UNIDIMENSIONAL VERTICAL SCALING OF MIXED FORMAT TESTS

IN THE PRESENCE OF ITEM FORMAT EFFECT

by

Debra White Moore

B.M.Ed., University of North Carolina, Greensboro, 1989

M.A., University of Pittsburgh, 2006

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Research Methodology

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

School of Education

This dissertation was presented

by

Debra W. Moore

It was defended on

July 8, 2015

and approved by

Clement A. Stone, Professor, Psychology in Education

Feifei Ye, Assistant Professor, Psychology in Education

Levent Kirisci, Professor, Pharmaceutical Sciences and Psychiatry

Dissertation Advisor: Suzanne Lane, Professor, Psychology in Education

UNIDIMENSIONAL VERTICAL SCALING OF MIXED FORMAT TESTS

IN THE PRESENCE OF ITEM FORMAT EFFECT

Debra White Moore, PhD

University of Pittsburgh, 2015

The purpose of this study was to contribute to the existing body of evidence on vertically scaling mixed format tests by examining the impact of item format effect in conjunction with specific configurations of common item sets on two of the most popular calibration methods under test specification and scaling scenarios likely to exist in practice. In addition to advice for practical application provided by the investigation, this study also explored the impact of explicitly modeling the vertical scale factor when simulating data compared to a traditional model for in which the underlying vertical scale is implied.

Using a CINEG data collection design, six grade level tests, consisting of 61 items, were created with a 9:1 ratio of multiple-choice to constructed-response items and two different sets of 14 mixed format items designated as common items. Ability distributions for 2000 students per grade level were generated with the mean ability for successive grade levels increasing at varying increments to simulate grade level separation along with four covariance structures that reflected varying degrees of correlation to simulate item format effects.

Under a 3PL-GRM model combination, expected scores were calculated with recovery of expected score used as evaluation criteria. Ability and item parameters were estimated using the MLE proficiency estimator in MULTILOG and transformation constants were calculated in

STUIRT using the Stocking-Lord linking method. The performance of separate and pairwise concurrent calibration was then examined by calculating average BIAS and average RMSE values across 100 replications.

While the results of the study provided evidence that item format effects, vertical scaling method, and separation between grade levels significantly impacted the vertical scales, influence of these variables was often in combination with one another. The major findings were (1) pairwise concurrent calibration holistically performed better compared to separate calibration; (2) moderate to large item format effects were more likely to bias resultant vertical scales; (3) a large separation between grade levels resulted in a more biased vertical scale; and (4) explicitly modeling the vertical scaling factor during data generation influenced mean RMSE values more significantly than mean BIAS values.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

ACKNOWLEDGEMENTS

1.0   INTRODUCTION

1.1   STATEMENT OF THE PROBLEM

1.1.1   Background

In educational contexts, it is often desirable to examine growth in student achievement across time, particularly within large-scale assessment programs. However, students learn so much from one year to the next, even within the same domain, that a single test designed to evaluate growth over time would be inappropriate since items covering early grade content would be too easy for upper grade students and items covering upper grade content would be too difficult for early grade students (Kolen & Brennan, 2004). Typically, to accommodate this, a series of tests are constructed where each test is appropriate for a specific grade level and all tests are linked to the same score scale in a process known as vertical scaling (Kolen & Brennan, 2004).

As pressure mounts in large scale assessment programs to measure higher-level thinking processes, the use of mixed item format tests is increasing (Li, Lissitz, & Yang, 1999; Muraki, Hombo, & Lee, 2000; Sykes & Yen, 2000; Lane & Stone, 2006). Mixed item format tests are characterized by the use of different item types on the same test. Typically, this is a series of multiple-choice items which are dichotomously scored and a set of constructed-response items that are polytomously scored. Multiple-choice items can sample a broad range of the content

1

domain being measured while constructed-response items can elicit more complex cognitive processes related to the content domain (Linn, 1995; Martinez, 1999) with scoring of answers based on the quality of the answer rather than simply whether the answer is correct or incorrect (Muraki, Hombo, & Lee, 2000). The use of both item formats on the same test allows the test developer to combine the strengths of each item format while compensating for the weaknesses inherent in each item type (Martinez, 1999).

Vertical scaling of a dichotomously scored test is a complex process in which different choices can lead to different outcomes and, consequently, different interpretations of student growth (Tong & Kolen, 2007). No single combination of methodologies has been found to be better than another and inconsistencies abound in the literature (Harris, 2007). While some of the issues surrounding the use of vertical scales with tests that consist exclusively of dichotomously scored items are applicable to all tests regardless of item type, vertically scaling mixed format tests poses its own set of concerns (Meng, 2007; Cao, 2008; Kim, Walker, & McHale, 2010). It is clear that with the increased use of vertically scaled tests in large scale assessment and accountability programs, as well as value-added systems, and the stakes attached to the results of these endeavors, scaling decisions need to be carefully considered (Kolen & Tong, 2010). However, research is needed that examines the use of mixed item format tests in a vertical scale.

1.1.2 The Value of Mixed Format Tests

The cognitive and psychometric investigation of comparisons between multiple-choice and constructed-response item formats have centered on three major areas; practical concerns such as time and cost, content coverage, and cognitive processes being tapped (Cao, 2008). In general, scoring multiple-choice items is inexpensive compared to the time and cost of scoring

2

constructed-response items and multiple-choice items can cover a broad range of content quickly, especially compared to the amount of time students must invest in answering constructed-response items (Linn, 1995; Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998). In addition, the scoring of constructed-response items relies on the development of extensive scoring rubrics and use of raters who must be trained (Lane & Stone, 2006). Even with training, scoring of constructed-response items still has some degree of subjectivity and measurement error (Tate, 1999; Tate, 2000). Also, due to the time required to construct an answer, only a small number of constructed-response items can be used during any single test administration. However, an assessment consisting of only a limited number of constructed-response items often leads to inadequate domain sampling (Linn, 1995; Dunbar, Koretz, & Hoover, 1991).

While multiple-choice items can be written in such a way as to elicit divergent thinking, constructed-response items are capable of eliciting a broader range of complex cognitive processes (Martinez, 1999). In some cases, not using constructed-response items can lead to content under-representation as some higher order cognitive processes cannot be tapped with multiple-choice items (Messick, 1995). Additionally, test taking skills help students eliminate improbable answer choices to multiple-choice items potentially leading to construct-irrelevance and inflated scores (Burton, 2001).Thus, combining both item formats into the same test makes use of the strengths of both item types while offsetting the weaknesses, potentially offering a better understanding of student learning and growth (Martinez, 1999; Cao, 2008).

1.1.3 Attractiveness of IRT for Scoring Mixed Format Tests

Combining both item formats on a single test, however, necessarily brings into question how the test will be scored. Weighting polytomously scored items equally with dichotomously scored items on the same test may not represent their importance in the overall test specification plan (Li, Lissitz, & Yang, 1999). The potential weighting options; weighting by time required for each item type, equally weighting each item type, and weighting by raw score points for each item type, all have substantial limitations (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998). Weighting by the amount of time a student takes to complete each item results in constructed-response items having greater weight than multiple-choice items and, consequently, unpredictable results in total score for individual students depending on their relative performance on the different item types (Wainer & Thissen, 1993).

The lower reliabilities typically associated with constructed-response items place limitations on the use of equal weighting and weighting by raw score points (Wainer & Thissen, 1993). In the case of equal weighting, the process inefficiently makes use of the measurement accuracy offered by the different item types in that the procedure treats each point as if it were equal, even though multiple-choice items can involve guessing (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998). Item response theory (IRT) solves the weighting issue by creating a single scale from pattern scoring while providing statistically optimal weights (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998). This eliminates the need to calculate and justify separate weights for the multiple-choice and constructed-response items when reporting a total score (Thissen, Wainer, & Wang, 1994; Sykes & Yen, 2000; Kim & Lee, 2006).

1.1.4   Choosing an IRT Model

Item Response Theory (IRT) has become a popular scaling approach in the educational measurement field (Kim & Lee, 2006). IRT is a method of modeling an examinee's performance on a given test item as a function of the characteristics of that item and the examinee's latent ability (Hambleton & Swaminathan, 1985). It specifies the relationship between the latent ability and observed performance on an item and represents it mathematically as an item characteristic function (Hambleton & Swaminathan, 1985). Those examinees with a higher latent ability are expected to have a higher probability of answering a given item correctly when compared to examinees with a lower latent ability (Hambleton & Swaminathan, 1985). The use of IRT in large-scale educational assessment programs began with application to dichotomously scored items, was extended to include polytomously scored options, and now, is being applied to mixed item format tests (Kim & Lee, 2006). Several different models exist within the IRT paradigm. Briggs and Weeks (2009) suggest that choice of IRT model can be statistical, philosophical, or pragmatic. Statistically, they note that more complex models will fit the data better and provide more precise estimates of examinees ability.

Within the context of vertically scaling tests consisting of dichotomously scored items under the assumption of essential unidimensionality, there are three applicable IRT models; a one parameter logistic (1PL) or Rasch model, a two parameter logistic (2PL) model, and a three parameter logistic (3PL) model (Hambleton & Swaminathan, 1985). These models differ in the number of item parameters (difficulty, discrimination, guessing) estimated (Hambleton & Swaminathan, 1985). Historically, the Rasch model has not performed well in vertical scaling contexts. Slinde and Linn (1979), using artificially created groups based on ability level, found that the Rasch model did not perform well. They proposed that the poor performance was due to

the lack of a guessing parameter, especially since the results were better when the low ability group was not involved in cross-validation studies. Using a more traditional methodology, Loyd and Hoover (1980) found similar results. Specifically, they noted that the results of the vertical equating were not independent of the ability group used in the equating. They concluded that the 1PL model probably did not meet one or more of the underlying IRT model's assumptions. Holmes (1982), using data from two adjacent grade levels of the Comprehensive Test of Basic Skills (CTBS), found that the Rasch model did not perform equally well across all ability levels; performing poorly for low ability examinees. The author attributed this to not taking guessing into account when low ability examinees are linked based on difficult items. Becker and Forsythe (1992), using data from three subtests of the Iowa Test of Educational Development (ITED) for grades 9 through 12, found that grade-to-grade variability increased as grade level increased, especially with the Rasch scaling method. After a comprehensive literature review of these studies and others, Skaggs and Lissitz (1986) recommended using a 3PL model for vertical scaling dichotomously scored tests.

From a constructed-response item perspective and also under the assumption of essential unidimensionality, there are several possible models; the partial credit model (PCM), the generalized partial credit model (GPCM), the graded response model (GRM), and the nominal model (NM) (Ostini & Nering, 2006). The use of IRT to scale mixed format tests requires selecting a model for the multiple-choice items as well as selecting a model for the constructed-response items (Baker & Kim, 2004; Kim & Lee, 2006). In studies that examined the equating of mixed format tests, the 1PL model typically was paired with the PCM while the 2PL and 3PL models typically were paired with the GRM or the GPCM (Bastardi, 2000). Fitzpatrick, Link, Yen, Burket, Ito, and Sykes (1996), using data from several state assessment programs, applied

the 1PL-PCM combination and the 3PL-GPCM combination to several mixed format tests and found the GPCM alone or the 3PL-GPCM combination fit the data better. The authors attributed this to the differing discrimination estimates (slopes) between the multiple-choice and constructed-response items as well as the inclusion of a guessing parameter in the 3PL model. Sykes and Yen (2000) used data from two mixed format state proficiency tests to compare the 1PL-PCM combination and the 3PL-GPCM combination. They found the 1PL-PCM combination could spuriously inflate item information for the constructed-response items which led to underestimated standard errors. They concluded the poor model fit of the 1PL-PCM combination was due to the lack of guessing parameter and the assumption of equal discrimination parameters across multiple-choice and constructed-response items. Since the Rasch model pairing has been found to be insufficient except in certain circumstances (Sykes & Yen, 2000), the 3PL model paired with the GPCM or GRM appears to be the most promising combination for a vertical scale (Bastardi, 2000).

1.1.5 Dimensionality

These IRT models, however, come with several strong underlying assumptions including the assumption of unidimensionality. In other words, the instrument in question is measuring only one underlying latent trait (Hambleton & Swaminathan, 1985). Unfortunately, evidence suggests that multidimensionality may exist within mixed format tests due to differences in item formats (Kim & Kolen, 2006; Yao, 2008) or between content dimensions within tests (Reckase & Martineau, 2004) and between tests of different grade levels due to subtle (and not so subtle) shifts in content domain across years (Reckase & Martineau, 2004; Yen, 1995; Yen, 1996). This

leads to several ways multidimensionality potentially can impact results when vertically scaling mixed item format tests within the test and across grade levels.

Evidence indicates that there is the potential for multidimensionality to exist within and across tests. Less clear, however, is how to deal with that dimensionality in practice. Evidence exists showing that using a unidimensional IRT model to estimate data that are multidimensional can result in biased vertical scales (Béguin, Hanson, & Glas, 2000; Béguin, & Hanson, 2002). However, multidimensional vertical scaling procedures developed thus far are difficult to implement and often lead to results that are difficult to interpret (Eastwood, 2014). On the other hand, some studies have shown that using a multidimensional scaling method leads to smaller bias in the resultant scale than a unidimensional procedure specifically when the multidimensional model is the bifactor model (Li, 2006: Li & Lissitz, 2012; Gibbons et al., 2007).

1.1.6 Characteristics of Common Items

The method of data collection for vertical scaling tests has the potential to impact results as well. While there are three basic designs used to collect data for use in IRT vertical scaling, the common items non-equivalent group (CINEG) design, or some variation of it, is the most widely used design to collect data for large-scale assessment programs (Muraki, Hombo, & Lee, 2000; Peterson, 2010). In the CINEG design, an appropriate test is designed for each grade level with a predetermined number of common items overlapping adjacent grades. Each test is administered to the appropriate grade level and the scores on the common items between adjacent grade levels are used to place all items on a common score scale (Kolen & Brennan, 2004). Using the CINEG design requires the construction of a set of items common for adjacent grade levels, but the

specifications for creation of the common item set is not clear. Research into the construction of the common items for a test with dichotomously scored items has shown that the composition of the item set affects equating and scaling results (Yao, 2008).

In the case of horizontal equating, the current recommendation is for the set of common items to be representative of the content and statistical characteristics of the overall test (Kolen & Brennan, 2004). In addition, it is recommended that the items be placed in the same position on the tests being equated (Hagee & Kolen, 2011). However, in a vertical scaling context, adjacent tests are deliberately designed to be more difficult as the grade levels increase (Chin, Kim, & Nering, 2006). Thus, finding a set of items that are common to adjacent grade levels can be difficult (Kolen, 2007). Therefore, for vertical scaling, a recommendation for determining common items is to select items that provide a good representation of the domain overlap between the adjacent grades levels (Tong & Kolen, 2011). In any case, the argument has been made that the scaling results will be determined by the particular set of items chosen as common items, so the choice of common items is an important consideration when vertically scaling with the CINEG design (Patz & Yao, 2007; Kolen, 2007).

One challenge when constructing a representative common item set for a mixed format test is that there will be few constructed-response items compared to the number of multiple-choice items on the test, so there will be fewer constructed-response options to contribute to the set of common items (Muraki, Hombo, & Lee, 2000). Additionally, since constructed-response items tend to be more memorable because of their uniqueness compared to a large set of multiple-choice items, test security becomes an issue; particularly when laws require periodic release of test forms (Muraki, Hombo, & Lee, 2000). As a result, a common item set consisting solely of multiple-choice items often is used in practice when equating and scaling mixed format

tests (Kim & Lee, 2006). This practice, however, is not consistent with the current recommendations on representativeness.

A second challenge to constructing a set of common items for a vertical scaling scenario is deciding which of the adjacent grade levels should be used to supply the common items. As Tong and Kolen (2011) note, common items can be chosen from both of the adjacent grade levels, just the upper grade level, or just the lower level. Choosing from only the upper grade level means that students may not have had instruction on the content covered by the items while choosing from only the lower level means that students in the upper grade could find the items too basic (Tong & Kolen, 2011, Peterson, 2010).

Most of the research into characteristics of common item sets that could potentially impact equating results of mixed format tests has focused on the length of the common item set, the proportion of format types of the items used to construct the set, and differences in proficiency level between groups (Cao, 2008; Hagee & Kolen, 2011). However, the results are inconsistent and there is no clear recommendation for how to select items for the common item set within a vertical scaling scenario (Chin, Kim, & Nering, 2006).

1.1.7 IRT Scaling

The set of common items referenced above are then used to relate each examinee's score to the base score scale by some method of calibration (Kolen & Brennan, 2004). IRT item and ability parameters can be estimated using separate runs for each grade level or in one single simultaneous run for all grade levels (Kolen & Brennan, 2004). These processes are known as separate calibration and concurrent calibration, respectively.

When using separate calibration, item parameters and ability distributions are estimated individually for each grade level or group (Kolen, 2006). The resultant IRT parameter estimates from each separate run are within a linear transformation of each other because of the invariance property of IRT modeling (Lord, 1980; Kim & Lee, 2006). After choosing one grade level as the base grade level, linking coefficients are estimated for the parameters and used to perform a series of chained linear scale transformations (Kolen & Brennan, 2004). The linking coefficients can be estimated by one of several methods that are divided into two classes; moment methods and test characteristic curve (TCC) methods. The moment methods include the mean/mean (Loyd & Hoover, 1980) and the mean/sigma (Marco, 1977) methods and the TCC methods include the Stocking-Lord method (Stocking & Lord, 1983) and the Haebara method (Haebara, 1980).

When concurrent calibration is used, all item and ability parameters for all grade levels are estimated in one simultaneous run (Kolen & Brennan, 2004). Once the simultaneous run is complete, the parameter estimates are on the same scale across all grades and do not require further transformation (Kolen & Brennan, 2004).

More recently, a hybrid of the two calibration methods has been proposed; pairwise concurrent calibration (Karkee, Lewis, Hoskens, Yao, Huang, 2003; Meng, 2007). In pairwise concurrent calibration, adjacent non-overlapping grade levels are concurrently calibrated and the resulting estimates are then placed on the same scale by one of the linking methods used with separate calibration (Karkee et al, 2003). Because pairwise concurrent calibration uses concurrent calibration along with separate calibration but with fewer linkings, it is believed that the procedure will perform comparably to or better than either concurrent or separate calibration

11

in the presence of multidimensionality (Meng, 2007). However, evidence of this has been contradictory (Meng, 2007; Karkee et al, 2003).

Much research has been done comparing separate calibration and concurrent calibration in equating and scaling. Generally, concurrent calibration results in more stable estimates, presumably because this process makes use of as much of the available data as possible (Kolen & Brennan, 2004; Kim & Cohen, 1998). From a logistic standpoint, in the case of unidimensional IRT (UIRT), Kolen and Brennan (2004) have made the point that concurrent estimation is easier to implement than separate estimation because only one computer run is required, but suggest that a computer program capable of estimating parameters for multiple groups should be used to guard against biased estimates. Additionally, they note that due to a large number of 'not reached' items in examinee response strings, there can be convergent problems with this method of calibration.

However, some empirical evidence suggests that separate calibration could be more appropriate in certain situations, specifically when substantial multidimensionality exists between grade level tests (Béguin, Hanson, & Glas, 2000; Yao & Mao, 2008). Additionally, Kolen (2006) proposed that separate calibration allows the researcher to examine item parameter estimates for individual items so as to identify items that are behaving differently for adjacent grade levels. Prevailing opinion is that separate calibration has the potential to ameliorate the effect of the violation of the assumption of unidimensionality because each grade level is calibrated separately and then linked (Kolen & Tong, 2010). When using separate calibration, empirical evidence suggests that the test characteristic curve linking methods perform better than the moment linking methods (Béguin & Hanson, 2002; Kim & Cohen, 1998).

1.1.8 Grade Level Separation

Kolen and Brennan (2004) have made the argument that the characteristics of the groups used in horizontal equating of test forms can impact the results. When the groups used in the horizontal equating scenario are randomly equivalent and representative; the equating relationship seems to be group invariant (Cook, 2010). However, when there are large differences in mean ability level between groups, the equating relationship can fail (Harris & Hoover, 1987; Kolen & Brennan, 2004; Hanson & Béguin, 1999). In vertical scaling, the expectation is that mean ability levels will be different across grade levels (Chin, Kim, & Nering, 2006; Kolen, 2007). Therefore, it is reasonable that when large differences in mean ability levels between grades exist, vertical scaling results could be less accurate (Chin, Kim & Nering, 2006).

1.1.9 Summary

Tracking student achievement over time is one way schools can provide accountability data to relevant stakeholders. As the pressure mounts to measure complex thinking skills, schools are turning to mixed format tests because of their ability to sample a broad range of content quickly with multiple-choice items while capturing higher order thinking skills with constructed-response items. IRT-based vertical scaling is an attractive procedure for examining growth in student achievement over time and is accommodating of mixed format tests. However, there are practical issues, some within the practitioner's control and some not, which can cause the resultant vertical scales to be biased. Biased results can lead to incorrect inferences about student growth at the aggregate or individual level. Despite considerable research into vertical scaling,

inconsistencies exist in the literature and no single combination of methodologies has been emerged as best practice.

1.2   PURPOSE

The purpose of this simulation study is to examine the performance of calibration methods under different conditions of format effect and common item characteristics when vertically scaling mixed format tests under the assumption of unidimensionality. Using a 3PL-GRM UIRT model combination, the CINEG data collection design, different conditions of grade level separation and item format effect as reflected in the covariance structure of the ability parameter generation, the performance of separate calibration and pairwise concurrent calibration will be examined with two configurations of common item sets. In addition, data will be generated in the traditional way by assuming the existence of a vertical scale underlying the test data and in a non-traditional way in which the underlying vertical scale is explicitly modeled during the ability generation process.

1.3   RESEARCH QUESTIONS

There are five main research questions that will be addressed by this study:

> *Research Question #1*: Does item format effect impact the resultant vertical scale when scaling mixed format tests if unidimensionality is assumed?

> *Research Question #2*: Does the range in difficulty of the common items impact the resultant vertical scale when scaling mixed format tests in the presence of multidimensionality?

14

Research Question #3: Does the degree of separation in ability level between different grade levels impact the resultant vertical scale when scaling mixed format tests in the presence of item format effect?

Research Question #4: Which vertical scaling method (separate or pairwise concurrent) produces the most accurate vertical scale in the presence of item format when unidimensionality is assumed?

Research Question #5: Does explicitly modeling the otherwise assumed vertical scale underlying the test data influence the simulation?

1.4 SIGNIFICANCE OF STUDY

Beginning with No Child Left Behind (NCLB: U.S. Department of Education, 2001) and continuing with the Race to the Top program (RTT: U.S. Department of Education, 2009), a spotlight was focused on accountability in education. Under NCLB and continuing with RTT, the focus shifted from accountability based on static proficiency classifications to accountability through examining student growth in achievement (U.S. Department of Education, 2009). Within this context of assessment and accountability, mixed format tests offer advantages over single format tests with their ability to tap higher level thinking skills while still covering a broad range of content in a reasonable time frame. Additionally, IRT-based vertical scaling offers one method of examining such growth in student achievement over time and is appropriate for mixed format tests.

However, some characteristics unique to mixed format tests, groups being scaled, and techniques used to generate the scales can result in less than desirable vertical scales. For instance, when constructing vertical scales, assessment programs tend to assume unidimensionality within tests. Evidence, however, suggests that this is an unlikely scenario.

Simulation studies in which item format effects are manipulated within the context of vertical scaling have not been done despite the possibility that both of these effects may be present in operational tests currently being used. Additionally, how the most popular calibration methods will perform within the same context in the presence of item format effect is not certain.

Further, despite research already done on equating and scaling mixed format tests, there are still outstanding questions about how to choose common items when using a CINEG design. While the use of multiple-choice items only as common items when scaling a mixed format test is typical in practice, whether or not this is best practice is still in question. Additionally, research is inconsistent concerning how to construct the set of common items used for linking purposes and if large differences in mean ability level between grades will result in less desirable vertical scales.

As part of RTT, the Race to the Top Assessment (RTTA) initiative was founded with the intent to grant federal funding to groups willing to design assessments based on the Common Core State Standards (CCSS) that could be used to examine growth in student achievement from year to year (U.S. Department of Education, 2010). Two multi-state consortiums, The SMARTER Balanced Assessment Consortium (SMARTER, 2012) and the Partnership for Assessment of Readiness for College and Careers (PARCC, 2013), secured federal funding to develop assessments based on the CCSS. Both of these consortiums have designed assessments that use performance-based components (constructed-response) in addition to multiple-choice items (PARCC, 2013). In addition, **S**MARTER plans to use a vertical scale to measure student growth (SMARTER, 2012) and PARCC is considering the use of a vertical scale, as well (Kolen, 2011; Briggs, 2011).

While some research exists to guide these and other practitioners concerning best practices in vertical scaling, most of the research takes place in the context of horizontal equating or in vertical scaling of single format tests. The dearth of research pertaining to the vertical scaling of mixed format tests in general, as well as the impact of item format effect, degree of grade level separation, and the need to know the necessary characteristics for common items, call for additional research. This study will add to the existing body of evidence on vertically scaling mixed format tests by examining the impact of item format effect in conjunction with specific configurations of common item sets on the most popular calibration methods used for creating vertical scales. In addition to these practical applications, this study will also explore the explicit modeling of the vertical scale factor compared to the traditional model for generating data in which the underlying vertical scale is implied when simulating data for this type of study.

2.0   LITERATURE REVIEW

In this chapter, the literature on IRT-based vertical scaling is reviewed with an emphasis on the studies pertaining to mixed format tests. The section begins with a review of the literature specific to vertically scaling mixed format tests (Karkee, Lewis, Hoskens, Yao, and Haug, 2003; Yao & Mao, 2004; Meng, 2007). Given the small number of studies conducted in this area, the section will then review studies concerning equating of mixed format tests as they relate to the aspects highlighted previously and supplemented with studies concerning equating of single format tests when necessary, providing a rationale for the choice of factors of investigation in this study.

2.1   VERTICALLY SCALING MIXED FORMAT TESTS

Few studies have been done examining the vertical scaling of mixed format tests. Most of these studies have focused on the performance of the most popular calibration methods traditionally used for single format tests applied to the mixed format context. Karkee, Hoskens, Yao, and Haug (2003) examined the performance of separate, concurrent, and pairwise concurrent calibration using operational data from a statewide mathematics mixed format test administered to grades 5 through 10. The grade 5 test consisted of 69 items, 54 multiple-choice items and 15 constructed-response items, while the grade 6 through 10 tests consisted of 60 items in which 45

were multiple-choice items and 15 were constructed-response items. The authors randomly selected 10,000 responses from a possible 54,000 responses with 5,000 of those responses used for the calibration study and the other 5,000 responses used for a cross-validation study. A CINEG design, in which at least 20 common items of a mixed format nature, was used to create the vertical scale. Researchers chose ten of the common items from the core grade level test and the other 10 common items from the test for the grade level below the core test. The 3PL model was used to estimate the multiple-choice items, the 2-parameter partial credit was used to estimate the constructed-response items, and grade 7 (grade 7/8 for the pairwise concurrent calibration) was used as the base grade level.

Results were compared for non-convergence, model fit, differential item functioning (DIF), grade-to-grade growth, grade-to-grade variation, separation of grade level distribution, and expected versus observed performance. The authors found that separate calibration using the Stocking-Lord linking method produced consistently better results when compared to concurrent calibration and pairwise concurrent calibration in terms of model fit, convergence, and DIF. For grades 6 through 10, pairwise concurrent calibration performed better than concurrent calibration. Also, when parameter and ability estimates were compared, results were similar to those outlined above were reported across calibration methods. Grade-to-grade growth was similar across methods although differences did exist between methods for the end grade levels. Grade-to-grade variability was relatively flat for separate and concurrent calibration while pairwise concurrent calibration showed greater variability in grades 5 and 6.

Jodoin, Keller and Swaminathan (2003) investigated the impact of calibration method and proficiency estimator (ML, EAP, and raw scores) on ability estimates and classification consistency (4 levels) for a mixed format statewide mathematics test. The authors used real data

from three different years of the test that resulted in a matrix-sample external anchor equating design in which 5000 examinees took one of 12 difference external item blocks. Using the 3PL model for the multiple-choice items, the 2PL model for short answer items, and the GRM for open response items, vertical scales were generated by concurrent calibration, separate calibration and the mean/sigma linking method, and fixed common item parameter method (FCIP). The lowest year was used as the base level for all conditions.

To examine differences between calibration methods and estimators, three ability estimates were generated for each examinee; raw score, ML, and EAP. In addition, each examinee was classified into one of four 'proficiency categories' for each of the three different ability estimates. Cross-tabulations were performed to assess consistency of classification between the methods. ML ability estimates showed greater mean growth than EAP ability estimates. Also, the proportion of examinees classified as 'below proficient' by ML ability estimates decreases more quickly than the proportion of examinees classified in the same category by EAP ability estimates. With regard to calibration method, separate calibration with mean/sigma linking showed less growth in student achievement across years when compared with concurrent and FCIP calibration. Classification consistency for all calibration methods were highly correlated (> 0.94). When differences occurred between separate calibration and FCIP calibration, FCIP calibration consistently classified students in the next highest category comparatively. When differences occurred between concurrent calibration and FCIP calibration, concurrent calibration consistently classified students in the next highest category as well.

While ability estimates regardless of proficiency estimator were highly correlated with each other across calibration methods (< 0.091), concurrent and FCIP calibration classified more students in the middle categories when compared to separate calibration which classified more

students in the extreme categories (highest and lowest). This pattern of results was found for both EAP and LM estimators.

Yao and Mao (2004) investigated the performance of different calibration methods in the presence of multidimensionality when vertically scaling a mixed format test across five grade levels under the assumption of unidimensionality. Using item parameters from a large-scale mixed format writing assessment, they simulated multidimensional data with a set of 13-18 common items consisting of multiple-choice items only. Then, they compared the performance of concurrent, separate, and pairwise concurrent calibration in terms of accuracy of proficiency scores distributions using mean squared error of frequency over replications. Results showed that when data were simulated to be two dimensional, a unidimensional solution was sufficient and produced more stable proficiency score distributions over time; however, the two dimensional solution produced the smallest mean bias. Results for the three dimensional condition were similar. Although the three dimensional solution produced the smallest mean bias, the two dimensional solution produced comparable results and the one dimensional solution produced the largest mean bias. In terms of accuracy of calibration method, when the data was unidimensional, separate calibration was more accurate than concurrent calibration. However, when the data was multidimensional, concurrent calibration was more accurate than separate calibration; but the score distributions that resulted from the three different dimensional estimations varied little regardless of calibration method. The authors propose that this was due to the high correlation between items of different formats.

Meng (2007) examined the performance of separate, concurrent, semi-concurrent, and pairwise concurrent calibration methods when vertical scaling unidimensional data from a mixed format test across grades 3 through 8. Sample size, number of common items, format of the

common items, and the ratio of polytomously scored items to dichotomously scored items on the test were varied. Using item parameters from a large-scale reading assessment, eight 60-item study tests were created with different numbers of multiple-choice and constructed-response items for the entire test and for the common item sets. Multiple-choice items were estimated using the 3PL model and constructed-response items were estimated using the GPCM model. Grade 5 was used as the base grade level.

Results of each condition were evaluated using the root mean squared error (RMSE), standard error, and absolute bias for estimation of four criterion; proficiency score means, proficiency score standard deviations, patterns of student growth (effect sizes), and classification proportions across replications. Pairwise concurrent and semi-concurrent calibration were most accurate in estimating proficiency score means. Likewise, pairwise concurrent or semi-concurrent calibration methods were more accurate and stable in terms of recovering patterns of student growth, as measured by effect size, when compared to separate or concurrent calibration. For classification proportions, results were dependent on the classification level. For the lowest level, concurrent calibration was most accurate in reproducing proportion of classification, while pairwise concurrent calibration was most accurate for classification levels 2 and 4. Semi-concurrent calibration resulted in the most accurate estimates of the third classification level. Concurrent calibration produced the lowest error when recovering proficiency score standard deviations.

For all calibration methods, increasing sample size decreased error in estimating proficiency score means and standard deviations, effect sizes, and classification proportions. Additionally, using a mixed format common item set increased the accuracy of pairwise concurrent, semi-concurrent, and concurrent calibration methods across the four parameters of

evaluation. When using only multiple-choice items as common items, the smaller number of common items resulted in better estimates of proficiency score means and effect sizes across calibration methods, while the larger number of common items resulted in more accurate estimates of proficiency score standard deviations and classification proportions for the lowest two classification levels. The accuracy of classification proportions for the top two levels was dependent on calibration method used. While generally pairwise concurrent or semi-concurrent calibration resulted in more accurate estimates proficiency score means, standard deviations, effect sizes, and classification proportions, the author noted that increasing sample size, increasing the number of common items, using a mixed format common item set, and increasing the number of constructed-response items produced more accurate and stable results when vertically scaling a mixed item format test. Given that some of these parameters are out of the practitioner's hands, the author suggests using a mixed format common item set to produce the most accurate vertical scales.

2.1.1   Summary

Four published studies have examined the vertical scaling of mixed format tests. Making generalizations from these studies is difficult because overlap between them is small. All four studies used real item parameters from tests used in statewide large scale assessments and three of the four studies used the 'lowest' grade level as the base grade level to examine the performance of the most popular calibration methods for creating vertical scales. However, the number of common items, sample size, number of grade levels scaled, and the format of the common items differed between them. Results suggest that the number of common items, as well as the format of the common items, impacts the resultant vertical scale and that dimensionality

23

can make a difference under certain conditions. Additionally, performance of the different calibration methods was inconsistent across the studies and conditions.

2.2   DIMENSIONALITY

A chief concern when scaling mixed format tests is the potential for multidimensionality to bias results, especially since it is common practice to assume unidimensionality. Past research has shown IRT ability and item parameter estimation to be somewhat robust to violations of the assumption of unidimensionality under certain conditions. Simulating different conditions, Reckase (1979) investigated the impact of multidimensionality introduced when more than one underlying trait was being measured. Using 50-item tests, 5 of which were operational data sets and 5 were data sets simulated to represent specific factor structures and ability parameter distributions, he investigated the relative performance of the 1PL and 3PL model when estimating multidimensional data. In general, he found that the 3PL model picks one factor and discriminates among the ability continuum while ignoring the other factors. On the other hand, the 1PL model estimates ability parameters based on a sum of the factors. For the 3PL model, whether there was one dominant latent trait with a series of weakly correlated latent traits or two independent latent traits; the resulting ability parameter estimates were relatively stable as long as the first factor accounted for at least 10% of the total test variance. However, item parameter estimates were unstable under these conditions unless the dominant factor accounted for at least 20% of the total test variance. In addition, under the 3PL model, the discrimination factor was related to the size of the factor loadings on the dominant factor and the shape of the item difficulty parameter distribution.

Drasgow and Parsons (1983) conducted a simulation in which they used a higher order latent trait with a series of lower level factors that were correlated to varying degrees. They found that as long as the dominant trait was large enough (intercorrelations 0.46 or higher), item parameter estimates were unbiased. Likewise, Harrison (1986) found IRT item parameter estimation to be robust to violations of the unidimensionality assumption. Harrison manipulated several factors including; test length, number of common factors, distribution of item factor loadings, and the strength of the factor loadings on the common factor. Results showed that test developers can improve item and ability parameter estimation by increasing the number of relevant test items, increasing the number of common factors, and balancing the influences of the common factors.

Using GRE verbal scores, Dorans and Kingston (1985) examined the robustness of IRT equating to violations of the unidimensionality assumption. Factor analysis of GRE verbal scores found that two distinct verbal abilities were present; discrete verbal ability and reading comprehension ability. Calibrating all items together or calibrating based on item content, ability parameters and item discrimination parameters were estimated and compared. Also, IRT equating was performed across six conditions; equivalent group versus anchor-test, concurrent calibration or separate calibration then linking with the Stocking-Lord procedure, and how the old form was calibrated compared to how the new form was calibrated (both separate, both concurrent, old-concurrent with new-separate). Ability parameter estimates, item discrimination parameters, and raw-scores from the equating conditions were compared.

Correlations between the overall verbal ability estimates and the discrete verbal ability estimates ranged from .86 to .89 while correlations between the overall verbal ability estimates with the reading comprehension ability estimates ranged from .73 to .77. The authors concluded

that the test measured two distinct, but highly correlated abilities. Comparison of item discrimination parameter estimates showed similar results. Calibration by content resulted in a higher mean discrimination estimate than calibrating all items together and reading comprehension items had a greater difference in mean item discrimination parameter when compared to the simultaneous calibration of all items together. Raw-score to raw-score comparisons found that results of the equatings differed by no more than 10 scale points. The greatest differences in scale score comparisons occurred at the extremes of the scale score range. The authors concluded that dimensionality exerts its effect on equating through its impact on the estimation of the item discrimination parameter and that since the equatings were similar, that IRT equating may be robust to violations of unidimensionality like found in this dataset.

All of these studies examined the impact of within-grade multidimensionality that may exist when a test measures more than a single underlying ability. These studies also suggest that when one underlying dimension accounts for a substantial portion of the overall test variance or if the underlying dimensions being measured are strongly correlated, IRT equating is somewhat robust to violations of the unidimensionality assumption.

2.2.1 Item Format Effects

There is evidence that adding additional item formats increases the dimensional complexity of a test (Kim & Kolen, 2006; Yao, 2008). Several studies have investigated the dimensionality of mixed format tests with inconsistent results depending on the content and structure of the test. Evidence suggests that the extent to which the constructed-response and multiple-choice portions of tests measure the same content and skills, tests can be essentially unidimensional (Yao, 2008).

Bennett and colleagues (1990, 1991) investigated the dimensionality of the College Board's Advanced Placement (AP) Computer Science exam using factor analysis in two different studies. In the first study (Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990), they fit four different factor models to two sets of dichotomously scored items, polytomously scored items, and constrained polytomously scored items. They found a one factor solution was sufficient for one set of items, but the second set of items required a two factor, albeit highly correlated, solution. In the second study (Bennet, Rock, & Wang, 1991), the authors compared model fit indices for a one factor and a two factor solution. The two factor structure investigated was a one factor multiple-choice item and one factor constructed-response item solution. In order to analyze the entire test, they parceled the 50 multiple-choice items into five 10-item bundles creating a summed score from 0-10 for each bundle. While the one factor model did not fit particularly well according to the goodness-of-fit measure, it was the more parsimonious solution.

Thissen, Wainer, and Wang (1994) re-examined the same Advanced Placement Computer Science test data, but used a bifactor solution. In the bifactor model, a general multiple-choice factor and two associated constructed-response factors that were uncorrelated with the multiple-choice factor was fit to the data. The results suggested that the constructed-response items measured something unique compared to the multiple-choice items based on the small but significant loadings on the constructed-response factors. However, since the multiple-choice and constructed-response items loaded more heavily on the general factor than the single factors, the authors concluded that the structure was *essentially* unidimensional in that the different item types measured the same construct for most of the test. The authors then completed the same analysis for the Advanced Placement Chemistry examination and found

similar results. As the authors note, however, these tests were designed specifically to measure the same construct.

Perkhounkova and Dunbar (1999) applied DIMTEST to the language arts and mathematics tests of the ITBS and the Constructed-Response Supplement (CRS) to the ITBS for seventh and eighth graders. They investigated the structure of three researcher-created tests; multiple-choice items only test, constructed-response items only test, and a test that was a combination of both item types. On the language arts tests, the researchers found the constructed-response and mixed format test to be essentially unidimensional. While the multiple-choice test was considered to be essentially unidimensional, it was noteworthy that it was not 'strictly' unidimensional due to nuisance factors associated with content and item format. This phenomenon, however, was not consistent across the grade levels examined. For the mathematics test, the multiple-choice and mixed format tests were not found to be essentially unidimensional. Again, the results were not consistent across grade levels and differences were attributed to item format, content, and location of the item within the test. Additionally, the constructed-response test was determined to be dimensionally distinct from the multiple-choice test. Overall, the researchers concluded that using multiple-choice and constructed-response items on a single test may introduce complexity into the dimensional structure of the resultant mixed format test. Sykes, Hou, Hanson, & Wang (2002), using 35 multiple-choice and 10 constructed-response items from a field test of a mathematics state assessment for fifth graders, determined the structure to be multidimensional across item format with constructed-response items loading on two dimensions and multiple-choice items loading primarily on the second of two dimensions.

2.2.2 Multidimensional IRT Models

Despite evidence that unidimensional IRT (UIRT) models are somewhat robust to violations of the dimensionality assumption, continued concerns over the impact of multidimensionality on the estimation of item and ability parameters within educational assessment by UIRT models have led to the proposal of several multidimensional IRT (MIRT) models (Yon, 2006). Three MIRT models have been proposed; partially (or non) compensatory, compensatory, and, more recently, the bifactor model. Compensatory and partially compensatory models differ in the way that information from the underlying latent abilities is combined with the item characteristics to calculate the probability of a correct response (Reckase, 2009; Yon, 2006).

In a partially compensatory, sometimes called noncompensatory, MIRT model, a task is broken into component parts and a unidimensional model is used to estimate each component with the resultant probability of a correct response for the task being the product of these separate unidimensional probabilities (Reckase, 2009). The multiplicative nature of this model means that a low ability estimate on one dimension is partially compensated for by a higher ability estimate on another dimension (Yon, 2006). However, because of estimation difficulties associated with partially compensatory MIRT models and the fact that they have not been extended to polytomously scored items, most research into MIRT equating has been done with compensatory MIRT models (Reckase, 2009; Yon, 2006).

In a compensatory MIRT model, a linear combination of latent abilities is used with a logistic (or normal ogive) model to specify the probability of a correct response (Reckase, 2009). The linear combination approach allows for different combinations of ability estimates to combine for the same overall sum (Reckase, 2009). This means that a high ability in one domain can 'compensate' for a lower ability in another domain (Yon, 2006). As with UIRT, MIRT

comprises a series of models that mathematically express the interaction between persons and items (Reckase, 2009). For dichotomously-scored items, the compensatory MIRT model has been mathematically expressed as an extension of the 1PL UIRT model (Adams, Wilson, & Wang, 1997), the 2PL UIRT model (McKinley & Reckase, 1983) and the 3PL UIRT model (Reckase, 2009). Additionally, for polytomously-scored items, Muraki and Carlson (1995) extended the GRM to multidimensional space while Yao and Schwarz (2006) extended the GPCM and Adams, Wilson, and Wang (1997) extended the PCM.

Just as with UIRT models, results from MIRT model estimation need to be linked to create a vertical scale. Li and Lissitz (2000) argue that three indeterminacies must be resolved in order to equate MIRT data; rotational, unit, and origin. This means that axes of the new form need to be rotated to match the axes of the base form, units of identified dimensions on the new form need to be compressed or expanded to match the units of the base form, and origins of identified dimensions on the new form translated to match the origins of the base form (Yao, 2006). Several linking methods have been proposed in conjunction with compensatory MIRT models. While some equating methods are multidimensional extensions of UIRT equating methods, newer MIRT equating methods have been proposed that differentially address how to resolve these indeterminacies (Yon, 2006).

