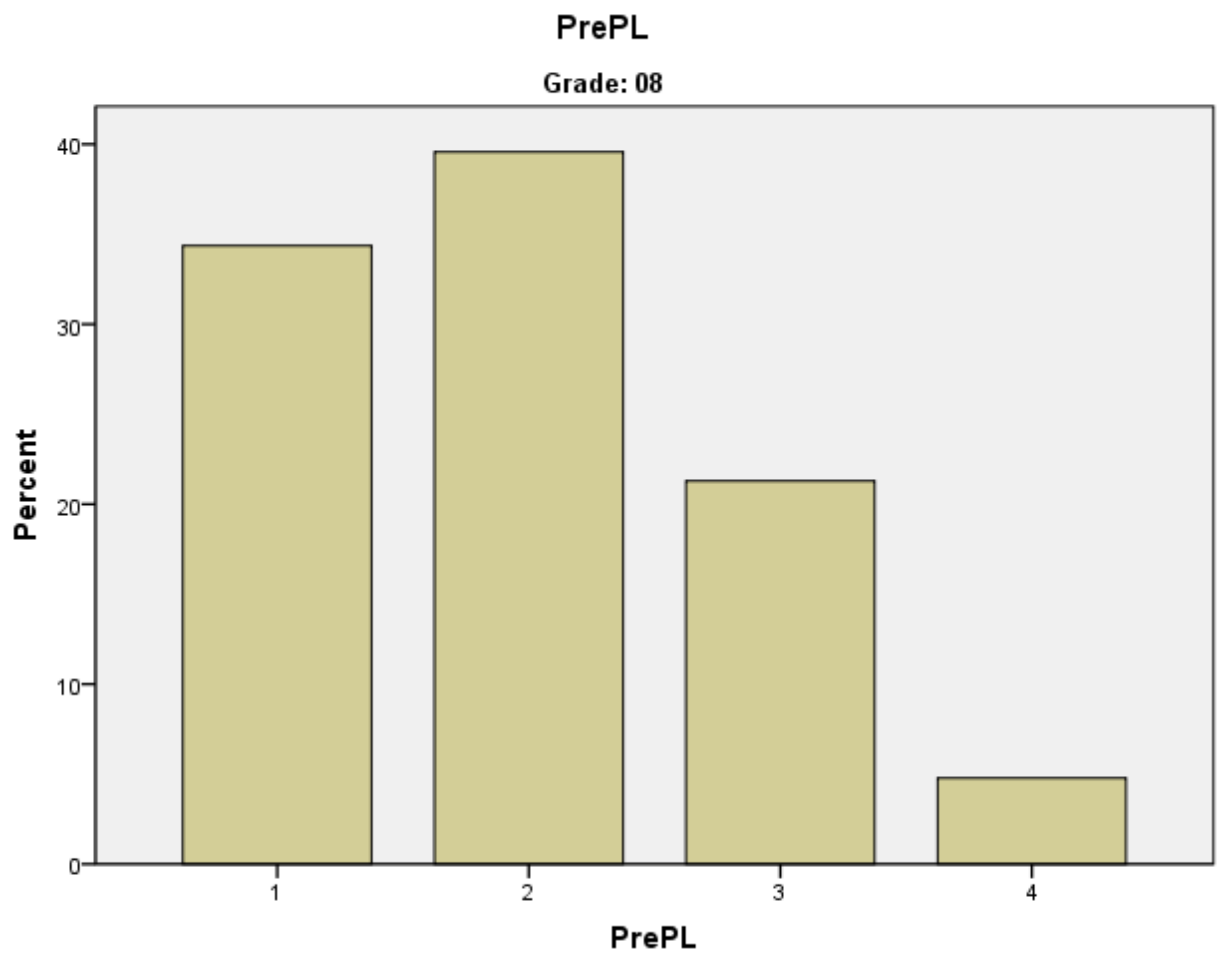


Figure 8 – Post-test Scale Score Distribution (Grade 9)