Recent years have seen a rediscovery of the bifactor model as a way of modeling multidimensionality in educational contexts (Reise, 2012). The bifactor model is characterized by a single general factor that accounts for common variance across all items and one or more group factors that account for common variance above and beyond the general factor among these groups of items (Reise, 2012). To apply a bifactor model to an educational assessment, each test item would be an indicator for both the general factor and one additional group factor

30

(Li & Lissitz, 2012). In the same educational assessment scenario from a covariance structure perspective, the covariance of the set of all items is explained by the general factor and the additional variance of sets of items within the overall set of items is explained by group factors (Reise, 2012). This requires that each item have a nonzero loading on the general factor and only one nonzero loading on the group factors and that the group factors be orthogonal to the general factor and to the other group factors in the model (Gibbons & Hedeker, 1992). Extending the bifactor model into the context of vertical scaling, the general factor, then, would reflect the overall vertical scale across all grade levels while the group factors would represent individual grade level dimensions (Li & Lissitz, 2012). Since items can load only on the general factor and only one additional grade level factor, regardless of how many grade levels are being scaled, estimating the bifactor model is no more computationally complex than estimating a two-dimensional MIRT model (Li & Lissitz, 2012). Also, the orthogonal nature of the relationship between the general factor and the grade level factors as well as the orthogonal relationship among grade level dimensions eases interpretation of the model, especially when compared to high-dimension MIRT models (Li & Lissitz, 2012).

In addition to the advantages of computational simplicity and ease of interpretation, bifactor models have been applied successfully to student achievement data in a variety of contexts; unidimensional data (Li & Lissitz, 2012; Gibbons & Hedecker, 1992), multidimensional data (Li & Lissitz, 2012; Reise, Morizot, & Hays, 2007, Gibbons & Hedecker, 1992), testlet applications (Rijmen, 2010), vertical scaling (Li & Lissitz, 2012), vertical scaling with construct shift (Li & Lissitz, 2012), and differential item functioning (Fukuhara & Kamata, 2011). The bifactor model also has been extended to the graded response model to accommodate

31

polytomously scored items (Gibbons, Bock, Hedecker, Weiss, Segawa, Bhaumik, Kupfer, Frank, Grochocinski, Stover, 2007).

Any of these MIRT models, theoretically, can account for the effects of both item format effects and shifts in content across grade levels if the correct model can be specified. However, in practice, assessment systems continue to rely on unidimensional IRT models for vertical scaling procedures for several reasons (Yon, 2006; Li & Lissitz, 2012). First, some MIRT models, (i.e., high-dimension, partially compensatory, partially compensatory polytomously score items) are difficult or not possible to estimate at this time (Reckase, 2009, Yon, 2006). Additionally, there are outstanding questions about how different multidimensional linking methods will behave within the context of vertical scaling and, as yet, no single multidimensional linking method has emerged as preferable in all situations (Yon, 2006). Also, despite the advantages of the bifactor model, not all researchers are convinced of its applicability to item response data. Chiefly, the underlying premise of the model that all items load on the general factor and only one grade level factor and that these factors are orthogonal to one another may not accurately reflect real-world item response data (Reise, 2012). While research continues in these areas, applicability of these MIRT models to state assessment programs currently is limited.

2.3   CHARACTERISTICS OF THE COMMON ITEMS

When scaling tests, performance on the common items is used to estimate the average amount of growth from year to year (Tong & Kolen, 2010). This places a considerable burden on the selection of common items. Evidence based on research in horizontally equating single and

mixed format tests suggests that common items at least be statistically and content representative of the entire test (Cao, 2008; Tong & Kolen, 2010). This presents a challenge when common items must span adjacent grade levels, especially since common items that are too hard or too easy for the intended population can result in skewed distributions and biased results (Peterson, 2010). Most research into the necessary characteristics of common item sets for mixed format tests have focused on whether it is necessary to include both item types or if a multiple-choice only common item block is sufficient. A few additional studies have examined the impact of varying ranges of difficulty levels for the common item set.

2.3.1   Characteristics of Common Items for Mixed Format Tests in Horizontal Equating

Using simulated and real data for both a horizontal and vertical scaling scenario, Li, Lissitz, and Yang (1999) investigated the effect of the proportion of score points from different item types when the proportion of dichotomously scored items was less than, equal to, or more than the polytomously scores items in the set of common items when linking a mixed format test. Using the item parameters from a fourth grade reading and writing assessment, the 3PL model for dichotomously scored items, and the GPCM for the polytomously scored items; the authors found that increasing the proportion of multiple-choice items in the common item set did not lead to less (or more) error in recovery of equating coefficients across all research conditions. The authors concluded that the characteristics of the item parameters for the common items may be more influential than the total number of common items in the set on the equating accuracy.

Tate (2000) simulated data designed to mimic a writing examination by correlating the multiple-choice items and constructed-response items at 0.6. He investigated the performance of the moment methods and the Stocking-Lord method when the common items were multiple-

choice only or mixed-format. The test consisted of 30 multiple-choice items and 10 constructed-response items with either a multiple-choice only common item set or a 2 constructed-response and 6 multiple-choice mixed item format set of common items. Using the 2PL and GRM models for estimation, he found that using multiple-choice only items as common items was satisfactory when the test was strictly unidimensional. However, linking performance was better when the proportion of constructed-response items to multiple-choice items within the common item set mimicked the proportion of item types in the overall test regardless of the presence of multidimensionality.

Using simulated data, Bastardi (2000) investigated the impact of test length, proportion of each item type on the test, anchor test length, sample size, and differences in ability distributions on the performance of concurrent and separate calibration. The 3PL model was used to estimate dichotomous items and the GRM model was used to estimate the polytomous items. Anchor tests were mixed format with two conditions of anchor test length, 10% and 20% of total test length, examined. The calculated RMSE across 50 replications between the estimated and true test curves for both the multiple-choice items and the constructed-response items were analyzed with an ANOVA. Results indicated that anchor test length had no effect on the accuracy of the linkings for multiple-choice items, but that longer anchor tests resulted in less bias in the accuracy of linkings for constructed-response items.

Sykes, Hou, Hanson, and Wang (2002) investigated the effect of multidimensional anchor item sets on resulting student scores. Using real data from a statewide fifth grade mathematics test, the authors modified the operational form of the test to contain the same set of items, but different items designated as anchor items. Four anchor item sets were designed to be content and statistically representative of the original test form with an average item difficulty

34

similar to the average item difficulty of the overall test form. Two sets were considered baseline common item sets and had items which loaded approximately equally on two factors (balanced). The other two common item sets had factor loadings which showed the items loading more heavily on one factor compared to the other factor (unbalanced). Using a 3PL model for the multiple-choice items, the GPCM for the polytomously scored items, and the Stocking-Lord linking method; the authors found that the balanced anchor item sets resulted in lower standard errors of equating than either unbalanced common item sets. Additionally, the authors caution that designing an anchor test to be a 'miniature version' of the overall test is not sufficient unless the dimensional structure of the test is also taken into account as well.

Kim and Lee (2006) also investigated the effect of the ratio of multiple-choice to constructed-response items in the common item set on horizontal equating results. The authors created three different types of mixed format tests in which the proportion of multiple-choice items to constructed-response items was varied. In addition, they also varied the proportion of multiple-choice items to constructed-response items in the common set. The tests and common item sets were identified as 'dichotomously scored dominant' or 'polytomously scored dominant' depending on which item type contributed the larger number of response categories. The calculated absolute bias and RMSE across 100 replications between the estimated and true test curves were compared. The authors determined that when linking tests, using common item sets that were the same 'dominant type' as the test resulted in smaller bias and mean squared error in item parameter estimation regardless of linking method performed.

Cao (2008) systematically investigated the effect of representative and non-representative common item sets on horizontal equating results. The author created a 54-item mixed format test with 18 common items and used the 3PL/GRM combination to examine the effect of

dimensionality and various conditions of statistical, format, and content representativeness when assuming unidimensionality. Two conditions of format representativeness were investigated; 8:1 ratio of multiple-choice to constructed-response item format with a 2:1 score point ratio anchor test and an all multiple-choice item anchor test. Three conditions of content representativeness were investigated; equally balanced between content areas, one content area under-represented, and one content area missing. Two conditions of statistical representativeness were investigated; average item difficulty similar to the test and average item difficulty set to 0.3 mean different from test. Absolute bias, RMSE, and classification consistency for expected score over 100 replications were calculated and analyzed with an ANOVA. Results from the unidimensional condition showed that anchor tests that were statistically representative of the overall test had equating bias closer to zero, smaller RMSE, and higher classification consistency. There were no significant differences in equating bias, RMSE, or classification consistency among the conditions of content representativeness and format representativeness. In the presence of multidimensionality, again, content representativeness had no significant impact on the equating bias, RMSEs, or classification consistency. However, under the multidimensional condition, statistical representativeness and format representativeness did impact the accuracy of equating results. Specifically, anchor tests that were statistically representative resulted in lower equating bias, RMSEs, and higher classification consistency. Additionally, as the degree of multidimensionality increased, format representativeness played a more influential role in producing more accurate equating results in terms of lower bias, RMSEs, and higher classification consistency.

2.4   CALIBRATION METHOD

The linking methods described in the introduction were developed within a dichotomous IRT framework (Muraki, Hombo, & Lee, 2000). As polytomous IRT models increased in popularity, these linking methods were extended to these models as well (Baker, 1992; Baker 1993, Kim & Cohen, 1995; Cohen & Kim, 1998). Recently, Kim and Lee (2006) systematically extended all four of the popular linking methods to a mixed item format context and examined the performance of each of them. Research investigating the relative performance of these linking methods, whether in a single format or a mixed format context, have yielded mixed results.

2.4.1   Horizontal Equating of Single Format Tests

Kim and Cohen (1998) compared separate calibration, concurrent calibration using marginal maximum likelihood estimation (MLE), and concurrent calibration using marginal maximum a posteriori estimation (EAP) with simulated data varying the number of common items (5, 10, 20, and 50) and the mean ability level between groups (N(0,1) and N(1,1)). Results of the calibrations were evaluated by examining the root mean square differences for item discriminations and item difficulties and the mean Eucledian distances between the estimated item parameters and the item parameters from the generating model. They found concurrent calibration produced more accurate item discrimination and item difficulty parameter estimates, except when the number of common items was very small. In this case, separate calibration produced more accurate item parameter estimates. In general, however, it was noted that the larger the number of common items, the more accurate the results regardless of estimation method used. When using separate calibration, they found the test characteristic methods

outperformed the mean/sigma method when the number of linking items was small, but all methods produced similar results when the number of linking items was large.

Béguin, Hanson, and Glas (2000), using a simulated two-dimensional test model, performed a concurrent and a separate estimation of parameters using a unidimensional 3PL model. The covariance and variance of the second dimension for each model (covariance of .5, .7, and .9 with a variance of 1.25, 1.49 and 1.81) and the mean proficiency level for the first dimension of Form B (0 or 1) were varied. The mean proficiency for all forms was 0 for the second dimension. Results of the calibrations were evaluated using the difference between the score distribution for Form B and the score distribution of the generating model and the difference equivalent score points from the observed score equating function compared to the equivalent scores points for the generating model. They concluded that under the equivalent groups design, separate calibration was more accurate than the concurrent estimation procedure. Additionally, the error was very large for concurrent and separate estimation under the nonequivalent groups design and the researchers concluded that the method was not appropriate for this design.

Hanson and Béguin (2002) conducted another study using the common item design in which the number of common items (10 or 20), sample size (1000 or 3000), and mean proficiency of the second test form were varied. In addition to comparing concurrent and separate calibration, the authors compared the performance of all four popular linking methods (mean/mean, mean/sigma, Stocking-Lord, and Haebara) used with separate calibration. Items from two forms of the ACT mathematics exam were used to create two 100-item exams with 20 common items. Mean squared errors and bias were calculated based on differences in the true score equating functions between form B and form A and in the estimated item characteristic

curves between form B and the true item characteristic curves. Results showed that concurrent estimation produced equating functions and text characteristic curves that were less biased compared to separate estimation. In addition, when using separate calibration, test characteristic curve methods performed better in regards to bias and mean squared error than the moment methods, but little difference in bias or mean squared error was found between the Stocking-Lord (SL) method and the Haebara method.

Kim and Cohen (2002) investigated calibration method in a polytomously scored only test under the graded response model using simulated data. Item parameters from a mathematics test were used to generate scores for a 30-item, 5 ordered category polytomously scored test. In addition to calibration method, sample size (300, 1000), group ability level (N(0,1) and N(1,1)), and length of common item set (5, 10, 30) were varied. RMSDs for item difficulty, item discrimination, and ability parameters were calculated. As an overall index of accuracy of item parameter recovery, the mean distance measure was also calculated. Across all conditions for item difficulty, item discrimination, and ability parameter estimates; concurrent calibration resulted in smaller RMSDs and MDMs than the same parameters estimates from the separate calibration runs. However, regardless of calibration method, RMSDs and MDMs decreased as the number of common items increased. Also, regardless of calibration method, RMSDs and MDMs increased for item difficulty parameters that were farther from the mean of the ability distribution.

2.4.2 Vertical Scaling of Single Format Tests

Chin, Kim, and Nering (2006) conducted a simulation study in which they examined factors that can affect the results of vertically scaling dichotomously scored tests; degree of grade level

overlap (effect size 0.5 or 1.0), number of grade levels being scaled (3, 4, or 5), length of the set of common items (20%, 30%, or 40%), difficulty range of the linking items (narrow or wide) and calibration method (concurrent, separate with mean/sigma). To perform the simulation, item parameters for 60 dichotomously scored items per test per grade level and ability parameters for 10,000 examinees per grade level were generated. Using the 3PL model for estimation and the lowest grade level as the base grade level, sixty replications across all conditions were performed. Root mean squared errors and bias were calculated for each item and ability parameter to evaluate estimation accuracy. While no method or design consistently performed better across all conditions of the study, in terms of calibration method, concurrent calibration appeared to be less affected by restriction of common item difficulty range and number of common items in terms of RMSE and bias of item and ability parameter recovery. This observation was consistent across conditions except when there was a large difference in mean ability level between grades, especially as the number of grade levels being calibrated together increased and the number of common items decreased. When the difficulty range of the common items was restricted, separate calibration with mean/sigma linking performed poorly in terms of item and ability parameter recovery.

Ito, Sykes, and Yao (2008) used real data from a dichotomous item only reading and a dichotomous item only mathematics test for grades K-9 to compare the relative performance of concurrent and grouped-concurrent calibration. Approximately 1700 students were chosen randomly from all students taking the operational tests. To perform grouped-concurrent calibration, they concurrently calibrated grade *groupings* (K1, 23, 456, 789) and then linked them using the Stocking-Lord procedure. This reduced the number of linkings necessary to put all grades on the same score scale. Using grade 4 or grade grouping 456 as the base grade level,

item parameters and ability parameters were estimated using a maximum likelihood estimator. Ability estimates were then transformed onto a common score scale to allow comparisons between calibration methods. Correlations between item parameter and transformed scale scores for concurrent versus grouped-concurrent calibrations were compared. Correlations for item difficulty parameters between scaling methods were greater than 0.97, but were more similar at the base grade level than at the extremes. Also, item difficulty parameters between scaling methods for reading were more comparable than item difficulty estimates between methods for mathematics. Likewise, a similar pattern of results were found when comparing item discrimination parameters between scaling methods. Scale score means were also comparable between scaling methods although it was noted that deviations between methods became more pronounced when moving away from the base grade level with an increase in variance in scale scores at the upper and lower grade levels when using concurrent calibration. The authors concluded that scaling method may matter more in mathematics than in reading.

Using simulated data, Smith, Finkelman, Nering, and Kim (2008) examined the impact of multidimensionality on linking methods for an all multiple-choice item test. With grade 5 as the base grade level, they vertically scaled grades 3 through 8 using a test consisting of 60 items with 45 operational items and 15 common items. To evaluate results, summed square differences between estimated test characteristic curves and true test characteristic curves, RMSEs between ability estimates and true ability parameters, and percent of correctly ordered examinees were calculated and compared across 10 replications. The authors found the performance of the linking methods to suffer in the presence of multidimensionality with the Haebara method having more error than the other methods examined. Compared to the unidimensional results, summed square differences were larger in the presence of multidimensionality, especially for the

Stocking-Lord and Haebara linking methods at the extreme grade levels. RMSEs appeared to be very similar for all linking methods, but were higher for the multidimensional case. In terms of percent of correctly ordered examinees, all methods produced similar results (roughly 88%). Overall the authors concluded that all linking methods perform more poorly the further the transformation from the base grade level and that multidimensionality impacts vertical scaling results.

2.4.3   Horizontal Equating of Mixed Format Tests

Kim and Lee (2006) extended the four most popular linking methods (mean/mean, mean/sigma, Stocking-Lord, Haebara) to a set of mixed format scenarios within the context of horizontal equating. They used simulated data (abilities and items) in which the ability levels between groups (equivalent or nonequivalent), sample size (500, 3000), and the proportion of multiple-choice items to constructed-response items in the common item set (10/10, 20/5, 30/2) were varied. The 3PL model was used to estimate the parameters for the multiple-choice items and the GPC model was used to estimate the constructed-response items. The average mean squared error (MSE) and average bias between the estimated category characteristic curve and the true category characteristic curve across 100 replications was used as evaluation criterion. With regard to calibration method, concurrent calibration was preferable to separate calibration regardless of linking method due to smaller bias and MSEs. Additionally, across linking methods, test characteristic curve methods produced lower MSEs and bias than the moment methods.   They note, however, that when using separate calibration; linking through the dominant item type (determined through number of response categories and reliability of the parameters estimates for that item type) resulted in smaller MSEs. Further, results using the

42

Haebara method usually had the lowest MSEs of the four linking methods. The authors also note that the Haebara method can handle any mixture of IRT models, while the Stocking-Lord method cannot handle the nominal response model.

In addition, several studies described previously have investigated the efficacy of concurrent versus separate calibration with mixed format tests also within the context of horizontal equating. In Li, Lissitz, and Yang (1999), the authors found concurrent calibration to be an unbiased estimator for mixed format tests, but results were substantially impacted by the proportion of different item types included in the common item set. Likewise, Tate's simulation study (2000) examined both horizontal and vertical equating and found that the moment methods and the Stocking-Lord method performed similarly with regard to recovery of true linking coefficients, but the performance of all linking methods suffered when the proportion of multiple-choice items to constructed-response items was not similar to the overall test in the presence of multidimensionality. Bastardi's study (2000), also found concurrent calibration resulted in smaller RMSEs compared to separate calibration and linking with the Stocking-Lord method. However, in Cao's (2008) simulation study, concurrent calibration was found to be influenced by the presence of multidimensionality. Specifically, bias and RMSE of the expected score increased with increasing multidimensionality.

Kim and Kolen (2006) investigated the impact of item format effects on the performance of linking methods. Simulating data to mimic different item formats effects, they compared concurrent calibration to separate calibration using both the TCC and moment linking methods. They simulated two conditions of item format effect (small and large), two types of mixed format tests (narrow and wide) as determined by the information functions, and three conditions of differing ability levels between groups (0, .5, and 1). The authors found concurrent calibration

outperformed separate calibration regardless of linking method, however, the differences in accuracy and robustness to the multidimensionality introduced by format effects was small. They also noted that the TCC linking methods produced more consistent and stable results when separate calibration was used.

2.5   GRADE LEVEL SEPARATION

Harris and Hoover (1987), using data from an administration of the ITBS for grade 3 through 8, determined that IRT vertical scaling within their specific context was not person-free, but rather was dependent on the ability level of the groups. This is problematic because in vertical scaling, it is assumed that the ability level will increase across grade levels (Chin, Kim, & Nering, 2006). Studies examining this aspect of vertical scaling within the context of mixed format tests are few, but several of the studies described previously have examined the impact of group ability distribution on horizontal equating and vertical scaling.

Within the context of a dichotomously scored test, Chin, Kim, and Nering (2006) determined that the degree of grade level ability overlap impacted the performance of calibration method. As the separation between grade levels increased, accuracy of recovery of item difficulty, item discrimination, and ability parameter decreased, regardless of whether separate or concurrent calibration was used. Using a polytomous item only test, Kim and Cohen (2002) examined the effect of differences in mean ability distributions on concurrent and separate calibration methods. In general, regardless of calibration method, RMSDs and bias were smaller when the ability distributions of both groups were well- matched to the distribution of the item difficulty parameters.

Within the context of horizontal equating of mixed format tests, Kim and Lee (2006) determined that bias and MSE between the observed category characteristic curve and the true category characteristic curve increased when linking non-equivalent groups as compared to linking equivalent groups. Likewise, Cao (2008) found that group ability distribution was the most influential of the factors investigated in their simulation study. Specifically, the equivalent groups condition outperformed the non-equivalent groups condition in estimation of the expected score consistently across all conditions of common item configuration. In the non-equivalent groups condition, the differences in mean ability level for the two groups was 0.5, which is the typical mean grade level separation for simulated vertical scaling studies.

2.6 OTHER CONSIDERATIONS

There are four additional factors that potentially can impact the results of the vertical scaling process; choice of base grade level, number of grade levels being vertically scaled, the proficiency estimator used, and the presence of construct shift. While not of particular importance in this study because they will be held constant during the simulation, a brief review of the literature investigating these aspects is given as justification for choices made.

2.6.1 Base Grade Level

When using the CINEG design and separate or pairwise concurrent calibration, one grade level must be designated the base grade level (Kolen & Brennan, 2004). Tong and Kolen (2007)

examined the effect of choosing a base grade level other than grade 3 for some of the replications (common items design with Stocking and Lord linking method) of their simulated data. The authors saw little effect of changing the base grade level. Hendrickson, Cao, Chin, and Li (2006) also found little effect in choice of base grade level for the scaling test design. However, Kim, Lee, and Kim (2008) found that the further the transformed grade level was from the base grade level, the more error was introduced with the estimation of the linking constants and thus, more bias is introduced into the vertical scale. Smith, Finkelman, Nering, and Kim (2008) found similar results when using the equivalent groups design and five popular linking methods.

2.6.2   Number of Grade Levels being Vertically Scaled

In practice, it is typical to link six grade levels (Karkee et al., 2008; Briggs & Weeks, 2009). However, Chin, Kim, and Nering (2006) note that as the number of grade levels being vertically scaled increases, additional error can be introduced. This can be either due to the long chain of linkings needed when using separate calibration or the number of parameters that must be estimated simultaneously when using concurrent calibration. Additionally, when using concurrent calibration, the potential for construct shift to introduce bias increases because of the change in content and complexity of operations between the elementary grades and the secondary grades (Yen, 1986).

2.6.3   Proficiency Estimation in Vertical Scaling

When performing IRT vertical scaling, a decision must be made about how to estimate ability (Kolen, 2006). In general, there are three common methods used to estimate proficiency

parameters, maximum likelihood (ML), maximum a posteriori (MAP), and expected (EAP) a posteriori (Briggs & Weeks, 2009; Tong & Kolen, 2007). The ML method maximizes the likelihood function of the examinee's response pattern to estimate ability level (Tong & Kolen, 2007). The MAP and EAP use a Bayesian approach to estimate ability with MAP being the mode and EAP being the mean of the posterior proficiency distribution (Tong & Kolen, 2007).

Jodoin, Keller and Swaminathan (2003), reviewed previously, found differences in resultant vertical scales based on choice of proficiency estimator. They found that MLE ability estimates consistently show larger mean growth than EAP ability estimates or raw scores. Also, the proportion of students classified as below proficient reduced more quickly across years when using ME ability estimates as compared to EAP ability estimates or raw scores. The authors concluded that choice of proficiency estimator made a difference in making inferences about student growth over time and in proficiency classification.

Tong and Kolen (2007) investigated the three IRT proficiency estimators mentioned above, as well as quadrature distribution (QD) estimation, in their study. They used real data from 4 content areas (vocabulary, reading, math, language) from a standardized test for grades 3 to 8 and 9 simulated datasets across 3 sample sizes (500, 2000, 8000) and three within-grade variability conditions (increasing, decreasing, constant). Concurrent calibration under the 3PL model was compared to Thurstone scaling for the scaling test design and the common item scaling design. Grade 3 was used as the base grade level for all scaling procedures and the grade level separation for the simulated study was set to the grade level separation found for the vocabulary test. In addition, EAP estimates were generated using both pattern scoring and summed scores (EAP_PS, EAP_SS). Mean and standard deviation estimates for the ability parameters, as well as Yen's effect size and horizontal distance (percentile differences on scale

47

scores for the same percentage between two distributions) were used to evaluate results of the scalings.

With regard to IRT scaling using the common item design, they found all estimators tended to produce similar ability estimates under the simulated data condition when the underlying assumptions of the IRT model were met. However, when the underlying assumptions of the IRT model were not met, EAP_PS, EAP_SS, and QD estimators tended to produce the most accurate ability estimates. For real data, all proficiency estimators and scoring methods produced mean ability estimates within 1 scale score point. Differences, however, were observed for standard deviation estimates and effect sizes with ML estimates producing somewhat different ability estimates from the other estimators; specifically, overestimated within-grade standard deviations and underestimated effect sizes.

In another study, Kolen and Tong (2010) investigated the performance of ML estimator, EAP estimation, summed score EAP (EAP_SS) estimation, and the test characteristic curve function (TCF) for four real data situations. The vertical scale used for this investigation was created using concurrent calibration of all items from a multiple-choice only vocabulary scaling test for grades 3 to 8 with the grade 3 set as the base grade level. Comparing the resultant proficiency estimates from each estimator, within-grade standard deviations for the ML and TCF estimators were greater than 1 for grade 3 while the standard deviations for the EAP and EAP_SS estimators were smaller than 1. Since this was not unexpected due to the behavior of Bayesian proficiency estimators, effect sizes were calculated as a measure of the magnitude of group differences. The Bayesian estimators consistently produced larger effect sizes, but all effect sizes decreased with increasing grade level regardless of estimator. As Kolen and Tong (2010) note, however, no estimator seems to be superior to the others, but it is clear that choice

48

of proficiency estimator effects vertical scale results and must be carefully considered and is most likely dependent on the content of the tests being vertically scaled.

2.6.4   Construct Shift

Yen (1986) raised the concern that as content changes across grades to include increasingly more complex operations and thought processes, multidimensionality will be introduced. This increasing complexity has been labeled *construct shift* in the vertical scaling literature and is of predominate concern in subject areas such as mathematics and science where substantial changes in content occur across grade levels (Yen 1986; Reckase & Martineau, 2004; Yon, 2006) or in cases where a large span of grade levels are being scaled (Yen, 1986). In fact, most research to date suggests that multidimensionality is the most likely reason for the variability in vertical scaling studies (Yon, 2006) and that practitioners should always assume that construct shift is present when vertically scaling achievement tests (Li & Lissitz, 2012).

Yen (1985) simulated data for a 30-item unidimensional and a 30-item multidimensional test to investigate the plausibility that when unidimensional models are applied to multidimensional data the resulting ability estimates would be a weighted combination of the two (or more) underlying traits. Additionally, Yen wanted to determine if the weights would be proportional to the relative representation of the traits on the test. She proposed that the scale shrinkage observed in previous vertical scaling studies was due to an increase in item complexity and, in fact, results of the simulation were successful in predicting the amount of scale shrinkage observed. From these results, Yen concluded that certain types of complex items introduce multidimensionality as a function of increasing item complexity and that some content areas are particularly susceptible to this type of bias. An example of this type of item complexity is a

mathematics test in which the lower grade level test focuses on the mathematical operation of addition while the upper grade level test focuses on using addition to solve multiplication problems.

Harris and Hoover (1987) re-examined data used by Loyd and Hoover (1980) from the Iowa Test of Basic Skills (ITBS) to investigate the accuracy of IRT vertical scaling methods for achievement data. They found the equatings to be dependent on group ability. One possible explanation given for this finding was the multidimensionality of the data since a factor analysis found the data was not unidimensional. Skaggs and Lissitz (1986) reviewed results from research on IRT test equating in an attempt to summarize existing evidence and provide direction for future research. When reviewing research related to IRT vertical scaling, they noted that the studies used a wide variety of tests and that when factor analyses were included, many of these studies showed more than one substantial dimension. While they acknowledge that interpretations across studies are often difficult because of the use of different tests, different samples, and different methods of judging the accuracy of results, they concluded that dimensionality has an impact on vertical scaling results and they recommend that a dimensionality analysis be done prior to performing vertical scaling.

Lin, Wei, and Lissitz (2007) used multi-group confirmatory factor analysis to investigate construct invariance across six grade levels of a state-wide mathematics assessment. Even though all grade-level tests covered the same five content strands and those content strands were uniformly represented in the tests, it was determined that construct shift existed across grade levels as shown by disparate factor loadings. More recently, Li and Lissitz (2012) compared the performance of multi-group concurrent calibration using a bifactor model and traditional UIRT estimation for data simulating various degrees of construct shift. The process was then applied to

real data from a statewide mathematics assessment. In the simulation study, item discrimination parameters were underestimated and person parameters and group mean parameter estimates were less accurate for the UIRT model compared to the bifactor model. In the real data study, minimal construct shift was detected and ability estimates from the best-fitting bifactor model and the corresponding UIRT model were highly correlated. The authors concluded that the degree of construct shift significantly affects the stability of parameter estimates from the vertical scaling process and that construct shift should always be assumed and investigated within vertical scaling contexts.

2.7   SUMMARY

Vertically scaling is a complex process with different methodological decisions (i.e., calibration method, characteristics of common items) often leading to different conceptualizations of student growth even when using the same data (Tong & Kolen, 2007). These procedural decisions are separate from characteristics inherent in achievement tests that are beyond the control of the practitioner (i.e., construct shift, item format effects, grade level separation) which have been shown to impact vertical scaling results. Evidence from studies on horizontal equating of single and mixed format tests and vertical scaling of single-format tests suggests that several factors can impact the resultant vertical scales: dimensionality, characteristics of the common items, format of the common items, calibration method, degree of grade level separation, base grade level, number of grade levels scaled, and proficiency estimator used. The impact of these methodological decisions and test characteristics are more uncertain in the context of vertically scaling mixed format tests.

51

The performance of calibration method is the most researched methodological issue in the vertical scaling literature. The results, while inconsistent, suggest that multidimensionality may play a role in the accuracy of these methods (Béguin, Hanson, & Glas, 2000; Béguin, & Hanson, 2002). Evidence suggests that unless mixed format tests are carefully designed, item format effects can exist (Kim & Kolen, 2006; Yao, 2008). Therefore, it is probable that this type of multidimensionality can interact with calibration method to produce less than desirable vertical scales. In addition to dimensionality, evidence also suggests grade level separation will affect the performance of calibration method in that the more disparate the mean ability level between grades, the less accurate the performance of calibration method (Chin, Kim, & Nering, 2006).

The choice of common items is important in vertical scaling because they allow differences in grade levels to be disentangled (Tong & Kolen, 2010). This, in turn, determines the inferences concerning the average amount of growth for the year. Evidence suggests that statistical representativeness is the most important factor in determining common items (Cao, 2008). However, in the context of vertical scaling; whether the upper grade level, the lower grade level or both adjacent grade levels should be statistically represented is uncertain (Peterson, 2010). The choice of common items will most likely be confounded by grade level separation in that the larger the grade level separation the more difficult to find statistically representative common items (Chin, Kim, and Nering, 2006). Also in question is the need for format representativeness in the common items when vertically scaling mixed format tests. The evidence on this topic is inconsistent, but is most likely related to the dimensionality present in the overall test. When the test is essentially unidimensional, the format representativeness of the

common items compared to the test seems irrelevant, but in the presence of multidimensionality, it becomes important.

Few published studies examine vertically scaling mixed format tests. Due to the different factors investigated and models used, generalizations across the studies are difficult. Karkee, et al., (2003) found separate calibration performed best in regards to model fit, convergence, and DIF with pairwise concurrent calibration also out performing concurrent calibration. However, grade-to-grade growth was similar across calibration methods even though grade-to-grade variability was greater for a subset of grade levels when pairwise calibration was used. Jodoin, Keller, and Swaminathan (2003) found that even though classifications of student proficiency were highly correlated regardless of calibration method, the use of proficiency estimators did impact the classification of students. Yao and Mao (2004) determined that if the underlying dimensions are highly correlated, unidimensional calibration methods with multidimensional data are sufficient and result in stable proficiency estimates, but result in larger mean bias. Finally, Meng (2007) found pairwise concurrent and semi-concurrent calibration to be most accurate in estimating proficiency score means, standard deviations, and classification proportions and recovering student growth patterns. However, using common items that included both formats increased accuracy for all calibration methods except separate calibration. In addition, classification consistency was calibration method dependent, but also seemed to be dependent on the number of common items, the format of the common items, and the number of constructed-response items on the overall test.

This study will extend the existing evidence by examining the performance of popular calibration methods in the presence of item format effects across six grade levels. In addition, it will also add to the body of evidence concerning the choice of the common items and the impact

of different disparate mean ability levels between grades in the context of mixed format test vertical scaling. Finally, this study will examine differences in data generation that occur due to explicit inclusion, rather than implicit assumption, of a separate factor representing the vertical scale (Table 1).

Table 1. Summary of factors of investigation for the most closely related studies

Study	Equating/ Scaling	Test Format	Model Used	Calibration	Format Effect	Ability Generation Model	Ability Level Difference	Common Item	Base Grade Level	Number of Levels
Karkee, Hoskins, Yao, & Haug (2003)	Scaling	Mixed	3PL/ PCM	Con/Sep/ Pair				MF	Middle	6
Jodoin, Keller, & Swaminathan (2003)	Scaling	Mixed	3PL/ 2PL/ GRM	Con/Sep/ FCIP					Lowest	3
Yao & Mao (2004)	Scaling	Mixed	3PL/ GPCM	Con/Sep/ Pair		2-dim 3-dim		MC	Lowest	5
Meng (2007)	Scaling	Mixed	3PL/ GPCM	Con/Sep/ Pair/Semi- Con				MF/MC	Lowest	6
Kim & Kolen (2006)	Equating	Mixed		Con/Sep	X	2-dim	X	MF/MC	N/A	N/A
Chin, Kim, & Nering (2006)	Scaling	Single	3PL	Con/Sep			X	MC	Lowest	3,4,5
Cao (2008)	Equating	Mixed	3PL/ GRM	Con/Sep			X	MF/MC	N/A	N/A
Li & Lissitz (2012)	Scaling	Single	2PL	Con/ Bifactor		Bifactor		MC	Lowest	3
Current Study	Scaling	Mixed	3PL/ GRM	Sep/Pair	X	2-dim 3-dim	X	MF	Middle	6

3.0 METHODOLOGY

The main purpose of this study is to examine the performance of two popular calibration methods to vertically scale a series of 6 mixed item format tests under different conditions of format effect and common item configurations. The design of this study is meant to imitate a situation in which the operational test is complex dimensionally; however, the vertical scaling is done under the assumption of unidimensionality.

This chapter is divided into six sections; general test configuration, factors of investigation, ability parameter generation, item parameter description, student response generation, and evaluation criteria. The general test configuration section outlines the factors that remain constant for this study while the factors of investigation section provides an overview of the manipulated factors in this study and rationale for the levels chosen. The ability parameter section outlines the models for generating item format effects within a grade level test that will be simulated in this study. Presented here are the general formulas for the covariance structure needed to accomplish this, as well as the method for simulating different mean grade level separation conditions. The item parameter section gives item parameter specifications for each grade level test, as well as the different common item configurations. Next, the student response generation section gives the process of using the simulated ability and item parameters to create the student response data. The process by which the vertical scales are generated across

conditions by the two different calibration methods is described next. Finally, the evaluation criteria are discussed.

3.1 TEST CONFIGURATION

Response data for 2000 students per grade level was generated for a series of tests across six grade levels. This sample size was determined based on studies performed by Hanson and Béguin (2002) and Kim and Lee (2004) showing this to be an adequate sample size to produce stable estimates of equating coefficients within a unidimensional situation. Each test was constructed to include a total of 61 items divided into two format groups; 54 multiple-choice items and 7 four-ordered category constructed-response items. This gives an overall item ratio between dichotomously scored items to polytomously scored items of approximately 9:1, a score point ratio between item formats of approximately 2:1, and a possible total score of 75. The ratios are consistent with studies examining the equating and scaling of mixed format tests as well as mixed format tests currently used for some state assessment programs (Cao, 2008).

Item parameters were generated using specific criteria based on prior simulation studies and guidelines determined while validating response files. Since the nonequivalent groups common item design (CINEG) is the data collection method most commonly used in vertical scaling, it was the data collection method simulated in this study. Toward that end, common items between adjacent grade levels were established. The number of common items was 14 (23% of total number of items) which was intended to mirror the same proportion of dichotomous items to polytomous items as the overall test and was based on the current recommendation made by Kolen and Brennan (2004) for the minimum number of common items

for vertical scaling (see Figure 1). In addition, the mean item difficulty level of the common items and overall test were preserved for the tests constructed. For simplicity, base grade level, proficiency estimator, format of the common item set, and number of grade levels being scaled remained constant throughout the study and construct shift was assumed to be negligible across grade levels.

Figure 1. Vertical scaling test design

Grade	MC	CR	MC	CR	MC	CR	MC	CR	MC	CR	MC	CR
5	42	5										
C56	12	2	12	2								
6			30	3								
C67			12	2	12	2						
7					30	3						
C78					12	2	12	2				
8							30	3				
C89							12	2	12	2		
9									30	3		
C910									12	2	12	2
10											42	5
Total	54	7	54	7	54	7	54	7	54	7	54	7
	61		61		61		61		61		61	

3.2 FACTORS OF INVESTIGATION

Literature examining the vertical scaling of mixed format tests is small and results concerning performance of the most popular calibration methods are inconsistent. Evidence from these studies and others investigating calibration method when scaling single format tests (e.g. Béguin, Hanson, & Glas, 2000) suggests that calibration method is impacted by item format effects, characteristics of the common items used for linking, and differences in ability levels between adjacent grade levels. The two most widely used calibration methods for single format tests are concurrent calibration and separate calibration (Kolen & Brennan, 2004).

There is evidence, however, that concurrent calibration can have convergence problems when the number of grade levels being scaled is large and/or the separation between grade level distributions is large (Chin, Kim, & Nering, 2008). On the other hand, separate calibration has been theorized to perform better in the presence of some types of multidimensionality because it does not simultaneously calibrate all grade levels (Kolen & Brennan, 2004). However, separate calibration is often associated with larger linking error because the number of linkings, and potential for error, increases as the number of grade levels being scaled increases (Kolen & Brennan, 2004). A compromise calibration method is the pairwise concurrent method that simultaneously calibrates adjacent grade level pairs and then links the pairs. Theoretically, this calibration method could decrease linking error compared to separate calibration because of the decrease in the number of linkings performed (Karkee, et al. 2003). In addition, the method theoretically could increase precision by using more than one grade level of information, similar to concurrent calibration (Kolen & Brennan, 2004). In both the Meng (2007) and Karkee et al. (2003) studies separate and/or pairwise concurrent calibration was found to produce more accurate parameter estimates or more stable proficiency scores within the context of vertical scaling mixed format tests. Given the degree of multidimensionality that will be introduced into the data and the number of grade levels being scaled, only separate and pairwise concurrent calibration methods with the Stocking-Lord linking method will be examined in this study.

Secondly, evidence suggests that multidimensionality can be introduced when vertically scaling mixed format tests because of the inclusion of different item formats on a single test (e.g. Kim & Kolen, 2006). While high correlations between item format dimensions produce bias similar to that found in a unidimensional context, lower correlations between these item format dimensions seem to introduce bias in the vertical scale. Additionally, relatively high or low

correlations between these format factors tended to be subject specific. Four levels of item format effect were investigated ranging from none to large. The correlation values associated with the small and moderate item format effect conditions have been shown to exist in operational tests while the largest item format effect value was chosen as an extreme contrast.

Third, a review of the literature also suggests that construction of the set of common items used in the CINEG scaling design is an important consideration in vertical scaling. From research done in horizontal equating of tests, it has been found that the common items need to be statistically and content representative of the overall test (Tong & Kolen, 2010). How to construct a set of common items that are statistically representative of a set of tests being scaled when the tests are from adjacent grade levels with the expectation that the upper grade level test will be more difficult than the lower grade level test is still in question (Peterson, 2010). To explore this question of common item construction, two sets of common items were created that differed in boundary of item difficulty level. For the narrow condition, the item difficulty was bounded by the mean ability level of the lower grade level and the mean ability level of the upper grade level. In essence, a selection of the hardest of the lower grade level items combined with a selection of the easiest of the upper grade level items. For the expanded condition, item difficulty was bounded by 0.5 standard deviations lower than the mean of the lower grade level and 0.5 standard deviations higher than the mean ability level for the upper grade level. These conditions are based on the study done by Chin, Kim, and Nering (2006) in which they found that performance of the calibration methods can be confounded by restriction of item difficulty range.

Fourth, in the context of vertical scaling it is expected that the mean ability level of successive grade levels will increase as grade level increases. Studies examining horizontal equating with groups of different mean ability levels have found equating results to be negatively

impacted by these differences (e.g. Hanson & Béguin, 1999). In addition, Chin, Kim and Nering (2006) found vertical scales resulting from different calibration methods with a single format test to be impacted by grade level separation. For this reason, two conditions of grade level simulation based on work by Chin, Kim and Nering (2006) are examined in this study; small and large.

Finally, simulation studies investigating vertical scaling have traditionally generated data in which the vertical scale underlying the process is assumed. In other words, the vertical scale hypothesized to exist underlying the test scores is not explicitly modeled in the data generation process. More recent studies investigating the bifactor model as a potential method of estimating a vertical scale have explicitly modeled the vertical scale in the data generation process (Li & Lissitz, 2013; Koepfler, 2012). In these studies, the error values produced under the bifactor model have been smaller than those produced under the traditional method. However, the data was generated and scaled under the same kind of model; not generated under the multidimensional model and scaled under the assumption of unidimensionality.

In summary, five factors were manipulated in this study: 1) degree of multidimensionality between item formats within test – 4 levels, 2) range of item difficulty values for the common item set – 2 levels, 3) amount of separation between grade levels – 2 levels, 4) calibration method – 2 levels and 5) method of data generation – 2 levels (Table 2). The four conditions of degree of format effect were crossed with the two conditions of common item difficulty range and the two conditions of grade level separation to create 16 different vertical scaling scenarios. Each scenario utilized data generated using a 2-dimensional matrix and a 3- dimensional matrix and then scaled by both calibration methods for a total of 64 vertical scaling scenarios. Likewise, the unidimensional data was scaled with both calibration methods

using both common item sets and grade level separation conditions for an additional 6 conditions. Therefore, this study examines a total of 70 vertical scaling scenarios.

Table 2. Fixed versus manipulated factors

Fixed	Manipulated
Number of grade levels scaled - 6	Degree of item format effect – none (0.95), small (0.80), moderate (0.50), large (0.20)
Base grade level - 7 or 7/8	
Sample size – 2000 per grade	Degree of grade separation – small (0.2), large (0.5)
Test length - 61 items	Calibration method - separate, pairwise
Percent MC to CR items – 90/10	Range in item difficulty level of common items – narrow (mean of adjacent grades), expanded (0.5 SD of mean of adjacent grades)
Data collection design - CINEG	
Number of common items - 14	
Format of common items - mixed	Method of data generation – 2-dimensional, 3-dimensional
IRT model - 3PL and GRM	
Linking method – Stocking-Lord	
Proficiency estimator - MLE	

3.3   ABILITY PARAMETER GENERATION

Two different ability generation models were used; one in which the underlying vertical scale was assumed and one in which it was explicitly modeled (Figure 2). For both data generation models, two separate factors were needed to represent multiple choice items and constructed-response items. For the traditional data generation model, these two dimensions were all that were needed and the creation of covariance matrices was straight-forward. In order to model the underlying vertical scale, a 3-factor model was needed: one to model the underlying vertical scale, one to model the multiple-choice items, and one to model the constructed-response items.