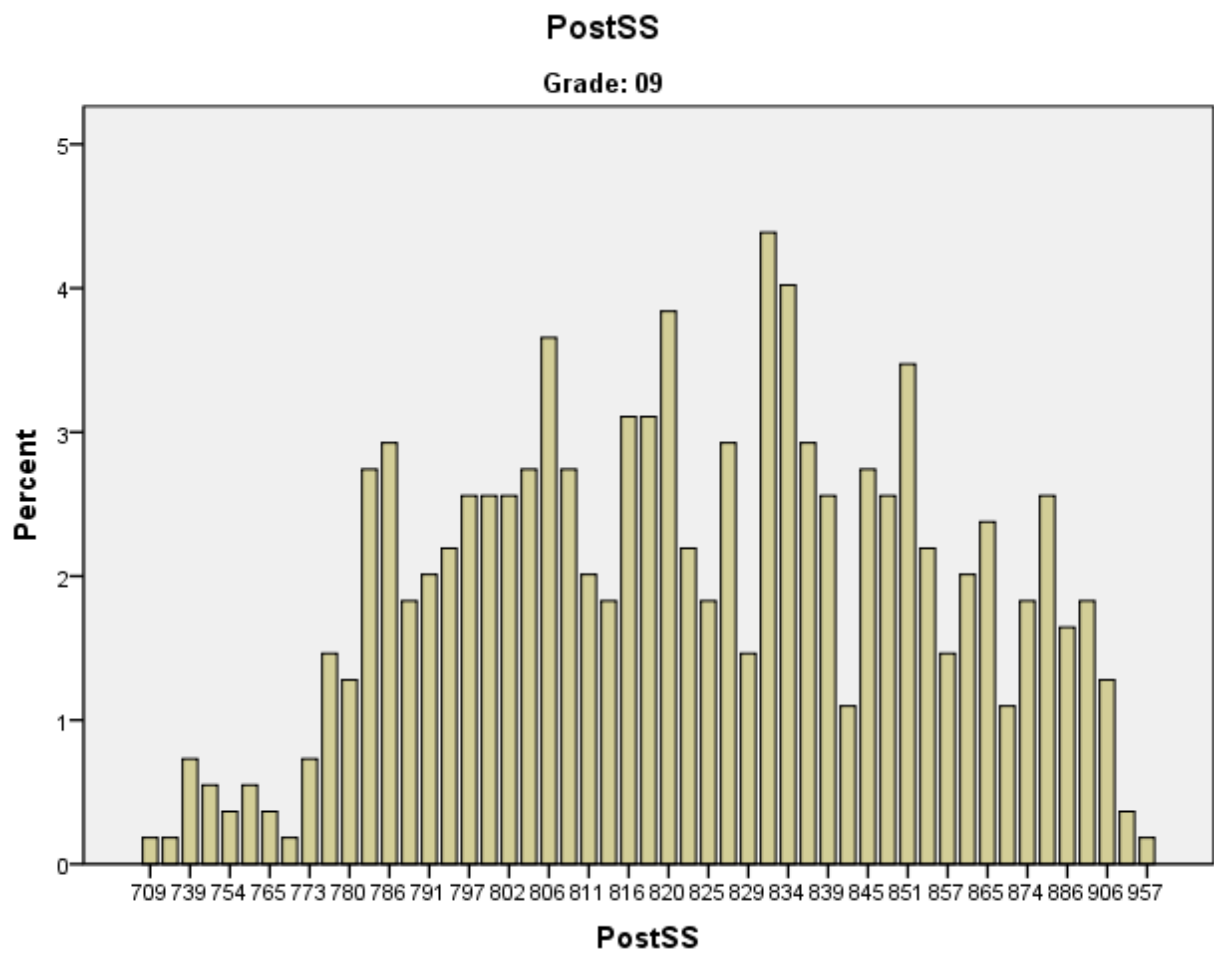


Figure 9 – Post-test Performance Level Frequencies (Grade 9)

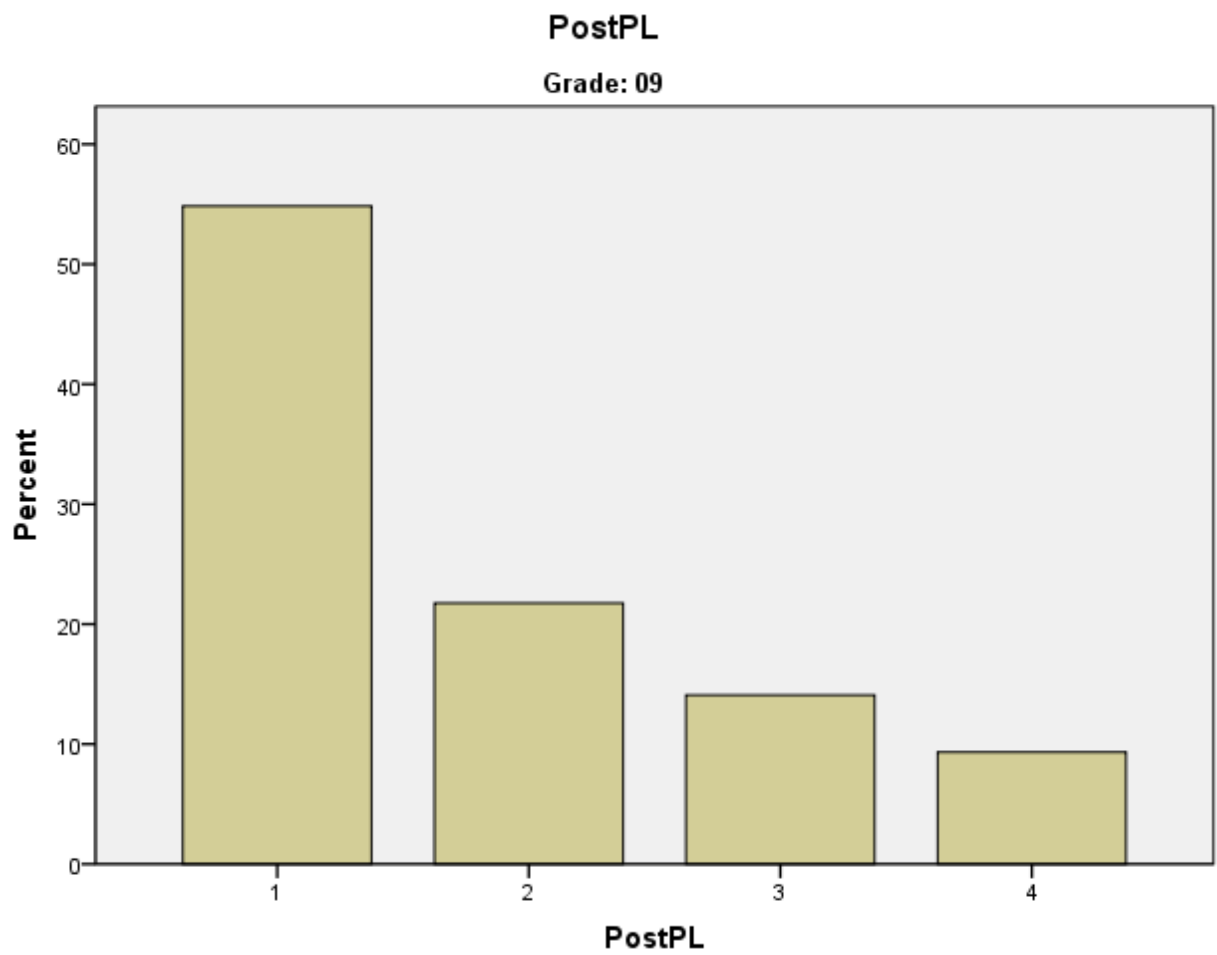


Figure 10 – Pre-test Scale Score Distribution (Grade 9)

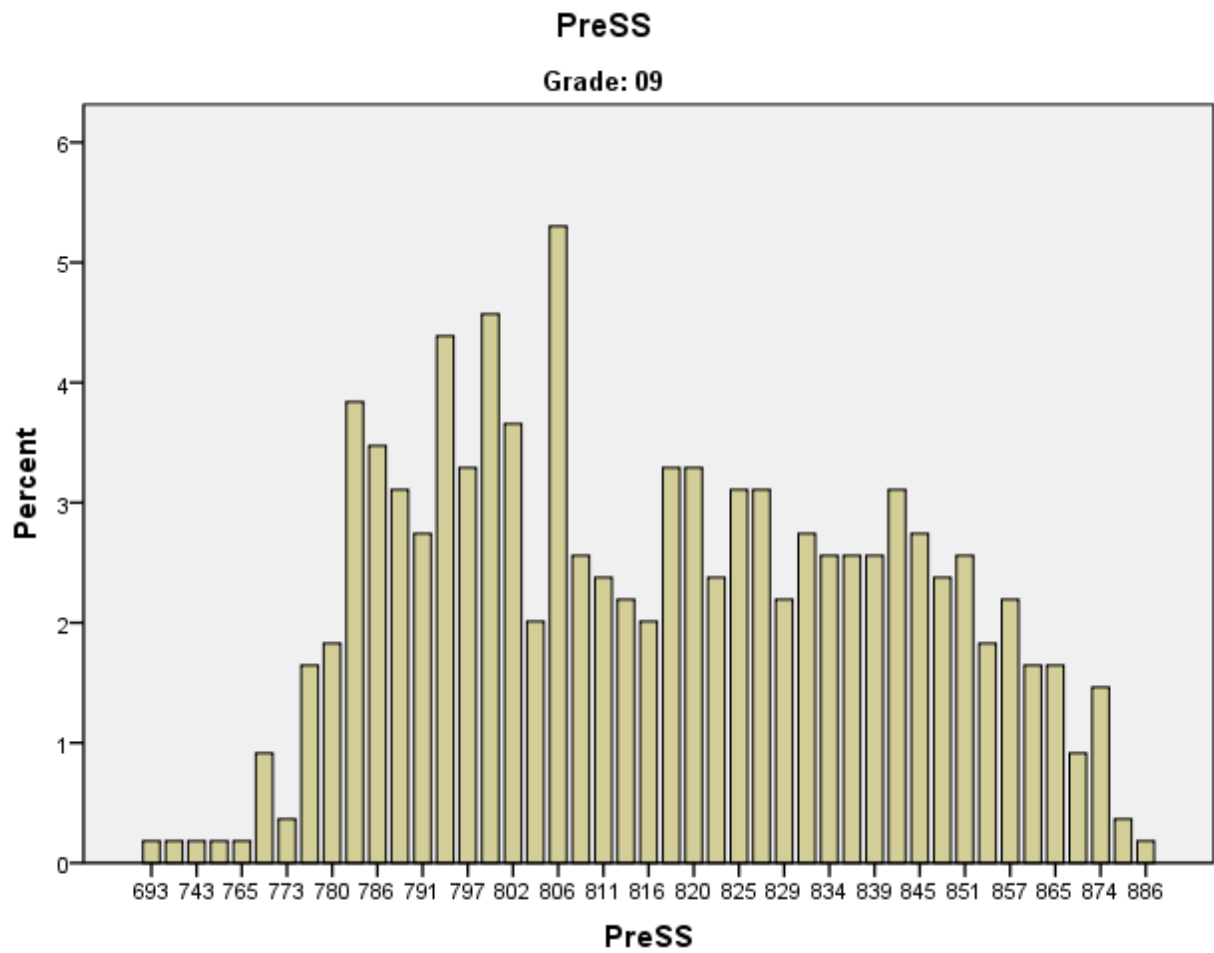


Figure 11 – Pre-test Performance Level Frequencies (Grade 9)

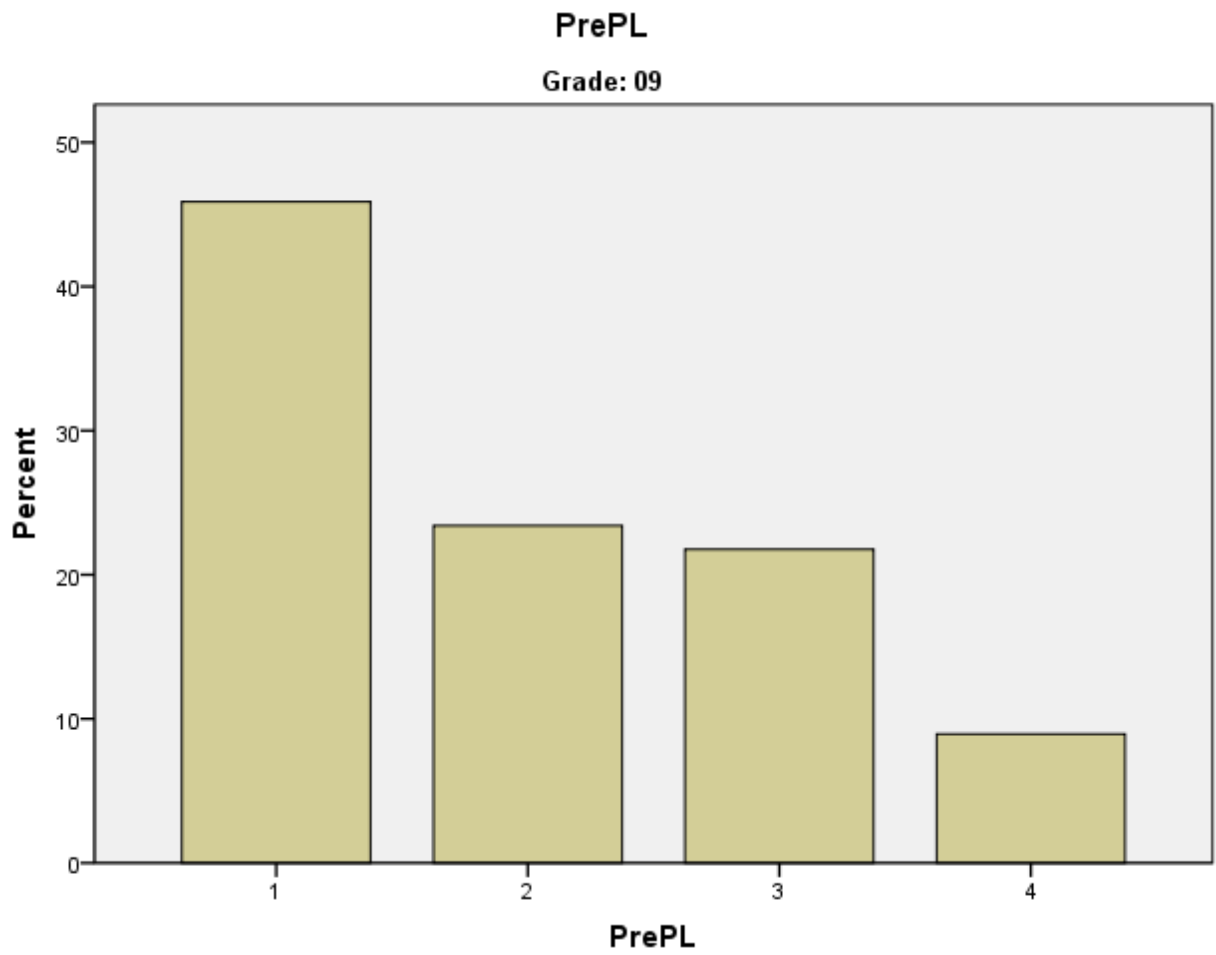


Figure 12 – Post-test Scale Score Distribution (Grade 10)