Figure 2. Data generation models
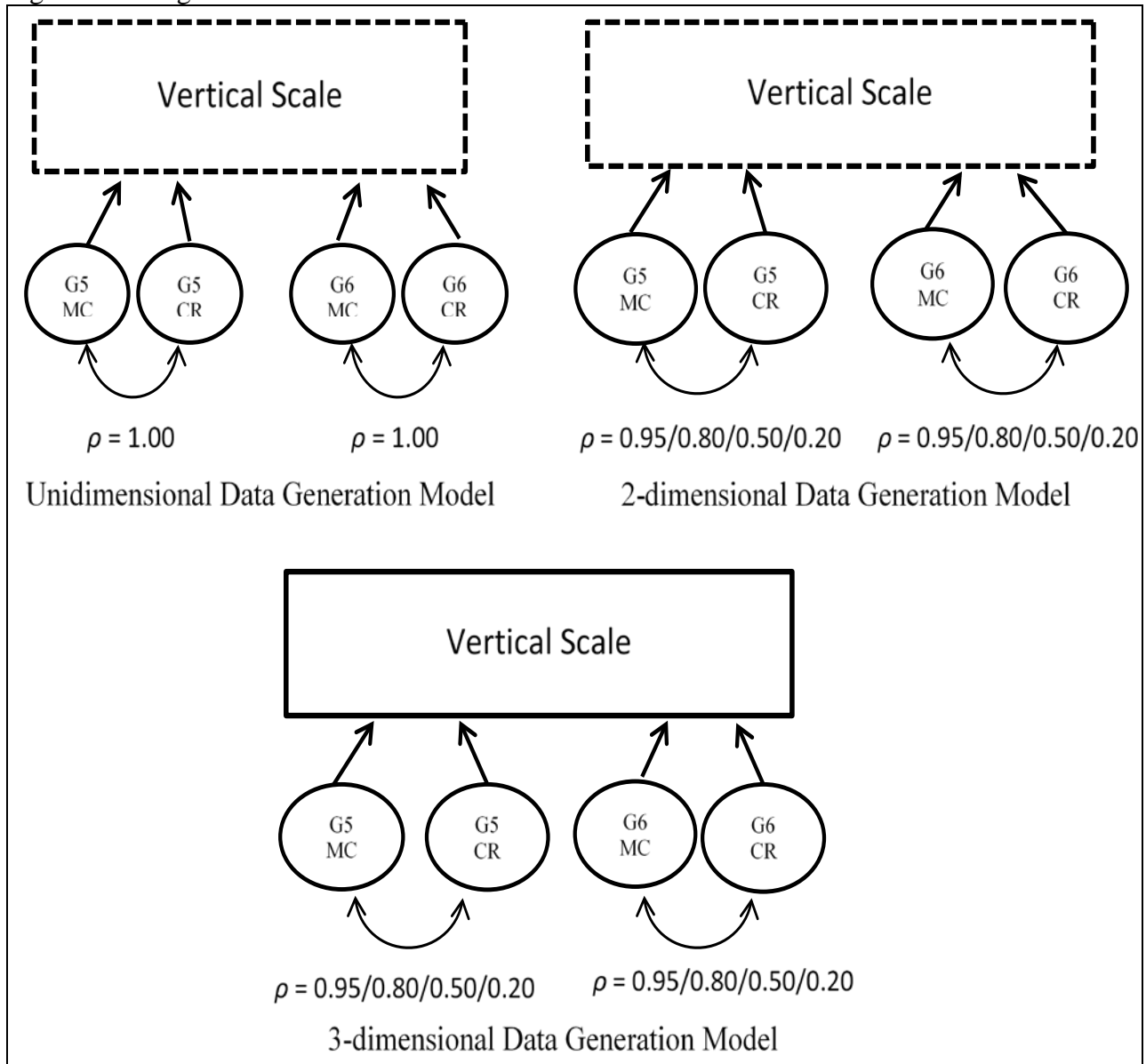
3.3.1   Simulation of Item Format Effects

Item format effects were simulated through manipulation of the covariance structure during the ability parameter generation process. To produce item format effects, it was presumed that two separate ability parameters influenced a student's responses to items on the mixed format test (Kim & Kolen, 2006). One of these abilities was specific to responses for only multiple-choice

items and the other was specific to responses for only constructed-response items. So, two grade

level item format ability factors were generated for each grade level, a multiple-choice ability

factor (θ_{MC}) and a constructed-response ability factor (θ_{CR}). The covariance value between the

ultiple-choice item format factor and the constructed-response item format factor was determined

by varying the correlation between these two ability factors while maintaining a variance of 1 for

both of the item format factors.

To simulate the four conditions of item format effects, these two item format ability

dimensions were correlated at 0.95 (no item format effect), 0.80 (small item format effect), or

0.50 (moderate item format effect) based on research by Traub (1993) and used in Kim and

Kolen (2006). An additional level, 0.20 (large item format effect), was included to provide

extreme contrast. The item format effect was assumed to be constant across all six grade levels

for each respective condition.

3.3.2   Simulation of Unidimensional Ability Distribution

The unidimensional ability distribution was generated using the RANDNORMAL call in SAS

using a mean outlined for the grade level separation conditions and a variance of 1. The seed

value used to generate these unidimensional ability distributions also was used across the

multidimensional data generation models.

3.3.3 Simulation of Two – and Three-Dimensional Ability Distributions

The following general covariance structures were established to be used in the generation of the

ability distributions, $\Sigma_{\theta_s} = \begin{bmatrix} \sigma_{\theta_{MC}}^2 & cov_{\theta_F} \\ cov_{\theta_F} & \sigma_{\theta_{CR}}^2 \end{bmatrix}$ or $\Sigma_{\theta_s} = \begin{bmatrix} \sigma_{\theta_{VS}}^2 & cov_{\theta_{VSMC}} & cov_{\theta_{VSCR}} \\ cov_{\theta_{VSMC}} & \sigma_{\theta_{MC}}^2 & cov_{\theta_F} \\ cov_{\theta_{VSCR}} & cov_{\theta_F} & \sigma_{\theta_{CR}}^2 \end{bmatrix}$, where

$\sigma_{\theta_{VS}}^2$ is the grade level vertical scale ability variance, $\sigma_{\theta_{MC}}^2$ is the grade specific multiple-choice

ability variance, $\sigma_{\theta_{CR}}^2$ is the grade specific constructed-response ability variance, cov_{θ_F} is the

covariance between the item format abilities, and $cov_{\theta_{VSMC}}$ and $cov_{\theta_{VSCR}}$ are the covariances

between the respective item format ability and the vertical scale ability. Then, eight covariance

matrices were constructed in which the two conditions of data generation were crossed with the

four conditions of item format effects (Table 3). Covariance parameters were calculated using

the formula, $cov_{\theta_F} = r * \sigma_{\theta_F} \sigma_{\theta_F}$.

Table 3. Covariance structures used for ability parameter generation

	2 - Dimensional	3 - Dimensional
Essentially Unidimensional	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.95 \\ 0.95 & 1 \end{bmatrix}$	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.95 & 0.95 \\ 0.95 & 1 & 0.95 \\ 0.95 & 0.95 & 1 \end{bmatrix}$
Small Format Effect	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.80 \\ 0.80 & 1 \end{bmatrix}$	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.80 & 0.80 \\ 0.80 & 1 & 0.80 \\ 0.80 & 0.80 & 1 \end{bmatrix}$
Moderate Format Effect	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.50 \\ 0.50 & 1 \end{bmatrix}$	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.50 & 0.50 \\ 0.50 & 1 & 0.50 \\ 0.50 & 0.50 & 1 \end{bmatrix}$
Large Format Effect	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.20 \\ 0.20 & 1 \end{bmatrix}$	$\Sigma_{\theta_S} = \begin{bmatrix} 1 & 0.20 & 0.20 \\ 0.20 & 1 & 0.20 \\ 0.20 & 0.20 & 1 \end{bmatrix}$

3.3.4 Simulation of Grade Level Separation

The two conditions of grade level separation were specified by varying the effect size difference between the mean of the ability distributions of adjacent grade levels (Chin, Kim, & Nering, 2006). A small grade level separation was defined as a Cohen's $d = 0.2$ and a large grade level separation was defined as a Cohen's $d = 0.5$ (Chin, Kim, & Nering, 2006). The degree of grade level separation was assumed to be constant across the six grade levels. Individual vertical scale and item format ability levels were drawn from a multivariate normal distribution with a mean equal to the grade level separation condition assuming the following format, $\mu_{\theta_S} = \begin{bmatrix} \mu_{\theta_{MC}} \\ \mu_{\theta_{CR}} \end{bmatrix}$ or

$\mu_{\theta_S} = \begin{bmatrix} \mu_{\theta_{VS}} \\ \mu_{\theta_{MC}} \\ \mu_{\theta_{CR}} \end{bmatrix}$ where $\mu_{\theta_{VS}}$ is the mean of the grade level vertical scale ability factor, $\mu_{\theta_{MC}}$ is the mean of the grade level multiple-choice ability factor and $\mu_{\theta_{CR}}$ is the mean of the grade level constructed-response ability with s = 5, 6, 7, 8, 9, or 10 depending on the grade level. The mean structure for the vertical scale and the grade specific ability factors for the small grade level separation condition were increased by 0.20 while the means for the large grade level separation were increased by 0.50 (Table 4).

Table 4: Mean structure by grade level separation and data generation model

	Grade 5	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
2-dimensional Small separation	$\begin{bmatrix} -0.4 \\ -0.4 \end{bmatrix}$	$\begin{bmatrix} -0.2 \\ -0.2 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.4 \\ 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}$
3-dimensional Small separation	$\begin{bmatrix} -0.4 \\ -0.4 \\ -0.4 \end{bmatrix}$	$\begin{bmatrix} -0.2 \\ -0.2 \\ -0.2 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.4 \\ 0.4 \\ 0.4 \end{bmatrix}$	$\begin{bmatrix} 0.6 \\ 0.6 \\ 0.6 \end{bmatrix}$
2-dimensional Large separation	$\begin{bmatrix} -1 \\ -1 \end{bmatrix}$	$\begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}$
3-dimensional Large separation	$\begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}$	$\begin{bmatrix} -0.5 \\ -0.5 \\ -0.5 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1.5 \\ 1.5 \\ 1.5 \end{bmatrix}$

For the 2-dimensional ability generation, both means needed to be equal since each ability was used to calculate the probability of a correct response for its respective item type. For the 3-dimensional ability generation, this was not necessarily the case. However, in order for the data generation models to be as comparable as possible, the means were kept equivalent across all three factors. Also, as outlined above, grade level separation was assumed to be constant across the span of grade levels being vertically scaled.

3.3.5 Generation and Validation of Ability Parameters

Ability parameters then were drawn from a multivariate normal distribution, $N_2 \sim (\mu_{\theta_S}, \Sigma_{\theta_S})$ or $N_3 \sim (\mu_{\theta_S}, \Sigma_{\theta_S})$, with a mean corresponding to the grade level separation condition for each respective grade and a covariance structure outlined above for each item format effect and data generation model. Ability distributions were generated within PROC IML in SAS. These ability distributions were validated through factor analysis to confirm the multidimensional structure and the calculation of means and variances of and correlations between latent abilities.

3.4   ITEM PARAMETER GENERATION

Originally, item parameters from a large statewide mathematics assessment were manipulated to construct the grade level tests needed for this study. After attempting to validate the student response files generated using these created tests, it was determined that the item discrimination parameters needed to be within a more restricted range to ensure the simulated student response

files would behave as predicted. As a result, item parameters for fifty-four multiple-choice items and seven four-category constructed-response items were simulated for each grade level test.

3.4.1 Generation of Multiple-Choice Items

For the 3-dimensional data generation model, item discrimination parameters for 54 items, designated for the vertical scale factor, were sampled from a uniform distribution between the values of 1.2 and 2.2. This created a set of item discrimination parameters that was used as the item discrimination parameters across all grade levels. For simplicity, item discriminations parameters for item format specific factor in the 3-dimensional ability generation for all multiple choice items were set to 1. This value was chosen because it was believed to be sufficiently high in discrimination, but lower than the lowest possible value for an item discrimination factor for the vertical scale factor. This second set of item discrimination parameters were not used in the 2-dimensional ability generation. Additionally, the guessing parameters for each multiple-choice item were set to 0.25 for all grade levels for simplicity.

Item difficulty parameters for the 42-48 (depending on grade level) multiple-choice items were sampled from a normal distribution with a mean approximately equal to the grade level mean ability for the small grade level separation (see Table 4 for means) and a standard deviation of 1 (Li & Lissitz, 2012). In addition, the item difficulty parameters were restricted to a range between -2.00 and 2.00 and adjusted for the mean of the ability distribution for the small grade level separation for each grade level. For example, for grade 5, item difficulty values were sampled from a normal distribution with a mean of approximately -0.40, a standard deviation of 1, and restricted to a range approximately from -2.4 to 1.6. The additional items needed to complete the test for each grade level were drawn from the adjacent grade level test and were of

appropriate difficulty to serve as common items. Item parameters for all multiple-choice items for each test by grade level are listed in Tables A1 – A12.

3.4.2 Generation of Constructed-Response Items

Each grade level test required 7 four-category constructed response items. As with the multiple-choice items, item discrimination parameters for 7 constructed response items for the vertical scale factor was sampled from a uniform distribution between the values of 1.2 and 2.2. This set of item discrimination parameters were used for all grade levels. Again, for simplicity, item discriminations parameters for the item format factor in the 3-dimensional ability generation were set to 1. Next, three category threshold values were generated for five constructed-response items. First, the range of difficulty values determined for the multiple-choice items was divided into three approximately equal segments. Then, from within each segment, 5 item category thresholds were drawn from a uniform distribution bounded by the segment values. For example, for grade 5, the range from -2.4 to 1.6 was divided into 3 approximately equal segments (-2.4 to -1.1, -1.1 to 0.25, and 0.25 to 1.6). Then, 5 category thresholds were drawn from the first region, 7 category thresholds were drawn from the second region, and 5 category thresholds were drawn from the last region. The additional threshold values needed to complete the entire set of 7 constructed-response items were drawn from adjacent grade levels and designated as common items. Item parameters for the constructed-response items for each grade level are listed in Tables A1 – A12.

3.4.3   Selection of Common Items

With respect to the studies investigating the necessary characteristics of the common items when multidimensionality is present (e.g. Cao, 2008), two sets of mixed format common items were chosen from the simulated items used to design the grade level tests as outlined in sections 3.4.1.and 3.4.2. The two sets of common items were chosen to investigate the importance of statistical representativeness of the common item set compared to the overall test. Two conditions of statistical representativeness were used; narrow and expanded. These conditions differed in the range of difficulty level for the items. The narrow range condition contained items with item difficulty values bounded by the mean ability level of the lower adjacent grade level and the mean ability level of the upper adjacent grade level. The expanded range condition contained items with item difficulty values bounded by 0.5 standard deviations below the mean ability level of the lower adjacent grade level and 0.5 standard deviations above the mean ability level of the upper adjacent grade level. For example, for grade 5/6 common items, the *narrow* condition consisted of items with difficulty values ranging from -0.40 to -0.20 while the *expanded* condition consisted of items with difficulty values ranging from -0.90 to 0.30. These conditions were chosen to help explore the issue of how to choose common items that are statistically representative of adjacent grade levels which are necessarily of differing ability levels (Chin, Kim & Nering, 2006; Tong & Kolen, 2010; Peterson, 2010). Finally, both common item sets were compared for overall mean difficulty level to ensure they were as comparable as possible. The anchor item set conditions are summarized below.

The *narrow* condition consisted of both multiple-choice and constructed-response items and had an item difficulty range bounded by the mean difficulty level of the upper adjacent grade level and the lower adjacent grade level. To construct this common item set, six multiple-choice

items and one constructed-response item from each grade level in the appropriate item difficulty range were chosen. The category threshold parameter for the middle score value possible was used to determine whether or not the item fell within the necessary item difficulty range.

The *expanded* condition had an item difficulty range bounded by 0.50 standard deviations above the mean difficulty level for the upper adjacent grade level and 0.50 standard deviations below the mean difficulty level for the lower adjacent grade level. To construct this common item set, three multiple-choice items from each adjacent grade level were removed and replaced with the three multiple-choice items in the appropriate expanded item difficulty range being careful to add three additional items from the lower expanded region and three additional items from the expanded upper region. The same constructed-response items as used in the *narrow* condition were used in both common item sets.

3.4.4   Creation of Grade Level Tests

Using an iterative process, sets of multiple-choice items, constructed-response items, and common items for all conditions were combined into a complete grade level test such that the mean item difficulty levels matched the condition of the small grade level separation in the study design and the mean item difficulty levels of the common item sets were approximately equal across conditions (Table 4). Additionally, common items were placed in the same position across the grade level tests for which they were used to scale to ameliorate any potential item location effects. A list of items for each grade level test and common item set is included in Tables A1-A12.

Table 5. Summary of means for each grade level test and common item set

	Grade 5	5/6	Grade 6	6/7	Grade 7	7/8	Grade 8	8/9	Grade 9	9/10	Grade 10
Narrow	-0.65		-0.21		0.06		0.30		0.64		0.83
		-0.36		-0.14		0.14		0.44		0.64	
Expanded	-0.65		-0.21		0.06		0.30		0.64		0.83
		-0.30		-0.16		0.15		0.43		0.66	

3.5   GENERATION OF STUDENT RESPONSES

The ability parameter distributions generated in section 3.3.6 were applied to the grade level tests constructed in section 3.4 to generate probabilities of a correct response. The decision rules outlined below were then applied to the individual probabilities and student responses were generated and expected scores were calculated for each of the 2000 participants per grade level per condition.

3.5.1   Generating Student Responses for Uni- and Two-Dimensional Models

3.5.1.1 Multiple-Choice Items

Even though some ability distributions were generated under what is termed a 2-dimensional model, there is only one dimension associated with each item format type. Therefore, student responses for both the unidimensional and two-dimensional multiple-choice items are simulated under the unidimensional case. Under this 3PL unidimensional model, the probability of a correct response for person i to multiple choice item j is (Hambleton & Swaminathan, 1985):

$$P_{ij} \;=\; P_j\left(\theta_i\right) = P(\theta_i | a_j, b_j, c_j) \;=\; c_j \; + \left(1 + c_j\right) \frac{\exp\left[Da_j\left(\theta_i - b_j\right)\right]}{1 + \exp\left[Da_j\left(\theta_i - b_j\right)\right]} \qquad (Eq.1)$$

where θ_{iMC} is the grade specific ability or grade specific multiple-choice ability generated under the uni- or two-dimensional model, a_{jVS} is the item discrimination parameter associated with the vertical scale factor, and b_j is the item difficulty parameter.

Student responses were generated for multiple-choice items by applying equation 1 to the multiple-choice item parameters appropriate for each grade level test and ability parameters generated under either the unidimensional or 2-dimensioanl models to determine the probability (P_{ij}) that examinee i with ability θ_{iMC} would correctly answer item j. These probabilities were compared to a uniform random number (R) in the range $[0, 1]$ and student responses (U_{ij}) for each item j were coded according to the following rule (Kim & Kolen, 2008), $U_{ijk} = \begin{cases} 0, P_{ij} \leq R \\ 1, P_{ij} > R \end{cases}$.

3.5.1.2 Constructed-Response Items

Under the GRM unidimensional model, the probability that examinee i will score at or above category k on item j is (Samejima, 1969):

$$P^*_{ijk} \; = \; P^*(\theta_i | a_j, b_{jk}) = \begin{cases} 1 & k = 1 \\ \dfrac{\exp[Da_j(\theta_i - b_{jk})]}{1 + \exp[Da_j(\theta_i - b_{jk})]} & 2 < k \leq K_j \\ 0 & k > K_j \end{cases} \qquad (Eq.\,2)$$

where category $k = 1, 2,…K$, θ_{iCR} is the grade specific constructed-response ability, and a_{jVS} is the item discrimination parameter associated the vertical scale factor.

To generate student responses for constructed-response items, equation 2 was applied to the constructed-response item parameters appropriate for each grade level test and ability parameters generated under either the unidimensional or 2-dimensional models to calculate the conditional probability (P^*_{ijk}) that a response from examinee i with θ_{iCR} would fall within or above category k. This conditional probability (P^*_{ijk}) was compared to a uniform random number (R) in

the range [0, 1] and student responses (U_{ijk}) were coded by comparing P^*_{ijk} to R with the

following rule (Kim & Kolen, 2008; Cao, 2008), $U_{ijk} = \begin{cases} 0, & P^*_{i1} \leq R < 1 \\ 1, & P^*_{i2} \leq R < P^*_{i1} \\ 2, & P^*_{i3} \leq R < P^*_{i2} \\ 3, & 0 \leq R < P^*_{i3} \end{cases}$.

Calculating the expected score for a constructed-response item requires two steps. The first is calculating the conditional probability of a correct response for each threshold and then calculating a category response function, which is the difference between two adjacent categories. The category response function, P_{ijk}, was calculated using the following formula (Cao, 2008):

$$P_{ijk} = P_{jk}(\theta_i) = P^*_{jk}(\theta_i) - P^*_{j(k+1)}(\theta_i) \qquad (Eq.3)$$

3.5.2 Generating Student Responses for Three-Dimensional Model

3.5.2.1 Multiple-Choice Items

For the three-dimensional model, the vertical scale ability and item discrimination parameter as well as the item format specific ability and item discrimination parameter both contribute to the probability of a correct response. Therefore, a multidimensional probability of a correct response is needed to estimate the 3-dimensional model. Under the 3PL multidimensional model, the probability of a correct response for person i to multiple choice item j is (Gibbon & Hedeker, 1992):

$$P(X_j = 1 \mid \theta_i a_j B_j c_j) = c_j + \frac{1 - c_j}{1 + \exp[-(a_{jVS}\theta_{iVS} + a_{jMC}\theta_{iMC} + B_j)]} \qquad (Eq.4)$$

where θ_{iVS} is the grade specific vertical scale ability, θ_{iMC} is the grade specific multiple-choice ability, a_{jVS} is the item discrimination parameter associated with the vertical scale factor, a_{jMC} is the item discrimination parameter associated with the grade specific multiple-choice factor, and B_j is the scalar parameter related to multidimensional difficulty. B_j was calculated using the following formula (Reckase, 2009):

$$B_j = -b_j \sqrt{a_{jVS}^2 + a_{jMC}^2} \qquad (Eq. 5)$$

Student responses were generated for multiple-choice items by applying equation 4 to the multiple-choice item parameters appropriate for each grade level test and ability parameters generated under the 3-dimensional model to determine the probability (P_{ij}) that examinee i with ability θ_{iMC} would correctly answer item j. These probabilities were compared to a uniform random number (R) in the range [0, 1] and student responses (U_{ij}) for each item j were coded according to the following rule (Kim & Kolen, 2008), $U_{ijk} = \begin{cases} 0, P_{ij} \leq R \\ 1, P_{ij} > R \end{cases}$.

3.5.2.2 Constructed-Response Items

Under the GRM multidimensional model, the probability that examinee i will score at or above category k on item j is (Samejima, 1969; Gibbons et. al, 2010):

$$P^*(X_j \geq k \mid \theta_i a_j B_{jk}) = \begin{cases} \dfrac{1}{1} & k = 1 \\ \dfrac{1}{1 + \exp[-(a_{jVS}\theta_{iVS} + a_{jS}\theta_{iCR} + B_{jk})]} & 2 < k \leq K_j \\ 0 & k > K_j \end{cases} \qquad (Eq. 6)$$

where category $k = 1, 2,...K$, θ_{iVS} is the vertical scale ability, θ_{iCR} is the grade specific constructed-response ability, a_{jVS} is the item discrimination parameter associated with the vertical scale factor, a_{jCR} is the item discrimination parameter associated with the grade specific item

format factor, and B_{jk} is the scalar parameter related to multidimensional difficulty and is calculated using the following formula (Reckase, 2009):

$$B_{jk} = -b_{jk} \sqrt{a_{jVS}^2 + a_{jCR}^2} \qquad (Eq.\,7)$$

To generate student responses for constructed-response items, equation 6 was applied to the chosen constructed-response item parameters appropriate for each grade level test and ability parameters generated to calculate the conditional probability (P^*_{ijk}) that a response from examinee i with θ_{iCR} would fall within or above category k. This conditional probability (P^*_{ijk}) was compared to a uniform random number (R) in the range [0, 1] and student responses (U_{ijk}) were coded by comparing P^*_{ijk} to R with the following rule (Kim & Kolen, 2008; Cao, 2008),

$$U_{ijk} = \begin{cases} 0, & P^*_{i1} \leq R < 1 \\ 1, & P^*_{i2} \leq R < P^*_{i1} \\ 2, & P^*_{i3} \leq R < P^*_{i2} \\ 3, & 0 \leq R < P^*_{i3} \end{cases}$$ Again, the category response function, P_{ijk}, was calculated using

the following formula (Cao, 2008) for the multidimensional case:

$$P_{ijk} = P_{jk}(\theta_i) = P^*_{jk}(\theta_i) - P^*_{j(k+1)}(\theta_i) \qquad (Eq.\,8)$$

3.5.3 Validation of Student Response Files

Student response files were validated using factor analysis and through parameter recovery. A factor analysis was used to confirm the multidimensional structure and correlations between latent abilities of grade level response files. In order to assess parameter recovery, an additional response file using an orthogonal covariance structure consistent with the bifactor model was created. Parameters estimated using this response file were compared to the original test parameters for accuracy.

3.6   GENERATING THE VERTICAL SCALES

Student responses generated in section 3.5.1 and 3.5.2 for each grade level test constructed in section 3.4, were vertically scaled using each of two different calibration methods. (Figure 3) Calibrations were performed with MULTILOG (Thissen, 1991). Separate calibration and pairwise concurrent calibration both required an additional linking step to complete the scaling process. In these cases, the Stocking-Lord linking method was performed using STUIRT (Kim & Kolen, 2004).

Figure 3. Graphic representation of calibration methods used in study
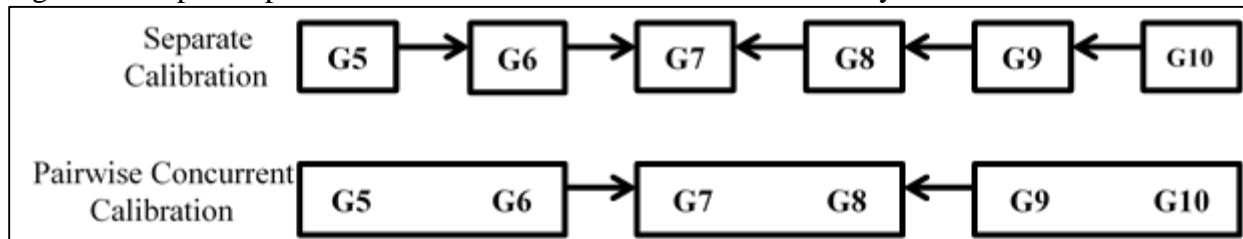
3.6.1   Separate Calibration

To perform separate calibration, item parameters and ability parameters were estimated for student response data for each grade individually using the default settings in MULTILOG (Thissen, 1991). After the separate calibration for each grade level was complete, Stocking-Lord transformation constants were calculated using STUIRT (Kim & Kolen, 2004) for the narrow common item set. Item parameters and ability estimates were transformed to the grade 7 scale. Then, this procedure was repeated for the expanded common item set condition. Each anchor item set condition was replicated 100 times.

3.6.2 Pairwise Concurrent Calibration

To perform pairwise concurrent calibration, response files for adjacent grade level pairs (5-6, 7-8, 9-10) were combined. Items not taken by a grade level were coded as missing. For each adjacent grade level pair, one grade level was designated as the base grade level for the concurrent calibration. In the 5-6 grade level, grade 6 was designated the base grade level; in the 7-8 grade level pair, grade 7 was designated as the base grade level; and in the 9-10 grade level pair, grade 9 was designated as the base grade level. Item and ability parameters were estimated for each grade level pair using concurrent calibration in MULTILOG (Thissen, 1991). Once the three concurrent calibrations were complete, transformation constants to place the 5-6 grade level pair and the 9-10 grade level pair for the narrow common item set condition were calculated using STUIRT (Kim & Kolen, 2004). These two grade level pairs then were placed on the 7-8 grade level pair scale. This procedure was repeated for the expanded common item set condition and each anchor item set condition was replicated 100 times.

3.7   EVALUATION CRITERIA

While vertical scaling is used to evaluate growth over time, there is no generally accepted growth model or definition of growth (Tong & Kolen, 2007). Camilli (1993) makes the point that linear growth is dependent on the meaning assigned to the intervals along the scale based on the transformation that is chosen. So, instead of being able to compare the results of vertical scaling to an absolute criterion, results from vertical scales are often compared to each other and themselves as a means of determining their appropriateness (Becker & Forsythe, 1992). In a

simulation study, however, an absolute criterion does exist because the true parameters are known (Harris & Crouse, 1993). Thus, the criterion for evaluating scaling results in simulation studies can focus on the accuracy of the recovery of true parameters (Harris & Crouse, 1993).

3.7.1 Determining the Accuracy of Scaling Results

Because the results of each scaling method produce unique vertical scales, comparison of expected scores was chosen to preclude the additional transformation necessary to make direct comparisons of ability estimates across vertical scaling techniques. Therefore, to determine the accuracy of the scaling results, examinee expected total scores were calculated using the population parameters (Cao, 2008, p. 58),

$$E(X_k \,|\theta_k\,) = \sum_{i=1}^{n_{MC}} P_i\,(\theta_k\,) + \sum_{i=1}^{n_{CR}} \sum_{j=1}^{J} u_{ij}\; P_{ij}\,(\theta_k\,) \qquad (Eq.\,9)$$

and compared to the expected total scores calculated using the estimated and transformed parameters (Cao, 2008, p. 58),

$$E(X_k \,|\hat{\theta}_k\,) = \sum_{i=1}^{n_{MC}} P_i\,\left(\hat{\theta}_k\,\right) + \sum_{i=1}^{n_{CR}} \sum_{j=1}^{J} u_{ij}\; P_{ij}\,\left(\hat{\theta}_k\,\right) \qquad (Eq.\,10)$$

where θ_k is the true expected score for individual k, $\hat{\theta}_k$ is the estimated expected score for individual k, $P_i\,(\theta_k\,)$ and $P_i\,(\tilde{\theta}_k\,)$ are calculated using equation 1 or 4 using true ability parameters and observed ability parameters, respectively, u_{ij} is the category score and $P_{ij}(\theta_k\,)$ and $P_{ij}\hat{\theta}_k$ are calculated using equation 2 or 5 using the true ability parameters and the observed ability parameters, respectively.

Two summary statistics were calculated, by grade level, to evaluate the accuracy of the expected scores from the scaling results; BIAS and root mean squared error (RMSE). Each summary statistic was calculated and averaged across the 100 replications.

BIAS reflects the average difference between the estimated parameter value and the true parameter value over replications and gives an indication of the accuracy and direction of the scaling results (Cao, 2008; Li & Lissitz, 2012). Thus, the smaller the bias, the more accurate the method was in recovering the true expected scores. BIAS was calculated using the following formula (Cao, 2008):

$$BIAS_r(E(X_k \,|\hat{\theta}_k)) = \frac{\sum_{r=1}^{R} E(X_k \,|\hat{\theta}_k)_r - E(X_k | \theta_k)}{2000} \qquad (Eq.11)$$

$$average\ BIAS_r(E(X_k \,|\hat{\theta}_k)) = \frac{BIAS_r}{R} \qquad (Eq.12)$$

where R is the number of replications, r is the replication, $E(X_k | \theta_k)$ is the true expected score for individual k, and $E(X_k \,|\hat{\theta}_k)_r$ is the estimated expected score for individual at the rth replication.

RMSE reflects the overall accuracy in estimating parameters with smaller RMSEs indicating the calibration method was more accurate in recovering the expected score (Meng, 2007). RMSE was calculated using the following formula (Cao, 2008):

$$RMSE_r\ (E(X_k \,|\hat{\theta}_k)) = \sqrt{\frac{1}{R}\sum_{r=1}^{R} \left(E(X_k \,|\hat{\theta}_k)_r - E(X_k | \theta_k)\right)^2} \qquad (Eq.13)$$

$$average\ RMSE_r(E(X_k \,|\hat{\theta}_k)) = \frac{RMSE_r}{R} \qquad (Eq.14)$$

where R is the number of replications, r is the replication, $E(X_k|\theta_k)$ is the true expected score or individual k, and $E(X_k|\hat{\theta}_k)_r$ is the estimated expected score for individual at the rth replication.

3.7.1   Comparing Results

First, tables and plots for average BIAS and average RMSE values were created to summarize differences in conditions of interest. Then, to determine if any one condition was more influential than another, two five-way mixed model ANOVAs were performed using the PROC MIXED procedure in SAS, by grade level, using the 100 replications as observations and mean BIAS and mean RMSE values as dependent variables. Since the same dataset was scaled using both calibration methods, vertical scaling method was treated as a within-subjects factor while all other variables were treated as between-subject factors. All factors were assumed to be fixed. Next, all main effects and two-, three-, four-way, and five-way interactions in the model were estimated. All possible pairwise comparisons were tested using the LSMEANS command with a Scheffé adjustment. Finally, Cohen's d was calculated for each comparison from the t-values and df in the LSMEANS output using the following formula (Rosenthal & Rosnow, 1991), $2t/\sqrt{df}$. Only significant results with at least a small effect size (0.20) were examined (Cohen, 1998).

4.0   RESULTS

This chapter describes the results of the simulation study outlined in Chapter 3. Seventy vertical

scales were simulated for grades 5, 6, 7, 8, 9, and 10. BIAS and RMSE were calculated for each

of the conditions with respect to recovery of the expected score. In addition to tables and graphs

which explore the descriptive characteristics of the mean BIAS and mean RMSE results for the

simulation, two mixed ANOVAs were performed using the PROC MIXED procedure in SAS for

each grade level. Using the 100 replications as observations and mean BIAS and mean RMSE

values as dependent variables; two levels of vertical scaling method (separate, pairwise

concurrent), two levels of common item set configuration (narrow, expanded), four levels of

format effect (FE: none, small, moderate, large), two levels of grade level separation (small,

large), two data generation methods (2-dimensional, 3-dimensional), and the traditional

unidimensional case served as the independent variables. All factors were assumed to be fixed

with calibration method treated as a within-subjects factor and all other variables treated as

between-subject factors. All main effects and interactions were estimated and all pairwise

comparisons were tested with a Scheffé adjustment. Cohen's d was calculated for each

comparison based on values from the LSMEANS output and only significant results with an

effect size of 0.20 or greater were reported.

The results chapter describes grade level results by trends in performance of the vertical

scaling configurations by condition. Next, ANOVA test results are discussed followed by a

comparison of significant differences in the four calibration and common item set configurations. Finally, a summary of results for each grade level as well as an overall summary of results is presented. Supporting tables for all results are found in Appendix B and Appendix C.

4.1   GRADE LEVEL RESULTS

4.1.1   Grade 5 Results

4.1.1.1 Trends by Condition

<u>Average BIAS</u>

Generally, mean BIAS values increased as degree of item format effect increased regardless of common item set, grade level separation, or data generation method for separate calibration (Table 5). For pairwise concurrent calibration, trends were less obvious because of inconsistencies between replications for the mean estimated expected score. In general, mean BIAS values for all item format effect conditions that resulted from data generated under the 3-dimensional model tended to be similar to one another. Additionally, scores resulting from the small grade level separation condition tended to be smaller than those resulting from the large grade level separation values regardless of common item set, vertical scaling method, item format effect condition, or data generation method. Scores from data generated under the 3-dimensional model generally resulted in smaller mean values than those from data generated under the 2-dimensional model or unidimensional model for the small grade level separation condition. The pattern was reversed somewhat for the large grade level separation condition under the narrow common item set condition. Of note, however, is that scores resulting under the

2-dimensional data generation procedure for the small grade level separation tended to be underestimated, however, as format effect increased, mean BIAS values became increasingly more positive. This means that for the no and small format effect conditions, mean BIAS scores were more evenly split between replications that tended to be over- and underestimated resulting in values that could be spuriously low. For the 3-dimensional data generation condition, however, the opposite was true. Scores tended to be overestimated, but as format effect increased, mean BIAS values for each replication became increasingly more positive. This means that for this data generation condition, values for the moderate and large format effect conditions could be spuriously low. However, this phenomenon was not observed for the large grade level condition for which mean BIAS values per replication were consistently negative for all conditions. This phenomenon also did not impact the mean RMSE values (Figures 4 and 5).

Trends in vertical scaling method and common item set were more difficult to characterize because scores were not consistently over- or under-estimated across methods. For the large grade level separation condition, scores were consistently overestimated regardless of common item set, vertical scaling method, item format condition, or data generation method. This was not the case for the small grade level separation. For this condition, separate calibration underestimated the expected score for all vertical scaling, item format effect, common item set, and data generation conditions. Pairwise concurrent calibration, on the other hand, overestimated scores generated with the 3-dimensional and unidimensional models as well as the small item format effect condition under the 2-dimensional model regardless of common item set or degree of item format effect. However, other scores for the 2-dimensional model (no, moderate, and large item format effect), regardless of common item set, were overestimated.

84

Average RMSE

Mean RMSE values increased as degree of item format effect increased, regardless of common item set, grade level separation, vertical scaling method, or data generation method. Additionally, scores from the small grade level separation condition resulted in smaller mean RMSE values compared to scores from the large grade level separation condition, regardless of common item set, vertical scaling method, item format effect condition, or data generation method. Data generated under the 3-dimensional model generally resulted in smaller mean values than those resulting from data generated under the 2-dimensional or unidimensional models regardless of condition. Also, pairwise concurrent calibration produced smaller average RMSE values than separate calibration, regardless of common item set and data generation method. Finally, mean values for common item set were similar regardless of vertical scaling method, grade level separation, item format effect condition, and data generation method except those resulting from data generated under the 3-dimensional model for separate calibration with the expanded common item set which tended to be slightly smaller (Figures 6 and 7).

Table 6. Grade 5 mean BIAS and RMSE by condition

Scaling Method	Common Item Set	Model	Format Effect	Mean BIAS		Mean RMSE	
				Small Separation	Large Separation	Small Separation	Large Separation
Separate	Narrow	Unidim	None	0.53	-1.57	4.04	7.70
		2-Dim	None	0.54	-1.56	4.04	7.52
			Small	0.53	-1.68	4.20	7.68
			Moderate	1.16	-2.02	5.19	9.23
			Large	4.52	-2.58	7.91	9.89
		3-Dim	None	0.02	-1.72	3.52	5.41
			Small	0.04	-1.88	3.57	5.63
			Moderate	0.08	-2.62	3.77	6.61
			Large	0.16	-3.63	4.10	7.79
	Expanded	Unidim	None	0.59	-1.56	4.07	7.75
		2-Dim	None	0.51	-1.53	4.00	7.55
			Small	0.58	-1.70	4.26	7.78
			Moderate	1.15	-2.02	5.18	9.18
			Large	4.54	-2.70	7.97	10.13
		3-Dim	None	0.03	-1.54	3.49	5.34
			Small	0.07	-1.94	3.66	5.75
			Moderate	0.09	-2.55	3.75	6.50
			Large	0.15	-3.39	4.08	7.41
Pairwise	Narrow	Unidim	None	0.16	-0.75	3.56	4.60
		2-Dim	None	0.13	-0.76	3.55	4.53
			Small	0.14	-0.79	3.79	4.63
			Moderate	0.29	-0.99	4.52	5.43
			Large	1.88	-1.16	5.88	6.81
		3-Dim	None	-0.08	-0.69	3.26	3.77
			Small	-0.06	-0.84	3.30	3.84
			Moderate	-0.07	-1.08	3.45	4.05
			Large	-0.06	-1.37	3.66	4.31
	Expanded	Unidim	None	0.15	-0.75	3.61	4.60
		2-Dim	None	0.13	-0.74	3.60	4.54
			Small	0.11	-0.81	3.77	4.62
			Moderate	0.33	-0.99	4.55	5.44
			Large	1.93	-1.16	5.92	6.83
		3-Dim	None	-0.08	-0.73	3.27	3.67
			Small	-0.08	-0.84	3.38	3.78
			Moderate	-0.06	-0.97	3.46	3.95
			Large	-0.08	-1.30	3.67	4.24

Figure 4. Grade 5 average BIAS for small grade level separation
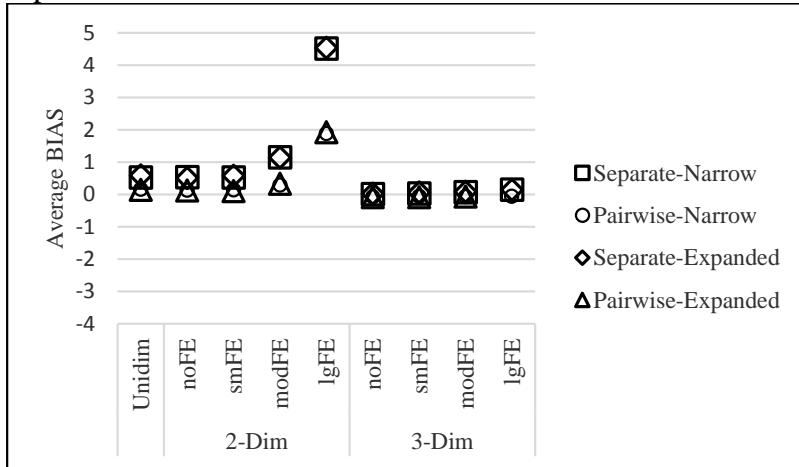
Figure 6. Grade 5 average RMSE for small grade level separation
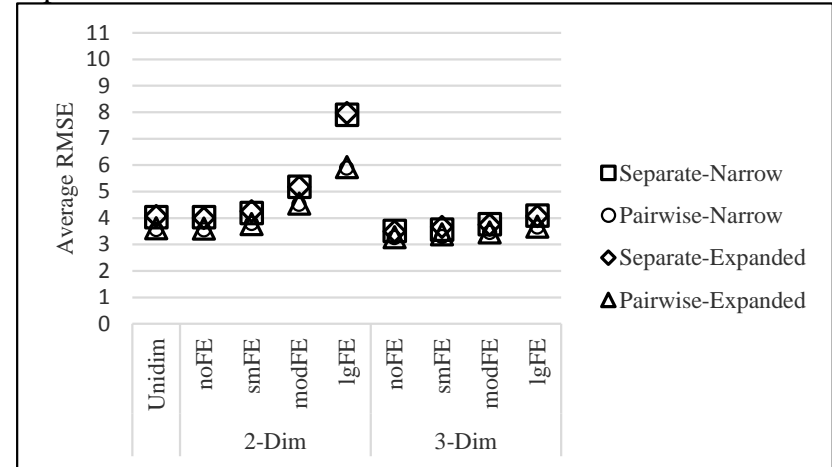
Figure 5. Grade 5 average BIAS for large grade level separation
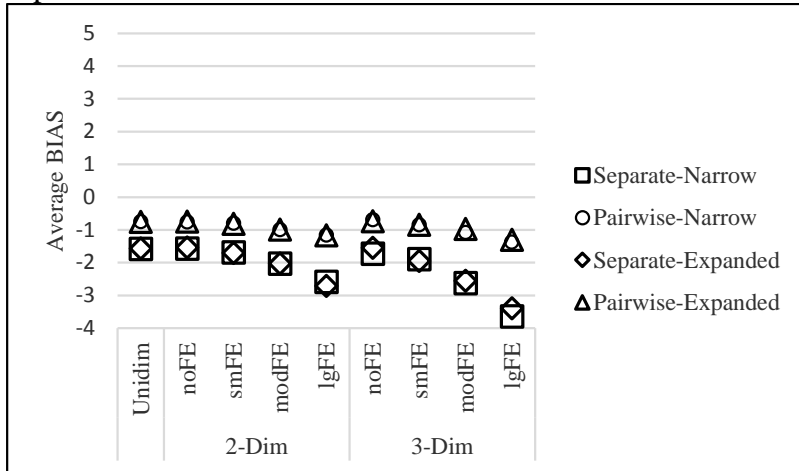
Figure 7. Grade 5 average RMSE for large grade level separation

87
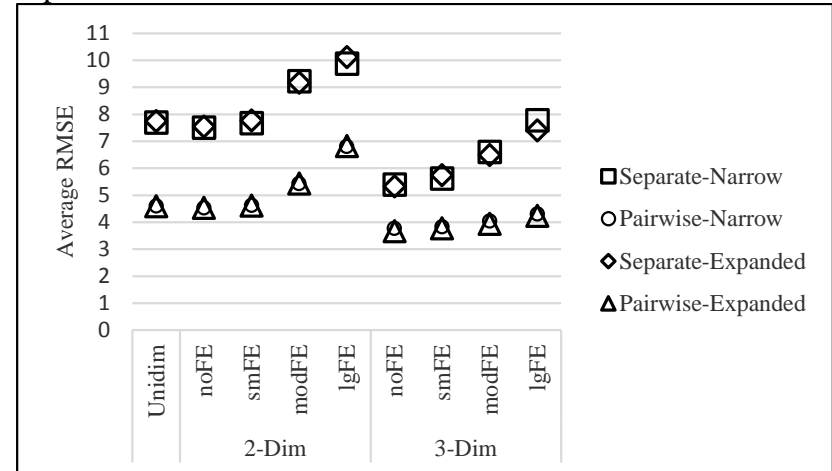
4.1.1.2 ANOVA Results

Two separate five-way mixed ANOVAs using mean BIAS and mean RMSE values as the dependent variable were performed on the results of the 100 replications for grade 5. Most interactions and main effects were significant for these models (Table B1 and B2). Using the LSMEANS command in SAS, pairwise comparisons were examined and effect sizes were calculated. Additionally, the difference matrix produced was used to examine relevant comparisons between vertical scaling configurations by condition to determine the scaling method(s) that resulted in the most accurate vertical scales for each condition.