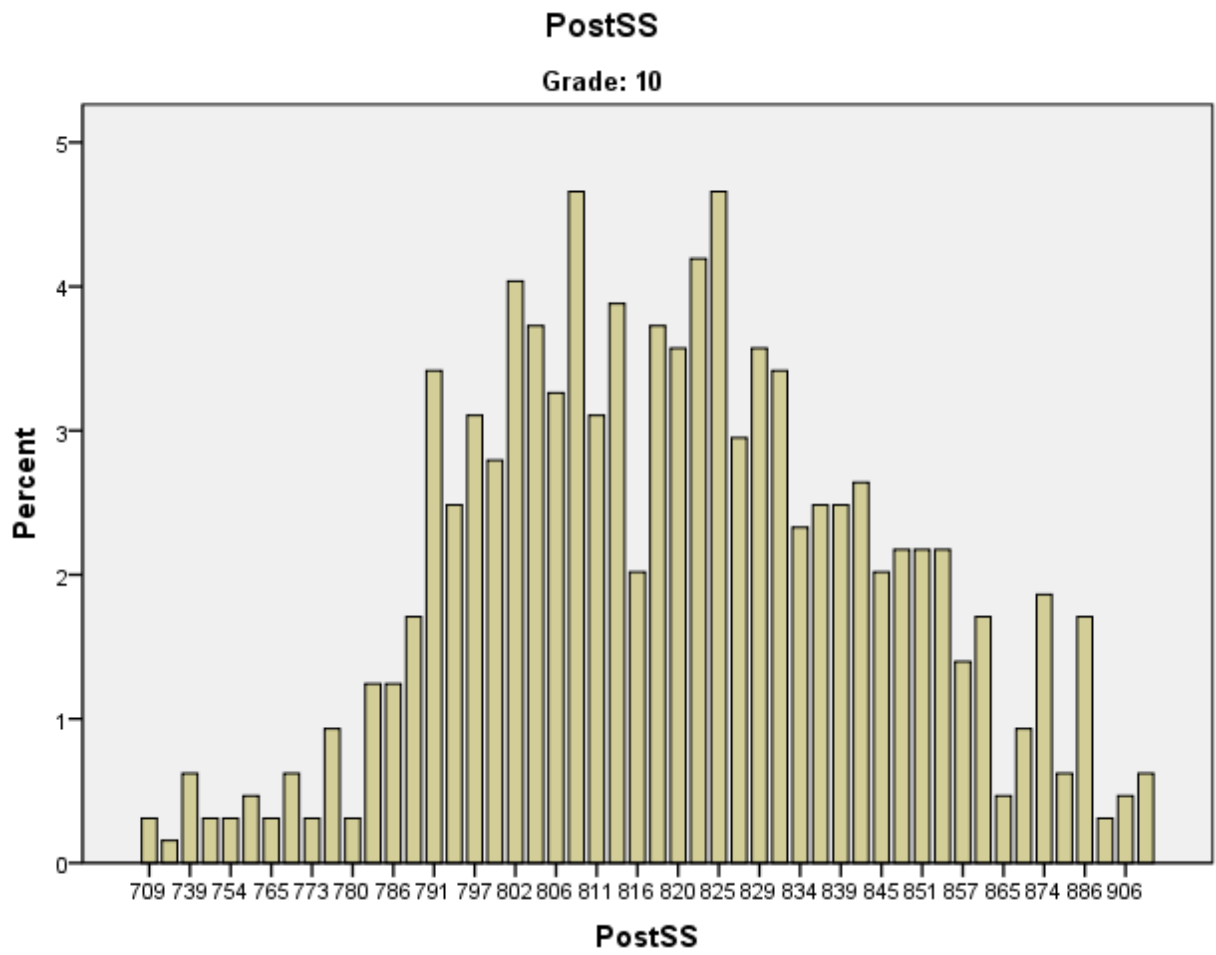


Figure 13 – Post-test Performance Level Frequencies (Grade 10)

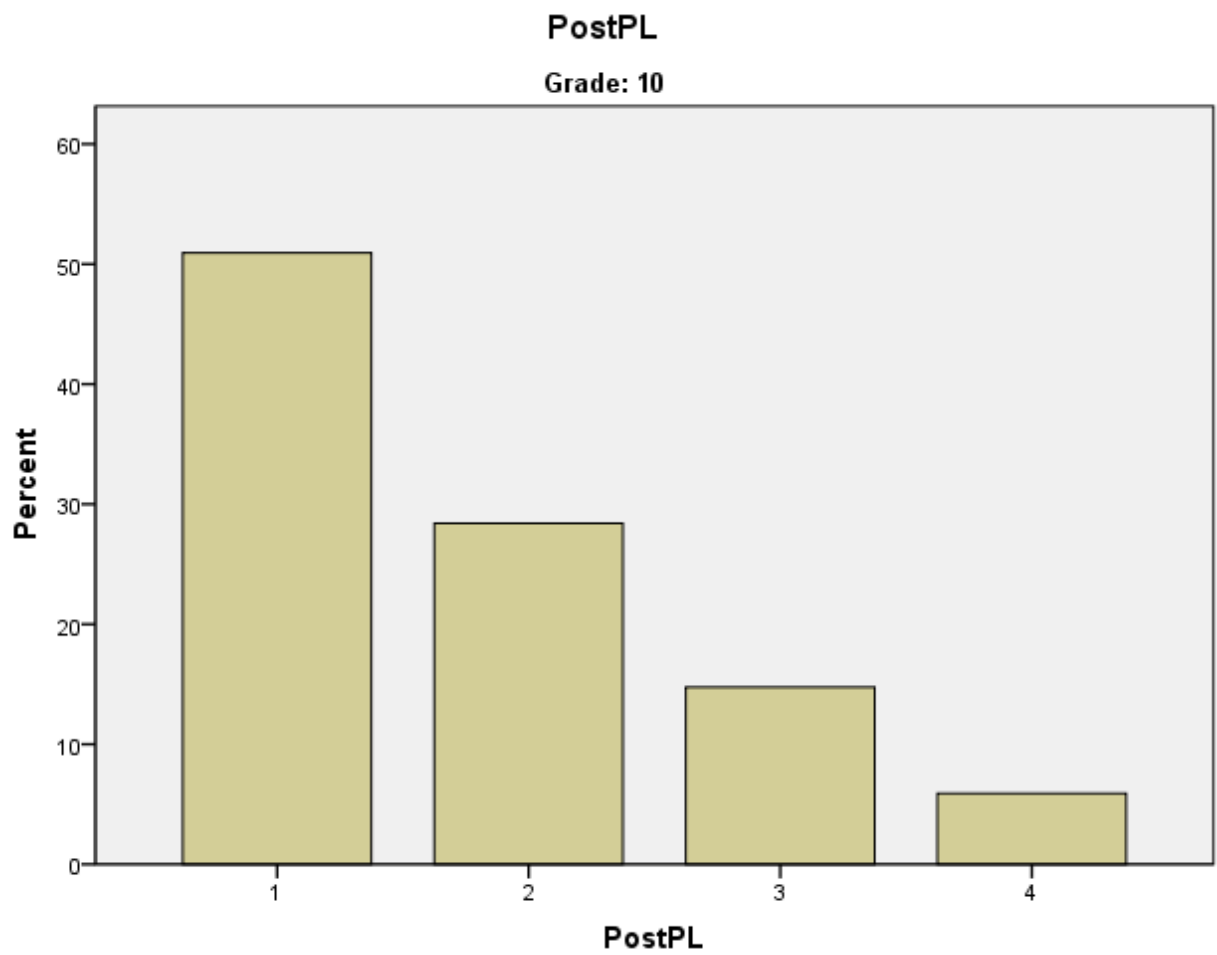


Figure 14 – Pre-test Scale Score Distribution (Grade 10)

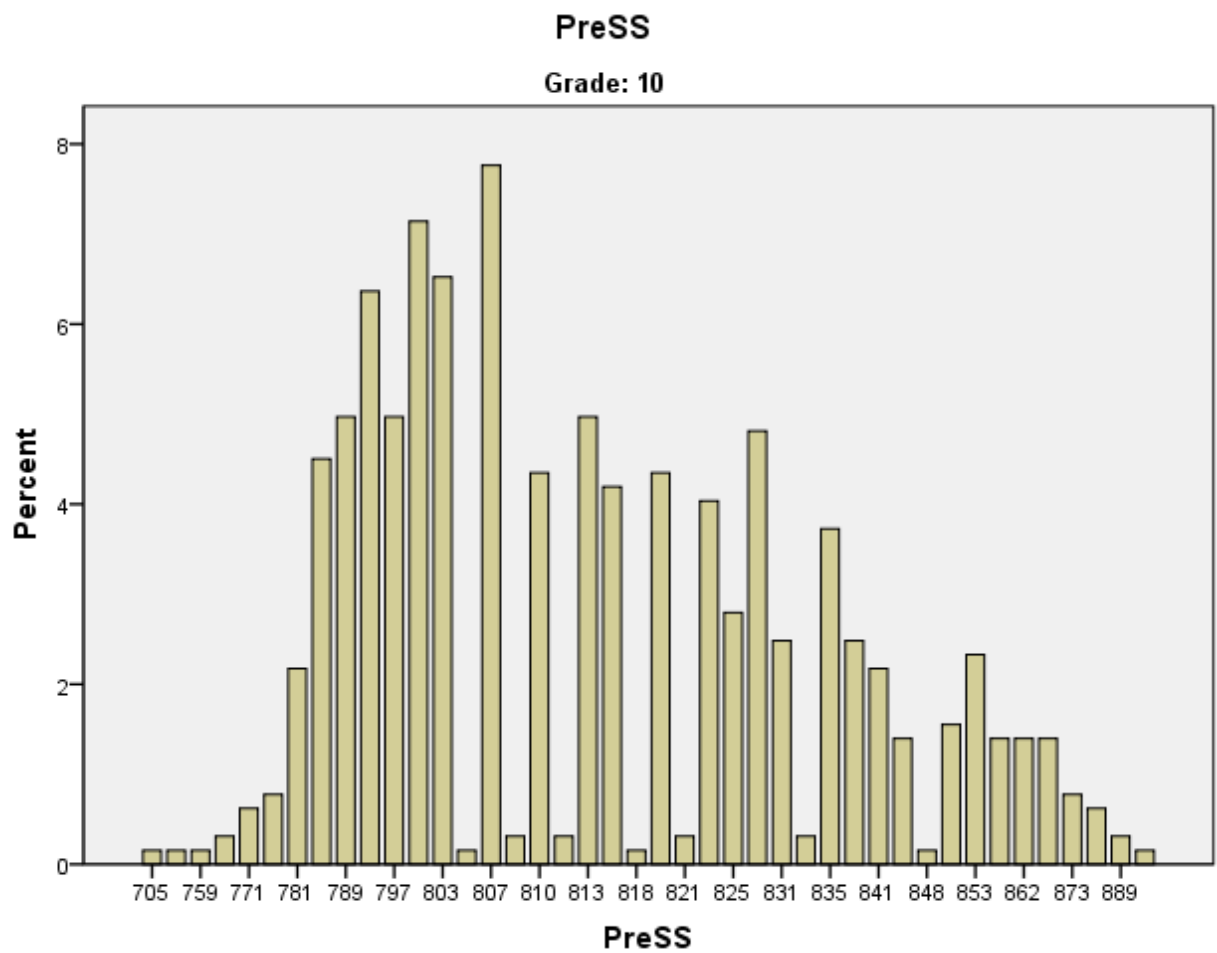
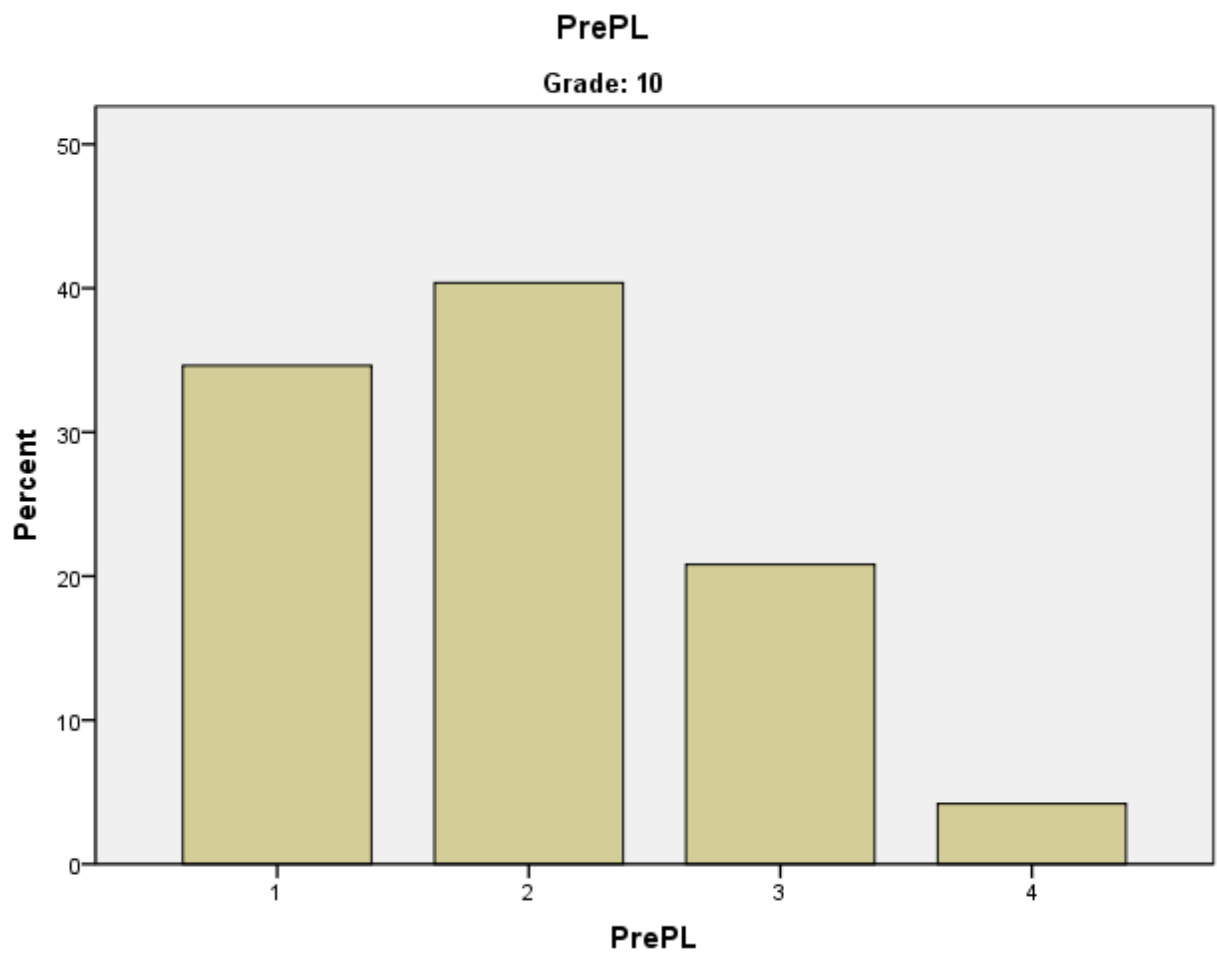