Item Format Effect

Scaling results for item format effect conditions were dependent on grade level separation, vertical scaling method, and data generation method (Table C1). Effect sizes for mean BIAS values ranged from 0.37 to 2.59 and effect sizes for mean RMSE values ranged from 0.36 to 1.51.

Under the 2-dimensional model for the small grade level separation, regardless of common item set and vertical scaling method, the large item format effect produced significantly larger mean BIAS values than any other item format effect and the moderate item format effect produced significantly larger values compared to the small format effect condition. In addition, under separate calibration, the moderate item format effect condition was significantly larger than the no item format effect condition. Under the large grade level separation, results depended upon vertical scaling method also. Mean BIAS values under separate calibration were significantly larger for the large item format effect condition compared to the no, small, and moderate item format effect conditions. However, there was no significant difference in mean BIAS values under pairwise concurrent calibration. Under the 3-dimensional model for the small

grade level separation condition, there were no significant differences among levels of item format effect regardless of vertical scaling method and common item set. For the large grade level separation condition, however, regardless of common item set, mean BIAS values for the large item format effect condition were significantly larger than both the no and small item format effect conditions and values for the moderate item format effect condition were significantly larger than both the no and small item format effect conditions.

Under the 2-dimensional model for the large grade level separation condition, regardless of common item set, grade level separation, and vertical scaling method, the large item format effect produced significantly larger mean RMSE values when compared to the no, small, and moderate format effect conditions and the moderate item format effect produced significantly larger values compared to the no item format conditions. Additionally, under separate calibration, the moderate item format effect condition was significantly larger than the small item format effect condition. Again, under the 3-dimensional model for the small grade level separation condition, there was no significant item format effect regardless of vertical scaling method and common item set. For the large grade level separation, the large item effect condition was significantly larger than the no, small, and moderate item format effect conditions and the moderate item format effect condition was significantly larger than the no item format effect condition, but only under separate calibration.

Vertical Scaling Method

Significant differences in average BIAS values were dependent on item format effect, grade level separation, and data generation model. Average BIAS and average RMSE values were significantly larger for separate calibration compared to pairwise concurrent calibration under the large grade level separation condition regardless of item format effect, common item set, or data

generation method (Table C2). Effect sizes for mean BIAS values ranged from 0.53 to 1.69 while effect sizes for mean RMSE values ranged from 0.90 to 1.55.

Under the small grade level separation condition, results depended on the method of data generation. There was no significant difference between separate and pairwise concurrent calibration when the data was generated under the 3-dimensional or unidimensional data generation model. Under the 2-dimensional data generation model, however, average BIAS values were significantly larger than those for pairwise concurrent calibration for the small, moderate, and large item format effect conditions regardless of common item set configuration. Likewise, average RMSE values for the large grade level separation were significantly larger for separate calibration compared to pairwise concurrent calibration, regardless of item format effect condition. For the small grade level separation condition, the only significant difference in average RMSE values was for the large item format effect condition.

Common Item Set Configuration

There was no significant difference between average BIAS or average RMSE values for the narrow and expended common item sets regardless of item format condition, vertical scaling method, grade level separation condition, or data generation method (Table C3).

Grade Level Separation

Average BIAS and average RMSE values were significantly larger for the large grade level separation condition under the separate calibration condition for all item format effect conditions and common item set configuration regardless of how the data was generated (Table C4). Effect sizes for mean BIAS values ranged from 0.80 to 4.65 and effect sizes for mean RMSE values ranged from 0.35 to 1.57.

Under pairwise concurrent calibration, average BIAS and RMSE values for the large grade level separation condition were significantly larger than the small grade level separation values for all item format effect conditions when data was generated under the unidimensional and 2-dimensional models. Additionally, average BIAS values were significantly larger for the large grade level separation condition for the large and moderate item format effect conditions under pairwise concurrent calibration for data generated under the 3-dimensional model while for the same was true for mean RMSE values for data generated under the 2-dimensional model. For data generated under the 3-dimensional and unidimensional models, there was no significant difference in average RMSE values between grade level separation conditions for pairwise concurrent calibration.

Data Generation Method

Significant differences in data generation methods were dependent upon item format effect and grade level separation. In general, including the vertical scale factor in the data generation process resulted in lower mean BIAS and mean RMSE values for most conditions (Table C5). Effect sizes for mean BIAS values ranged from 0.34 to 2.82 and for mean RMSE values ranged from 0.82 to 1.51.

For data generated under the unidimensional and 2-dimensional models, average BIAS values for separate calibration were significantly larger than those resulting from data generated under the 3-dimensional model, but only for the moderate and large item format effect conditions regardless of common item set configuration. There was, however, no significant difference in average BIAS values among data generation models for pairwise concurrent calibration under the large grade level separation condition. In addition, there was no significant difference in

average BIAS values for the no item format effect condition regardless of grade level separation, vertical scaling method, and common item set.

Significant differences in average RMSE values were dependent on item format effect and grade level separation. In general, average RMSE values were larger for traditional data generation methods compared to results generated under the 3-dimensional model. For the small grade level separation, mean RMSE values were significantly larger for the moderate and large item format effect conditions regardless of common item set configuration. However, for the large grade level separation, this was true for all item format effect conditions. In addition, the mean RMSE values were significantly smaller under the 3-dimensional data generation model compared to the unidimensional model, but mean RMSE values were not significantly different between the 2-dimensional and unidimensional data generation models.

4.1.1.3 Performance of Vertical Scaling Configurations

Of more practical importance, perhaps, is which vertical scaling configuration performs best under the different conditions of format effect simulated in this study. Using the pairwise comparisons matrix produced by the LSMEANS procedure in SAS, a table was constructed to summarize the overall performance of each of the four vertical scaling configurations; separate calibration with narrow common item set (sn), separate calibration with expanded common item set (se), pairwise concurrent calibration with narrow common item set (pn), and pairwise concurrent calibration with expanded common item set (pe) (Table 6). For the small grade level separation condition, few significant differences between vertical scaling configurations exist. Of those, all but one favored pairwise concurrent calibration over separate calibration. Even though the common item set configuration was rarely significant, the narrow common item set produced smaller average values. For the large grade level separation, pairwise concurrent calibration

92

produced the smallest average values for all conditions. Additionally, average BIAS values were smallest when using the expanded common item set, however, no common item set configuration consistently produced the smallest average RMSE values.

Table 7. Grade 5 vertical scaling methods by accuracy

	Data Generation	Format Effect	Small Separation	Large Separation
BIAS	Unidim	None	pn=sn=se=pe	pe=pn<se<sn
	2-Dim	None	pe=pn<se=sn	pe=pn<se=sn
		Small	pe<sn=se<pn	pe=pn<se=sn
		Moderate	pn=pe<se=sn	pn=pe<sn=se
		Large	pn<pe<sn=se	pe=pn<sn=se
	3-Dim	None	sn=se=pe=pn	pe=pn<se=sn
		Small	sn=se<pn=pe	pe<pn<sn=se
		Moderate	pe=pn=sn=se	pe=pn<se=sn
		Large	pn=pe<se=sn	pe=pn<se=sn
RMSE	Unidim	None	pn=pe=sn=se	pe=pn<sn=se
	2-Dim	None	pe=pn=se=sn	pn=pe<sn=se
		Small	pn=sn=se=pe	pn=pe<sn=se
		Moderate	pn=pe=se=sn	pn=pe<se=sn
		Large	pn=pe<sn=se	pn=pe<sn=se
	3-Dim	None	pn=pe=se=sn	pn=pe<sn=se
		Small	pn=pe=sn=se	pe=pn<sn=se
		Moderate	pn=pe=se=sn	pe=pn<se=sn
		Large	pn=pe=se=sn	pe=pn<se=sn

sn=separate - narrow; se=separate - expanded; pn=pairwise – narrow; pe=pairwise - expanded
< significantly smaller   = not significantly different

4.1.1.4 Summary of Grade 5 Results

Based on the ANOVA results; item format effect, vertical scaling method, and grade level separation had a statistically significant impact on the scaling results for this grade level. Common item set configuration did not have a significant impact on the scaling results. In general, a large separation between grade levels when using separate calibration resulted in the largest mean BIAS and mean RMSE values, especially for the largest item format effect conditions. In cases where significant differences between vertical scaling configurations resulted, pairwise concurrent produced smaller average values. There was rarely a significant

93

difference for common item set, however, the narrow common item set produced smaller average values for the small grade level separation while the expanded common item set produced smaller average values for the large grade level separation.

Important to note is the relative proportion of replications with positive and negative mean BIAS values. For this grade level, there was a large proportion of replications with positive mean BIAS values. This resulted in an overall positive average BIAS, regardless of grade level separation or item format effect condition. This relative proportion of positive replications increased as item format effect increased and resulted in spuriously small average BIAS values when the relative proportions were essentially equal. This phenomenon was particularly evident for the 3-dimensional model, regardless of calibration method, and for the 2-dimensional model under pairwise concurrent calibration as well as for the small and/or moderate item format effect conditions for separate calibration. However, this phenomenon was eliminated for mean RMSE values because the calculation and these values revealed a more expected pattern and magnitude.

Finally, significant differences were found between data generation methods. Including the vertical scale factor in the data generation model resulted in smaller average RMSE values for all conditions, although not always significant, and smaller average BIAS values for the small grade level separation conditions. On the other hand, values for average BIAS under the large grade level separation condition, when significant, favored the traditional data generation method.

4.1.2 Grade 6 Results

4.1.2.1 Trends by Condition

Average BIAS

Mean BIAS values generally increased as degree of item format effect increased, regardless of common item set, grade level separation, and vertical scaling method (Table 8). As with grade 5, however, trends for the small grade level separation were less clear because of the inconsistent estimation of the observed expected score. For separate calibration, scores were estimated such that the resulting mean BIAS values were both under- and overestimated, however, the proportion of under- to overestimated values differed by item format effect condition. This resulted in mean values for the no and small item format effect conditions being predominately underestimated (as expected) while values for the large item format effect condition were consistently overestimated. However, for the moderate item format effect, replications were split roughly 50/50 resulting in spuriously low mean BIAS values. This inconsistent estimation was observed for values generate under the 3-dimensional model and for the large grade level separation, but the mean BIAS value for each replication was predominately negative for both. Mean RMSE values were not impacted by this phenomenon.

Additionally, in general, mean BIAS values for all item format effect conditions tended to be similar to one another rather than following a particular pattern when using data generated under the 3-dimensional model. Additionally, small grade level separation values were smaller than large grade level separation values for most common item set, vertical scaling, item format effect, and data generation conditions. However, scores produced for the large item format effect condition under the small grade level separation condition for the 2-dimensional model

generation resulted in larger mean BIAS values compared to the large grade level separation condition for that same data generation model (Figures 8 and 9).

Average RMSE

Mean RMSE values increased as degree of item format effect increased, regardless of common item set, grade level separation, vertical scaling method, or data generation method. Additionally, values from the small grade level separation condition were smaller than those form the large grade level separation values regardless of common item set, vertical scaling method, item format effect condition, and data generation method. Scores from data generated under the 3-dimensional model were smaller than scores generated from data under the 2-dimensional or unidimensional models. In general, however, both calibration methods produced scores with similar average RMSE values regardless of common item set under the 3-dimensional model generation; but smaller values than resulted from the data generated under the 2-dimensional model. Values resulting from scores generated under the traditional unidimensional and 2-dimensional essentially unidimensional model generation were, in general, comparable while the essentially unidimensional values resulting from the 3-dimensional model generation were smaller comparatively (Figures 10 and 11).

Table 8. Grade 6 mean BIAS and RMSE by condition

Scaling Method	Common Item Set	Model	Format Effect	Mean BIAS		Mean RMSE	
				Small Separation	Large Separation	Small Separation	Large Separation
Separate	Narrow	Unidim	None	-0.03	-0.53	3.73	4.53
		2-Dim	None	-0.03	-0.54	3.73	4.52
			Small	-0.11	-0.59	3.90	4.54
			Moderate	-0.03	-0.67	4.70	5.54
			Large	0.86	-0.65	5.71	6.84
		3-Dim	None	-0.13	-0.47	3.32	3.69
			Small	-0.14	-0.54	3.38	3.82
			Moderate	-0.16	-0.72	3.51	4.12
			Large	-0.17	-0.92	3.80	4.71
	Expanded	Unidim	None	-0.02	-0.51	3.73	4.51
		2-Dim	None	-0.06	-0.56	3.72	4.53
			Small	-0.07	-0.62	3.94	4.63
			Moderate	-0.01	-0.67	4.60	5.85
			Large	0.87	-0.59	5.75	6.87
		3-Dim	None	-0.13	-0.50	3.31	3.74
			Small	-0.14	-0.54	3.42	3.83
			Moderate	-0.15	-0.68	3.54	4.03
			Large	-0.17	-0.86	3.83	4.95
Pairwise	Narrow	Unidim	None	-0.03	-0.55	3.74	4.65
		2-Dim	None	-0.07	-0.54	3.68	4.54
			Small	-0.10	-0.59	3.88	4.62
			Moderate	0.01	-0.62	4.56	5.39
			Large	0.88	-0.79	5.72	6.73
		3-Dim	None	-0.07	-0.49	3.32	3.82
			Small	-0.11	-0.54	3.35	3.92
			Moderate	-0.15	-0.70	3.51	4.10
			Large	-0.16	-0.89	3.84	4.36
	Expanded	Unidim	None	-0.03	-0.55	3.71	4.65
		2-Dim	None	-0.05	-0.57	3.71	4.59
			Small	-0.06	-0.58	3.88	4.60
			Moderate	0.04	-0.62	4.66	5.61
			Large	0.88	-0.76	5.75	6.73
		3-Dim	None	-0.14	-0.52	3.31	3.83
			Small	-0.11	-0.52	3.44	3.89
			Moderate	-0.15	-0.64	3.51	4.00
			Large	-0.18	-0.84	3.89	4.42

Figure 8. Grade 6 average BIAS for small grade level separation

Figure 10. Grade 6 average RMSE for small grade level separation
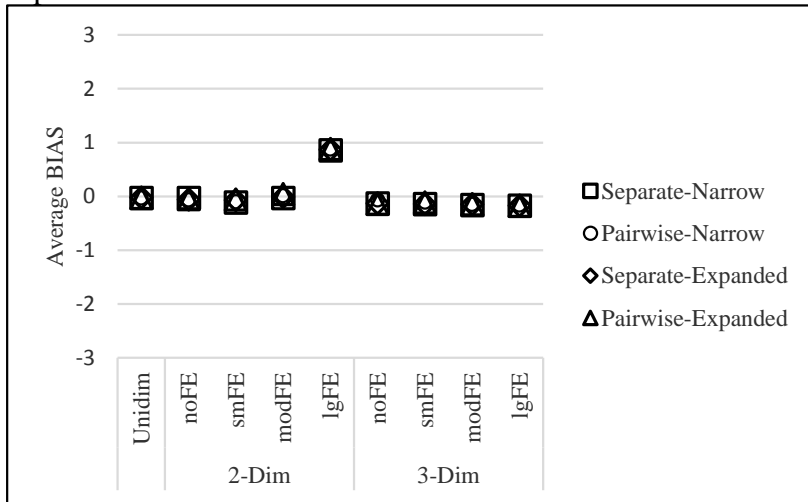
Figure 9. Grade 6 average BIAS for large grade level separation
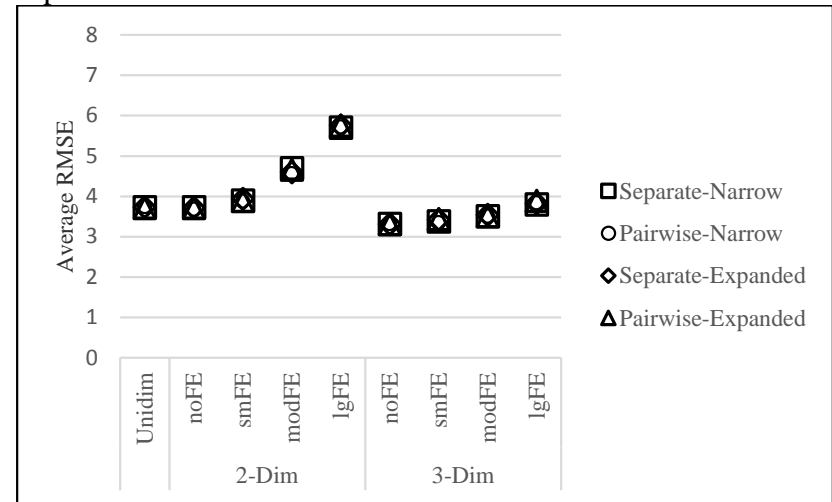
Figure 11. Grade 6 average RMSE for large grade level separation
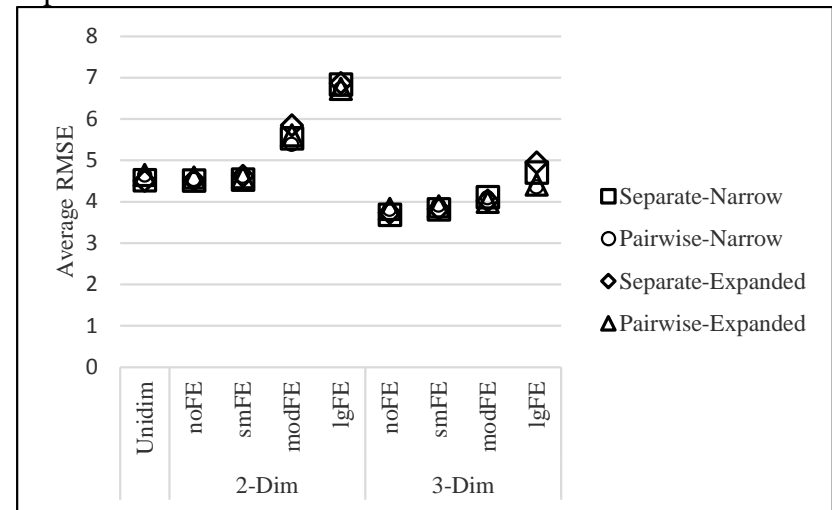
4.1.2.2 ANOVA Results

Again, two separate five-way mixed ANOVAs using BIAS and RMSE values as the dependent variable were performed on the results of the 100 replications for this grade level. Most interactions and main effects were significant for the model (Table B3 and B4). Using the LSMEANS command in SAS, pairwise comparisons for the two- and three-way interactions were examined and effect sizes were calculated. Additionally, the difference matrix produced was used to examine relevant comparisons between vertical scaling configurations by condition to determine the scaling method(s) that resulted in the most accurate vertical scales for each condition.

Item Format Effect

Mean BIAS values for item format effect were dependent on grade level separation, vertical scaling method, and data generation method (Table C6). Effect sizes for mean BIAS values ranged from 0.47 to 2.05 while effect sizes for mean RMSE values ranged from 0.33 to 0.76.

For the 2-dimensional data generation model, mean BIAS values were significantly larger for the large item format effect condition compared to the no, small, and moderate item format effect conditions under the small grade level separation condition regardless of scaling method. For the large grade level separation condition there was no significant difference in mean BIAS values. For the 3-dimensional generation model, there was no significant difference in mean BIAS values between item format effect conditions when using either vertical scaling method for the small grade level separation condition. However, for the large grade level separation condition, the large item format effect condition had mean BIAS values significantly larger than the no and small item format effect conditions.

Average RMSE values for the 2-dimensional data generation model under the small grade level separation were significantly higher for the large item format effect compared to the other item format effect conditions. Under the large grade level separation condition, while the large item format effect was, again, significantly larger than the other item format effect conditions, additionally, the moderate item format effect conditions had significantly higher mean RMSE values than the no and small item format effect condition under the expanded common item set configuration while under the narrow common item set configuration the moderate item format effect condition was larger than only the no item format effect condition. For the 3-dimensional model data generation, there was no significant difference in mean RMSE values for any item format effect conditions under the small grade level separation condition. Under the large grade level separation condition, regardless of common item set, the large item format effect condition had higher mean values than the no item format effect condition, but with the expanded common item set, the large item format effect condition was also larger than the small item format effect condition. For pairwise concurrent calibration under the large grade level separation there was no significant difference in mean RMSE values between item format effect conditions.

Vertical Scaling Method

There was also no significant difference in mean BIAS or mean RMSE values for any study conditions for this grade level (Table C7).

Common Item Set Configuration

There was no significant difference in mean BIAS or mean RMSE values between the narrow and expanded common item set configurations regardless of item format effect, vertical scaling method, grade level separation, and data generation method (Table C8).

<u>Grade Level Separation</u>

Regardless of how scores were scaled, average BIAS values were significantly larger for the large grade level separation condition for all item format effect conditions for all common item set configuration and data generation methods (Table C9). Effect sizes for these values ranged from 0.50 to 2.44.

Average RMSE values were significantly larger for the large grade level separation condition compared to the small grade level separation condition only for the large item format effect condition under the 2-dimensional data generation model and for separate calibration conditions under the 3-dimensional data generation model. Effect sizes for these values were small and ranged from 0.33 to 0.41.

<u>Data Generation Method</u>

Average BIAS values were significantly larger under the 2-dimensional model for the large item format effect condition regardless of grade level separation or common item set configuration when using separate calibration. (Table C10). While this was also true when using pairwise concurrent calibration under the small grade level separation conditions, under the large grade level separation condition, it was not true. Effect sizes for these values ranged from 0.39 to 1.53. However, average RMSE values were significantly larger under the traditional data generation model for both the moderate and large item format effect conditions regardless of grade level separation, scaling method, and common item set configuration. Effect sizes for mean RMSE values ranged from 0.34 to 0.77.

4.1.2.3 Performance of Vertical Scaling Configurations

Using the pairwise comparisons matrix produced by the LSMEANS procedure, a table was constructed to summarize the overall performance of each of the four vertical scaling

101

configurations to determine which vertical scaling method produced the most accurate results (Table 9). For the small grade level separation, mean BIAS values tended to be smallest when using separate calibration regardless of common item set while mean RMSE values tended to be smallest when using pairwise concurrent calibration; although these differences were rarely significant. For the large grade level separation, whether examining mean BIAS or mean RMSE, pairwise concurrent calibration tended to produce the smallest values although, again, differences were rarely significant. Concerning the common item set configuration, mean BIAS values tended to be smallest when using the expanded common item set while mean RMSE values tended to be smallest when using the narrow common item set. These differences, however, were not significant.

Table 9. Grade 6 vertical scaling methods by accuracy

	Data Generation	Format Effect	Small Separation	Large Separation
BIAS	Unidim	None	se=sn<pn=pe	pe<pn<se=sn
	2-Dim	None	sn=se=pn=pe	sn=se=pn=pe
		Small	se=sn<pe<pn	pe<pn<sn=se
		Moderate	pn=se=pe=sn	pe=pn=se=sn
		Large	sn=se=pe=pn	se=sn=pe=pn
	3-Dim	None	se=pn=sn=pe	sn=se=pn=pe
		Small	se=sn=pn=pe	pe<pn<sn=se
		Moderate	pe=se=pn=sn	pe=se=pn=sn
		Large	pe=pn=sn=se	pe=se=pn=sn
RMSE	Unidim	None	pn=sn=se=pe	pe=pn=se=sn
	2-Dim	None	pe=pn=se=sn	sn=se=pn=pe
		Small	sn=se=pn=pe	sn=se=pn=pe
		Moderate	pn=se=pe-sn	pn=sn=pe=se
		Large	sn=se=pn=pe	pn=pe=sn=se
	3-Dim	None	pe=pn=se=sn	sn=se=pn=pe
		Small	pn=sn=se=pe	pe=pn=sn=se
		Moderate	pe=se=sn=pn	pe=pn=se=sn
		Large	sn=se=pn=pe	pn=pe=sn=se

sn=separate - narrow; se=separate - expanded; pn=pairwise – narrow; pe=pairwise - expanded
< significantly smaller   = not significantly different

4.1.2.4 Summary of Grade 6 Results

Based on the ANOVA results item format effect, vertical scaling method and grade level separation had a statistically significant impact on the scaling results for this grade level. As with grade 5, common item set configuration did not have a significant impact on the scaling results for this grade level. In general, a large separation between grade levels resulted in the largest mean BIAS and mean RMSE values, especially for the moderate and large item format effect conditions. However, average BIAS values were less impacted by item format effect when using pairwise concurrent calibration when a large separation in grade levels was present. Significant differences between vertical scaling configurations were rare, but grade level separation had a significant impact on mean BIAS values with a large separation producing larger mean values. Mean RMSE values were less impacted by grade level separation except for the largest item format effect conditions.

The same phenomenon observed for grade 5 in which the relative proportion of replications with positive and negative mean BIAS values changed across item format effect conditions was also observed for grade 6. While the average BIAS values for grade 6 overall were negative, the relatively equal proportions of negative and positive replications was observed for the 2-dimensional condition regardless of scaling method and resulted in spuriously low mean BIAS values, especially for the moderate item format effect condition. Again, the mean RMSE values were not affected by this issue and revealed a more expected pattern and magnitude.

Few statistically significant differences were found between data generation methods. Where significant differences existed, they were generally for the moderate and/or large item format effect conditions with the 3-dimensional data generation model producing smaller values.

103

Including the vertical scale factor in the data generation model resulted in smaller average RMSE values for all conditions, although not always significant, and smaller average BIAS values for the small grade level separation conditions. Smaller average BIAS values under the large grade level separation condition, when significant, favored the traditional data generation method.

4.1.3 Grade 7 Results

4.1.3.1 Trends by Condition

Average BIAS

Mean BIAS values generally increased as degree of item format effect increased, regardless of common item set, grade level separation, or vertical scaling method (Table 10). Scores resulting from data generated under the 2-dimensional and 3-dimensional models were similar across vertical scaling methods, common item sets, and grade level separation. In addition, all methods of scaling and data generation consistently overestimated the expected scores. All data generation models resulted in similar average BIAS values regardless of condition (Figures 12 and 13).

Average RMSE

Likewise, mean RMSE values generally increased as degree of item format effect increased, regardless of common item set, grade level separation, or vertical scaling method for both grade level separation conditions. While both calibration methods produced similar results, scores resulting from the 3-dimensional model generation had smaller mean RMSE values than those resulting from data generated under the 2-dimensional model. Also, the essentially unidimensional scores generated with the 3-dimensional model resulted in comparatively smaller

mean values compared to the scores generated under the traditional unidimensional and 2-dimensional essentially unidimensional models (Figure 14 and 15).

Table 10. Grade 7 mean BIAS and mean RMSE by condition

Scaling Method	Common Item Set	Model	Format Effect	Mean BIAS		Mean RMSE	
				Small Separation	Large Separation	Small Separation	Large Separation
Separate	Narrow	Unidim	None	-0.18	-0.18	3.65	3.65
		2-Dim	None	-0.18	-0.19	3.66	3.66
			Small	-0.23	-0.23	3.82	3.82
			Moderate	-0.32	-0.32	4.55	4.54
			Large	-0.29	-0.29	5.47	5.46
		3-Dim	None	-0.17	-0.18	3.25	3.25
			Small	-0.19	-0.18	3.30	3.29
			Moderate	-0.20	-0.20	3.42	3.42
			Large	-0.25	-0.24	3.63	3.62
	Expanded	Unidim	None	-0.18	-0.17	3.63	3.64
		2-Dim	None	-0.18	-0.18	3.66	3.64
			Small	-0.24	-0.26	3.81	3.80
			Moderate	-0.31	-0.31	4.55	4.52
			Large	-0.28	-0.28	5.46	5.47
		3-Dim	None	-0.18	-0.17	3.25	3.24
			Small	-0.18	-0.19	3.29	3.30
			Moderate	-0.23	-0.21	3.44	3.43
			Large	-0.22	-0.25	3.61	3.63
Pairwise	Narrow	Unidim	None	-0.18	-0.18	3.64	3.65
		2-Dim	None	-0.18	-0.20	3.65	3.66
			Small	-0.24	-0.26	3.81	3.82
			Moderate	-0.36	-0.42	4.61	4.54
			Large	-0.38	-0.50	5.47	5.45
		3-Dim	None	-0.15	-0.18	3.25	3.25
			Small	-0.18	-0.18	3.29	3.29
			Moderate	-0.21	-0.21	3.41	3.42
			Large	-0.26	-0.27	3.63	3.62
	Expanded	Unidim	None	-0.17	-0.17	3.63	3.64
		2-Dim	None	-0.18	-0.18	3.66	3.64
			Small	-0.25	-0.28	3.81	3.80
			Moderate	-0.35	-0.41	4.55	4.51
			Large	-0.37	-0.49	5.46	5.45
		3-Dim	None	-0.17	-0.16	3.24	3.24
			Small	-0.17	-0.19	3.29	3.30
			Moderate	-0.21	-0.22	3.42	3.43
			Large	-0.24	-0.28	3.61	3.63

Figure 12. Grade 7 average BIAS for small grade level separation
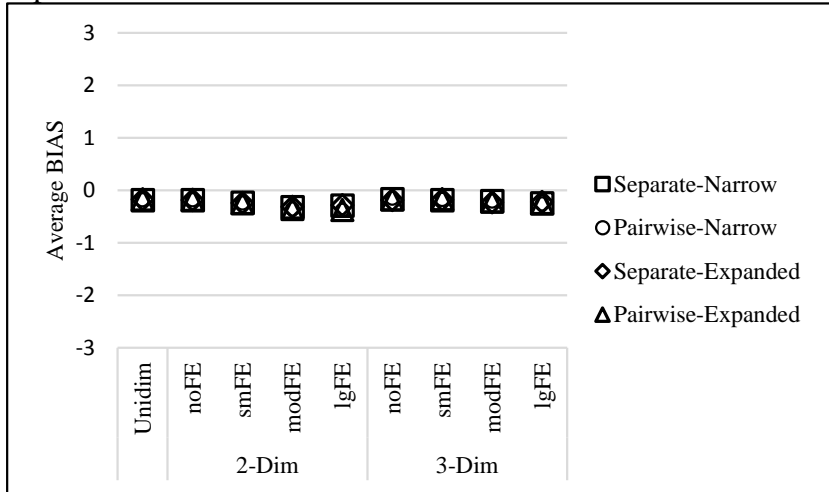
Figure 14. Grade 7 average RMSE for small grade level separation
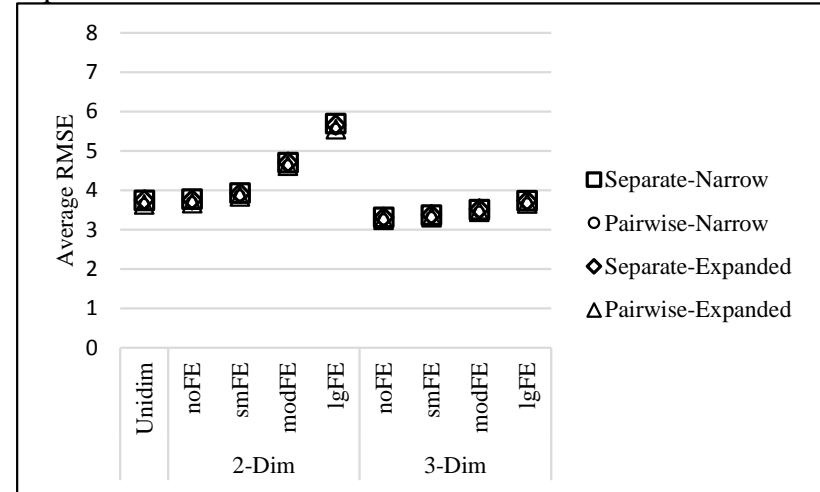
Figure 13. Grade 7 average BIAS for large grade level separation
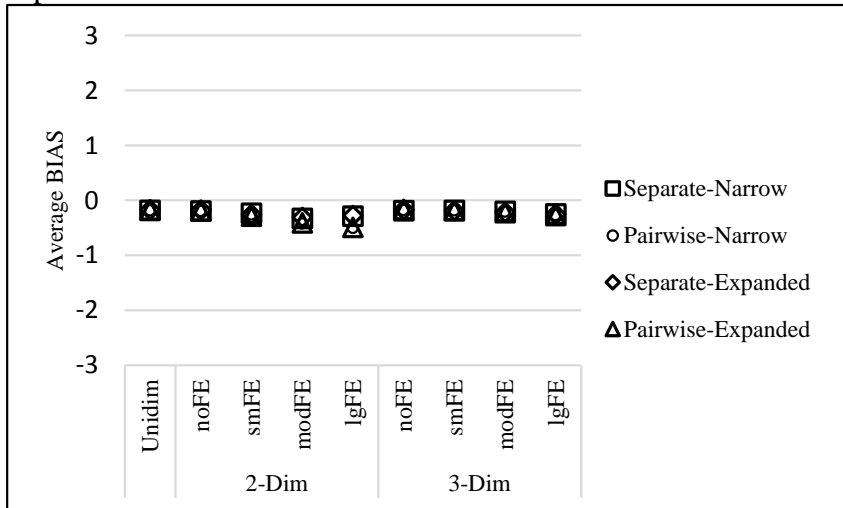
Figure 15. Grade 7 average RMSE for large grade level separation
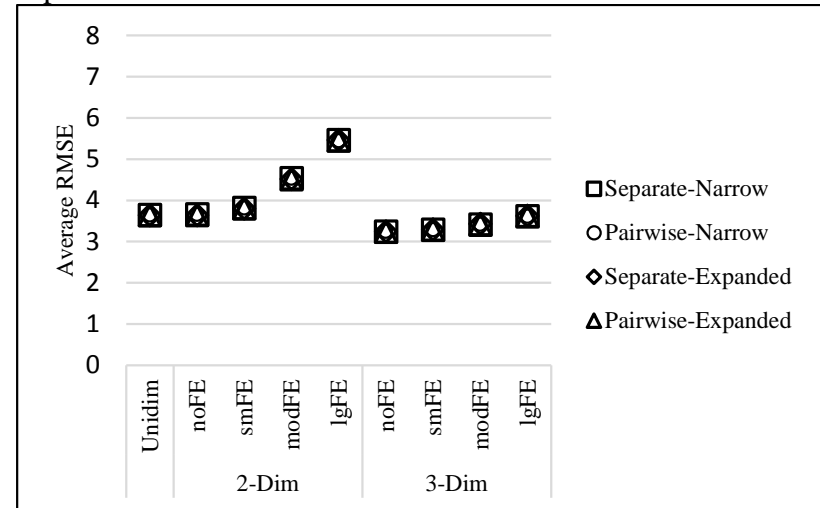
4.1.3.2 ANOVA Results

Again, two separate five-way mixed ANOVAs using BIAS and RMSE values as the dependent variable were performed on the 100 replications for grade 7. While only about half of the interactions and main effects were significant for the BIAS model, only a handful of interactions and main effects were significant for the RMSE model (Table B5 and B6). Using the LSMEANS command in SAS, pairwise comparisons were examined and effect sizes were calculated. Additionally, the difference matrix produced was used to examine relevant comparisons between vertical scaling configurations by condition to determine the scaling method(s) that resulted in the most accurate vertical scales for each condition.

<u>Item Format Effect</u>

Mean BIAS values for the item format effect were dependent on vertical scaling method and data generation method (Table C11). Effect sizes for these values ranged from 0.37 to 0,81. Under the 2-dimensional data generation model, regardless of grade level separation or common item set configuration, there was no significant difference among item format effect condition when using separate calibration. When using pairwise concurrent calibration, mean BIAS values for the large and moderate item format effect conditions were significantly larger than the no item format effect condition as well as the small item format effect condition under the large grade level separation. However, there was no significant difference in mean BIAS values among item format effect conditions regardless of scaling method, grade level separation, or common item set configuration under the 3-dimensional data generation model.

Mean RMSE values were significantly larger for the large item format effect condition compared to the no, small, and moderate item format effect conditions and were larger for the moderate item format effect compared to the no and small item format effect conditions for all

vertical scaling methods, common item set configurations, grade level separation conditions, and data generation models. In addition, for the 2-dimensional data generation method, the moderate item format effect condition was significantly larger than the small item format effect condition regardless of any other condition as well. Effect sizes for these values ranged from 0.36 to 4.33.

Vertical Scaling Method

There was no significant difference in mean BIAS and mean RMSE values between scaling methods regardless of item format effect, grade level separation, common item set configuration, or data generation method except for the large grade level separation condition under the 2-dimensional generation model for which the large item format effect condition with pairwise concurrent calibration was significantly smaller (Table C12). Effect size for these two significant results was 0.57.

Common Item Set Configuration

There was no significant difference in mean BIAS or mean RMSE values between common item set configurations regardless of item format effect, scaling method, grade level separation, or data generation model (Table C13).

Grade Level Separation

There was also no significant difference in mean BIAS or mean RMSE values between the small and large grade level separation conditions regardless of item format effect, scaling method, common item set, or data generation model (Table C14).

Data Generation Method

There was no significant difference in mean BIAS values among the data generation models for the small grade level separation and the large grade level separation with separate calibration regardless of item format effect condition, common item set configuration, or grade level

separation (Table C15). However, for the large grade level separation with pairwise concurrent calibration, the 3-dimensional model produced significantly smaller mean BIAS values comparatively for the moderate and large item format effect conditions regardless of common item set and grade level separation condition. Effect sizes for mean BIAS values ranged from 0.48 to 0.61.

The 3-dimensional data generation model produced significantly smaller mean RMSE values compared to both the 2-dimensional and unidimensional data generation models regardless of item format effect condition, vertical scaling method, common item set configuration, and grade level separation condition, but values produced under the unidimensional model were not significantly different from those generated under the 2-dimensional model. Effect sizes for these differences ranged from 0.93 to 4.39.

4.1.3.3 Performance of Vertical Scaling Configurations

Using the pairwise comparisons matrix, a summary table was produced (Table 11). Mean BIAS values tended to be smaller under the large grade level separation when using separate calibration, although differences were not significant. On the other hand, mean RMSE values tended to be smaller, although not significant, when using pairwise concurrent calibration with the expanded common item set regardless of item format effect condition, grade level separation and data generation model. Also, the expanded common item set tended to produce the smallest mean BIAS and mean RMSE values for the unidimensional data generation model. Again, there was no significant difference between vertical scaling configurations.

Table 11. Grade 7 vertical scaling methods by accuracy

	Data Generation	Format Effect	Small Separation	Large Separation
BIAS	Unidim	None	pe=se=pn=sn	se=pe=sn=pn
	2-Dim	None	pe=pn=se=sn	se=pe=sn=pn
		Small	sn=se=pn=pe	sn=pn=se=pe
		Moderate	se=sn=pe=pn	se=sn=pe=pn
		Large	se=sn=pe=pn	se=sn<pe=pn
	3-Dim	None	pn=sn=pe=se	pe=se=pn=sn
		Small	pe=pn=se=sn	sn=pn=se=pe
		Moderate	sn=pn=pe=se	sn=pn=se=pe
		Large	se=pe=sn=pn	sn=se=pn=pe
RMSE	Unidim	None	pe=se=pn=sn	pe=se=pn=sn
	2-Dim	None	pn=sn=pe=se	pe=se=pn=sn
		Small	pe=se=pn=sn	pe=se=pn=sn
		Moderate	pe=sn=se=pn	pe=se=pn=sn
		Large	pe=se=pn=sn	pe=pn=sn=se
	3-Dim	None	pe=se=pn=sn	pe=pn=sn=se
		Small	pe=se=pn=sn	pn=sn=pe=se
		Moderate	pn=sn=pe=se	pn=sn=pe=se
		Large	pe=se=pn=sn	pn=sn=pe=se

sn=separate - narrow; se=separate - expanded; pn=pairwise – narrow; pe=pairwise - expanded
< significantly smaller   = not significantly different

4.1.3.4 Summary of Grade 7 Results

Based on the ANOVA results, item format effect had a statistically significant impact on the vertical scaling results for this grade level. Item format effect, common item set, and vertical scaling method did not have a significant impact on grade 7 results. In general, mean RMSE values were significantly larger for the moderate and/or large item format effect condition. Mean BIAS values, on the other hand, were less likely to be impacted by item format effect. Significant differences between vertical scaling configurations did not really exist, but mean RMSE values were smaller when using pairwise concurrent calibration with the expanded common item set.

However, significant differences were found between data generation methods. Including the vertical scale factor in the data generation model resulted in significantly smaller average

RMSE values for most conditions and significantly smaller average BIAS values for the largest item format effect conditions.

4.1.4 Grade 8 Results

4.1.4.1 Trends by Condition

<u>Average BIAS</u>

Mean BIAS values generally increased as degree of item format effect increased, regardless of common item set, grade level separation, and vertical scaling method. However, the large item format effect condition under the 2-dimensional data generation model for pairwise concurrent calibration had smaller than expected mean BIAS values (Table 12). Mean BIAS values produced from separate calibration were smaller for the small grade level separation condition compared to the large grade level separation condition regardless of common item set and data generation method. However, under pairwise concurrent calibration, mean BIAS values were similar across data generation methods and common item set; but values for the large grade level separation condition tended to be slightly smaller than the mean values resulting from the small grade level separation condition. In addition, all methods of scaling and data generation consistently produced average scores that were predominately overestimated. As with grade 6, smaller item format effect conditions produced replications with negative BIAS values while the larger item format effect conditions produced replications in which the mean BIAS values were increasingly positive. This was especially true for the large item format effect condition and produced average BIAS scores across replications that were spuriously low. This phenomenon, however, did not impact the mean RMSE values.

The unidimensional as well as the essentially unidimensional conditions for both the 2-dimensional and 3-dimensional model data generation resulted in similar average BIAS values regardless of condition (Figures 16 and 17).

Average RMSE

Likewise, mean RMSE values generally increased as degree of item format effect increased, regardless of common item set, grade level separation, and vertical scaling method. While values across common item sets, scaling methods, and data generation methods were similar; pairwise concurrent calibration tended to produce smaller mean values, especially for the large item format effect conditions, and mean values resulting from the 3-dimensional model generation were smaller than those resulting from data generated under the 2-dimensional model. Again, scores from the essentially unidimensional condition generated under the 3-dimensional model resulted in comparatively smaller mean RMSE values compared to those generated under the traditional unidimensional and 2-dimensional essentially unidimensional models (Figures18 and 19).

Table 12. Grade 8 mean BIAS and mean RMSE by condition

Scaling Method	Common Item Set	Model	Format Effect	Mean BIAS		Mean RMSE	
				Small Separation	Large Separation	Small Separation	Large Separation
Separate	Narrow	Unidim	None	-0.20	-0.66	3.74	4.03
		2-Dim	None	-0.21	-0.69	3.78	4.04
			Small	-0.27	-0.73	3.93	4.21
			Moderate	-0.32	-0.87	4.71	4.93
			Large	-0.30	-1.47	5.69	6.08
		3-Dim	None	-0.20	-0.50	3.32	3.40
			Small	-0.22	-0.52	3.38	3.46
			Moderate	-0.24	-0.57	3.51	3.58
			Large	-0.26	-0.79	3.74	3.87
	Expanded	None	None	-0.19	-0.68	3.76	4.01
		2-Dim	None	-0.22	-0.72	3.76	4.07
			Small	-0.25	-0.74	3.92	4.20
			Moderate	-0.30	-0.89	4.72	4.95
			Large	-0.30	-1.36	5.67	6.02
		3-Dim	None	-0.20	-0.50	3.31	3.41
			Small	-0.19	-0.51	3.38	3.45
			Moderate	-0.24	-0.61	3.53	3.62
			Large	-0.26	-0.77	3.73	3.86
Pairwise	Narrow	Unidim	None	-0.18	-0.15	3.64	3.57
		2-Dim	None	-0.19	-0.17	3.69	3.59
			Small	-0.26	-0.17	3.86	3.78
			Moderate	-0.28	-0.16	4.62	4.54
			Large	-0.20	-0.02	5.55	5.45
		3-Dim	None	-0.18	-0.17	3.26	3.15
			Small	-0.19	-0.16	3.31	3.20
			Moderate	-0.20	-0.16	3.45	3.34
			Large	-0.21	-0.15	3.66	3.52
	Expanded	Unidim	None	-0.17	-0.17	3.65	3.57
		2-Dim	None	-0.21	-0.18	3.67	3.59
			Small	-0.24	-0.20	3.85	3.79
			Moderate	-0.27	-0.15	4.63	4.55
			Large	-0.20	-0.02	5.55	5.46
		3-Dim	None	-0.18	-0.16	3.25	3.15
			Small	-0.18	-0.16	3.32	3.19
			Moderate	-0.20	-0.15	3.46	3.32
			Large	-0.21	-0.15	3.66	3.52

Figure 16. Grade 8 average BIAS for small grade level separation
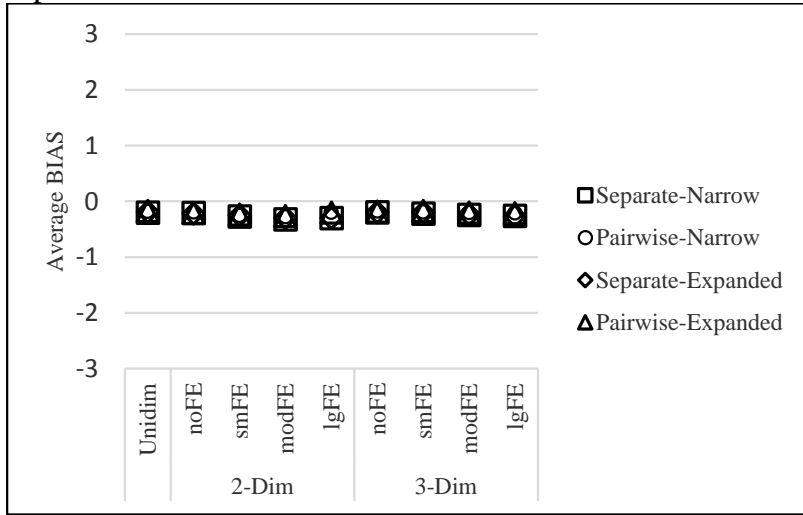
Figure 18. Grade 8 average RMSE for small grade level separation
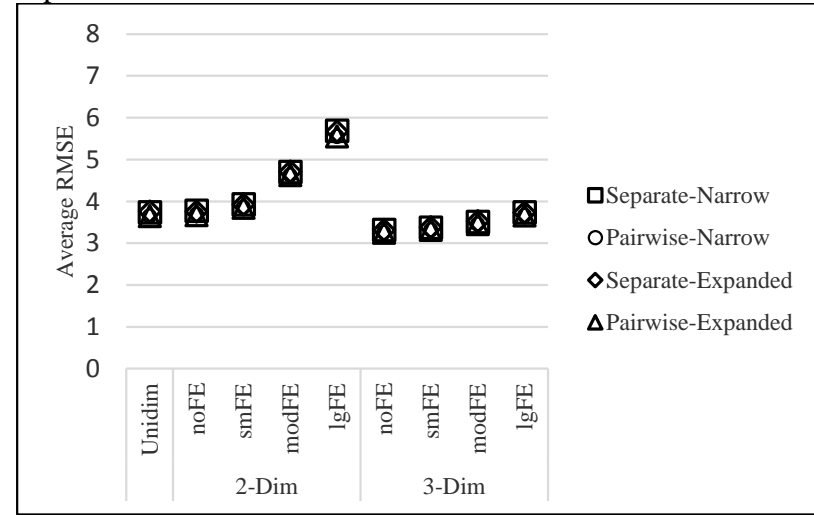
Figure 17. Grade 8 average BIAS for large grade level separation
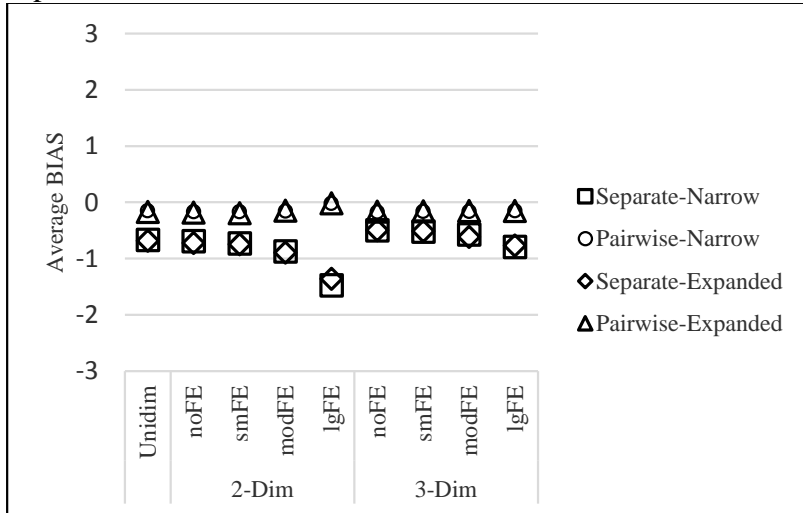
Figure 19. Grade 8 average RMSE for large grade level separation
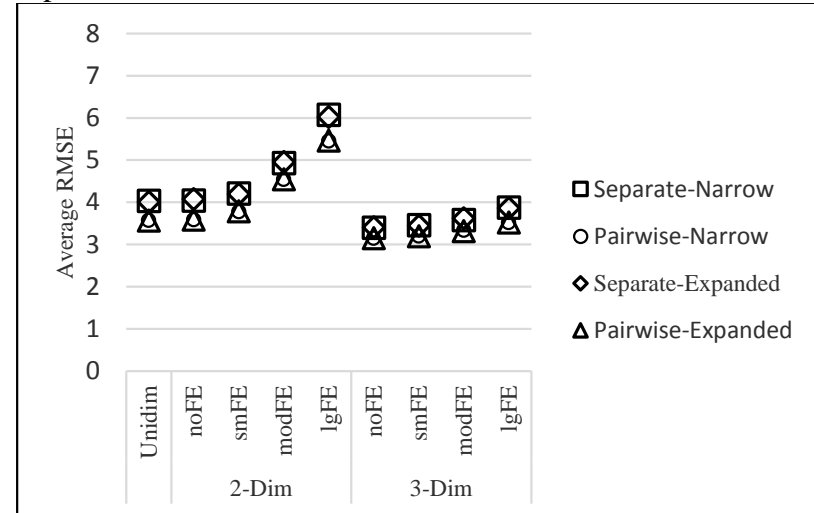
4.1.4.2 ANOVA Results

Two separate five-way mixed ANOVAs using BIAS and RMSE values as the dependent variable were performed on the results of the 100 replications for this grade level. For this grade level, most interactions and main effects were significant for the BIAS model, while only a handful of inteactions and main effects were significant for the RMSE values (Table B7 and B8). Again, pairwise comparisons were examined using the LSMEANS command in SAS and effect sizes were calculated. The difference matrix produced was used to examine relevant comparisons between scaling configurations to determine the scaling method(s) that resulted in the most accurate vertical scales for each condition.

<u>Item Format Effect</u>

Mean BIAS values were influenced by grade level separation, vertical scaling method, and data generation model, however, values were not significantly different for any item format effect conditions under the small grade level separation condition (Table C16). Effect sizes ranged from 0.28 to 1.52 for significant results. For the large grade level separation under separate calibration regardless of data generation model and common item set configuration, the large item format effect was significantly larger than the no, small, and moderate item format effect conditions. Additionally, the moderate item format effect condition was significantly larger than the no and small item format effect conditions under the 2-dimensional data generation model. For pairwise concurrent calibration, the large item format effect condition produced larger mean values than the no and small item format effect conditions, but only under the 2-dimensional data generation model. There was no significant difference among item format effect conditions for pairwise concurrent calibration under the 3-dimensional model.

Mean RMSE values under the 2-dimensional data generation model were significantly larger for both the moderate and large item format effect conditions compared to the no and small item format effect conditions and for the larger item format effect condition compared to the moderate item format effect condition. Values under the 3-dimensional data generation, however, were only significantly larger for the large item format effect compared to the no and small item format effect conditions. Effect sizes for these values ranged from 0.68 to 4.34.

Vertical Scaling Method

There was no significant difference in mean BIAS or mean RMSE values between separate and pairwise concurrent for any conditions under the small grade level separation calibration (Table C17). Under the large grade level separation condition, however, pairwise concurrent calibration produced significantly smaller mean values compared to those produced with separate calibration for all conditions regardless of item format effect, common item set configuration, and data generation model. Effect sizes for mean BIAS values ranged from 0.98 to 2.28 while effect sizes for mean RMSE values ranged from 0.52 to 1.33.

Common Item Set Configuration

There was no significant difference in mean BIAS or mean RMSE values between common item set configurations regardless of item format effect, scaling method, grade level separation, or data generation model (Table C18).

Grade Level Separation

Mean BIAS values were influenced by vertical scaling method with significantly larger values for the large grade level separation condition for separate calibration regardless of item format effect, common item set, and data generation conditions (Table C19). Effect sizes for mean BIAS values ranged from 0.58 to 2.28. Likewise, mean RMSE values were significantly larger

for the large grade level separation condition under separate calibration regardless of item format effect and common item set condition, but only for the 2-dimensional and unidimensional data generation models. Effect sizes for these values ranged from 0.48 to 0.82. There was no significant difference in mean values between grade level separation conditions for pairwise concurrent calibration condition and for separate calibration under the 3-dimensional data generation model regardless of condition.

Data Generation Method

Mean BIAS values were significantly larger for data generated under the traditional models for separate calibration with a large grade level separation condition for all item format effect and common item set conditions (Table C20). Mean RMSE values, on the other hand, were significantly larger under the traditional generation models for all scaling, item format effect, grade level separation, and common item set conditions. Effect sizes for mean BIAS values ranged from 0.33 to 1.33 and effect sizes for mean RMSE values ranged from 0.82 to 4.70.

4.1.4.3 Performance of Vertical Scaling Configurations

Based on the pairwise comparisons matrix, the overall performance of each scaling configurations was examined (Table 13). Significant differences in performance of vertical scaling configurations existed for the large grade level separation condition for this grade. In all cases, pairwise concurrent calibration produced significantly smaller mean BIAS and mean RMSE values. Even though not significant, pairwise concurrent calibration also produced the smallest mean BIAS and mean RMSE values for the small grade level separation.

Table 13. Grade 8 vertical scaling methods by accuracy

	Data Generation	Format Effect	Small Separation	Large Separation
BIAS	Unidim	None	pe=pn=se=sn	pn=pe<sn=se
	2-Dim	None	pn=pe=sn=se	pn=pe<sn=se
		Small	pe=se=pn=sn	pn=pe<sn=se
		Moderate	pe=pn=se=sn	pe=pn<sn=se
		Large	pe=pn=se=sn	pe=pn<se=sn
	3-Dim	None	pn=pe=se=sn	pn=pe<se=sn
		Small	pe=pn=se=sn	pn=pe<se=sn
		Moderate	pn=pe=se=sn	pe=pn<sn=se
		Large	pn=pe=se=sn	pe=pn<se=sn
RMSE	Unidim	None	pn=pe=sn=se	pe=pn=se=sn
	2-Dim	None	pe=pn=se=sn	pn=pe<sn=se
		Small	pe=pn=se=sn	pn=pe<se=sn
		Moderate	pn=pe=sn=se	pn=pe<sn=se
		Large	pn=pe=se=sn	pn=pe<se=sn
	3-Dim	None	pe=pn=se=sn	pn=pe<sn=se
		Small	pn=pe=sn=se	pe=pn<se=sn
		Moderate	pn=pe=sn=se	pe=pn<sn=se
		Large	pn=pe=se=sn	pn=pe<se=sn

sn=separate - narrow; se=separate - expanded; pn=pairwise – narrow; pe=pairwise - expanded
< significantly smaller   = not significantly different

4.1.4.4 Summary of Grade 8 Results

Based on the ANOVA results, item format effect, vertical scaling method, and grade level separation had a statistically significant impact on the scaling results for this grade level. Common item set had no impact on the scaling results. In general, mean BIAS and mean RMSE values were significantly larger for the large grade level separation especially when using separate calibration. Also, mean BIAS and mean RMSE values increased as item format effect condition increased with significant differences between conditions for the largest item format effects. As observed for grades 5 and 6, under pairwise concurrent calibration, a large enough proportion of replications resulted in positive mean BIAS values compared to negative such that some item format effect conditions had spuriously small values. Mean RMSE values for these conditions, however, were more aligned with the expected pattern and magnitude.

Additionally, significant differences were found between data generation methods. Including the vertical scale factor in the data generation model resulted in significantly smaller average BIAS and RMSE values for most conditions, especially the large grade level separation and separate calibration conditions.

4.1.5 Grade 9 Results

4.1.5.1 Trends by Condition

<u>Average BIAS</u>

Mean BIAS values generally increased as degree of item format effect increased for the small grade level separation condition regardless of common item set or vertical scaling method. However, for the large grade level separation condition, the moderate item format effect condition under the 3-dimensional data generation model and the small item format effect condition under the 2-dimensional data generation model had smaller than expected mean BIAS values when scaled with pairwise concurrent calibration (Table 14). This grade level was also impacted by the phenomenon observed for grades 5 and 6 except it occurred for the large grade level separation. Under this condition for pairwise concurrent calibration, mean BIAS values fell into two (unidimensional and 2-dimensional) or three (3-dimensional) 'distributions' for each item format effect condition; one positive and one negative along with a third centered at zero when applicable. For the smaller item format effect conditions, each replication tended to produce distributions with BIAS values with large proportions of both negative and positive values. As item format effect increased, the proportion of replications with negative BIAS values increased. This means that the smaller item format effect conditions could be spuriously high with the larger item format effect conditions more accurately portrayed, however, due to the

presence of the more neutral distribution for the 3-dimensional model, the larger item format effect conditions could still be spuriously high. Again, average RMSE values were not impacted by this phenomenon.

Average values produced by both calibration methods were smaller for the small grade level separation condition compared to the large grade level separation condition regardless of common item set and data generation method. However, for the large grade level separation condition, mean BIAS values resulting from the 3-dimensional model generation tended to be slightly smaller than those resulting from the 2-dimensional model generation. Although all methods of scaling and data generation consistently overestimated the expected score, pairwise concurrent calibration produced smaller average BIAS values for the large grade level separation condition while values were similar across scaling methods for the small grade level separation condition. Finally, the traditional unidimensional and the essentially unidimensional condition under the 2-dimensional model resulted in similar average values regardless of condition. However, mean values resulting from data generated under the 3-dimensional model tended to be smaller (Figures 20 and 21).

<u>Average RMSE</u>

Mean RMSE values generally increased as degree of item format effect increased, regardless of common item set, grade level separation, vertical scaling method, and data generation method. While average values across common item sets, vertical scaling methods, and data generation methods are similar; those resulting from data generated under the 3-dimensional model generation were smaller than those resulting from the 2-dimensional model generation and mean values under the small grade level separation condition were smaller than those under the large grade level separation condition. Again, scores generated under the essentially unidimensional

condition for the 3-dimensional model resulted in comparatively smaller mean RMSE values compared to the scores generated under the traditional unidimensional and 2-dimensional essentially unidimensional conditions (Figures 22 and 23).

Table 14. Grade 9 mean BIAS and mean RMSE by condition

Scaling Method	Common Item Set	Model	Format Effect	Mean BIAS		Mean RMSE	
				Small Separation	Large Separation	Small Separation	Large Separation
Separate	Narrow	Unidim	None	-0.12	-1.51	4.03	4.71
		2-Dim	None	-0.11	-1.52	4.02	4.71
			Small	-0.15	-1.62	4.17	4.86
			Moderate	-0.24	-2.10	5.07	5.56
			Large	-0.26	-3.67	6.05	7.17
		3-Dim	None	-0.20	-1.01	3.47	3.67
			Small	-0.23	-1.09	3.54	3.71
			Moderate	-0.27	-1.30	3.72	3.87
			Large	-0.31	-1.78	3.94	4.27
	Expanded	Unidim	None	-0.11	-1.48	4.05	4.68
		2-Dim	None	-0.13	-1.49	4.04	4.67
			Small	-0.16	-1.63	4.17	4.85
			Moderate	-0.26	-2.17	4.97	5.66
			Large	-0.26	-3.52	6.03	7.01
		3-Dim	None	-0.22	-1.01	3.50	3.68
			Small	-0.22	-1.08	3.52	3.72
			Moderate	-0.28	-1.33	3.70	3.91
			Large	-0.31	-1.74	3.93	4.26
Pairwise	Narrow	Unidim	None	-0.12	-1.10	4.06	4.77
		2-Dim	None	-0.12	-0.88	4.04	4.77
			Small	-0.16	-0.78	4.20	4.84
			Moderate	-0.29	-2.13	4.97	5.65
			Large	-0.41	-3.70	5.97	7.13
		3-Dim	None	-0.19	-0.46	3.45	3.68
			Small	-0.22	-0.45	3.50	3.79
			Moderate	-0.26	-0.92	3.67	4.11
			Large	-0.31	-1.16	3.87	4.31
	Expanded	Unidim	None	-0.11	-1.24	4.06	4.72
		2-Dim	None	-0.13	-1.02	4.07	4.69
			Small	-0.17	-0.59	4.18	4.82
			Moderate	-0.31	-1.43	4.96	5.50
			Large	-0.40	-3.79	6.00	7.16
		3-Dim	None	-0.20	-0.74	3.42	3.58
			Small	-0.21	-0.78	3.49	3.66
			Moderate	-0.27	-0.83	3.65	3.91
			Large	-0.32	-0.99	3.93	4.20

Figure 20. Grade 9 average BIAS for small grade level separation

Figure 21. Grade 9 average BIAS for large grade level separation
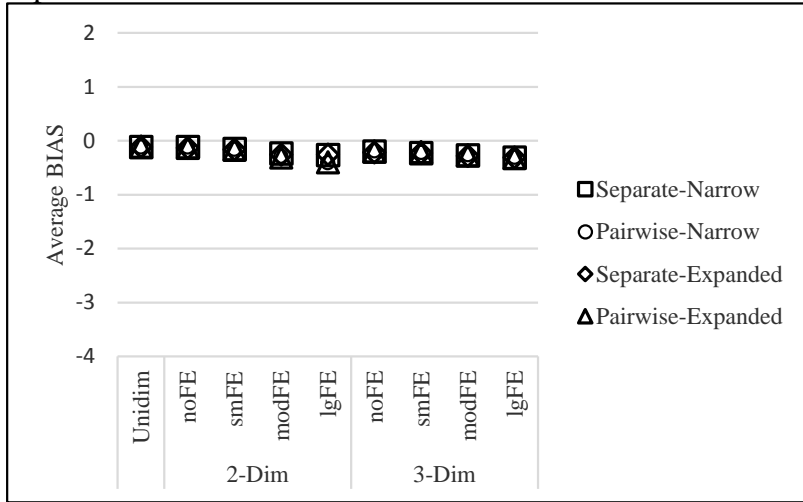
Figure 22. Grade 9 average RMSE for small grade level separation
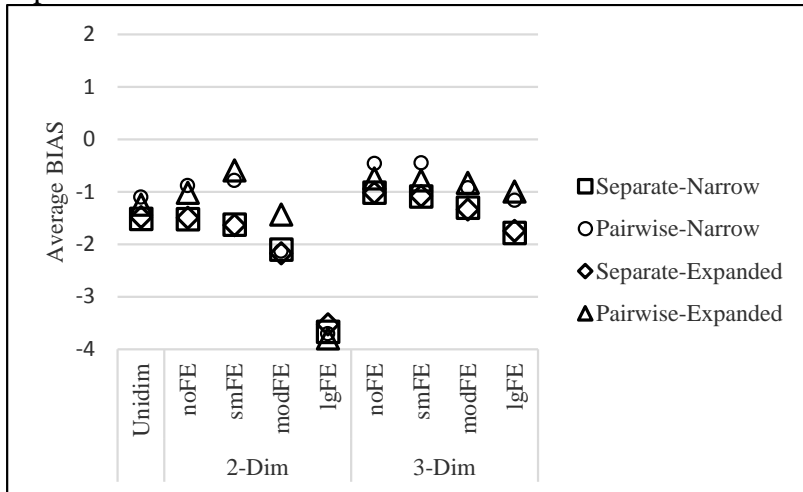
Figure 23. Grade 9 average RMSE for large grade level separation

122
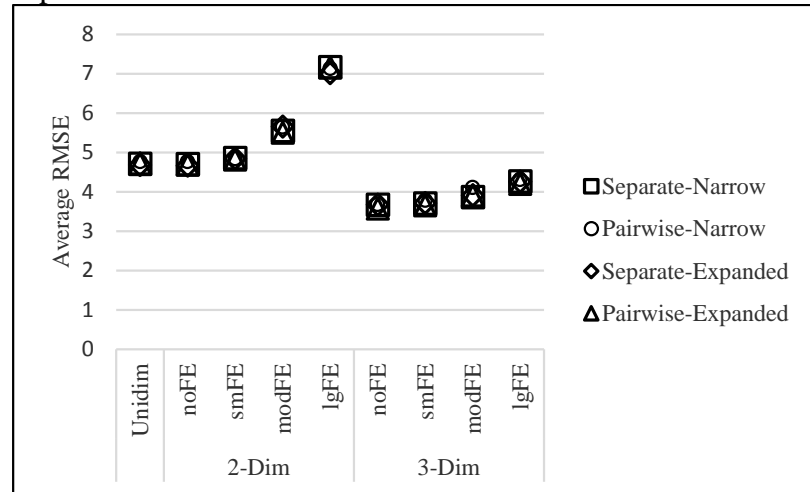
4.1.5.2 ANOVA Results

Again, two separate five-way mixed ANOVAs using BIAS and RMSE values as the dependent variable were performed on the results of the 100 replications for this grade. Most interactions and main effects for this grade level were significant for both the BIAS and RMSE models (Table B9 and B10). Pairwise comparisons were examined, effect sizes were calculated, and the difference matrix produced was used to examine relevant comparisons between vertical scaling configurations by condition to determine the scaling method(s) that resulted in the most accurate vertical scales for each condition.

Item Format Effect

Significant differences in item format effect condition were observed for both data generation models and both grade level separation conditions (Table C21). Effect sizes for mean BIAS values ranged from 0.23 to 1.33 while effect sizes for mean RMSE values ranged from 0.39 to 1.82. Mean BIAS values were only significantly different for the large grade level separation conditions only regardless of data generation model. For separate calibration, regardless of model, the large item format effect condition was significantly larger than the no and small item format effect conditions. For pairwise concurrent calibration, the large item format effect condition was significantly larger than the small and no item format effect condition and the moderate item format effect condition was significantly larger than the no and small item format effect condition, regardless of model, for the narrow common item set configuration. Additionally, under the 2-dimensional model for the pairwise concurrent calibration with expanded common item set, the large item format effect condition was significantly larger than the moderate, small, and no item format effect conditions and the moderate item format effect condition was significantly larger than the small and no item format effect conditions.

Under the 2-dimensional model, mean RMSE values for the large item format effect condition was significantly larger than those for the moderate, small, and no item format effect conditions and the moderate item format effect condition was significantly larger than the small and no item format effect conditions regardless of any other condition. However, under the 3-dimensional model, there was no significant difference among item format effect conditions for the small grade level separation, but for the large grade level separation, the large item format effect condition was significantly larger than the no and small item format effect conditions regardless of vertical scaling method.

Vertical Scaling Method

There was no significant difference in mean BIAS or mean RMSE values between calibration methods except for the small item format effect condition for the small grade level separation under the 2-dimensional data generation model (Table C22). Effect sizes for these significant results were small and ranged from 0.37 to 0.43.

Common Item Set Configuration

There was no significant difference in mean BIAS or mean RMSE values between common item set configurations regardless of item format effect, vertical scaling method, grade level separation, and data generation method (Table C23).

Grade Level Separation

Differences in grade level separation were dependent on vertical scaling method and data generation method (Table C24). Effect sizes for significant mean BIAS results ranged from 0.33 to 1.41. There was a significant difference in mean BIAS values between grade level separation conditions for separate calibration regardless of data generation model, item format effect condition, and common item set configuration. Additionally, for pairwise concurrent calibration,

mean BIAS values were significantly larger for the large grade level separation condition for the large item format effect conditions regardless of data generation model or common item set configuration. Additionally, mean BIAS values for the large grade level separation condition for the moderate item format effect condition for the 2-dimensional data generation model were significantly larger.

Mean RMSE values were significantly larger for the larger grade level separation condition under separate calibration regardless of data generation model, item format effect condition or common item set configuration. For pairwise concurrent calibration, mean RMSE values for the large grade level separation condition were significantly larger than those for the small grade level separation condition regardless of item format effect and common item set configuration, but only for the 2-dimensional and unidimensional data generation models. Effect sizes for these significant results ranged from 0.36 to 0.85.

Data Generation Method

Differences in mean BIAS values for data generation model were dependent on grade level separation (Table C25). While there was no significant difference in mean BIAS values between data generation models for the small grade level separation, data generated under the 2-dimensional model resulted in significantly larger values for both the moderate and large item format effect conditions for the large grade level separation. Effect sizes for these results ranged from 1.24 to 2.18. Mean RMSE values were significantly larger under the 2-dimensional data generation model for all item format effect conditions regardless of common item set and scaling method. Additionally, the 3-dimensional data generation model produced significantly smaller mean values compared to the unidimensional model regardless of condition. Effect sizes for these values ranged from 0.40 to 2.18.

4.1.5.3 Performance of Vertical Scaling Configurations

The summary table produced from the pairwise comparisons matrix shows that significant mean BIAS differences existed between most vertical scaling configurations for the large grade level separation with pairwise concurrent calibration resulting in smaller values (Table 15). In addition, although not significant, under the 3-dimensional model, mean BIAS values were smallest when using pairwise concurrent calibration with the small grade level separation. However, mean BIAS values were smallest when using separate calibration under the traditional data generation model with the small grade level separation. Again, although not significant, mean BIAS and mean RMSE values were smallest when using the expanded common item set regardless of item format effect, scaling method, and data generation model.

Table 15. Grade 9 vertical scaling methods by accuracy

	Data Generation	Format Effect	Small Separation	Large Separation
	Unidim	None	se=pe=sn=pn	pn=pe=se=sn
		None	sn=pn=se=pe	pn=pe<se=sn
	2-Dim	Small	sn=pn=se=pe	pe=pn<sn=se
		Moderate	sn=se=pn=pe	pe=sn=pn=se
BIAS		Large	se=sn=pe=pn	se=sn=pn=pe
		None	pn=pe=sn=se	pn=pe<sn=se
	3-Dim	Small	pe=pn=se=sn	pn=pe<se=sn
		Moderate	pn=sn=pe=se	pn=pe<se=sn
		Large	pn=se=sn=pe	pn=pe<se=sn
	Unidim	None	sn=se=pe=pn	se=sn=pe=pn
		None	sn=se=pn=pe	se=pe=sn=pn
	2-Dim	Small	se=sn=pn=pe	pe=pn=se=sn
		Moderate	pe=pn=se=sn	pe=sn=pn=se
RMSE		Large	pn=pe=se=sn	se=pn=pe=sn
		None	pe=pn=sn=se	pe=sn=se=pn
	3-Dim	Small	pe=pn=se=sn	pe=sn=se=pn
		Moderate	pe=pn=se=sn	sn=se=pe=pn
		Large	pe=pn=se=sn	pe=se=sn=pn

sn=separate - narrow; se=separate - expanded; pn=pairwise – narrow; pe=pairwise – expanded
< significantly smaller   = not significantly different

4.1.5.4 Summary of Grade 9 Results

Based on the ANOVA results, item format effect and grade level separation had a statistically significant impact on the vertical scaling results for this grade level. Common item set and vertical scaling method did not have an impact on the scaling results. The largest mean RMSE values resulted from the two largest item format effect conditions under traditional data generation models regardless of common item set, grade level separation, or scaling method. The same was true for mean BIAS values, but only for the large grade level separation. Additionally, the large grade level separation condition resulted in larger mean BIAS and mean RMSE values for all item format effect conditions although only the largest item format effect conditions were impacted when using pairwise concurrent calibration. Opposite of the phenomenon observed for grades 5 and 6, a large proportion of replications for the no and small item format effect conditions resulted in positive mean BIAS values compared to the relatively high proportion of negative mean BIAS values for the moderate and large item format effect conditions. This resulted in spuriously small average values for the no and small item format effect conditions. Again, mean RMSE values were not impacted by this issue and showed an expected pattern and magnitude.

Data generation method had the largest significant impact on scaling results for the large grade level separation, especially for the largest item format effect conditions.

4.1.6   Grade 10 Results

4.1.6.1 Trends by Grade Level Separation

127

Average BIAS

Mean BIAS values generally increased as degree of item format effect increased for separate calibration under the large grade level separation condition regardless of common item set and vertical scaling method (Table 16). However, mean values were similar across all conditions under the small grade level separation condition for data generated under the 3-dimensional model. Again, for this grade level under pairwise concurrent calibration, mean BIAS values fell into the same two or three 'distributions'. The same pattern was observed with smaller item format effect conditions producing distributions with large proportions of both negative and positive BIAS values with proportions of the negative BIAS values increasing as item format effect increased. Again, due to the presence of the more neutral distribution for the 3-dimensional model, the larger item format effect conditions could still be spuriously high. As expected, average RMSE values were not impacted.

Pairwise concurrent calibration produced smaller average BIAS values compared to separate calibration and values resulting from both calibration methods were smaller for the small grade level separation condition compared to the large grade level separation condition regardless of common item set or data generation method. However, for the large grade level separation condition, mean BIAS values resulting from the 3-dimensional generation tended to be smaller than those resulting from the 2-dimensional generation while the opposite was true for the small grade level separation condition. Finally, the traditional unidimensional and essentially unidimensional condition under the 2-dimensional resulted in similar average values regardless of condition. However, values resulting from data generated under the 3-dimensional model tended to be larger for the small grade level separation and smaller for the large grade level separation than the other data generation models (Figures 24 and 25).

<u>Average RMSE</u>

As with grade 9, mean RMSE values generally increased as degree of item format effect increased, regardless of common item set, grade level separation, vertical scaling method, and data generation method for both grade level separation conditions. Mean RMSE values resulting from data generated under the 3-dimensional model were smaller than those resulting from data generated under the 2-dimensional model and values under the small grade level separation condition were smaller than those under the large grade level separation condition across common item sets, scaling methods, and data generation methods. Also, the essentially unidimensional values generated under the 3-dimensional model resulted in comparatively smaller average values compared to the average values generated under the traditional unidimensional and 2-dimensional essentially unidimensional conditions (Figures 26 and 27).

Table 16. Grade 10 mean BIAS and mean RMSE by condition

Scaling Method	Common Item Set	Model	Format Effect	Mean BIAS		Mean RMSE	
				Small Separation	Large Separation	Small Separation	Large Separation
Separate	Narrow	Unidim	None	-0.01	-2.38	4.14	4.96
		2-Dim	None	-0.05	-2.34	4.10	4.92
			Small	-0.01	-2.48	4.23	5.12
			Moderate	0.00	-2.95	4.96	5.83
			Large	0.26	-4.36	5.86	7.46
		3-Dim	None	-0.24	-1.46	3.51	3.58
			Small	-0.26	-1.51	3.54	3.61
			Moderate	-0.29	-1.85	3.73	3.81
			Large	-0.33	-2.50	3.98	4.32
	Expanded	Unidim	None	-0.01	-2.29	4.15	4.89
		2-Dim	None	-0.03	-2.32	4.15	4.93
			Small	-0.04	-2.43	4.20	5.08
			Moderate	-0.01	-3.11	4.91	6.02
			Large	0.26	-4.23	5.86	7.31
		3-Dim	None	-0.29	-1.46	3.54	3.56
			Small	-0.27	-1.55	3.55	3.63
			Moderate	-0.29	-1.92	3.74	3.88
			Large	-0.33	-2.44	3.93	4.26
Pairwise	Narrow	Unidim	None	-0.05	-1.35	3.93	4.48
		2-Dim	None	-0.09	-1.02	3.91	4.49
			Small	-0.06	-0.75	4.05	4.59
			Moderate	-0.03	-2.19	4.77	5.31
			Large	0.29	-3.27	5.67	6.52
		3-Dim	None	-0.22	-0.50	3.35	3.31
			Small	-0.24	-0.46	3.40	3.41
			Moderate	-0.25	-1.01	3.55	3.71
			Large	-0.26	-1.20	3.75	3.86
	Expanded	Unidim	None	-0.05	-1.28	3.93	4.41
		2-Dim	None	-0.06	-1.22	3.93	4.43
			Small	-0.07	-0.51	4.03	4.57
			Moderate	-0.02	-1.34	4.76	5.23
			Large	0.31	-3.38	5.70	6.57
		3-Dim	None	-0.23	-0.54	3.33	3.20
			Small	-0.25	-0.90	3.39	3.27
			Moderate	-0.25	-0.92	3.55	3.50
			Large	-0.29	-1.01	3.77	3.73

Figure 24. Grade 10 average BIAS for small grade level separation
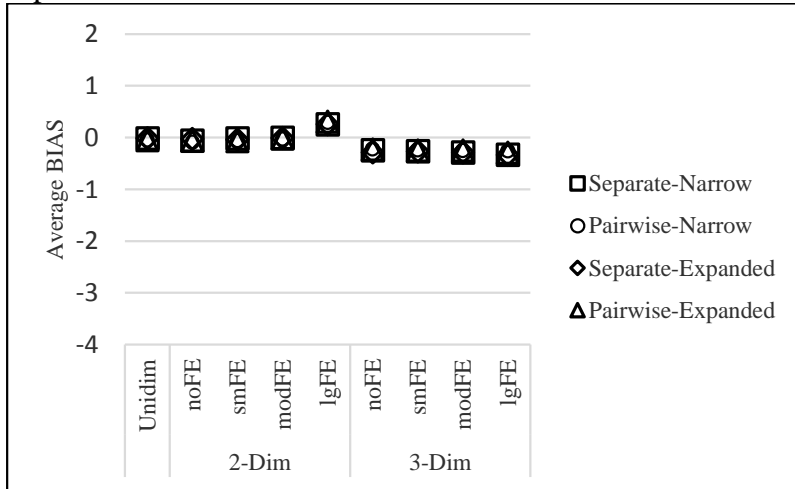
Figure 26. Grade 10 average RMSE for small grade level separation
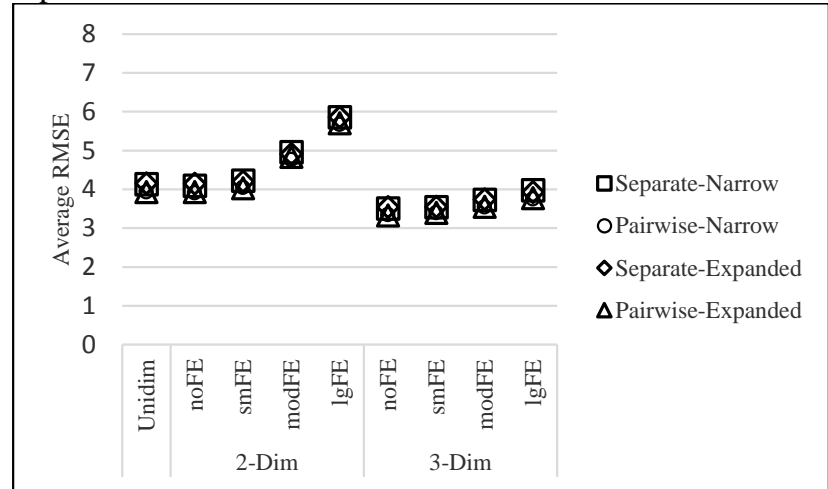
Figure 25. Grade 10 average BIAS for large grade level separation
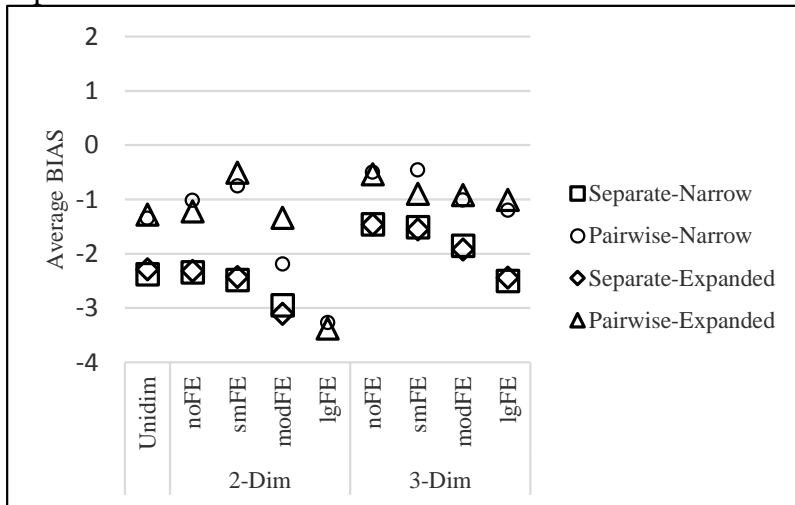
Figure 27. Grade 10 average RMSE for large grade level separation
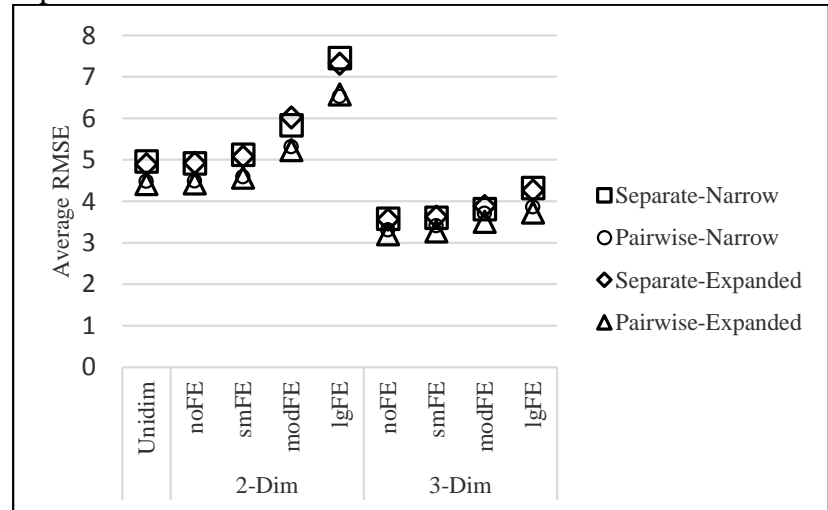
4.1.6.2 ANOVA Results

As with all previous grade levels, the two separate five-way mixed ANOVAs performed using BIAS and RMSE values as the dependent variable resulted in mostly significant interactions and main effects for both BIAS and RMSE models (Table B11 and B12). Again, pairwise comparisons were examined using the LSMEANS command in SAS and effect sizes were calculated. Again, the difference matrix produced was used to examine relevant comparisons between vertical scaling configurations.

Item Format Effect

Significant differences in mean BIAS and mean RMSE values were dependent upon grade level separation, common item set configuration, and data generation method (Table C26). Effect sizes for mean BIAS values ranged from 0.35 to 0.92 and effect sizes for mean RMSE values ranged from 0.30 to 1.88.

The only mean BIAS values that were significantly different under the 2-dimensional data generation model were those for the large grade level separation. Under this condition, mean BIAS values for the large item format effect were significantly larger compared to the no, small, and moderate item format effect conditions as well as the moderate item format effect compared to the small and no item format effect conditions for pairwise concurrent calibration with the narrow common item set configuration. There was no significant difference among item format effect conditions under the 3-dimensional data generation model for this grade level.

Mean RMSE values under the 2-dimensional data generation model were significantly larger for the large and moderate item format effect conditions compared to the no and small item format effect conditions as well as the large item format effect condition compared to the moderate item format effect condition regardless of vertical scaling method, common item set

132

configuration, and grade level separation. Under the 3-dimensional data generation model, significant differences were observed for the large grade level separation condition. Regardless of scaling method, the large item format effect condition was significantly larger than the no and small item format effect conditions and for separate calibration with the narrow common item set configuration, the large item format effect condition was also significantly larger than the moderate item format effect condition.

Vertical Scaling Method

Significant differences in mean BIAS values for vertical scaling method were dependent on grade level separation (Table C27). Mean BIAS values were significantly larger for separate calibration regardless of common item configuration, item format effect, or data generation method for the large grade level separation condition only. Effect sizes for these significant results ranged from 0.24 to 0.62. Likewise, mean RMSE values were significantly larger for separate calibration regardless of common item configuration, item format effect for the large grade level separation; but only for the traditional data generation methods. Effect sizes for mean RMSE values were small and ranged from 0.35 to 0.43.

Common Item Set Configuration

There was no significant difference in mean BIAS or mean RMSE values between common item set configurations regardless of item format effect, vertical scaling method, grade level separation, and data generation method (Table C28).

Grade Level Separation

Significant differences in mean BIAS values between grade level separation conditions were dependent on vertical scaling method (Table C29). Regardless of item format effect, common item set, and data generation method, the large grade level separation condition produced larger

mean BIAS values for the separate calibration condition. For pairwise concurrent calibration under the 2-dimensional data generation model, mean BIAS values for the large grade level separation condition were larger for the moderate and large item format effect conditions regardless of common item set configuration as well as for the unidimensional data generation model regardless of vertical scaling method. There was no significant difference between grade level separation conditions under the 3-dimensional generation model for pairwise concurrent calibration. Effect sizes for significant mean BIAS results ranged from 0.37 to 1.49.