Figure 15 – Pre-test Performance Level Frequencies (Grade 10)



Appendix B

Item Characteristic Curves by Group

Figure 16 – Item Characteristic Curves by Group

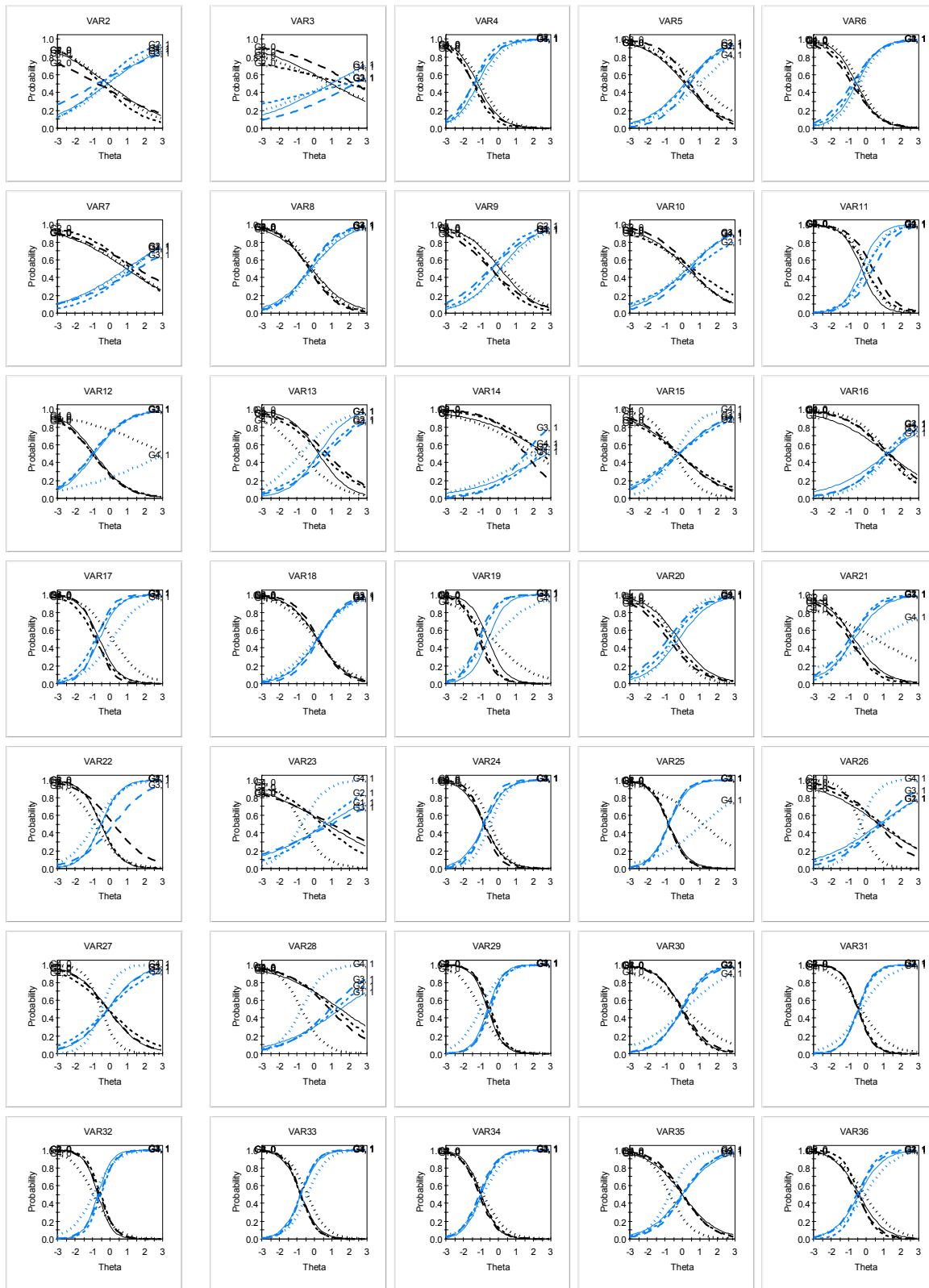
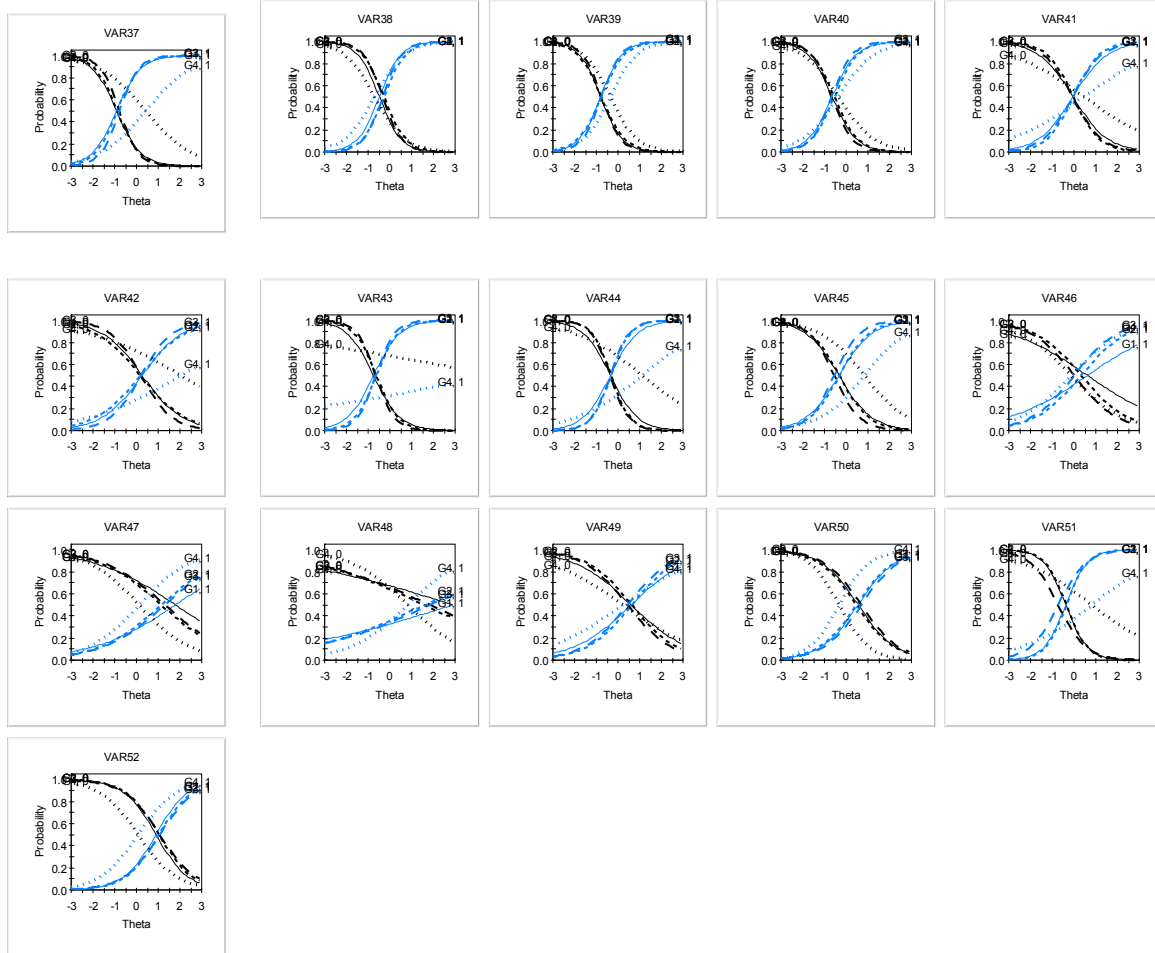