Mean RMSE values were not significantly different between grade level separation conditions under the 3-dimensional data model. However, under the unidimensional and 2-dimensional data generation model, mean RMSE values for the large grade level separation condition were significantly larger than those for the small grade level separation condition for all conditions of item format effect, vertical scaling method, and common item set configurations. Effect sozes for these significant values ranged from 0.35 to 1.18.

Data Generation Method

Mean BIAS values were significantly larger for the 2-dimensional data generation model compared to the 3-dimensional data generation model only for the large grade level separation condition for the moderate and large item format effect conditions regardless of common item set and vertical scaling method (Table C30). Effect sizes for these significant results ranged from 0.35 to 0.76. On the other hand, mean RMSE values were significantly larger for the 2-dimensional data generation model compared to the 3-dimensional data generation model for all item format effect conditions regardless of common item set, grade level separation, and vertical scaling method. Additionally, the unidimensional model produced larger mean RMSE values

than the 3-dimensional model, but similar mean RMSE values compared to the 2-dimensional model. Effect sizes for these results ranged from 0.41 to 2.32.

4.1.6.3 Performance of Vertical Scaling Configurations

Using the pairwise comparisons matrix, a summary table was created to examine each vertical scaling configuration (Table 17). Differences in vertical scaling configurations existed for the large grade level separation with pairwise concurrent calibration producing significantly smaller mean BIAS and mean RMSE values. With the exception of the data generated under the 2-dimensional model, pairwise concurrent calibration produced smaller, but not always significant, mean BIAS values. Also, although not significant, the expanded common item set configuration consistently produced smaller mean RMSE values under the large grade level separation condition. However, regardless of grade level separation, the narrow common item set usually produced the smallest mean BIAS values.

Table 17. Grade 10 vertical scaling methods by accuracy

	Data Generation	Format Effect	Small Separation	Large Separation
BIAS	Unidim	None	sn=se=pn=pe	pn=pe<se=sn
	2-Dim	None	se=sn=pe=pn	pn=pe<se=sn
		Small	sn=se=pn=pe	pe=pn<se=sn
		Moderate	sn=se=pn=pe	pe=pn <sn=se
		Large	se=sn=pe=pn	pn=pe<se=sn
	3-Dim	None	pn=pe=sn=se	pn=pe<sn=se
		Small	pn=pe=sn=se	pn<pe<sn=se
		Moderate	pn=pe=se=sn	pn<pe<sn=se
		Large	pn=pe=sn=se	pe=pn<se=sn
RMSE	None	None	pe=pn=sn=se	pe=pn=se=sn
	2-Dim	None	pn=pe=sn=se	pe=pn=sn=se
		Small	pe=pn=se=sn	pe=pn<se=sn
		Moderate	pn=pe=sn=se	pe=pn<sn=se
		Large	pn=pe=sn=se	pn=pe<se=sn
	3-Dim	None	pe=pn=sn=se	pe=pn=se=sn
		Small	pe=pn=sn=se	pe=pn=sn=se
		Moderate	pe=pn=se=sn	pe=pn=sn=se
		Large	pn=pe=se=sn	pe=pn<se=sn

sn=separate - narrow; se=separate - expanded; pn=pairwise – narrow; pe=pairwise - expanded
< significantly smaller   = not significantly different

4.1.6.4 Summary of Grade 10 Results

Based on the ANOVA results, item format effect, grade level separation, and vertical scaling method had a statistically significant impact on the vertical scaling results for this grade level. Common item set configuration did not have an impact on the resultant vertical scales. In general, the large grade level separation condition produced significantly larger mean BIAS and mean RMSE values, especially when using separate calibration and for the large item format condition. Additionally, the small and moderate item format effect conditions resulted in significantly larger mean RMSE values for both grade level separation conditions and under the traditional data generation method. As observed with grade 9, a large proportion of replications with positive mean BIAS values resulted in spuriously small values for both the no and small item format effect conditions under pairwise concurrent calibration regardless of data generation

model. Again, mean RMSE values did not exhibit this phenomenon and showed an expected pattern and magnitude.

Significant differences also existed between data generation methods with the 3-dimensional model generally resulting in smaller mean BIAS and mean RMSE values.

4.1.7   Overall Summary of Grade Level Results

Item format effect and grade level separation had a significant impact on all grade level results while vertical scaling method had a significant influence on scaling results for most grade levels. However, common item set did not significantly influence the scaling results for any grade level. Additionally, significant differences were observed between data generation methods.

Item Format Effects

Item format effect had a significant impact on the resultant vertical scales. In general, mean BIAS and mean RMSE values increased as item format effect increased for all grade levels, however, only the largest item format effect condition resulted in significantly larger mean values across grade levels. The moderate item format effect produced significantly larger mean values for several grade levels. Also, item format effect had more impact on mean RMSE values than on mean BIAS values. Item format effect was influenced by grade level separation, data generation model, and vertical scaling method.

Under the 2-dimensional data generation model for the small grade level separation regardless of scaling method, there was no significant difference among item format effect conditions for grades 8, 9, and 10, as well as grade 7 for separate calibration. For grades 5 and 6 for the small grade level separation for both scaling methods, the large item format effect condition had significantly larger mean BIAS values compared to all other format effect

137

conditions. In addition, for grade 5 under separate calibration the moderate item format effect was significantly larger than both the small and no item format effect conditions.

Under the 2-dimensional data generation model for the large grade level separation condition and separate calibration, there was no significant difference in mean BIAS values for grades 6 and 7. For grades 5, 8, 9 and 10 for the same configuration, the large item format effect was significantly larger than the small and no item format effect conditions. Additionally, for grades 5, 8, and 10, the large item format effect had significantly larger mean BIAS values than the moderate item format effect condition. For grade 8, the moderate item format effect had significantly larger mean BIAS values compared to the no and small item format effect conditions as well.

Under the 2-dimensional model, large grade level separation, and pairwise concurrent calibration, the large item format effect condition produced significantly larger mean BIAS values compared to the small and no item format effect conditions for grades 7, 8, 9, and 10. Additionally, for grades 9 and 10, the large item format effect condition had significantly larger values compared to the moderate item format effect condition. For grades 7 and 9, the moderate item format effect condition had significantly larger values than the small item format effect condition as well as for grade 10 with the narrow common item set configuration. Also, for grades 9 and 10 under the narrow common item set configuration, the moderate item format effect condition was significantly larger than the no item format effect condition.

Under the 3-dimensional data generation model, there was no significant difference in item format effect conditions for the small grade level condition for all grade levels regardless of scaling method. This was also true for the large grade level separation condition under pairwise concurrent calibration for grades 7, 8, and 10 and for separate calibration for grades 7 and 10.

138

For grades 5, 6, 8, and 9 under separate calibration, the large item format effect condition had significantly larger mean BIAS values than the no and small item format effect condition. Additionally, under separate calibration for grades 5, 8 and 9, the large item format effect condition had significantly larger mean values than the moderate item format effect condition and for grade 5 for the same configuration, the moderate item format effect condition had significantly larger mean values compared to the small and no item format effect conditions.

Under pairwise concurrent calibration, for grades 5, 8, and 9 with the narrow common item set configuration, the large item format effect condition was significantly larger than both the small and no item format effect conditions. In addition, for the grade 9 configuration, the moderate item format effect condition was significantly larger than the no and small item format effect conditions.

Under the 2-dimensional data generation model, mean RMSE values were significantly larger for the large item format effect compared to all other item format effect conditions and the moderate item format effect condition had significantly larger mean values compared to the no and small item format conditions for grades 7, 8, 9, and 10 regardless of common item set, grade level separation, and scaling method. The same pattern was observed for grade 5 except under pairwise concurrent calibration, the moderate item format effect condition was significantly higher for only the no item format condition. For grades 6 regardless of grade level separation and scaling method, the large item format effect condition had significantly larger values compared to the moderate, small, and no item format effect conditions although under the large grade level separation condition for pairwise concurrent calibration with the expanded common item set, the moderate item format effect condition had significantly larger values than the small and no item format effect conditions.

Under the 3-dimensional data generation model for the small grade level separation condition regardless of common item set or scaling method, there was no significant difference in item format effect conditions for grades 5, 6, 9, and 10.

For grade 8, regardless of grade level separation condition, common item set and scaling method, the large item format effect condition had significantly larger mean RMSE values compared to both the small and no item format effect conditions. For grade 7, regardless of grade level separation condition, common item set and scaling method, the larger item format effect condition had significantly larger mean RMSE than all other item format effect conditions and the moderate item format effect had significantly larger values than the no item format effect condition.

For grades 9 and 10 under the 3-dimensional model with the large grade level separation condition, the large item format effect also had significantly larger mean values compared to the no and small item format effect conditions and for grade 10 with the narrow common item set configuration, the mean RMSE value for the large item format effect condition was also significantly larger the moderate item format effect condition.

<u>Vertical Scaling Effects</u>

Across grade levels, calibration method had a significant impact on the vertical scaling results with pairwise concurrent calibration producing smaller mean BIAS and RMSE values. In addition, calibration method interacted with grade level separation condition. Specifically, mean BIAS values for grades 5, 8, and 10 for all item format effects for the large grade level separation condition were significantly smaller for pairwise concurrent calibration. In addition, mean BIAS values were significantly smaller for pairwise concurrent calibration for grade 5 under the 2-dimensional data generation model for the small, moderate, and large item format effect

conditions. Likewise, mean RMSE values for pairwise concurrent calibration were significantly smaller for grades 5, 8, and 10 for all item format effects for the large grade level separation. In addition, mean RMSE values were significantly smaller for pairwise concurrent calibration under the 2-dimensional data generation model for the large item format effect condition.

Common Item Set Effects

There was no significant impact of common item set configuration on the resultant vertical scales regardless of item format effect, vertical scaling method, grade level separation, or data generation model for any grade level.

Grade Level Separation Effects

In general, the large grade level separation produced significantly larger mean BIAS and mean RMSE values compared to the small grade level separation for each grade level. This was especially true when using separate calibration. While there was no significant difference in mean BIAS values between grade level separation conditions for the base grade level (grade 7), there was a significant difference in mean BIAS values for all item format effect conditions for all grade levels when using separate calibration which favored the small grade level separation condition. Additionally, the large grade level separation produced significantly larger mean BIAS values under pairwise concurrent calibration differing item format effect conditions for grades 5, 6, 9, and 10. For grade 6, this was true for all item format effect conditions for both data generation models; for grade 5 this was true for the unidimensional model and for all item format effect conditions for the 2-dimensional data generation model as well as the moderate and large item format effect conditions for the 3-dimensional data generation model; for grades 9 and 10 this was true for the unidimensional model as well as the moderate and large item format

141

effect conditions for the 2-dimensional data generation model as well as the large item format effect condition for grade 9 for the 3-dimensional data generation model.

There was no significant difference in mean RMSE values between grade level separation conditions for grade 7. There was, however, a significant difference in mean RMSE values between grade level separation conditions for the other grade levels depending on scaling method and data generation model. For grades 9 and 10, the small grade level condition produced smaller mean RMSE values compared to the large grade level separation condition for the undimensional and 2-dimensional data generation model for all item format effect conditions regardless of scaling method. For grade 8, this same pattern was true, but only for separate calibration. There was no significant difference in grade level separation conditions for pairwise concurrent calibration for this grade. This pattern was also true for grade 5, but only for the separate calibration condition as well as the 3-dimenisonal data generation model. For pairwise concurrent calibration for this grade level, significant differences were observed for the moderate and large item format effect conditions under the 2-diemsnional data generation model. For grade 6, the small grade level separation produced smaller mean RMSE values only for the large item format effect condition for the 2-dimensional model regardless of scaling method and for separate calibration under the 3-dimensional data generation model.

Data Generation Model Effects

Differences in results based on the method by which the data was generated were also observed. In general, including the vertical scale factor in the data generation model resulted in smaller mean values for most grade levels and item format effect conditions. However, the data generation model was influenced by grade level separation, scaling method, and item format effect condition.

Mean BIAS values were not significantly different among data generation models for the small grade level separation condition for grades 7, 8, 9, and 10. However, for grade 6 the 2-dimensional model produced larger mean BIAS values for the large item format effect condition. This was also true for grade 5, but the moderate item format effect condition also produced smaller mean BIAS values when using separate calibration. For the large grade level separation, the 3-dimensional model produced smaller mean BIAS values for both moderate and large item format effect conditions for grades 9 and 10; for grade 7 with pairwise concurrent calibration; for grade 5 with separate calibration; and for the large item format effect condition with separate calibration only for grade 6. For grade 8, under separate calibration, the 3-dimensional model produced smaller mean RMSE values for all item format effect conditions.

Mean RMSE values were significantly different between the 2-dimensional and the 3-dimensional models for all item format effect conditions regardless of scaling method and grade level separation, as well as between the unidimensional and 3-dimensional models for grades 7, 8, 9, and 10 as well as for the large grade level separation condition for grade 5. These significant differences favored the 3-dimensional model in all cases. For grade 6, regardless of grade level separation, as well as grade 5 for the small grade level separation condition; the 3-dimensional data generation model produced smaller mean RMSE values for the moderate and large item format effect conditions only. Also, there was no significant difference between the 3-dimensional essentially unidimensional and the unidimensional models unless noted.

5.0   SUMMARY AND DISCUSSION

This chapter provides a summary of the major findings and discusses the results of the study. It begins with a review of the study goals, followed by a review of the methodology used, and then discusses the major findings of the study by research question. It ends with a discussion of the limitations of this study and recommendations for future studies.

5.1.1   Review of Study Goals

The goal of this study was to investigate the impact of format effect on the performance of two popular methods of vertically scaling mixed item format tests, under plausible conditions of a real testing scenario. Secondary to this purpose was the examination of the influence of the common item set and grade level separation on these vertical scaling methods when scaling mixed item format tests. Third, ability data was generated under a traditional method in which the underlying vertical scale factor was implied and under a 3-dimensional method in which the underlying vertical scale factor was explicitly modeled. A simulation study was chosen because it allowed the systematic manipulation of format effect, common item set configuration, and grade level separation. Also, it allowed the comparison of observed results to a 'true' result such that absolute deviations could be explored.

5.1.2 Review of Study Methodology

In order to accomplish the goals of the study, six grade level tests consisting of 61 items each were generated with 90 percent of the items multiple-choice and 10 percent of the items constructed-response. This ratio is consistent with mixed item format tests currently in use. The multiple-choice items were generated under the 3PL model while the constructed-response items were generated under the GRM. Mean item difficulty for each grade level test was set slightly higher or lower (depending on whether the test was above or below the base grade level) than the small grade level separation mean ability level. Since the study was contextualized under the CINEG design, fourteen mixed format items (12 multiple-choice and 2 constructed-response) were designated as common items between adjacent grade levels. This number of common items is consistent with the recommendations for the minimum percentage of common items to use when vertical scaling tests. Two set of common items were chosen; one with a narrow range of item difficulties and one with an expanded range of item difficulties.

Ability distributions were generated for 2000 students per grade level. The mean ability for the general ability gradually increased at increments of 0.2 or 0.5 to simulate small and large grade level separation, respectively. These values were, in part, chosen so that the range of abilities across the span of 6 grade levels would be reasonable. Two different data generation methods were used; one in which the underlying vertical scale was implied but not modeled (2-dimensional) and one in which the underlying vertical scale was explicitly modeled (3-dimensional) along with a traditional unidimensional ability generation for comparison. During ability generation, the covariance structure was manipulated to simulate four degrees of format effect between a multiple-choice and a constructed-response factor.

Response files were then created for each of the 7 covariance conditions for each of the grade levels for both the small and large grade level separation conditions and true expected scores were calculated for each student across conditions and grade levels. Recovery of the expected score was chosen for evaluation purposes so that additional transformations would not be necessary to compare vertical scaling configurations. Item and ability parameters were estimated using the MLE proficiency estimator in MULTILOG. Then, STUIRT was used to calculate transformation constants for both separate and pairwise concurrent calibration for both common item set configurations. The Stocking-Lord linking method was used for all linkings and grade 7 (or 7-8 grade pair) was used as the base grade level for all configurations.

After expected scores were calculated for the observed parameters, BIAS and RMSE values were determined. Using expected score values as the dependent variables, a mixed ANOVA analysis was performed for each grade level and pairwise comparisons were calculated to examine differences in item format effect, calibration method, grade level separation, common item set configuration, and data generation model.

5.1.3   Major Findings by Research Question

This section describes the major results by research question.

Research Question #1: *Do item format effects impact the resultant vertical scale when scaling mixed format tests if unidimensionality is assumed?*

Evidence suggests that adding different item formats to a single test can increase the dimensional complexity of that test (Kim & Kolen, 2006; Yai, 2008). IRT, however, comes with the assumption that a test is measuring only one underlying dimension (Swaminathan & Hambleton, 1985). Fortunately, studies show that IRT is somewhat robust to violations of this assumption if

146

the dimensions being measured are highly correlated (e.g., Reckase, 1979; Dorans & Kingston, 1985). Also, if items of different formats on the same test measure the same content and skills (are highly correlated) a test can be considered 'essentially unidimensional' (Yao, 2008).

This study investigated the impact of item format effects between multiple-choice and constructed-response items represented by four different levels of correlation (0.95, 0.80, 0.50, 0.20). Results of the study were consistent with previous equating studies in that conditions with item format effects that were highly correlated resulted in expected scores for which the mean values were not significantly different from the unidimensional or essentially unidimensional conditions. However, item format effect often interacted with grade level separation, calibration method, and data generation model to influence scaling results for individual grade levels.

For the small grade level separation, significant differences in mean BIAS values were found for grades 5 and 6 regardless of scaling method and for grade 7 for pairwise concurrent calibration under the 2-dimensional model. For the large grade level separation, mean BIAS values were significantly larger for the larger item format effect conditions for grades 8, 9, and 10 regardless of scaling method. In addition, under the 3-dimensional model, significantly larger values were found for grades 5 and 6 regardless of scaling method and for grades 8 and 9 for pairwise concurrent calibration, but, again only for the largest item format effect conditions. Mean RMSE values were greatly impacted by item format effect condition. Under the 2-dimensional model regardless of grade level separation, significantly larger values for the largest two item format effect conditions were observed for all grade levels as well as for grades 8-10 for the large grade level separation under the 3-dimensional model. Also, this was true for grades 7-8 for the small grade level separation under the 3-dimensional model. Regardless, only the moderate and large item format effect conditions resulted in significantly different values across

147

grade levels. However, the small item format effect condition did not result in significantly higher mean BIAS or mean RMSE values compared to the no format effect condition for any grade level for any condition.

Differences in whether the large or the moderate and large item format effect resulted in significant differences could be attributed to the observation that as the item format effect increased the proportion of average BIAS values that were positive (underestimation of the observed expected score) increased. This pattern was evident in grades 8-10 as well as being particularly evident for grade 6. This shift in proportions resulted in an abnormally low average BIAS value for one particular item format effect condition for each grade level. For instance, for grade 6, it was the moderate item format effect that was close to 0 for some conditions and for grade 8, it was the large item format effect condition. When examining differences between item format effect conditions, this lead to small differences in many cases that were not statistically significant. Additionally, since grades 5 and 6 seemed to be the most adversely affected by this phenomenon, it raises the question of whether the choice of base grade level for the pair in pairwise concurrent calibration and/or the base grade level for the entire scale is influencing results. Mean RMSE values, however with the exception of grade 6, would not be as influenced by this estimation issue and, as a result, were more consistent in showing that item format effect condition was impacting the overall vertical scale regardless of scaling method, common item set configuration, and grade level separation.

Research Question #2: *Does the range in difficulty of the common items impact the resultant vertical scale when scaling mixed format tests in the presence of multidimensionality?*

The burden of the scaling process under the CINEG design rests on the set of common items used to link adjacent grade levels (Tong & Kolen, 2010). Evidence from horizontal equating studies suggests that common items should be statistically and content representative of the

148

whole test (Tong & Kolen, 2010). When considering mixed item format tests, evidence indicates that the common items should also be format representative (Li, Lissitz, & Yang, 1999; Bastardi, 2000; Kim & Lee, 2006). Since vertical scaling tests are expected to differ in difficulty as grade level increases, designing statistically representative common item sets between adjacent grade levels is challenging because it is not certain which test should be mirrored statistically to produce the most accurate scale (Peterson, 2010).

In this study, two sets of common item were designed to be format representative and simulate two types of statistical representativeness. Both sets of items were designed to include items from both adjacent grade level tests. The narrow set of common items included the most difficult items for the lower adjacent grade level and the easiest items for the upper adjacent grade level while the expanded set included items that covered a larger range of difficulty across the two adjacent tests. While previous studies have indicated differences in the performance of common item sets, this study resulted in no significant difference in performance between the sets of common items on scaling results for any grade level and common item set configuration rarely was significant in the overall ANOVA model.

Three issues related to the common item set configurations in this study are important to note. First, the common item sets were designed to be representative of the overall test both statistically and with respect to item format. However, the number of common items in each set was small because of the small number of test items. Between the narrow and expanded common item sets, eight of the fourteen common items were the same. This could ameliorate any common item set effect. Additionally, the number of constructed-response items within the common item sets was only two because of the proportion of multiple-choice to constructed-response items in the overall test. While this may not be unusual in an operational setting because of the

protectiveness surrounding these memorable items, it is probable that the small number of constructed-response items did not represent the item format effect suitably. Finally, while the number of common items reached the minimum recommendation for the separate calibration condition (Tong & Kolen, 2010), it did not reach that recommendation for format representativeness or proportion of items compared to total test items when calculating the linking constants for the final transformation of the pairwise concurrent calibration condition. At this point, the number of 'test items' was larger (108 as opposed to 61) due to the combining of the tests for concurrent calibration and the 14 common items was only 13% of the total number of items being linked between grade level pairs.

Research Question #3: Does the degree of separation in ability level between different grade levels impact the resultant vertical scale when scaling mixed format tests in the presence of multidimensionality when unidimensionality is assumed?

Studies examining differences in mean ability between groups within a horizontal equating or vertical scaling context consistently suggest that the larger the difference in mean ability between groups, the larger the bias in results (e.g. Hanson & Béguin, 1999; Chin, Kim & Nering, 2006). This is problematic since the expectation is that some difference in mean ability will exist between grade levels being vertically scaled (Chin, Kim & Nering, 2006). Additional evidence suggests that vertical scaling method can be differentially impacted by grade level separation (Chin, Kim, & Nering, 2006).

In this study two grade level separation conditions were investigated; small (0.2) and large (0.5). Consistent with other studies investigating differences in mean ability between groups being equated or scaled, grade level separation had a significant impact on the resultant vertical scales for most grade levels with greater bias present for a larger grade level separation.

150

However, grade level separation interacted with item format effect and calibration method to influence scaling results.

Overall, mean BIAS values for separate calibration appeared to be more sensitive to differences in separation between grade levels. On the other hand, mean BIAS values for pairwise concurrent calibration seemed to be differentially impacted by grade level separation. Grade levels below the base grade level (grades 5 and 6) showed significant differences between grade level separation conditions regardless of item format effect while only the largest item format effect conditions showed significant differences in grade level separation conditions for grade levels above the base grade level (grades 8, 9, and 10). On the other hand, all item format effect conditions showed significant differences in mean RMSE values between grade level separation conditions for grades 5, 8, 9, and 10 under separate calibration. In addition, this same pattern was also observed for grades 9 and 10 for pairwise concurrent calibration. However, unlike mean BIAS values, these grade levels are above the base grade level. For the grades below the base grade level grades (grades 5 and 6) only the largest item format effect conditions produced significant differences between grade level separation conditions. This was also true for the separate calibration condition for grade 6.

Important to note for this research question is the unusual distribution of average BIAS values resulting from pairwise concurrent calibration with the large grade level separation for grades 9 and 10. Depending on the data generation model, an examination of the average BIAS values from these conditions showed a two or three node discontinuous distribution. This phenomenon was not observed for separate calibration for these grades levels or for the other grade levels in the study for any conditions. The percentage of scores in the two or three nodes varied by item format effect, but generally was around 70/30 with the smaller percentage

151

observed for the more positive replications. Additionally, when three nodes were observed for some 3-dimensional conditions, the middle node was generally centered at zero and represented a larger percentage of scores compared to the highest distributional node.

Two issues could be influencing these results. First, there was a mismatch between the mean ability for grades 9 and 10 and the mean difficulty for the grade level tests for the large grade level separation condition. This mismatch could have made the grade level test too 'easy' for the simulated students for these grade levels leading to what potentially could be a ceiling effect in this data and instability in estimating the observed expected score. Second, as previously mentioned, the proportion of common items may not have been sufficient to estimate the linking constants between the grade 9/10 pair and the grade 7/8 pair which could potentially lead to instability in calculating the final observed expected scores. This issue could differentially impact whether or not pairwise concurrent calibration is influenced by grade level separation. Additionally, given that pairwise concurrent calibration for grades 5 and 6 showed differential effects for scaling method brings into question whether or not the choice of base grade level for the pair or for the overall scale is impacting results for this calibration method.

Research Question #4: Which vertical scaling method (separate or pairwise concurrent) produces the most accurate vertical scale in the presence of multidimensionality when unidimensionality is assumed?

The performance of different vertical scaling methods has been widely researched. Studies investigating calibration methods for horizontal equating of mixed item format tests suggests that concurrent calibration outperforms separate calibration (Cao, 2008; Kim & Lee, 2006; Bastardi, 2002; Li, Lissitz, & Yang, 1999). However, concurrent calibration often suffers from convergence problems when a large number of grade levels is scaled (Kolen & Brennan, 2006). Although inconsistent, studies of vertical scaling mixed item format tests seem to indicate that

either separate calibration or pairwise concurrent calibration perform better than concurrent calibration (Karkee, et. al, 2003; Meng, 2008). Additionally, studies show that grade level separation (Chin, Kim, & Nering, 2006), multidimensionality (Beguin & Hanson, 2002; Cao, 2008), and proportion of dichotomously- to polytomously- scored items (Li, Lissitz, & Yang, 1999) can influence calibration method.

Given the number of grade levels being scaled and the deliberate introduction of multidimensionality by the presence of item format effects within each test, only separate and pairwise concurrent calibration were examined in this study. Since previous studies produced inconsistent results concerning the performance of separate versus pairwise concurrent calibration within the context of mixed item format tests, results were not necessarily predictable. For this study, scaling method had a significant impact on the resultant vertical scales. When the results were significant, pairwise concurrent calibration produced smaller mean BIAS and RMSE values. However, scaling method interacted with grade level separation and data generation model to influence results.

Mean BIAS and mean RMSE values for calibration method were not significantly different for the small grade level separation condition for grades 6-10 regardless of item format effect, common item set configuration, or data generation model. For grade 5, however, under the 2-dimensional data generation model, pairwise concurrent calibration produced significantly smaller mean BIAS values. On the other hand, scaling method was significantly impacted by the large grade level separation. For grades 5, 8, and 10, separate calibration produced significantly larger mean BIAS and mean RMSE values for all item format effect conditions regardless of common item set configuration and data generation model. Of particular interest is that the grade levels producing significant differences were the non-base grade level for the concurrent

153

calibration part of the pairwise concurrent calibration. This would be consistent with the hypothesis that a reduction in the number of grade level linkings leads to smaller bias in the overall vertical scale because these grade levels were not linked but concurrently calibrated with their grade level pair (Kim, Lee, & Kim, 2008; Smith, Finkelman, Nering, & Kim, 2008).

Research Question #5: _Does explicitly modeling the otherwise assumed vertical scale underlying the test data influence the simulated results?_

The simulation of item format effects for this study necessitated the generation of two different but correlated ability parameters for each student. The traditional way of generating data for a vertical scaling study would be to do just that; generate two different but correlated ability parameters for scaling. In two recent studies (Li & Lissitz, 2012; Koepfler, 2014), a bifactor model was used to generate and scale simulated data. In both cases, bias values were smaller for the multidimensional data compared to the unidimensional model used for comparison.

For this study, two different data generation models were investigated in addition to the unidimensional case. One data generation model modeled the item format effect only while the other modeled the item format effect as well as the overall vertical scale. Data generation method did have a significant impact on the results of this study. Explicitly modeling the underlying vertical scale often resulted in significantly smaller bias in the final scaling outcome. In addition, data generation method interacted with item format effect, grade level separation, and vertical scaling method to influence results. In general, as grade level separation and item format effect increased, the 3-dimensional data generation model produced smaller bias in the final vertical scale.

Mean BIAS values under the small grade level separation were not significantly different between data generation models for grades 7-10. In addition under the same condition, for grades 5 and 6, the only significant difference between data generation models was for the largest item

154

format effect conditions. Values for the large grade level separation were more likely to be impacted by data generation model, however, it was generally only the largest item format effect conditions that were affected.

Average mean RMSE values, however, were significantly impacted by data generation model regardless of condition. For the large grade level separation condition, the 3-dimensional model produced smaller mean RMSE values compared to the 2-dimensional and unidimensional models regardless of item format effect conditions for all grades except grade 6. For grade 6, however, only the moderate and large item format effect conditions had significantly smaller values for the 3-dimensional model. Likewise, under the small grade level separation condition, all grades showed significantly smaller values for the 3-dimensional model regardless of item format effect condition except for grades 5 and 6. For these grade levels, only the moderate and large item format effects showed significant differences.

It is interesting to note that differences in data generation model for mean BIAS values tended to mimic the results from this study concerning differences in item format effect, grade level separation, scaling method, and common item set configuration. Common item set configuration did not impact results. However, the large grade level separation condition was more likely to be impacted, but only for the largest item format effect conditions and where significant, pairwise concurrent calibration resulted in smaller bias values compared to separate calibration.

Explicitly modeling the vertical scale (3-dimensional model) as opposed to just implying its existence (2-dimensional model) resulted in smaller mean RMSE values. This perhaps is not surprising since mean RMSE values speak specifically to the recovery of the underlying vertical scale. A bit unexpected, however, was the significant difference in mean RMSE values between

data generation models for the unidimensional case. While not the same scenario, a similar result is reported by Li & Lissitz (2012) in that data generated and recovered under a multidimensional model had smaller bias compared to unidimensional data. One potential issue was the choice to have the mean ability level for all factors be equal as opposed to the mean ability for the item format effect factors be 0 with only the mean ability for the vertical scale factor changing to reflect the increasing level of ability associated with the vertical scale.

5.1.4   Additional Comments

The results discussed above raised several questions about the choice of base grade level and choice of mean ability level for the factors in the 3-dimensional model. To see if these choices were differentially impacting the results of the study, three 'mini-investigations' were conducted.

5.1.4.1 Base Grade Level

Published evidence concerning choice of base grade level in vertical scaling is limited and inconsistent and there are no published studies examining the impact of base grade level in vertically scaling mixed item format tests. Of the four studies investigating vertically scaling mixed format tests, three used the lowest grade level as the base grade level while the fourth used the middle grade level (see Table 1). Additionally, when examining pairwise concurrent calibration, a grade level for each of the grade level pairs for the concurrent estimation must be chosen. For the three studies examining pairwise concurrent calibration, all three used the lowest grade level as the base grade level for the concurrent pair estimation.

In this study, because grade 7 was chosen as the base grade level for the overall vertical scale, grade 6 was chosen as the base grade level for the grade 5/6 pair because it was closer to

156

grade 7. For grades 5 and 6, however, replications tended to increase in proportion of underestimated mean values as item format effect increased. This trend towards increasingly underestimated mean values for each replication observed in grades 5 and 6 was also observed for grade 8, but to a lesser degree, with grade 6 being the grade level most adversely impacted by the observed phenomenon.

To investigate if this choice of base grade level for the 5/6 pair was influential in the observed results, thirty replications were performed in which the base grade level for the pair was changed to grade 5. This was done for the 2-dimensional data generation model under the large grade level separation with the narrow common item set because the phenomenon tended to be more pronounced for the large grade level separation and for the 2-dimensional model generation. Results of this mini-investigation confirmed that the choice of base grade level was differentially impacting the estimation of mean BIAS and mean RMSE values for this pair of grade levels. While the new base grade level resulted in greatly increased mean BIAS and mean RMSE values for both grade levels, the mean BIAS value for each replication was consistently negative meaning the scores were consistently underestimated across both grade levels.

Since changing the base grade level for the grade 5/6 pair impacted the results and because the mean values showed a large increase in bias for the below base grade level pair, the choice of base grade level for the entire scale was questioned especially in relationship to pairwise concurrent calibration. Therefore, a second mini-investigation in which the base grade level of the overall vertical scale was changed from grade 7 to grade 5 was conducted. This investigation used the 2-dimensional model for the small grade level separation with the narrow common item set. Results showed a predictable gradual increase in mean BIAS and mean RMSE value as grade level increased. Compared to the mean values resulting from grade 7 as the base

grade level, grades 5 and 6 showed decreased mean BIAS and mean RMSE values while grades 7-10 showed increased mean BIAS and mean RMSE values, especially for grades 8-10. This is consistent with previous studies that hypothesized that bias in the vertical scale was influenced by the number of linkings between each grade level and the base grade level (Kim, Lee, & Kim, 2010). Additionally, overall mean values for each replication became increasingly positive as grade level increased rather than increasingly positive across item format effects within grade level as observed in the current study.

5.1.4.2 Three-Dimensional Data Generation Model

One design decision that could have impacted the results regarding the 3-dimensional model, and the focus of the third mini-investigation, was the choice to keep the mean ability for the item format factors the same as the mean for the vertical scale factor. To conduct this investigation, the 3-dimensional data generation model under pairwise concurrent calibration for the large grade level separation with the narrow common item set was used and the mean ability parameters for the item format effect conditions were changed to 0 from their respective values. Base grade level of the overall vertical scale and base grade level for the grade 5/6 pair was the same as in the original study.

Results of this mini-investigation showed a reduction in the number of replications resulting in a positive mean BIAS value for grades 9 and 10, although the largest item format effect conditions for both grade levels still had a substantial proportion of positive replications. Additionally, the mean BIAS increased for grades 5, 6, 8, 9, and 10 while it stayed the same for grade 7. While mean RMSE values still increased with increasing item format effect conditions for each grade level, overall the mean values for each item format effect condition were smaller across grade levels and remarkably similar across grades 5, 6, 9, and 10. While how this was

interacting with grade level tests that had an in increase in mean difficulty level is unclear, these results suggest that choice of mean ability level for the item format effect factors could be impacting the results and may be able to ameliorate the multiple separate distributions seen in the results for some conditions in this study. Since the mean RMSE values were even smaller for this mini-investigation, it does not appear that this was impacting the significant difference between the 3-dimensional model and the unidimensional model observed in this study.

Two other design decisions that could have impacted the results but were not the subject of 'mini-investigations' were the propagation of the item format effect correlations throughout the covariance matrices as well as the use of 1 as the discrimination parameter for all items for the item format effect factors.

5.1.5 Limitations and Suggestion for Further Study

While considering the results for this study a few issues should be kept in mind. First, choices were made about which variables to manipulate and the levels of each variable to use based on plausible, or when possible, vertical scaling contexts commonly used in practice. Grade level tests were created using one of the two most common ratios of multiple-choice to constructed-response items with the number of common items slightly above the recommended minimum percentage of items. In addition, the format of the common items matched the format and ratio of item type for the overall test. However, this resulted in only 2 constructed-response items being used in the common item set. While this may be advantageous in actual practice since constructed-response items tend to be memorable and, from a test security perspective, the fewer of those items used in the common item set the better; it does raise the question of how format representative does the common item set need to be to detect a format effect. It could be that a

159

larger percentage of constructed-response items need to be used in the common item set so that the format effect is adequately represented in the linking items.

Second, mean item difficulty for each grade level test was designed to mimic the mean ability for each successive grade. Since two grade level separations were investigated, the mean item difficulty for each test is necessarily closer to one grade level separation configuration than the other. This ultimately resulted in a poor match between the mean ability of the grade level and the mean item difficulty for the test, as well as the range of item difficulty for the common items, for those grade levels farthest from the base grade level under the large grade level separation condition.

Third, simulation studies are important because they allow the direct manipulation of variables in specific ways and the comparison of the observed results to a true result. However, the situations are often contrived and even when great effort is made to be as realistic as possible, real world contexts are often more complex. For instance, item format effect was held constant across grade levels when in practice dimensional complexities most likely do not occur this way. It is much more likely that format effect is minimal between some grade levels and larger between other grade levels. This is most likely the case for grade level separation as well. For instance, it is highly likely that grade level separation is larger for lower grade levels and smaller for upper grade levels rather than constant across grade levels.

Fourth, for simplicity's sake, many variables suggested by other studies to possibly produce inconsistencies in vertical scaling results were held constant, namely; length of test, proportion of multiple-choice to constructed-response items, base grade level, number of grade levels being scaled, proficiency estimator, and construct shift. Whether results would change if the test were longer and/or the proportion of multiple-choice to constructed-response items were

160

8:2 instead of 9:1 and/or if there is a maximum number of grade levels that should be scaled are all legitimate questions that need to be explored. Also, changing the base grade from the central grade level to the lowest grade level, without changing the increasing mean ability for each grade, would also remove the additional potential confound of the base grade level having a mean ability of zero and/or changing the base grade level for the grade level pairs when using pairwise concurrent calibration would be worthy of further investigation. Preliminary results of two mini-investigations in this study suggest that both of these base grade level issues could differentially impact resultant vertical scales.

Finally, when constructing the mean vectors for the 3-dimensional model, the decision was made to keep all three factor means for a grade level the same. The other option would be to increase the mean of the vertical scale factor only and keep the mean of the item format factors for each grade level as zero. The mini-investigation conducted in this study seemed to suggest that this could impact scaling results. Also, for simplicity it was decided to use a constant of 1 for the item discrimination parameters for the item format effect factors. This may or may not be a realistic representation of the relationship between the vertical scale factor and the item format effect factors for a given test. Finally, the item format effect correlations were propagated throughout the matrix to fill in the correlation between the vertical scale factor and the item format factors. This also may not be a realistic representation of this relationship. In fact, a factor analysis of one real data set showed a high correlation between the vertical scale factor and the multiple-choice factor and a much lower correlation between the vertical scale factor and the constructed-response factor. More investigation with real data sets could illuminate a more likely correlational structure.

5.1.6 Final Summary

Vertical scaling is a complex process in which different design decisions and practical issues lead to different resultant scales and, by extension, different conceptualizations of student growth. To date, empirical evidence is inconsistent and no one set of decisions has been established as best practice.

In this study, the performance of pairwise concurrent and separate calibration on vertically scaling mixed format tests was compared in the presence of item format effect using test specifications likely to exist in practice. While the results of the study provide evidence that item format effects, vertical scaling method, and separation between grade levels can significantly impact resultant vertical scales; the influence of these variables is often in combination with one another. While interactions sometimes made it difficult to draw generalizations for some grade levels, results seemed to indicate that: pairwise concurrent calibration holistically performed better compared to separate calibration; moderate to large item format effects were more likely to bias resultant vertical scales; and a large separation between grade levels resulted in more biased vertical scales.

Explicitly modeling the vertical scaling factor did not always impact mean BIAS values, but mean RMSE values were influenced by the inclusion of the vertical scale factor. Although, the same patterns of increased bias for separate calibration, a large grade level separation, and moderate or large item format effects were evident for both data generation models; significant differences were not as likely to occur under the 3-dimensional data generation model.

As with any other simulation, the extent to which the results of this study can be generalized to other contexts are limited to situations in which the conditions are similar to the ones used in the study. While the results of this study indicated that further research into

vertically scaling mixed format tests is clearly warranted, it did reiterate that test characteristics need to be examined prior to attempting to vertically scale mixed format tests under the assumption of unidimensionality.