Figure 16 (cont'd)



Appendix C

IRT Calibration Values by Group

Table 14 – IRT Calibration Values by Group


	Census apar	Census bpar	Gr8 apar	Gr8 bpar	Gr9 apar	Gr9 bpar	Gr10 apar	Gr10 bpar
Item1	0.68	-0.14	0.59	-0.13	0.78	-0.48	0.44	-0.72
Item2	0.37	0.87	0.44	0.97	0.2	1.83	0.43	2.27
Item3	1.46	-1.01	1.56	-1.18	1.73	-1.43	1.42	-1.4
Item4	0.78	0.96	0.91	0.13	0.91	0.24	1.2	0.43
Item5	1.55	-0.37	1.53	-0.55	1.41	-0.67	1.25	-0.81
Item6	0.53	1.07	0.53	0.92	0.68	1.23	0.47	1.57
Item7	1.15	-0.08	0.96	-0.17	1.15	-0.28	1.24	-0.31
Item8	0.9	0.24	0.98	0.1	1.03	-0.39	0.8	-0.43
Item9	0.76	0.2	0.78	0.28	0.59	0.58	0.9	0.45
Item10	1.77	0.03	1.9	-0.27	1.51	0	1.49	0.4
Item11	0.38	3.11	1.07	-0.88	0.99	-1	1.05	-1.01
Item12	0.89	-0.78	1.18	0.25	0.79	0.42	0.82	0.72
Item13	0.52	2.01	0.49	2.78	0.7	2.45	1	1.55
Item14	1.38	-0.44	0.73	-0.24	0.64	-0.22	0.78	-0.17
Item15	1.01	1.31	0.6	1.22	0.91	1.09	0.81	1.33
Item16	1.2	0.12	1.77	-0.61	1.58	-0.84	2.19	-0.75
Item17	1	0.07	1.22	0.18	1.16	0.17	1.39	0.25
Item18	0.88	-0.31	1.93	-0.6	1.99	-1	2	-1.13
Item19	1.47	-0.5	1.09	-0.21	1.05	-0.55	1.03	-0.82
Item20	0.46	0.56	1.15	-0.68	1.48	-0.82	1.16	-0.96
Item21	1.29	-0.87	1.77	-0.53	1.71	-0.5	1	0.2
Item22	1.41	-0.73	0.49	0.7	0.69	0.44	0.41	1
Item23	1.97	-0.46	1.63	-0.82	1.76	-0.78	2.18	-0.84
Item24	0.69	1.18	1.79	-0.76	2.05	-0.76	1.93	-0.81
Item25	1.97	-0.32	0.58	0.75	0.64	0.93	0.87	0.68
Item26	2.1	-0.51	1.06	-0.12	0.79	-0.11	1.02	-0.14
Item27	1.53	-0.74	0.56	1.48	0.7	1.2	0.81	0.88
Item28	1.56	-1.03	2.28	-0.61	2.2	-0.45	2.35	-0.54
Item29	0.77	0.15	1.41	-0.05	1.46	-0.07	1.28	0.01
Item30	1.25	-0.28	2.24	-0.47	2.13	-0.43	2.3	-0.47
Item31	1.55	-0.95	2.51	-0.69	2.43	-0.53	2.22	-0.6
Item32	1.78	-0.41	1.95	-0.77	2.17	-0.79	2.23	-0.76
Item33	1.58	-0.86	2.07	-0.99	1.85	-1.02	1.95	-1.09
Item34	1.59	-0.71	1.03	-0.01	1.15	-0.02	1.24	0.06
Item35	1.22	-0.08	1.42	-0.4	2.1	-0.4	1.62	-0.52
Item36	0.94	0.43	1.79	-0.92	1.94	-0.88	2.28	-0.78
Item37	1.44	-0.7	2.01	-0.49	1.99	-0.27	2.26	-0.34
Item38	1.62	-0.35	2.02	-0.8	1.84	-0.77	2.25	-0.76
Item39	1.21	-0.37	1.8	-0.65	1.86	-0.54	2.12	-0.66