APPENDIX A

Table A1. Item parameters for grade 5 – narrow CI set

Item Number	a1	a2	b	c	d1	d2	d3
1	1.46	1.00	-0.97	0.25			
2	1.69	1.00	0.06	0.25			
3	1.26	1.00	-0.52	0.25			
4	1.72	1.00	-1.34	0.25			
5	2.18	1.00	0.10	0.25			
6	1.85	1.00	-0.49	0.25			
7	1.80	1.00	-0.85	0.25			
8	1.31	1.00	-0.86	0.25			
9	2.00	1.00	-1.05	0.25			
10	1.44	1.00	-0.52	0.25			
11	1.71	1.00	-1.98	0.25			
12	2.05	1.00	-0.88	0.25			
13	1.52	1.00	-0.93	0.25			
14	1.21	1.00	-0.01	0.25			
15	1.94	1.00	-1.31	0.25			
16	2.10	1.00	-1.76	0.25			
17	1.81	1.00	-1.73	0.25			
18	1.83	1.00	-1.28	0.25			
19	1.79	1.00	-1.48	0.25			
20	2.00	1.00	-0.67	0.25			
21	1.90	1.00	-0.62	0.25			
22	1.55	1.00	-1.86	0.25			
23	2.12	1.00	-0.51	0.25			
24	2.15	1.00	-0.70	0.25			
25	1.53	1.00	-0.76	0.25			
26	1.95	1.00	0.17	0.25			
27	1.90	1.00	-1.01	0.25			
28	1.37	1.00	0.24	0.25			
29	1.53	1.00	0.31	0.25			
30	1.29	1.00	0.06	0.25			
31	1.25	1.00	-1.53	0.25			
32	1.39	1.00	-1.41	0.25			
33	1.52	1.00	-0.19	0.25			
34	1.96	1.00	-1.15	0.25			

Item Number	a1	a2	b	c	d1	d2	d3
35	1.72	1.00	-1.86	0.25			
36	2.18	1.00	-1.57	0.25			
37	1.22	1.00	-0.56	0.25			
38	1.61	1.00	-0.19	0.25			
39	1.31	1.00	-1.25	0.25			
40	2.08	1.00	-1.04	0.25			
41	1.42	1.00	-0.59	0.25			
42	1.35	1.00	-0.58	0.25			
43*	1.84	1.00	-0.23	0.25			
44*	1.38	1.00	-0.20	0.25			
45*	1.26	1.00	-0.26	0.25			
46*	1.43	1.00	-0.31	0.25			
47*	2.14	1.00	-0.25	0.25			
48*	1.83	1.00	-0.41	0.25			
49*	1.30	1.00	-0.46	0.25			
50*	2.00	1.00	-0.18	0.25			
51*	1.87	1.00	-0.16	0.25			
52*	1.81	1.00	-0.42	0.25			
53*	2.04	1.00	-0.44	0.25			
54*	1.50	1.00	-0.18	0.25			
55*	2.02	1.00			-1.85	-0.85	0.02
56*	1.61	1.00			-1.50	-0.75	-0.19
57	1.56	1.00			-1.57	-0.74	-0.33
58	1.52	1.00			-1.90	-0.78	-0.35
59	1.46	1.00			-1.83	-0.99	-0.20
60	1.96	1.00			-1.95	-0.96	0.17
61	1.84	1.00			-1.65	-0.81	-0.24

*Common items between grade 5 and grade 6

Table A2. Item parameters for grade 5 – expanded CI set

Item Number	a1	a2	b	c	d1	d2	d3
1	1.46	1.00	-0.97	0.25			
2	1.69	1.00	0.06	0.25			
3	1.26	1.00	-0.52	0.25			
4	1.72	1.00	-1.34	0.25			
5	2.18	1.00	0.10	0.25			
6	1.85	1.00	-0.49	0.25			
7	1.80	1.00	-0.85	0.25			
8	1.31	1.00	-0.86	0.25			
9	2.00	1.00	-1.05	0.25			
10	1.44	1.00	-0.52	0.25			
11	1.71	1.00	-1.98	0.25			
12	2.05	1.00	-0.88	0.25			
13	1.52	1.00	-0.93	0.25			
14	1.21	1.00	-0.01	0.25			
15	1.94	1.00	-1.31	0.25			
16	2.10	1.00	-1.76	0.25			
17	1.81	1.00	-1.73	0.25			
18	1.83	1.00	-1.28	0.25			
19	1.79	1.00	-1.48	0.25			
20	2.00	1.00	-0.67	0.25			
21	1.90	1.00	-0.62	0.25			
22	1.55	1.00	-1.86	0.25			
23	2.12	1.00	-0.51	0.25			
24	2.15	1.00	-0.70	0.25			
25	1.84	1.00	-0.23	0.25			
26	1.38	1.00	-0.20	0.25			
27	1.26	1.00	-0.26	0.25			
28	1.43	1.00	-0.31	0.25			
29	2.14	1.00	-0.25	0.25			
30	1.83	1.00	-0.41	0.25			
31	1.25	1.00	-1.53	0.25			
32	1.39	1.00	-1.41	0.25			
33	1.52	1.00	-0.19	0.25			
34	1.96	1.00	-1.15	0.25			
35	1.72	1.00	-1.86	0.25			
36	2.18	1.00	-1.57	0.25			
37	1.22	1.00	-0.56	0.25			
38	1.61	1.00	-0.19	0.25			
39	1.31	1.00	-1.25	0.25			
40	2.08	1.00	-1.04	0.25			
41	1.42	1.00	-0.59	0.25			
42	1.35	1.00	-0.58	0.25			
43*	1.53	1.00	-0.76	0.25			
44*	1.95	1.00	0.17	0.25			
45*	1.90	1.00	-1.01	0.25			
46*	1.37	1.00	0.24	0.25			
47*	1.53	1.00	0.31	0.25			
48*	1.29	1.00	0.06	0.25			
49*	1.30	1.00	-0.46	0.25			
50*	2.00	1.00	-0.18	0.25			
51*	1.87	1.00	-0.16	0.25			
52*	1.81	1.00	-0.42	0.25			
53*	2.04	1.00	-0.44	0.25			
54*	1.50	1.00	-0.18	0.25			
55*	2.02	1.00			-1.85	-0.85	0.02
56*	1.61	1.00			-1.50	-0.75	-0.19
57	1.56	1.00			-1.57	-0.74	-0.33
58	1.52	1.00			-1.90	-0.78	-0.35
59	1.46	1.00			-1.83	-0.99	-0.20
60	1.96	1.00			-1.95	-0.96	0.17
61	1.84	1.00			-1.65	-0.81	-0.24

*Common items between grade 5 and grade 6

Table A3. Item parameters for grade 6 – narrow CI set

Item Number	a1	a2	b	c	d1	d2	d3
1^	1.42	1.00	-0.19	0.25			
2^	1.44	1.00	-0.20	0.25			
3^	1.71	1.00	-0.18	0.25			
4^	2.00	1.00	-0.21	0.25			
5^	1.50	1.00	-0.02	0.25			
6^	2.04	1.00	-0.12	0.25			
7^	1.80	1.00	-0.01	0.25			
8^	1.52	1.00	-0.13	0.25			
9^	1.69	1.00	-0.12	0.25			
10^	1.94	1.00	-0.03	0.25			
11^	1.81	1.00	0.04	0.25			
12^	1.83	1.00	0.04	0.25			
13	2.18	1.00	-0.68	0.25			
14	2.18	1.00	0.33	0.25			
15	1.69	1.00	-1.04	0.25			
16	1.72	1.00	-1.35	0.25			
17	1.85	1.00	0.09	0.25			
18	1.31	1.00	-0.09	0.25			
19	2.05	1.00	0.11	0.25			
20	1.79	1.00	0.49	0.25			
21	2.00	1.00	-0.99	0.25			
22	1.55	1.00	-1.28	0.25			
23	2.12	1.00	-0.75	0.25			
24	2.15	1.00	-0.91	0.25			
25	1.53	1.00	-0.76	0.25			
26	1.95	1.00	0.17	0.25			
27	1.90	1.00	-1.01	0.25			
28	1.37	1.00	0.24	0.25			
29	1.53	1.00	0.31	0.25			
30	1.29	1.00	0.06	0.25			
31	1.25	1.00	0.36	0.25			
32	1.39	1.00	-0.53	0.25			
33	1.52	1.00	-0.62	0.25			
34	1.96	1.00	0.52	0.25			
35	1.72	1.00	-0.73	0.25			
36	2.18	1.00	0.44	0.25			
37	1.22	1.00	-0.79	0.25			
38	1.61	1.00	-1.45	0.25			
39	1.31	1.00	-0.84	0.25			
40	2.08	1.00	-0.14	0.25			
41	1.90	1.00	0.74	0.25			
42	1.35	1.00	0.44	0.25			
43*	1.84	1.00	-0.23	0.25			
44*	1.38	1.00	-0.20	0.25			
45*	1.26	1.00	-0.26	0.25			
46*	1.43	1.00	-0.31	0.25			
47*	2.14	1.00	-0.25	0.25			
48*	1.83	1.00	-0.41	0.25			
49*	1.30	1.00	-0.46	0.25			
50*	2.00	1.00	-0.18	0.25			
51*	1.87	1.00	-0.16	0.25			
52*	1.81	1.00	-0.42	0.25			
53*	2.04	1.00	-0.44	0.25			
54*	1.50	1.00	-0.18	0.25			
55*	2.02	1.00			-1.85	-0.85	0.02
56*	1.61	1.00			-1.50	-0.75	-0.19
57	1.46	1.00			-1.35	-0.59	0.76
58	1.96	1.00			-1.14	-0.30	0.21
59	1.84	1.00			-1.20	-0.28	0.21
60^	1.46	1.00			-0.95	-0.12	0.89
61^	1.52	1.00			-1.28	-0.66	0.68

*Common items between grade 5 and grade 6
^ Common items between grade 6 and grade 7

Table A4. Item parameters for grade 6 – expanded CI set

Item Number	a1	a2	b	c	d1	d2	d3
1^	1.42	1.00	-0.19	0.25			
2^	1.44	1.00	-0.20	0.25			
3^	1.71	1.00	-0.18	0.25			
4^	2.00	1.00	-0.21	0.25			
5^	1.50	1.00	-0.02	0.25			
6^	2.04	1.00	-0.12	0.25			
7^	1.25	1.00	0.36	0.25			
8^	1.39	1.00	-0.53	0.25			
9^	1.52	1.00	-0.62	0.25			
10^	1.96	1.00	0.52	0.25			
11^	1.72	1.00	-0.73	0.25			
12^	2.18	1.00	0.44	0.25			
13	2.18	1.00	-0.68	0.25			
14	2.18	1.00	0.33	0.25			
15	1.69	1.00	-1.04	0.25			
16	1.72	1.00	-1.35	0.25			
17	1.85	1.00	0.09	0.25			
18	1.31	1.00	-0.09	0.25			
19	2.05	1.00	0.11	0.25			
20	1.79	1.00	0.49	0.25			
21	2.00	1.00	-0.99	0.25			
22	1.55	1.00	-1.28	0.25			
23	2.12	1.00	-0.75	0.25			
24	2.15	1.00	-0.91	0.25			
25	1.84	1.00	-0.23	0.25			
26	1.38	1.00	-0.20	0.25			
27	1.26	1.00	-0.26	0.25			
28	1.43	1.00	-0.31	0.25			
29	2.14	1.00	-0.25	0.25			
30	1.83	1.00	-0.41	0.25			
31	1.80	1.00	-0.01	0.25			
32	1.52	1.00	-0.13	0.25			
33	1.69	1.00	-0.12	0.25			
34	1.94	1.00	-0.03	0.25			
35	1.81	1.00	0.04	0.25			
36	1.83	1.00	0.04	0.25			
37	1.22	1.00	-0.79	0.25			
38	1.61	1.00	-1.45	0.25			
39	1.31	1.00	-0.84	0.25			
40	2.08	1.00	-0.14	0.25			
41	1.90	1.00	0.74	0.25			
42	1.35	1.00	0.44	0.25			
43*	1.53	1.00	-0.76	0.25			
44*	1.95	1.00	0.17	0.25			
45*	1.90	1.00	-1.01	0.25			
46*	1.37	1.00	0.24	0.25			
47*	1.53	1.00	0.31	0.25			
48*	1.29	1.00	0.06	0.25			
49*	1.30	1.00	-0.46	0.25			
50*	2.00	1.00	-0.18	0.25			
51*	1.87	1.00	-0.16	0.25			
52*	1.81	1.00	-0.42	0.25			
53*	2.04	1.00	-0.44	0.25			
54*	1.50	1.00	-0.18	0.25			
55*	2.02	1.00			-1.85	-0.85	0.02
56*	1.61	1.00			-1.50	-0.75	-0.19
57	1.46	1.00			-1.35	-0.59	0.76
58	1.96	1.00			-1.14	-0.30	0.21
59	1.84	1.00			-1.20	-0.28	0.21
60^	1.46	1.00			-0.95	-0.12	0.89
61^	1.52	1.00			-1.28	-0.66	0.68

*Common items between grade 5 and grade 6
^ Common items between grade 6 and grade 7

Table A5. Item parameters for grade 7 – narrow CI set

Item Number	a1	a2	b	c	d1	d2	d3
1	1.42	1.00	-0.19	0.25			
2	1.44	1.00	-0.20	0.25			
3	1.71	1.00	-0.18	0.25			
4	2.00	1.00	-0.21	0.25			
5	1.50	1.00	-0.02	0.25			
6	2.04	1.00	-0.12	0.25			
7	1.80	1.00	-0.01	0.25			
8	1.52	1.00	-0.13	0.25			
9	1.69	1.00	-0.12	0.25			
10	1.94	1.00	-0.03	0.25			
11	1.81	1.00	0.04	0.25			
12	1.83	1.00	0.04	0.25			
13	1.26	1.00	0.24	0.25			
14	1.72	1.00	0.91	0.25			
15	1.84	1.00	0.57	0.25			
16	1.38	1.00	-0.76	0.25			
17	1.26	1.00	-0.81	0.25			
18	2.05	1.00	0.20	0.25			
19	1.30	1.00	0.43	0.25			
20	2.00	1.00	-0.69	0.25			
21	1.87	1.00	-0.67	0.25			
22	1.21	1.00	-0.12	0.25			
23	1.55	1.00	0.38	0.25			
24	2.15	1.00	-0.63	0.25			
25	1.53	1.00	-0.53	0.25			
26	1.95	1.00	0.38	0.25			
27	1.90	1.00	-0.28	0.25			
28	1.37	1.00	-0.16	0.25			
29	1.53	1.00	0.48	0.25			
30	1.29	1.00	0.67	0.25			
31	1.25	1.00	0.36	0.25			
32	1.39	1.00	-0.53	0.25			
33	1.52	1.00	-0.62	0.25			
34	1.96	1.00	0.52	0.25			
35	1.72	1.00	-0.73	0.25			
36	2.18	1.00	0.44	0.25			
37	1.22	1.00	0.26	0.25			
38	1.61	1.00	-0.33	0.25			
39	1.31	1.00	0.34	0.25			
40	2.08	1.00	-0.91	0.25			
41	1.42	1.00	0.25	0.25			
42	1.35	1.00	0.32	0.25			
43*	1.43	1.00	0.12	0.25			
44*	2.14	1.00	0.19	0.25			
45*	1.83	1.00	0.21	0.25			
46*	1.85	1.00	0.09	0.25			
47*	1.80	1.00	-0.01	0.25			
48*	1.31	1.00	-0.09	0.25			
49*	1.94	1.00	-0.03	0.25			
50*	2.10	1.00	-0.02	0.25			
51*	1.81	1.00	0.04	0.25			
52*	1.83	1.00	0.08	0.25			
53*	1.79	1.00	0.18	0.25			
54*	2.00	1.00	0.21	0.25			
55*	1.61	1.00			-0.65	0.37	0.96
56*	1.56	1.00			-1.06	-0.49	0.01
57	2.02	1.00			-0.59	0.17	0.94
58	1.96	1.00			-0.69	-0.32	0.69
59	1.84	1.00			-0.87	0.00	0.88
60	1.46	1.00			-0.95	-0.12	0.89
61	1.52	1.00			-1.28	-0.66	0.68

*Common items between grade 7 and grade 8

Table A6. Item parameters for grade 7 – expanded CI set

Item Number	a1	a2	b	c	d1	d2	d3
1	1.42	1.00	-0.19	0.25			
2	1.44	1.00	-0.20	0.25			
3	1.71	1.00	-0.18	0.25			
4	2.00	1.00	-0.21	0.25			
5	1.50	1.00	-0.02	0.25			
6	2.04	1.00	-0.12	0.25			
7	1.25	1.00	0.36	0.25			
8	1.39	1.00	-0.53	0.25			
9	1.52	1.00	-0.62	0.25			
10	1.96	1.00	0.52	0.25			
11	1.72	1.00	-0.73	0.25			
12	2.18	1.00	0.44	0.25			
13	1.26	1.00	0.24	0.25			
14	1.72	1.00	0.91	0.25			
15	1.84	1.00	0.57	0.25			
16	1.38	1.00	-0.76	0.25			
17	1.26	1.00	-0.81	0.25			
18	2.05	1.00	0.20	0.25			
19	1.30	1.00	0.43	0.25			
20	2.00	1.00	-0.69	0.25			
21	1.87	1.00	-0.67	0.25			
22	1.21	1.00	-0.12	0.25			
23	1.55	1.00	0.38	0.25			
24	2.15	1.00	-0.63	0.25			
25	1.43	1.00	0.12	0.25			
26	2.14	1.00	0.19	0.25			
27	1.83	1.00	0.21	0.25			
28	1.85	1.00	0.09	0.25			
29	1.80	1.00	-0.01	0.25			
30	1.31	1.00	-0.09	0.25			
31	1.80	1.00	-0.01	0.25			
32	1.52	1.00	-0.13	0.25			
33	1.69	1.00	-0.12	0.25			
34	1.94	1.00	-0.03	0.25			

Item Number	a1	a2	b	c	d1	d2	d3
35	1.81	1.00	0.04	0.25			
36	1.83	1.00	0.04	0.25			
37	1.22	1.00	0.26	0.25			
38	1.61	1.00	-0.33	0.25			
39	1.31	1.00	0.34	0.25			
40	2.08	1.00	-0.91	0.25			
41	1.42	1.00	0.25	0.25			
42	1.35	1.00	0.32	0.25			
43*	1.53	1.00	-0.53	0.25			
44*	1.95	1.00	0.38	0.25			
45*	1.90	1.00	-0.28	0.25			
46*	1.37	1.00	-0.16	0.25			
47*	1.53	1.00	0.48	0.25			
48*	1.29	1.00	0.67	0.25			
49*	1.94	1.00	-0.03	0.25			
50*	2.10	1.00	-0.02	0.25			
51*	1.81	1.00	0.04	0.25			
52*	1.83	1.00	0.08	0.25			
53*	1.79	1.00	0.18	0.25			
54*	2.00	1.00	0.21	0.25			
55*	1.61	1.00			-0.65	0.37	0.96
56*	1.56	1.00			-1.06	-0.49	0.01
57	2.02	1.00			-0.59	0.17	0.94
58	1.96	1.00			-0.69	-0.32	0.69
59	1.84	1.00			-0.87	0.00	0.88
60	1.46	1.00			-0.95	-0.12	0.89
61	1.52	1.00			-1.28	-0.66	0.68

*Common items between grade 7 and grade 8

Table A7. Item parameters for grade 8 – narrow CI set

Item Number	a1	a2	b	c	d1	d2	d3
1^	1.26	1.00	0.22	0.25			
2^	1.72	1.00	0.29	0.25			
3^	2.18	1.00	0.35	0.25			
4^	1.84	1.00	0.27	0.25			
5^	2.00	1.00	0.29	0.25			
6^	1.44	1.00	0.36	0.25			
7^	1.71	1.00	0.39	0.25			
8^	2.05	1.00	0.45	0.25			
9^	1.30	1.00	0.35	0.25			
10^	2.00	1.00	0.24	0.25			
11^	1.52	1.00	0.45	0.25			
12^	1.26	1.00	0.42	0.25			
13	1.46	1.00	0.51	0.25			
14	1.90	1.00	0.64	0.25			
15	1.21	1.00	0.89	0.25			
16	1.25	1.00	0.87	0.25			
17	1.90	1.00	-0.28	0.25			
18	1.37	1.00	-0.16	0.25			
19	1.87	1.00	-0.73	0.25			
20	1.81	1.00	1.13	0.25			
21	2.04	1.00	0.29	0.25			
22	1.50	1.00	0.15	0.25			
23	1.55	1.00	-0.57	0.25			
24	2.12	1.00	-0.63	0.25			
25	1.53	1.00	-0.53	0.25			
26	1.95	1.00	0.38	0.25			
27	1.90	1.00	-0.28	0.25			
28	1.37	1.00	-0.16	0.25			
29	1.53	1.00	0.48	0.25			
30	1.29	1.00	0.67	0.25			
31	1.69	1.00	0.08	0.25			
32	2.15	1.00	0.78	0.25			
33	1.39	1.00	0.49	0.25			
34	1.52	1.00	0.23	0.25			
35	1.96	1.00	1.38	0.25			
36	1.72	1.00	0.10	0.25			
37	2.18	1.00	1.42	0.25			
38	1.22	1.00	0.42	0.25			
39	1.31	1.00	-0.65	0.25			
40	2.08	1.00	-0.09	0.25			
41	1.42	1.00	0.62	0.25			
42	1.35	1.00	-0.65	0.25			
43*	1.43	1.00	0.12	0.25			
44*	2.14	1.00	0.19	0.25			
45*	1.83	1.00	0.21	0.25			
46*	1.85	1.00	0.09	0.25			
47*	1.80	1.00	-0.01	0.25			
48*	1.31	1.00	-0.09	0.25			
49*	1.94	1.00	-0.03	0.25			
50*	2.10	1.00	-0.02	0.25			
51*	1.81	1.00	0.04	0.25			
52*	1.83	1.00	0.08	0.25			
53*	1.79	1.00	0.18	0.25			
54*	2.00	1.00	0.21	0.25			
55*	1.61	1.00			-0.65	0.37	0.96
56*	1.56	1.00			-1.06	-0.49	0.01
57	2.02	1.00			-0.44	0.56	1.04
58	1.52	1.00			-0.53	0.46	0.86
59	1.84	1.00			-0.41	0.59	0.99
60^	1.46	1.00			-0.38	0.31	1.03
61^	1.96	1.00			-0.40	0.17	1.09

*Common items between grade 7 and grade 8
^ Common items between grade 8 and grade 9

171

Table A8. Item parameters for grade 8 – expanded CI set

Item Number	a1	a2	b	c	d1	d2	d3
1^	1.46	1.00	0.51	0.25			
2^	1.90	1.00	0.64	0.25			
3^	1.21	1.00	0.89	0.25			
4^	1.25	1.00	0.87	0.25			
5^	1.90	1.00	-0.28	0.25			
6^	1.37	1.00	-0.16	0.25			
7^	1.71	1.00	0.39	0.25			
8^	2.05	1.00	0.45	0.25			
9^	1.30	1.00	0.35	0.25			
10^	2.00	1.00	0.24	0.25			
11^	1.52	1.00	0.45	0.25			
12^	1.26	1.00	0.42	0.25			
13	1.26	1.00	0.22	0.25			
14	1.72	1.00	0.29	0.25			
15	2.18	1.00	0.35	0.25			
16	1.84	1.00	0.27	0.25			
17	2.00	1.00	0.29	0.25			
18	1.44	1.00	0.36	0.25			
19	1.87	1.00	-0.73	0.25			
20	1.81	1.00	1.13	0.25			
21	2.04	1.00	0.29	0.25			
22	1.50	1.00	0.15	0.25			
23	1.55	1.00	-0.57	0.25			
24	2.12	1.00	-0.63	0.25			
25	1.43	1.00	0.12	0.25			
26	2.14	1.00	0.19	0.25			
27	1.83	1.00	0.21	0.25			
28	1.85	1.00	0.09	0.25			
29	1.80	1.00	-0.01	0.25			
30	1.31	1.00	-0.09	0.25			
31	1.69	1.00	0.08	0.25			
32	2.15	1.00	0.78	0.25			
33	1.39	1.00	0.49	0.25			
34	1.52	1.00	0.23	0.25			
35	1.96	1.00	1.38	0.25			
36	1.72	1.00	0.10	0.25			
37	2.18	1.00	1.42	0.25			
38	1.22	1.00	0.42	0.25			
39	1.31	1.00	-0.65	0.25			
40	2.08	1.00	-0.09	0.25			
41	1.42	1.00	0.62	0.25			
42	1.35	1.00	-0.65	0.25			
43*	1.53	1.00	-0.53	0.25			
44*	1.95	1.00	0.38	0.25			
45*	1.90	1.00	-0.28	0.25			
46*	1.37	1.00	-0.16	0.25			
47*	1.53	1.00	0.48	0.25			
48*	1.29	1.00	0.67	0.25			
49*	1.94	1.00	-0.03	0.25			
50*	2.10	1.00	-0.02	0.25			
51*	1.81	1.00	0.04	0.25			
52*	1.83	1.00	0.08	0.25			
53*	1.79	1.00	0.18	0.25			
54*	2.00	1.00	0.21	0.25			
55*	1.61	1.00			-0.65	0.37	0.96
56*	1.56	1.00			-1.06	-0.49	0.01
57	2.02	1.00			-0.44	0.56	1.04
58	1.52	1.00			-0.53	0.46	0.86
59	1.84	1.00			-0.41	0.59	0.99
60^	1.46	1.00			-0.38	0.31	1.03
61^	1.96	1.00			-0.40	0.17	1.09

*Common items between grade 7 and grade 8
^ Common items between grade 8 and grade 9

Table A 9. Item parameters for grade 9 – narrow CI set

Item Number	a1	a2	b	c	d1	d2	d3
1^	1.26	1.00	0.22	0.25			
2^	1.72	1.00	0.29	0.25			
3^	2.18	1.00	0.35	0.25			
4^	1.84	1.00	0.27	0.25			
5^	2.00	1.00	0.29	0.25			
6^	1.44	1.00	0.36	0.25			
7^	1.71	1.00	0.39	0.25			
8^	2.05	1.00	0.45	0.25			
9^	1.30	1.00	0.35	0.25			
10^	2.00	1.00	0.24	0.25			
11^	1.52	1.00	0.45	0.25			
12^	1.26	1.00	0.42	0.25			
13	1.46	1.00	0.51	0.25			
14	1.90	1.00	0.64	0.25			
15	1.21	1.00	0.89	0.25			
16	1.25	1.00	0.87	0.25			
17	1.90	1.00	-0.28	0.25			
18	1.37	1.00	-0.16	0.25			
19	1.87	1.00	0.98	0.25			
20	1.81	1.00	1.74	0.25			
21	2.04	1.00	1.98	0.25			
22	1.50	1.00	1.81	0.25			
23	1.31	1.00	-0.39	0.25			
24	1.21	1.00	-0.43	0.25			
25	1.94	1.00	-0.43	0.25			
26	2.10	1.00	1.52	0.25			
27	1.81	1.00	0.19	0.25			
28	2.15	1.00	0.99	0.25			
29	1.53	1.00	0.51	0.25			
30	1.95	1.00	0.38	0.25			
31	1.69	1.00	1.80	0.25			
32	1.39	1.00	-0.01	0.25			
33	1.52	1.00	1.44	0.25			
34	1.96	1.00	0.78	0.25			
35	1.72	1.00	1.11	0.25			
36	2.18	1.00	1.06	0.25			
37	1.22	1.00	0.11	0.25			
38	1.61	1.00	0.22	0.25			
39	1.31	1.00	0.35	0.25			
40	2.08	1.00	1.14	0.25			
41	1.42	1.00	0.94	0.25			
42	1.35	1.00	0.84	0.25			
43*	1.38	1.00	0.44	0.25			
44*	1.26	1.00	0.42	0.25			
45*	1.43	1.00	0.50	0.25			
46*	2.14	1.00	0.51	0.25			
47*	1.83	1.00	0.63	0.25			
48*	1.85	1.00	0.53	0.25			
49*	1.83	1.00	0.48	0.25			
50*	1.79	1.00	0.49	0.25			
51*	2.00	1.00	0.49	0.25			
52*	1.90	1.00	0.57	0.25			
53*	1.55	1.00	0.62	0.25			
54*	2.12	1.00	0.64	0.25			
55*	1.56	1.00			-0.15	0.25	1.15
56*	1.52	1.00			-0.20	0.31	1.46
57	2.02	1.00			-0.47	0.61	1.04
58	1.61	1.00			0.09	0.57	1.07
59	1.84	1.00			-0.04	0.33	1.05
60^	1.46	1.00			-0.38	0.31	1.03
61^	1.96	1.00			-0.40	0.17	1.09

*Common items between grade 7 and grade 8
^ Common items between grade 8 and grade 9

173

Table A10. Item parameters for grade 9 – expanded CI set

Item Number	a1	a2	b	c	d1	d2	d3
1^	1.46	1.00	0.51	0.25			
2^	1.90	1.00	0.64	0.25			
3^	1.21	1.00	0.89	0.25			
4^	1.25	1.00	0.87	0.25			
5^	1.90	1.00	-0.28	0.25			
6^	1.37	1.00	-0.16	0.25			
7^	1.71	1.00	0.39	0.25			
8^	2.05	1.00	0.45	0.25			
9^	1.30	1.00	0.35	0.25			
10^	2.00	1.00	0.24	0.25			
11^	1.52	1.00	0.45	0.25			
12^	1.26	1.00	0.42	0.25			
13	1.26	1.00	0.22	0.25			
14	1.72	1.00	0.29	0.25			
15	2.18	1.00	0.35	0.25			
16	1.84	1.00	0.27	0.25			
17	2.00	1.00	0.29	0.25			
18	1.44	1.00	0.36	0.25			
19	1.87	1.00	0.98	0.25			
20	1.81	1.00	1.74	0.25			
21	2.04	1.00	1.98	0.25			
22	1.50	1.00	1.81	0.25			
23	1.31	1.00	-0.39	0.25			
24	1.21	1.00	-0.43	0.25			
25	1.94	1.00	-0.43	0.25			
26	2.10	1.00	1.52	0.25			
27	1.81	1.00	0.19	0.25			
28	2.15	1.00	0.99	0.25			
29	1.53	1.00	0.51	0.25			
30	1.95	1.00	0.38	0.25			
31	1.69	1.00	1.80	0.25			
32	1.39	1.00	-0.01	0.25			
33	1.52	1.00	1.44	0.25			
34	1.96	1.00	0.78	0.25			

Item Number	a1	a2	b	c	d1	d2	d3
35	1.72	1.00	1.11	0.25			
36	2.18	1.00	1.06	0.25			
37	1.83	1.00	0.48	0.25			
38	1.79	1.00	0.49	0.25			
39	2.00	1.00	0.49	0.25			
40	1.90	1.00	0.57	0.25			
41	1.55	1.00	0.62	0.25			
42	2.12	1.00	0.64	0.25			
43*	1.38	1.00	0.44	0.25			
44*	1.26	1.00	0.42	0.25			
45*	1.43	1.00	0.50	0.25			
46*	2.14	1.00	0.51	0.25			
47*	1.83	1.00	0.63	0.25			
48*	1.85	1.00	0.53	0.25			
49*	1.22	1.00	0.11	0.25			
50*	1.61	1.00	0.22	0.25			
51*	1.31	1.00	0.35	0.25			
52*	2.08	1.00	1.14	0.25			
53*	1.42	1.00	0.94	0.25			
54*	1.35	1.00	0.84	0.25			
55*	1.56	1.00			-0.15	0.25	1.15
56*	1.52	1.00			-0.20	0.31	1.46
57	2.02	1.00			-0.47	0.61	1.04
58	1.61	1.00			0.09	0.57	1.07
59	1.84	1.00			-0.04	0.33	1.05
60^	1.46	1.00			-0.38	0.31	1.03
61^	1.96	1.00			-0.40	0.17	1.09

*Common items between grade 7 and grade 8
^ Common items between grade 8 and grade 9

174

Table A11. Item parameters for grade 10 – narrow CI set

Item Number	a1	a2	b	c	d1	d2	d3
1	1.46	1.00	1.58	0.25			
2	1.69	1.00	1.92	0.25			
3	1.26	1.00	1.05	0.25			
4	1.72	1.00	1.69	0.25			
5	2.18	1.00	0.22	0.25			
6	1.84	1.00	1.28	0.25			
7	1.80	1.00	1.67	0.25			
8	1.31	1.00	0.17	0.25			
9	2.00	1.00	0.15	0.25			
10	1.44	1.00	1.37	0.25			
11	1.71	1.00	1.06	0.25			
12	2.05	1.00	0.17	0.25			
13	1.30	1.00	0.32	0.25			
14	2.00	1.00	0.88	0.25			
15	1.87	1.00	0.28	0.25			
16	1.81	1.00	0.21	0.25			
17	2.04	1.00	0.09	0.25			
18	1.50	1.00	0.29	0.25			
19	1.52	1.00	-0.14	0.25			
20	1.21	1.00	1.04	0.25			
21	1.94	1.00	1.10	0.25			
22	2.10	1.00	1.33	0.25			
23	1.81	1.00	1.10	0.25			
24	2.15	1.00	1.14	0.25			
25	1.53	1.00	1.91	0.25			
26	1.95	1.00	0.88	0.25			
27	1.90	1.00	0.25	0.25			
28	1.37	1.00	0.39	0.25			
29	1.53	1.00	0.74	0.25			
30	1.29	1.00	-0.25	0.25			
31	1.25	1.00	1.43	0.25			
32	1.39	1.00	0.63	0.25			
33	1.52	1.00	1.17	0.25			
34	1.96	1.00	0.92	0.25			

Item Number	a1	a2	b	c	d1	d2	d3
35	1.72	1.00	1.07	0.25			
36	2.18	1.00	0.87	0.25			
37	1.22	1.00	0.11	0.25			
38	1.61	1.00	0.22	0.25			
39	1.31	1.00	0.35	0.25			
40	2.08	1.00	1.14	0.25			
41	1.42	1.00	0.94	0.25			
42	1.35	1.00	0.84	0.25			
43*	1.38	1.00	0.44	0.25			
44*	1.26	1.00	0.42	0.25			
45*	1.43	1.00	0.50	0.25			
46*	2.14	1.00	0.51	0.25			
47*	1.83	1.00	0.63	0.25			
48*	1.85	1.00	0.53	0.25			
49*	1.83	1.00	0.48	0.25			
50*	1.79	1.00	0.49	0.25			
51*	2.00	1.00	0.49	0.25			
52*	1.90	1.00	0.57	0.25			
53*	1.55	1.00	0.62	0.25			
54*	2.12	1.00	0.64	0.25			
55*	1.56	1.00			-0.15	0.25	1.15
56*	1.52	1.00			-0.20	0.31	1.46
57	2.02	1.00			-0.27	1.15	1.49
58	1.61	1.00			-0.24	0.69	1.79
59	1.46	1.00			-0.07	1.03	1.65
60	1.96	1.00			0.03	0.95	1.54
61	1.84	1.00			0.37	1.02	1.68

*Common items between grade 9 and grade 10

Table A12. Item parameters for grade 10 – expanded CI set

Item Number	a1	a2	b	c	d1	d2	d3
1	1.46	1.00	1.58	0.25			
2	1.69	1.00	1.92	0.25			
3	1.26	1.00	1.05	0.25			
4	1.72	1.00	1.69	0.25			
5	2.18	1.00	0.22	0.25			
6	1.84	1.00	1.28	0.25			
7	1.80	1.00	1.67	0.25			
8	1.31	1.00	0.17	0.25			
9	2.00	1.00	0.15	0.25			
10	1.44	1.00	1.37	0.25			
11	1.71	1.00	1.06	0.25			
12	2.05	1.00	0.17	0.25			
13	1.30	1.00	0.32	0.25			
14	2.00	1.00	0.88	0.25			
15	1.87	1.00	0.28	0.25			
16	1.81	1.00	0.21	0.25			
17	2.04	1.00	0.09	0.25			
18	1.50	1.00	0.29	0.25			
19	1.52	1.00	-0.14	0.25			
20	1.21	1.00	1.04	0.25			
21	1.94	1.00	1.10	0.25			
22	2.10	1.00	1.33	0.25			
23	1.81	1.00	1.10	0.25			
24	2.15	1.00	1.14	0.25			
25	1.53	1.00	1.91	0.25			
26	1.95	1.00	0.88	0.25			
27	1.90	1.00	0.25	0.25			
28	1.37	1.00	0.39	0.25			
29	1.53	1.00	0.74	0.25			
30	1.29	1.00	-0.25	0.25			
31	1.25	1.00	1.43	0.25			
32	1.39	1.00	0.63	0.25			
33	1.52	1.00	1.17	0.25			
34	1.96	1.00	0.92	0.25			
35	1.72	1.00	1.07	0.25			
36	2.18	1.00	0.87	0.25			
37	1.83	1.00	0.48	0.25			
38	1.79	1.00	0.49	0.25			
39	2.00	1.00	0.49	0.25			
40	1.90	1.00	0.57	0.25			
41	1.55	1.00	0.62	0.25			
42	2.12	1.00	0.64	0.25			
43*	1.38	1.00	0.44	0.25			
44*	1.26	1.00	0.42	0.25			
45*	1.43	1.00	0.50	0.25			
46*	2.14	1.00	0.51	0.25			
47*	1.83	1.00	0.63	0.25			
48*	1.85	1.00	0.53	0.25			
49*	1.22	1.00	0.11	0.25			
50*	1.61	1.00	0.22	0.25			
51*	1.31	1.00	0.35	0.25			
52*	2.08	1.00	1.14	0.25			
53*	1.42	1.00	0.94	0.25			
54*	1.35	1.00	0.84	0.25			
55*	1.56	1.00			-0.15	0.25	1.15
56*	1.52	1.00			-0.20	0.31	1.46
57	2.02	1.00			-0.27	1.15	1.49
58	1.61	1.00			-0.24	0.69	1.79
59	1.46	1.00			-0.07	1.03	1.65
60	1.96	1.00			0.03	0.95	1.54
61	1.84	1.00			0.37	1.02	1.68

*Common items between grade 9 and grade 10

176

APPENDIX B

Table B1. Grade 5 ANOVA results for BIAS

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	2611.13	<.0001
FE	3	297	281.48	<.0001
DG*FE	3	297	1819.95	<.0001
SEP	1	99	39475.50	<.0001
DG*SEP	2	198	959.58	<.0001
FE*SEP	3	297	4607.46	<.0001
DG*FE*SEP	3	297	1129.13	<.0001
CI	1	99	23.23	<.0001
DG*CI	2	198	0.23	0.80
FE*CI	3	297	6.28	0.00
DG*FE*CI	3	297	8.22	<.0001
SEP*CI	1	99	12.22	0.00
DG*SEP*CI	2	198	13.27	<.0001
FE*SEP*CI	3	297	0.32	0.81
DG*FE*SEP*CI	3	297	2.63	0.05
VS	1	99	883.15	<.0001
DG*VS	2	198	716.14	<.0001
FE*VS	3	297	51.36	<.0001
DG*FE*VS	3	297	258.04	<.0001
SEP*VS	1	99	14060.90	<.0001
DG*SEP*VS	2	198	329.42	<.0001
FE*SEP*VS	3	297	610.03	<.0001
DG*FE*SEP*VS	3	297	90.03	<.0001
CI*VS	1	99	5.66	0.02
DG*CI*VS	2	198	5.33	0.01
FE*CI*VS	3	297	11.03	<.0001
DG*FE*CI*VS	3	297	1.83	0.14
SEP*CI*VS	1	99	11.64	0.00
DG*SEP*CI*VS	2	198	11.05	<.0001
FE*SEP*CI*VS	3	297	2.35	0.07
DG*FE*SEP*CI*VS	3	297	7.27	0.00

Table B2. Grade 5 ANOVA results for RMSE

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	5531.15	<.0001
FE	3	297	3066.90	<.0001
DG*FE	3	297	717.77	<.0001
SEP	1	99	12557.90	<.0001
DG*SEP	2	198	127.76	<.0001
FE*SEP	3	297	102.84	<.0001
DG*FE*SEP	3	297	132.38	<.0001
CI	1	99	0.92	0.34
DG*CI	2	198	5.81	0.00
FE*CI	3	297	5.85	0.00
DG*FE*CI	3	297	1.90	0.13
SEP*CI	1	99	9.89	0.00
DG*SEP*CI	2	198	5.89	0.00
FE*SEP*CI	3	297	0.40	0.75
DG*FE*SEP*CI	3	297	0.87	0.46
VS	1	99	14093.10	<.0001
DG*VS	2	198	364.92	<.0001
FE*VS	3	297	161.48	<.0001
DG*FE*VS	3	297	4.69	0.00
SEP*VS	1	99	7176.80	<.0001
DG*SEP*VS	2	198	118.64	<.0001
FE*SEP*VS	3	297	62.95	<.0001
DG*FE*SEP*VS	3	297	154.06	<.0001
CI*VS	1	99	0.03	0.87
DG*CI*VS	2	198	1.03	0.36
FE*CI*VS	3	297	0.12	0.95
DG*FE*CI*VS	3	297	4.53	0.00
SEP*CI*VS	1	99	5.62	0.02
DG*SEP*CI*VS	2	198	3.98	0.02
FE*SEP*CI*VS	3	297	0.34	0.80
DG*FE*SEP*CI*VS	3	297	1.25	0.29

Table B3. Grade 6 ANOVA results for BIAS

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	898.62	<.0001
FE	3	297	263.23	<.0001
DG*FE	3	297	1338.15	<.0001
SEP	1	99	16615.40	<.0001
DG*SEP	2	198	560.57	<.0001
FE*SEP	3	297	2664.80	<.0001
DG*FE*SEP	3	297	672.03	<.0001
CI	1	99	52.61	<.0001
DG*CI	2	198	4.24	0.02
FE*CI	3	297	14.71	<.0001
DG*FE*CI	3	297	6.05	0.00
SEP*CI	1	99	35.40	<.0001
DG*SEP*CI	2	198	21.18	<.0001
FE*SEP*CI	3	297	2.05	0.11
DG*FE*SEP*CI	3	297	5.02	0.00
VS	1	99	18.67	<.0001
DG*VS	2	198	13.11	<.0001
FE*VS	3	297	83.42	<.0001
DG*FE*VS	3	297	11.32	<.0001
SEP*VS	1	99	706.79	<.0001
DG*SEP*VS	2	198	142.08	<.0001
FE*SEP*VS	3	297	392.65	<.0001
DG*FE*SEP*VS	3	297	60.96	<.0001
CI*VS	1	99	12.61	0.00
DG*CI*VS	2	198	0.82	0.44
FE*CI*VS	3	297	14.20	<.0001
DG*FE*CI*VS	3	297	3.76	0.01
SEP*CI*VS	1	99	24.93	<.0001
DG*SEP*CI*VS	2	198	12.45	<.0001
FE*SEP*CI*VS	3	297	3.26	0.02
DG*FE*SEP*CI*VS	3	297	1.85	0.14

Table B4. Grade 6 ANOVA results for RMSE

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	2012.88	<.0001
FE	3	297	1282.18	<.0001
DG*FE	3	297	317.13	<.0001
SEP	1	99	974.55	<.0001
DG*SEP	2	198	28.41	<.0001
FE*SEP	3	297	73.92	<.0001
DG*FE*SEP	3	297	6.85	0.00
CI	1	99	5.08	0.03
DG*CI	2	198	2.38	0.10
FE*CI	3	297	1.37	0.25
DG*FE*CI	3	297	2.96	0.03
SEP*CI	1	99	1.47	0.23
DG*SEP*CI	2	198	4.11	0.02
FE*SEP*CI	3	297	0.45	0.72
DG*FE*SEP*CI	3	297	2.15	0.09
VS	1	99	50.51	<.0001
DG*VS	2	198	11.30	<.0001
FE*VS	3	297	5.35	0.00
DG*FE*VS	3	297	3.39	0.02
SEP*VS	1	99	99.48	<.0001
DG*SEP*VS	2	198	13.78	<.0001
FE*SEP*VS	3	297	23.89	<.0001
DG*FE*SEP*VS	3	297	5.79	0.00
CI*VS	1	99	0.03	0.86
DG*CI*VS	2	198	1.50	0.23
FE*CI*VS	3	297	1.17	0.32
DG*FE*CI*VS	3	297	0.70	0.55
SEP*CI*VS	1	99	7.38	0.01
DG*SEP*CI*VS	2	198	2.42	0.09
FE*SEP*CI*VS	3	297	0.15	0.93
DG*FE*SEP*CI*VS	3	297	0.47	0.70

Table B5. Grade 7 ANOVA results for BIAS

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	649.15	<.0001
FE	3	297	664.77	<.0001
DG*FE	3	297	144.06	<.0001
SEP	1	99	40.71	<.0001
DG*SEP	2	198	9.78	<.0001
FE*SEP	3	297	10.07	<.0001
DG*FE*SEP	3	297	4.20	0.01
CI	1	99	1.43	0.23
DG*CI	2	198	1.10	0.34
FE*CI	3	297	2.44	0.06
DG*FE*CI	3	297	4.15	0.01
SEP*CI	1	99	0.15	0.70
DG*SEP*CI	2	198	0.55	0.58
FE*SEP*CI	3	297	3.82	0.01
DG*FE*SEP*CI	3	297	1.42	0.24
VS	1	99	159.26	<.0001
DG*VS	2	198	82.07	<.0001
FE*VS	3	297	79.86	<.0001
DG*FE*VS	3	297	37.03	<.0001
SEP*VS	1	99	38.14	<.0001
DG*SEP*VS	2	198	10.07	<.0001
FE*SEP*VS	3	297	8.89	<.0001
DG*FE*SEP*VS	3	297	7.48	<.0001
CI*VS	1	99	0.09	0.77
DG*CI*VS	2	198	0.00	1.00
FE*CI*VS	3	297	0.11	0.96
DG*FE*CI*VS	3	297	0.03	0.99
SEP*CI*VS	1	99	0.05	0.82
DG*SEP*CI*VS	2	198	0.04	0.96
FE*SEP*CI*VS	3	297	0.12	0.95
DG*FE*SEP*CI*VS	3	296	0.11	0.95

Table B6. Grade 7 ANOVA results for RMSE

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	74837.40	<.0001
FE	3	297	38699.50	<.0001
DG*FE	3	297	16949.80	<.0001
SEP	1	99	1.28	0.26
DG*SEP	2	198	3.36	0.04
FE*SEP	3	297	2.14	0.10
DG*FE*SEP	3	297	2.57	0.05
CI	1	99	7.94	0.01
DG*CI	2	198	3.76	0.03
FE*CI	3	297	0.01	1.00
DG*FE*CI	3	297	2.79	0.04
SEP*CI	1	99	1.40	0.24
DG*SEP*CI	2	198	1.00	0.37
FE*SEP*CI	3	297	1.73	0.16
DG*FE*SEP*CI	3	297	0.50	0.68
VS	1	99	0.42	0.52
DG*VS	2	198	0.17	0.85
FE*VS	3	297	0.65	0.58
DG*FE*VS	3	297	0.89	0.45
SEP*VS	1	99	0.81	0.37
DG*SEP*VS	2	198	0.91	0.40
FE*SEP*VS	3	297	0.51	0.67
DG*FE*SEP*VS	3	297	0.85	0.47
CI*VS	1	99	0.75	0.39
DG*CI*VS	2	198	0.19	0.82
FE*CI*VS	3	297	0.87	0.46
DG*FE*CI*VS	3	297	0.41	0.75
SEP*CI*VS	1	99	0.60	0.44
DG*SEP*CI*VS	2	198	0.20	0.82
FE*SEP*CI*VS	3	297	0.97	0.41
DG*FE*SEP*CI*VS	3	296	0.48	0.70