Table 14 (cont'd)

	Census apar	Census bpar	Gr8 apar	Gr8 bpar	Gr9 apar	Gr9 bpar	Gr10 apar	Gr10 bpar
Item40	0.57	0.45	1.26	-0.05	1.66	-0.04	1.48	-0.07
Item41	0.47	2.05	1.08	0.26	0.97	0.18	1.39	0.22
Item42	0.16	4.63	1.64	-0.75	1.98	-0.62	2.33	-0.65
Item43	0.65	1.12	1.6	-0.35	2.25	-0.32	2.29	-0.36
Item44	0.99	0.84	1.32	-0.34	1.46	-0.32	1.71	-0.55
Item45	0.83	-0.14	0.54	0.61	0.93	0.33	1	0.01
Item46	0.85	0.13	0.53	1.79	0.7	1.15	0.7	1.31
Item47	0.76	0.82	0.25	2.89	0.37	1.59	0.36	1.85
Item48	0.56	0.22	0.75	0.6	0.96	0.64	0.98	0.43
Item49	1.44	-0.32	1.17	0.55	1.14	0.47	1.15	0.69
Item50	0.61	0.82	2	-0.39	2.06	-0.36	1.54	-0.73
Item51	1.14	0.08	1.36	0.9	1.3	1.04	1.2	1.07

REFERENCES

REFERENCES

- Asparouhov & Muthen (2012). Multiple group multilevel analysis. Web Note 16.
www.statmodel.com
- Barrett, P (2011). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences*, 42, 815-824.
- Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A., & Long, J. S., Eds. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. v. D. Ong, Manfred H (Ed.), *Oxford handbook of methods in positive psychology*. Series in positive psychology (pp. 153-175). New York, NY: Oxford University Press.
- Browne, M. W., (2001) An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36, 111–150.
- Castellano, K. E., & Ho, A. D. (2012). [*A Practitioner's Guide to Growth Models*](#). 
[castellano_and_ho_-_practitioners_guide_to_growth.pdf](#).
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255.
- Cook, D.A. & Beckman, T.J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine* (2006) 119, 166.e7-166.e16.
- CCSSO ASR Working Group (2009). *Guide to United States Department of Education Growth Model Pilot Program 2005-2008*. Retrieved from the Council of Chief State School Officers Web site:
http://www.ccsso.org/Resources/Publications/Guide_to_United_States_Department_of_Education_Growth_Model_Pilot_Program_2005-2008.html
- Council of Chief State School Officers (2009). *Guide to United States Department of Education Growth Model Pilot Program 2005-2008*.
http://www.ccsso.org/Resources/Publications/Guide_to_United_States_Department_of_Education_Growth_Model_Pilot_Program_2005-2008.html
- Enders, C.K., & Tofghi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, 15, 75-95.

- Gorsuch, Richard L. (1983), *Factor Analysis* 2nd ed., Hillsdale, NJ: Erlbaum
- Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37, 827–838.
- Hayduk, L., Cummings, G. G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! Testing! One, two three – Testing the theory in structural equation models! *Personality and Individual Differences*, 42, 841-50.
- Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*, 1, 179-188.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453.
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409-426.
- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, 36(3), 347-387.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.
- Kenny, D.A. (2012). <http://davidakenny.net/cm/fit.htm>
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2011). *The performance of RMSEA in models with small degrees of freedom*. Unpublished paper, University of Connecticut.
- Kenny, D. A., , & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, 10, 333-351.
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130-149.
- Martineau, J. A. (2006) Distorting value added: the use of longitudinal, vertically scaled student achievement data for growth-based value-added accountability. *Journal of Educational and Behavioral Statistics*, 31(1), 35-62.
- Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, 13, 127-43.
- Meredith, W. (1993). Measurement Invariance, Factor-Analysis and Factorial Invariance. *Psychometrika*, 58(4), 525-543.

- Meredith, W. & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement invariance. *Psychometrika*, 57(2), 289-311.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13-24.
- Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.
- Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, 30(4), 577-605.
- Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, 2(3), 248-260.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297-334.
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered- Categorical Measures. *Multivariate Behavioral Research*, 39(3), 479-515.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6-12.
- Muthén, L.K. and Muthén, B.O. (1998-2012). *Mplus User's Guide*. Seventh Edition. Los Angeles, CA: Muthén & Muthén.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49, 115-132.
- Muthén, B., & Asparouhov, T. (2002). Latent Variable Analysis With Categorical Outcomes: Multiple-Group And Growth Modeling In Mplus. . *Mplus Web Notes: No. 4 Version 5*. Retrieved from <http://www.statmodel.com/download/webnotes/CatMGLong.pdf> website:
- Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5-8, 13, 24.
- Perie, M, Marion, S., & Gong, B. (June, 2007). A framework for considering interim assessments. Paper presented at the National Conference on Large-scale Assessment Conference sponsored by the Council of Chief State School Officers, Nashville, Tennessee.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.

- Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice*, 17(2), 13-16.
- Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics*, 29, 117-120.
- Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, 20, 643-671.
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15, 201-292.
- SteenKamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research* 25, 78-90.
- Tavakol, M. & Dennick, R. (2011). Making sense of Cronbach's Alpha. *International Journal of Medical Education*. 2011; 2:53-55
- Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*. American Psychological Association.
- Thurstone, L. L. (1947). *Multiple-factor analysis: a development and expansion of The Vectors of Mind*. Chicago, IL: University of Chicago Press.
- Tofighi, D., & Enders, C.K. (2007). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317-341). Greenwich, CT: Information Age.
- USED, 2005. PRESS RELEASES.Secretary Spellings Announces Growth Model Pilot, Addresses Chief State School Officers' Annual Policy Forum in Richmond Archived Information. <http://www2.ed.gov/policy/elsec/guid/secletter/051121.html>
- USED, 2006. <http://www2.ed.gov/admins/lead/account/growthmodel/cc.doc>
- Williams, L.J. & O'Boyle, E. (2011). The myth of global fit indices and alternatives for assessing latent variable relations. *Organizational Research Methods*, 14(2), 350-369.
- Wright, B.D., & Stone, M.H. (1979). *Best Test Design*. Chicago, IL: MESA Press