Table B7. Grade 8 ANOVA results for BIAS

Effect	*df*1	*df*2	*F*-value	*p*-value
DG	2	198	508.59	<.0001
FE	3	297	323.07	<.0001
DG*FE	3	297	22.41	<.0001
SEP	1	99	5330.76	<.0001
DG*SEP	2	198	183.45	<.0001
FE*SEP	3	297	176.39	<.0001
DG*FE*SEP	3	297	44.66	<.0001
CI	1	99	0.70	0.40
DG*CI	2	198	0.09	0.92
FE*CI	3	297	2.90	0.04
DG*FE*CI	3	297	2.78	0.04
SEP*CI	1	99	2.47	0.12
DG*SEP*CI	2	198	0.63	0.53
FE*SEP*CI	3	297	6.12	0.00
DG*FE*SEP*CI	3	297	1.62	0.19
VS	1	99	10748.10	<.0001
DG*VS	2	198	472.37	<.0001
FE*VS	3	297	573.74	<.0001
DG*FE*VS	3	297	146.32	<.0001
SEP*VS	1	99	8610.72	<.0001
DG*SEP*VS	2	198	426.97	<.0001
FE*SEP*VS	3	297	388.45	<.0001
DG*FE*SEP*VS	3	297	98.83	<.0001
CI*VS	1	99	0.74	0.39
DG*CI*VS	2	198	0.65	0.52
FE*CI*VS	3	297	3.20	0.02
DG*FE*CI*VS	3	297	0.96	0.41
SEP*CI*VS	1	99	0.07	0.80
DG*SEP*CI*VS	2	197	0.55	0.58
FE*SEP*CI*VS	3	297	3.77	0.01
DG*FE*SEP*CI*VS	3	297	0.74	0.53

Table B8. Grade 8 ANOVA results for RMSE

Effect	*df*1	*df*2	*F*-value	*p*-value
DG	2	198	74699.90	<.0001
FE	3	297	35317.40	<.0001
DG*FE	3	297	14885.30	<.0001
SEP	1	99	373.14	<.0001
DG*SEP	2	198	234.90	<.0001
FE*SEP	3	297	7.96	<.0001
DG*FE*SEP	3	297	3.33	0.02
CI	1	99	0.39	0.53
DG*CI	2	198	0.33	0.72
FE*CI	3	297	2.49	0.06
DG*FE*CI	3	297	0.39	0.76
SEP*CI	1	99	0.59	0.44
DG*SEP*CI	2	198	2.80	0.06
FE*SEP*CI	3	297	1.27	0.29
DG*FE*SEP*CI	3	297	0.41	0.75
VS	1	99	6947.61	<.0001
DG*VS	2	198	199.51	<.0001
FE*VS	3	297	49.27	<.0001
DG*FE*VS	3	297	9.96	<.0001
SEP*VS	1	99	2891.42	<.0001
DG*SEP*VS	2	198	105.73	<.0001
FE*SEP*VS	3	297	20.14	<.0001
DG*FE*SEP*VS	3	297	3.15	0.03
CI*VS	1	99	0.30	0.58
DG*CI*VS	2	198	0.93	0.39
FE*CI*VS	3	297	3.57	0.01
DG*FE*CI*VS	3	297	0.78	0.51
SEP*CI*VS	1	99	0.37	0.54
DG*SEP*CI*VS	2	197	1.02	0.36
FE*SEP*CI*VS	3	297	1.17	0.32
DG*FE*SEP*CI*VS	3	297	0.35	0.79

Table B9. Grade 9 ANOVA results for BIAS

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	644.74	<.0001
FE	3	297	766.97	<.0001
DG*FE	3	297	257.76	<.0001
SEP	1	99	7567.85	<.0001
DG*SEP	2	198	731.94	<.0001
FE*SEP	3	297	534.37	<.0001
DG*FE*SEP	3	297	212.11	<.0001
CI	1	99	0.11	0.74
DG*CI	2	198	2.39	0.09
FE*CI	3	297	1.75	0.16
DG*FE*CI	3	297	3.63	0.01
SEP*CI	1	99	0.40	0.53
DG*SEP*CI	2	198	2.70	0.07
FE*SEP*CI	3	297	1.82	0.14
DG*FE*SEP*CI	3	297	4.45	0.00
VS	1	99	272.50	<.0001
DG*VS	2	198	18.10	<.0001
FE*VS	3	297	17.28	<.0001
DG*FE*VS	3	297	35.06	<.0001
SEP*VS	1	99	324.51	<.0001
DG*SEP*VS	2	198	8.49	0.00
FE*SEP*VS	3	297	10.16	<.0001
DG*FE*SEP*VS	3	297	25.72	<.0001
CI*VS	1	99	0.00	1.00
DG*CI*VS	2	198	1.86	0.16
FE*CI*VS	3	297	3.16	0.03
DG*FE*CI*VS	3	297	6.17	0.00
SEP*CI*VS	1	99	0.00	0.95
DG*SEP*CI*VS	2	198	1.77	0.17
FE*SEP*CI*VS	3	297	2.98	0.03
DG*FE*SEP*CI*VS	3	297	6.54	0.00

Table B10. Grade 9 ANOVA results for RMSE

Effect	$df1$	$df2$	F-value	p-value
DG	2	198	15513.60	<.0001
FE	3	297	5817.73	<.0001
DG*FE	3	297	2188.89	<.0001
SEP	1	99	4350.53	<.0001
DG*SEP	2	198	514.63	<.0001
FE*SEP	3	297	80.08	<.0001
DG*FE*SEP	3	297	31.07	<.0001
CI	1	99	10.02	0.00
DG*CI	2	198	0.11	0.90
FE*CI	3	297	0.38	0.77
DG*FE*CI	3	297	0.26	0.85
SEP*CI	1	99	6.10	0.02
DG*SEP*CI	2	198	0.51	0.60
FE*SEP*CI	3	297	0.99	0.40
DG*FE*SEP*CI	3	297	0.68	0.56
VS	1	99	0.04	0.84
DG*VS	2	198	1.21	0.30
FE*VS	3	297	0.03	0.99
DG*FE*VS	3	297	4.36	0.01
SEP*VS	1	99	8.03	0.01
DG*SEP*VS	2	198	0.70	0.50
FE*SEP*VS	3	297	2.00	0.11
DG*FE*SEP*VS	3	297	2.49	0.06
CI*VS	1	99	2.97	0.09
DG*CI*VS	2	198	2.83	0.06
FE*CI*VS	3	297	3.70	0.01
DG*FE*CI*VS	3	297	0.39	0.76
SEP*CI*VS	1	99	7.06	0.01
DG*SEP*CI*VS	2	198	1.10	0.33
FE*SEP*CI*VS	3	297	3.59	0.01
DG*FE*SEP*CI*VS	3	297	1.77	0.15

Table B11. Grade 10 ANOVA results for BIAS

Effect	*df*1	*df*2	*F*-value	*p*-value
DG	2	198	271.82	<.0001
FE	3	297	279.45	<.0001
DG*FE	3	297	54.93	<.0001
SEP	1	99	8713.18	<.0001
DG*SEP	2	198	808.09	<.0001
FE*SEP	3	297	411.32	<.0001
DG*FE*SEP	3	297	136.51	<.0001
CI	1	99	0.08	0.78
DG*CI	2	198	3.33	0.04
FE*CI	3	297	1.20	0.31
DG*FE*CI	3	297	3.62	0.01
SEP*CI	1	99	0.24	0.63
DG*SEP*CI	2	198	2.22	0.11
FE*SEP*CI	3	297	1.05	0.37
DG*FE*SEP*CI	3	297	5.06	0.00
VS	1	99	1020.36	<.0001
DG*VS	2	198	1.51	0.22
FE*VS	3	297	6.10	0.00
DG*FE*VS	3	297	20.17	<.0001
SEP*VS	1	99	1005.90	<.0001
DG*SEP*VS	2	198	2.24	0.11
FE*SEP*VS	3	297	7.21	0.00
DG*FE*SEP*VS	3	297	21.57	<.0001
CI*VS	1	99	0.00	0.96
DG*CI*VS	2	198	2.35	0.10
FE*CI*VS	3	297	3.06	0.03
DG*FE*CI*VS	3	297	5.66	0.00
SEP*CI*VS	1	99	0.02	0.88
DG*SEP*CI*VS	2	198	2.00	0.14
FE*SEP*CI*VS	3	297	3.20	0.02
DG*FE*SEP*CI*VS	3	297	5.62	0.00

Table B12. Grade 10 ANOVA results for RMSE

Effect	df1	df2	F-value	p-value
DG	2	198	17155.70	<.0001
FE	3	297	5145.47	<.0001
SEP	3	297	1776.87	<.0001
CI	1	99	3711.23	<.0001
DG*FE	2	198	1166.67	<.0001
DG*SEP	3	297	105.30	<.0001
DG*CI	3	297	27.07	<.0001
FE*SEP	1	99	8.35	0.00
FE*CI	2	198	2.62	0.08
SEP*CI	3	297	1.08	0.36
DG*FE*SEP	3	297	0.63	0.60
DG*FE*CI	1	99	9.15	0.00
DG*SEP*CI	2	198	0.97	0.38
FE*SEP*CI	3	297	0.74	0.53
DG*FE*SEP*CI	3	297	0.56	0.64
VS	1	99	1623.25	<.0001
VS*DG	2	198	39.12	<.0001
VS*FE	3	297	16.93	<.0001
VS*SEP	3	297	1.79	0.15
VS*CI	1	99	346.89	<.0001
VS*DG*FE	2	198	43.14	<.0001
VS*DG*SEP	3	297	16.05	<.0001
VS*DG*CI	3	297	5.14	0.00
VS*FE*SEP	1	99	2.61	0.11
VS*FE*CI	2	198	3.16	0.04
VS*SEP*CI	3	297	4.24	0.01
VS*DG*FE*SEP	3	297	0.38	0.76
VS*DG*FE*CI	1	99	7.37	0.01
VS* DG*SEP*CI	2	198	1.14	0.32
VS*FE*SEP*CI	3	297	4.96	0.00
VS*DG*FE*SEP*CI	3	297	1.95	0.12

APPENDIX C

Table C1. Grade 5 Comparisons of Item Format Effects

Ability Generation	Scaling Method	Common Item Set	Format Effect		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
2-Dim	Separate	Narrow	None	Small	0.01	0.12	-0.16	-0.16
				Moderate	-0.62*	0.46	-1.15*	-1.71*
				Large	-3.98*	1.02*	-3.87*	-2.37*
			Small	Moderate	-0.63*	0.34	-0.99*	-1.55*
				Large	-3.99*	0.90*	-3.71*	-2.21*
			Moderate	Large	-3.36*	0.56*	-2.72*	-0.66
		Expanded	None	Small	-0.07	0.17	-0.26	-0.23
				Moderate	-0.64*	0.49	-1.18*	-1.63*
				Large	-4.03*	1.17*	-3.97*	-2.58*
			Small	Moderate	-0.57*	0.32	-0.92*	-1.40*
				Large	-3.96*	1.00*	-3.71*	-2.35*
			Moderate	Large	-3.39*	0.68*	-2.79*	-0.95*
	Pairwise	Narrow	None	Small	-0.01	0.03	-0.24	-0.10
				Moderate	-0.16	0.23	-0.97*	-0.90*
				Large	-1.75*	0.40	-2.33*	-2.28*
			Small	Moderate	-0.15	0.20	-0.73	-0.80
				Large	-1.74*	0.37	-2.09*	-2.18*
			Moderate	Large	-1.59*	0.17	-1.36*	-1.38*
		Expanded	None	Small	0.02	0.07	-0.17	-0.08
				Moderate	-0.20	0.25	-0.95*	-0.90*
				Large	-1.80*	0.42	-2.32*	-2.29*
			Small	Moderate	-0.22	0.18	-0.78	-0.82
				Large	-1.82*	0.35	-2.15*	-2.21*
			Moderate	Large	-1.60*	0.17	-1.37*	-1.39*
3-Dim	Separate	Narrow	None	Small	-0.02	0.16	-0.05	-0.22
				Moderate	-0.06	0.90*	-0.25	-1.20*
				Large	-0.14	1.91*	-0.58	-2.38*
			Small	Moderate	-0.04	0.74*	-0.20	-0.98*
				Large	-0.12	1.75*	-0.53	-2.16*
			Moderate	Large	-0.08	1.01*	-0.33	-1.18*
		Expanded	None	Small	-0.04	0.40	-0.17	-0.41
				Moderate	-0.06	1.01*	-0.26	-1.16*
				Large	-0.12	1.85*	-0.59	-2.07*
			Small	Moderate	-0.02	0.61*	-0.09	-0.75
				Large	-0.08	1.45*	-0.42	-1.66*
			Moderate	Large	-0.06	0.84*	-0.33	-0.91*
	Pairwise	Narrow	None	Small	0.02	0.15	-0.04	0.07
				Moderate	-0.01	0.39	-0.19	-0.28
				Large	-0.02	0.68*	-0.40	-0.54
			Small	Moderate	0.01	0.24	-0.15	-0.21
				Large	0.00	0.53*	-0.36	-0.47
			Moderate	Large	-0.01	0.29	-0.21	-0.26
		Expanded	None	Small	0.00	0.11	-0.11	-0.11
				Moderate	-0.02	0.24	-0.19	-0.28
				Large	0.00	0.57*	-0.40	-0.57
			Small	Moderate	-0.02	0.13	-0.08	-0.17
				Large	0.00	0.46	-0.29	-0.46
			Moderate	Large	0.02	0.33	-0.21	-0.29

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C2. Grade 5 Comparisons of Vertical Scaling Method Effects

Ability Generation	Common Item Set	Format Effect	Scaling Method		Mean BIAS Small Separation Mean *diff*	Mean BIAS Large Separation Mean *diff*	Mean RMSE Small Separation Mean *diff*	Mean RMSE Large Separation Mean *diff*
Uni	Narrow		Separate	Pairwise	0.37	-0.99*	0.48	3.56*
	Expanded		Separate	Pairwise	0.43	-1.19*	0.46	3.98*
2-Dim	Narrow	None	Separate	Pairwise	0.41	-1.41*	0.43	3.69*
		Small	Separate	Pairwise	1.34*	-0.98*	0.09	3.15*
		Moderate	Separate	Pairwise	0.87*	-1.03*	0.67	3.80*
		Large	Separate	Pairwise	2.64*	-1.42*	2.03*	3.08*
	Expanded	None	Separate	Pairwise	0.38	-1.66*	0.40	3.40*
		Small	Separate	Pairwise	0.91*	-0.96*	-0.18	3.24*
		Moderate	Separate	Pairwise	-0.66*	-1.03*	0.63	3.74*
		Large	Separate	Pairwise	2.61*	-1.54*	2.05*	3.30*
3-Dim	Narrow	None	Separate	Pairwise	0.12	-1.36*	0.26	2.09*
		Small	Separate	Pairwise	0.18	-1.07*	0.27	1.86*
		Moderate	Separate	Pairwise	0.15	-1.54*	0.32	2.56*
		Large	Separate	Pairwise	0.22	-2.26*	0.44	3.48*
	Expanded	None	Separate	Pairwise	0.12	-1.49*	0.22	2.10*
		Small	Separate	Pairwise	0.31	-1.15*	0.28	1.97*
		Moderate	Separate	Pairwise	0.15	-1.58*	0.29	2.55*
		Large	Separate	Pairwise	0.23	-2.09*	0.41	3.17*

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C3. Grade 5 Comparisons of Common Item Effects

Ability Generation	Scaling Method	Format Effect	Common Item Set		Mean BIAS Small Separation Mean *diff*	Mean BIAS Large Separation Mean *diff*	Mean RMSE Small Separation Mean *diff*	Mean RMSE Large Separation Mean *diff*
Uni	Separate		Narrow	Expanded	-0.06	-0.01	-0.03	-0.05
	Pairwise		Narrow	Expanded	0.01	0.00	-0.05	0.00
2-Dim	Separate	None	Narrow	Expanded	0.03	-0.03	0.04	-0.03
		Small	Narrow	Expanded	-0.05	0.02	-0.06	-0.10
		Moderate	Narrow	Expanded	0.01	0.00	0.01	0.05
		Large	Narrow	Expanded	-0.02	0.12	-0.06	-0.24
	Pairwise	None	Narrow	Expanded	0.00	-0.02	-0.05	-0.01
		Small	Narrow	Expanded	-0.48	0.02	0.02	-0.01
		Moderate	Narrow	Expanded	-0.04	0.00	-0.03	-0.01
		Large	Narrow	Expanded	-0.05	0.00	-0.04	-0.02
3-Dim	Separate	None	Narrow	Expanded	-0.01	-0.18	0.03	0.07
		Small	Narrow	Expanded	-0.03	0.06	-0.09	-0.12
		Moderate	Narrow	Expanded	-0.01	-0.07	0.02	0.11
		Large	Narrow	Expanded	0.01	-0.24	0.02	0.38
	Pairwise	None	Narrow	Expanded	0.00	0.04	-0.01	0.10
		Small	Narrow	Expanded	0.02	0.00	-0.08	0.06
		Moderate	Narrow	Expanded	-0.01	-0.11	-0.01	0.10
		Large	Narrow	Expanded	0.02	-0.07	-0.01	0.07

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C4. Grade 5 Comparisons of Grade Level Separation Effects

Ability Generation	Scaling Method	Common Item Set	Format Effect	Grade Level Separation		Mean BIAS	Mean RMSE
						Mean *diff*	Mean *diff*
Uni	Separate	Narrow		Small	Large	2.10*	-3.66*
		Expanded		Small	Large	2.15*	-3.68*
	Pairwise	Narrow		Small	Large	0.91*	-1.04*
		Expanded		Small	Large	0.90*	-0.99*
2-Dim	Separate	Narrow	None	Small	Large	2.10*	-3.48*
			Small	Small	Large	2.21*	-3.48*
			Moderate	Small	Large	3.18*	-4.04*
			Large	Small	Large	7.10*	-1.98*
		Expanded	None	Small	Large	2.04*	-3.55*
			Small	Small	Large	2.28*	-3.52*
			Moderate	Small	Large	3.18*	-4.00*
			Large	Small	Large	7.24*	-2.16*
	Pairwise	Narrow	None	Small	Large	0.89*	-0.98*
			Small	Small	Large	0.93*	-0.84*
			Moderate	Small	Large	1.28*	-0.91*
			Large	Small	Large	3.04*	-0.93*
		Expanded	None	Small	Large	0.87*	-0.94*
			Small	Small	Large	0.92*	-0.82*
			Moderate	Small	Large	1.32*	-0.89*
			Large	Small	Large	3.09*	-0.91*
3-Dim	Separate	Narrow	None	Small	Large	1.74*	-1.89*
			Small	Small	Large	1.92*	-2.06*
			Moderate	Small	Large	2.70*	-2.84*
			Large	Small	Large	3.79*	-3.69*
		Expanded	None	Small	Large	1.57*	-1.85*
			Small	Small	Large	2.01*	-2.09*
			Moderate	Small	Large	2.64*	-2.75*
			Large	Small	Large	3.54*	-3.33*
	Pairwise	Narrow	None	Small	Large	0.26	-0.51
			Small	Small	Large	0.67	-0.54
			Moderate	Small	Large	1.01*	-0.60
			Large	Small	Large	1.31*	-0.65
		Expanded	None	Small	Large	-0.04	0.40
			Small	Small	Large	0.55	-0.40
			Moderate	Small	Large	0.91*	-0.49
			Large	Small	Large	1.22*	-0.57

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C5. Grade 5 Comparisons of Data Generation Models

					Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
Scaling Method	Common Item Set	Format Effect	Ability Generation		Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Separate	Narrow	None	Uni	2-Dim	-0.01	-0.01	0.00	0.18
				3-Dim	0.51	0.15	0.52	2.29*
			2-Dim	3-Dim	0.52	0.16	0.52	2.11*
		Small	2-Dim	3-Dim	0.49	0.20	0.63	2.05*
		Moderate	2-Dim	3-Dim	1.08*	0.60*	1.42*	2.62*
		Large	2-Dim	3-Dim	4.36*	1.05*	3.81*	2.10*
	Expanded	None	Uni	2-Dim	0.08	-0.03	0.07	0.20
				3-Dim	0.50	-0.02	0.59	2.41*
			2-Dim	3-Dim	0.48	0.01	0.51	2.21*
		Small	2-Dim	3-Dim	0.51	0.24	0.60	2.03*
		Moderate	2-Dim	3-Dim	1.06*	0.53*	1.43*	2.68*
		Large	2-Dim	3-Dim	4.39*	0.69*	3.89*	2.72*
Pairwise	Narrow	None	Uni	2-Dim	0.03	-0.01	0.01	0.07
				3-Dim	0.24	-0.06	0.30	0.82*
			2-Dim	3-Dim	0.21	-0.07	0.29	0.76*
		Small	2-Dim	3-Dim	-0.20	0.05	0.49	0.79*
		Moderate	2-Dim	3-Dim	0.36	0.09	1.07*	1.38*
		Large	2-Dim	3-Dim	1.94*	0.21	2.22*	2.50*
	Expanded	None	Uni	2-Dim	0.02	-0.01	0.01	0.06
				3-Dim	0.23	-0.02	0.34	0.93*
			2-Dim	3-Dim	0.21	-0.01	0.33	0.87*
		Small	2-Dim	3-Dim	0.19	0.03	0.39	0.84*
		Moderate	2-Dim	3-Dim	0.39	-0.02	1.09*	1.49*
		Large	2-Dim	3-Dim	2.01*	0.14	2.25*	2.59*

* significant at *p* < 0.05 and Cohen's *d* > 0.20

Table C6. Grade 6 Comparisons of Item Format Effects

Ability Generation	Scaling Method	Common Item Set	Format Effect		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
2-Dim	Separate	Narrow	None	Small	0.08	0.05	-0.16	-0.02
				Moderate	0.00	0.13	-0.96	-1.02*
				Large	-0.90*	0.11	-1.98*	-2.32*
			Small	Moderate	-0.08	0.08	-0.80	-0.99
				Large	-0.97*	0.06	-1.82*	-2.30*
			Moderate	Large	-0.90*	-0.02	-1.01*	-1.30*
		Expanded	None	Small	0.01	0.06	-0.21	-0.11
				Moderate	-0.05	0.11	-0.88	-1.33*
				Large	-0.93*	0.04	-2.03*	-2.34*
			Small	Moderate	-0.06	0.05	-0.67	-1.22*
				Large	-0.94*	-0.02	-1.82*	-2.24*
			Moderate	Large	-0.89*	-0.07	-1.15*	-1.02*
	Pairwise	Narrow	None	Small	0.03	0.05	-0.20	-0.08
				Moderate	-0.08	0.08	-0.88	-0.85
				Large	-0.95*	0.25	-2.04*	-2.19*
			Small	Moderate	-0.11	0.08	-0.68	-0.77
				Large	-0.98*	0.20	-1.84*	-2.11*
			Moderate	Large	-0.87*	0.16	-1.15*	-1.33*
		Expanded	None	Small	0.01	0.01	-0.17	-0.01
				Moderate	-0.09	0.05	-0.94	-1.02*
				Large	-0.94*	0.19	-2.04*	-2.14*
			Small	Moderate	-0.10	0.05	-0.78	-1.02*
				Large	-0.94*	0.19	-1.87*	-2.14*
			Moderate	Large	-0.85*	0.13	-1.09*	-1.12*
3-Dim	Separate	Narrow	None	Small	0.01	0.07	-0.06	-0.13
				Moderate	0.03	0.25	-0.19	-0.43
				Large	0.04	0.45*	-0.48	-1.02*
			Small	Moderate	0.02	0.18	-0.13	-0.30
				Large	0.03	0.38*	-0.43	-0.89
			Moderate	Large	0.01	0.20	-0.29	-0.60
		Expanded	None	Small	0.01	0.04	-0.10	-0.09
				Moderate	0.02	0.18	-0.22	-0.29
				Large	0.04	0.36*	-0.51	-1.21*
			Small	Moderate	0.01	0.13	-0.12	-0.20
				Large	0.03	0.31*	-0.41	-1.12*
			Moderate	Large	0.02	0.18	-0.29	-0.92
	Pairwise	Narrow	None	Small	0.04	0.05	-0.03	-0.10
				Moderate	0.08	0.21	-0.19	-0.28
				Large	0.09	0.40*	-0.52	-0.54
			Small	Moderate	0.04	0.16	-0.15	-0.18
				Large	0.05	0.35*	-0.48	-0.44
			Moderate	Large	0.01	0.19	-0.33	-0.26
		Expanded	None	Small	-0.03	0.00	-0.12	-0.06
				Moderate	0.00	0.12	-0.20	-0.17
				Large	0.04	0.32*	-0.57	-0.59
			Small	Moderate	-0.04	0.12	-0.08	-0.11
				Large	-0.07	0.32*	-0.45	-0.53
			Moderate	Large	0.03	0.19	-0.37	-0.42

* significant at $p < 0.05$ and Cohen's *d* > 0.20

Table C7. Grade 6 Comparisons of Vertical Scaling Method Effects

Ability Generation	Common Item Set	Format Effect	Scaling Method		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Narrow		Separate	Pairwise	0.00	-0.02	0.01	-0.12
	Expanded		Separate	Pairwise	0.00	-0.04	0.02	-0.14
2-Dim	Narrow	None	Separate	Pairwise	0.04	0.00	0.05	0.02
		Small	Separate	Pairwise	0.01	0.00	0.02	-0.08
		Moderate	Separate	Pairwise	-0.04	-0.05	0.14	0.14
		Large	Separate	Pairwise	-0.02	0.14	-0.01	0.11
	Expanded	None	Separate	Pairwise	-0.01	0.01	0.01	-0.06
		Small	Separate	Pairwise	-0.01	0.04	0.06	0.03
		Moderate	Separate	Pairwise	-0.05	-0.05	-0.06	0.24
		Large	Separate	Pairwise	-0.01	0.17	0.00	0.14
3-Dim	Narrow	None	Separate	Pairwise	-0.06	0.02	0.00	-0.13
		Small	Separate	Pairwise	-0.03	0.00	0.03	-0.10
		Moderate	Separate	Pairwise	-0.01	-0.02	0.00	0.02
		Large	Separate	Pairwise	-0.01	-0.03	-0.04	0.35
	Expanded	None	Separate	Pairwise	0.01	0.02	0.00	-0.09
		Small	Separate	Pairwise	-0.03	-0.02	-0.02	-0.06
		Moderate	Separate	Pairwise	0.00	-0.04	0.03	0.03
		Large	Separate	Pairwise	0.01	-0.02	-0.06	0.53

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C8. Grade 6 Comparisons of Common Item Effects

Ability Generation	Scaling Method	Format Effect	Common Item Set		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Separate		Narrow	Expanded	0.01	-0.02	0.00	0.02
	Pairwise		Narrow	Expanded	0.00	0.00	-0.03	0.00
2-Dim	Separate	None	Narrow	Expanded	0.02	0.02	0.01	-0.01
		Small	Narrow	Expanded	-0.04	0.03	-0.04	-0.09
		Moderate	Narrow	Expanded	-0.02	0.00	0.10	-0.32
		Large	Narrow	Expanded	-0.01	-0.06	-0.04	-0.03
	Pairwise	None	Narrow	Expanded	0.01	-0.03	-0.03	-0.05
		Small	Narrow	Expanded	-0.02	-0.01	0.00	0.02
		Moderate	Narrow	Expanded	-0.04	0.00	-0.10	-0.22
		Large	Narrow	Expanded	-0.03	-0.03	-0.03	0.00
3-Dim	Separate	None	Narrow	Expanded	0.00	0.03	0.01	-0.05
		Small	Narrow	Expanded	0.00	0.00	-0.04	-0.01
		Moderate	Narrow	Expanded	0.00	-0.04	-0.03	0.09
		Large	Narrow	Expanded	-0.01	-0.06	-0.03	-0.24
	Pairwise	None	Narrow	Expanded	0.07	-0.03	0.01	-0.01
		Small	Narrow	Expanded	0.00	-0.02	-0.09	0.03
		Moderate	Narrow	Expanded	0.00	-0.06	0.00	0.10
		Large	Narrow	Expanded	-0.02	-0.05	-0.05	-0.06

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C9. Grade 6 Comparisons of Grade Level Separation Effects

Ability Generation	Scaling Method	Common Item set	Format Effect	Grade Level Separation		Mean BIAS Mean *diff*	Mean RMSE Mean *diff*
Uni	Separate	Narrow		Small	Large	0.50*	-0.80
		Expanded		Small	Large	0.49*	-0.88
	Pairwise	Narrow		Small	Large	0.51*	-0.91
		Expanded		Small	Large	0.52*	-0.94
2-Dim	Separate	Narrow	None	Small	Large	0.50*	-0.78
			Small	Small	Large	0.48*	-0.65
			Moderate	Small	Large	0.64*	-0.84
			Large	Small	Large	1.51*	-1.13*
		Expanded	None	Small	Large	0.50*	-0.80
			Small	Small	Large	0.55*	-0.69
			Moderate	Small	Large	0.65*	-1.25*
			Large	Small	Large	1.47*	-1.11*
	Pairwise	Narrow	None	Small	Large	0.47*	-0.86
			Small	Small	Large	0.49*	-0.74
			Moderate	Small	Large	0.63*	-0.83
			Large	Small	Large	1.67*	-1.01*
		Expanded	None	Small	Large	0.52*	-0.88
			Small	Small	Large	0.52*	-0.72
			Moderate	Small	Large	0.65*	-0.95
			Large	Small	Large	1.64*	-0.98
3-Dim	Separate	Narrow	None	Small	Large	0.34*	-0.37
			Small	Small	Large	0.40*	-0.44
			Moderate	Small	Large	0.56*	-0.61
			Large	Small	Large	0.75*	-0.91
		Expanded	None	Small	Large	0.36*	-0.42
			Small	Small	Large	0.40*	-0.41
			Moderate	Small	Large	0.53*	-0.49
			Large	Small	Large	0.69*	-1.12*
	Pairwise	Narrow	None	Small	Large	0.42*	-0.50
			Small	Small	Large	0.43*	-0.57
			Moderate	Small	Large	0.54*	-0.60
			Large	Small	Large	0.73*	-0.52
		Expanded	None	Small	Large	0.38*	-0.52
			Small	Small	Large	0.41*	-0.45
			Moderate	Small	Large	0.50*	-0.49
			Large	Small	Large	0.69*	-0.54

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C10. Grade 6 Comparisons of Data Generation Models

Scaling Method	Common Item Set	Format Effect	Ability Generation		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Separate	Narrow	None	Uni	2-Dim	0.00	0.01	0.00	0.01
				3-Dim	0.10	-0.06	0.41	0.84
			2-Dim	3-Dim	0.10	-0.07	0.42	0.83
		Small	2-Dim	3-Dim	0.03	-0.05	0.52	0.72
		Moderate	2-Dim	3-Dim	0.13	0.05	1.19*	1.42*
		Large	2-Dim	3-Dim	1.03*	0.27*	1.91*	2.13*
	Expanded	None	Uni	2-Dim	0.03	0.05	0.01	-0.02
				3-Dim	0.11	-0.01	0.42	0.77
			2-Dim	3-Dim	0.07	-0.06	0.41	0.79
		Small	2-Dim	3-Dim	0.07	-0.07	0.52	0.80
		Moderate	2-Dim	3-Dim	0.14	0.01	1.07*	1.82*
		Large	2-Dim	3-Dim	1.04*	0.27*	1.93*	1.92*
Pairwise	Narrow	None	Uni	2-Dim	0.04	-0.01	0.06	0.11
				3-Dim	0.04	-0.06	0.42	0.83
			2-Dim	3-Dim	0.00	-0.05	0.36	0.72
		Small	2-Dim	3-Dim	0.01	-0.01	0.53	0.70
		Moderate	2-Dim	3-Dim	0.16	0.08	1.06*	1.29*
		Large	2-Dim	3-Dim	1.04*	0.10	1.88*	2.37*
	Expanded	None	Uni	2-Dim	0.02	-0.02	0.00	0.06
				3-Dim	0.11	-0.03	0.41	0.82
			2-Dim	3-Dim	0.09	0.05	0.40	0.76
		Small	2-Dim	3-Dim	0.05	-0.06	0.44	0.71
		Moderate	2-Dim	3-Dim	0.18	0.02	1.14*	1.61*
		Large	2-Dim	3-Dim	1.06*	0.08	1.86*	2.30*

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C11. Grade 7 Comparisons of Item Format Effects

Ability Generation	Scaling Method	Common Item Set	Format Effect		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
2-Dim	Separate	Narrow	None	Small	0.05	0.04	-0.16*	-0.15*
				Moderate	0.13	0.13	-0.80*	-0.88*
				Large	0.11	0.10	-1.82*	-1.80*
			Small	Moderate	0.08	0.09	-0.73*	-0.72*
				Large	0.06	0.06	-1.66*	-1.65*
			Moderate	Large	-0.03	-0.04	-0.93*	-0.92*
		Expanded	None	Small	0.06	0.08	-0.15*	-0.16*
				Moderate	0.13	0.13	-0.89*	-0.87*
				Large	0.10	0.10	-1.80*	-1.82*
			Small	Moderate	0.07	0.05	-0.74*	-0.72*
				Large	0.04	0.02	-1.65*	-1.66*
			Moderate	Large	-0.03	-0.03	-0.91*	-0.95*
	Pairwise	Narrow	None	Small	0.06	0.06	-0.16*	-0.15*
				Moderate	0.18*	0.23*	-0.95*	-0.87*
				Large	0.20*	0.31*	-1.82*	-1.79*
			Small	Moderate	0.12	0.17*	-0.79*	-0.72*
				Large	0.14*	0.25*	-1.66*	-1.63*
			Moderate	Large	0.02	0.08	-0.86*	-0.91*
		Expanded	None	Small	0.07	0.10	-0.15*	-0.16*
				Moderate	0.17*	0.22*	-0.89*	-0.87*
				Large	0.19*	0.31*	-1.80*	-1.81*
			Small	Moderate	0.11	0.12	-0.74*	-0.71*
				Large	0.12	0.21*	-1.65*	-1.65*
			Moderate	Large	0.12	0.09	-0.91*	-0.93*
3-Dim	Separate	Narrow	None	Small	0.02	0.00	-0.04	-0.04
				Moderate	0.04	0.02	-0.16*	-0.17*
				Large	0.08	0.06	-0.38*	-0.37*
			Small	Moderate	0.01	0.02	-0.12	-0.13
				Large	0.06	0.06	-0.34*	-0.33*
			Moderate	Large	0.04	0.04	-0.22*	-0.20*
		Expanded	None	Small	-0.01	0.02	-0.04	-0.05
				Moderate	0.04	0.04	-0.19*	-0.19*
				Large	0.04	0.08	-0.37*	-0.39*
			Small	Moderate	0.05	0.03	-0.15	-0.13
				Large	0.04	0.06	-0.33*	-0.33*
			Moderate	Large	0.00	0.04	-0.18*	-0.20*
	Pairwise	Narrow	None	Small	0.02	0.01	-0.04	-0.04
				Moderate	0.06	0.03	-0.16*	-0.17*
				Large	0.11	0.09	-0.38*	-0.37*
			Small	Moderate	0.03	0.02	-0.12	-0.13
				Large	0.08	0.09	-0.34*	-0.33*
			Moderate	Large	0.05	0.07	-0.21*	-0.20*
		Expanded	None	Small	-0.01	0.02	-0.04	-0.05
				Moderate	0.04	0.06	-0.18*	-0.19*
				Large	0.07	0.11	-0.37*	-0.38*
			Small	Moderate	0.05	0.03	-0.14	-0.13
				Large	0.07	0.09	-0.32*	-0.33*
			Moderate	Large	0.03	0.05	-0.19*	-0.20*

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C12. Grade 7 Comparisons of Vertical Scaling Method Effects

Ability Generation	Common Item Set	Format Effect	Scaling Method		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Narrow		Separate	Pairwise	0.00	0.00	0.01	0.00
	Expanded		Separate	Pairwise	-0.01	0.00	0.00	0.00
2-Dim	Narrow	None	Separate	Pairwise	0.00	0.01	0.00	0.00
		Small	Separate	Pairwise	0.01	0.03	0.00	0.00
		Moderate	Separate	Pairwise	0.04	0.10	-0.06	0.01
		Large	Separate	Pairwise	0.09	0.22*	0.00	0.01
	Expanded	None	Separate	Pairwise	0.00	0.01	0.00	0.00
		Small	Separate	Pairwise	0.01	0.03	0.00	0.00
		Moderate	Separate	Pairwise	0.04	0.10	0.00	0.00
		Large	Separate	Pairwise	0.09	0.21*	0.00	0.02
3-Dim	Narrow	None	Separate	Pairwise	-0.01	-0.01	0.00	0.00
		Small	Separate	Pairwise	-0.01	0.00	0.00	0.00
		Moderate	Separate	Pairwise	0.01	0.01	0.00	0.00
		Large	Separate	Pairwise	0.01	0.03	0.00	0.00
	Expanded	None	Separate	Pairwise	-0.01	-0.01	0.00	0.00
		Small	Separate	Pairwise	-0.01	0.00	0.00	0.00
		Moderate	Separate	Pairwise	-0.01	0.01	0.01	0.00
		Large	Separate	Pairwise	0.02	0.03	0.00	0.00

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C13. Grade 7 Comparisons of Common Item Effects

Ability Generation	Scaling Method	Format Effect	Common Item Set		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Separate		Narrow	Expanded	0.00	-0.01	0.02	0.01
	Pairwise		Narrow	Expanded	-0.01	-0.01	0.01	0.01
2-Dim	Separate	None	Narrow	Expanded	0.00	-0.01	0.00	0.02
		Small	Narrow	Expanded	0.01	0.03	0.01	0.02
		Moderate	Narrow	Expanded	0.00	-0.01	0.00	0.02
		Large	Narrow	Expanded	-0.01	-0.01	0.02	0.00
	Pairwise	None	Narrow	Expanded	0.00	-0.01	0.00	0.02
		Small	Narrow	Expanded	0.01	0.03	0.01	0.01
		Moderate	Narrow	Expanded	-0.01	-0.02	0.06	0.02
		Large	Narrow	Expanded	-0.01	-0.01	0.02	0.00
3-Dim	Separate	None	Narrow	Expanded	0.02	-0.01	0.01	0.01
		Small	Narrow	Expanded	-0.01	0.00	0.01	0.00
		Moderate	Narrow	Expanded	0.02	0.01	-0.02	-0.01
		Large	Narrow	Expanded	-0.02	0.01	0.02	-0.01
	Pairwise	None	Narrow	Expanded	0.02	-0.01	0.01	0.01
		Small	Narrow	Expanded	-0.01	0.00	0.01	-0.01
		Moderate	Narrow	Expanded	0.00	0.02	-0.01	-0.01
		Large	Narrow	Expanded	-0.02	0.00	0.02	-0.01

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C14. Grade 7 Comparisons of Grade Level Separation Effects

Ability Generation	Scaling Method	Common Item set	Format Effect	Grade Level Separation		Mean BIAS Mean *diff*	Mean RMSE Mean *diff*
Uni	Separate	Narrow		Small	Large	0.00	0.00
		Expanded		Small	Large	-0.01	-0.01
	Pairwise	Narrow		Small	Large	0.00	-0.01
		Expanded		Small	Large	0.00	-0.01
2-Dim	Separate	Narrow	None	Small	Large	0.01	-0.01
			Small	Small	Large	0.00	0.00
			Moderate	Small	Large	0.01	0.00
			Large	Small	Large	0.00	0.01
		Expanded	None	Small	Large	-0.01	0.01
			Small	Small	Large	0.02	0.00
			Moderate	Small	Large	0.00	0.03
			Large	Small	Large	0.00	-0.01
	Pairwise	Narrow	None	Small	Large	0.02	-0.01
			Small	Small	Large	0.02	0.00
			Moderate	Small	Large	0.06	0.07
			Large	Small	Large	0.12	0.02
		Expanded	None	Small	Large	0.00	0.01
			Small	Small	Large	0.04	0.00
			Moderate	Small	Large	0.05	0.03
			Large	Small	Large	0.12	0.01
3-Dim	Separate	Narrow	None	Small	Large	0.02	0.00
			Small	Small	Large	0.00	0.00
			Moderate	Small	Large	0.00	0.00
			Large	Small	Large	0.00	0.01
		Expanded	None	Small	Large	-0.01	0.00
			Small	Small	Large	0.01	-0.01
			Moderate	Small	Large	-0.01	0.01
			Large	Small	Large	0.03	-0.01
	Pairwise	Narrow	None	Small	Large	0.03	0.00
			Small	Small	Large	0.01	0.00
			Moderate	Small	Large	0.00	0.00
			Large	Small	Large	0.01	0.01
		Expanded	None	Small	Large	-0.01	0.00
			Small	Small	Large	0.02	-0.01
			Moderate	Small	Large	0.01	0.00
			Large	Small	Large	0.04	-0.02

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C15. Grade 7 Comparisons of Data Generation Models

Scaling Method	Common Item Set	Format Effect	Ability Generation		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Separate	Narrow	None	Uni	2-Dim	0.00	0.01	-0.01	-0.02
				3-Dim	-0.02	0.00	0.39*	0.40*
			2-Dim	3-Dim	-0.02	-0.01	0.40*	0.42*
		Small	2-Dim	3-Dim	-0.05	-0.05	0.52*	0.53*
		Moderate	2-Dim	3-Dim	-0.11	-0.12	1.13*	1.12*
		Large	2-Dim	3-Dim	-0.04	-0.04	1.84*	1.84*
	Expanded	None	Uni	2-Dim	0.01	0.01	-0.02	0.00
				3-Dim	0.01	0.00	0.39*	0.40*
			2-Dim	3-Dim	0.00	-0.01	0.41*	0.40*
		Small	2-Dim	3-Dim	-0.06	-0.07	0.52*	0.51*
		Moderate	2-Dim	3-Dim	-0.09	-0.10	1.11*	1.09*
		Large	2-Dim	3-Dim	-0.06	-0.03	1.84*	1.84*
Pairwise	Narrow	None	Uni	2-Dim	0.00	0.01	-0.01	-0.02
				3-Dim	-0.03	0.00	0.39*	0.40*
			2-Dim	3-Dim	-0.03	-0.02	0.40*	0.42*
		Small	2-Dim	3-Dim	-0.06	-0.07	0.52*	0.53*
		Moderate	2-Dim	3-Dim	-0.15	-0.22*	1.19*	1.12*
		Large	2-Dim	3-Dim	-0.12	-0.23*	1.84*	1.83*
	Expanded	None	Uni	2-Dim	0.01	0.02	-0.02	0.00
				3-Dim	0.00	0.00	0.39*	0.40*
			2-Dim	3-Dim	-0.01	-0.02	0.41*	0.40*
		Small	2-Dim	3-Dim	-0.08	-0.10	0.52*	0.51*
		Moderate	2-Dim	3-Dim	-0.14	-0.18*	1.12*	1.09*
		Large	2-Dim	3-Dim	-0.13	-0.21*	1.85*	1.82*

* significant at $p < 0.05$ and Cohen's *d* > 0.20

Table C16. Grade 8 Comparisons of Item Format Effects

Ability Generation	Scaling Method	Common Item Set	Format Effect		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
2-Dim	Separate	Narrow	None	Small	0.06	0.04	-0.15	-0.17
				Moderate	0.11	0.18*	-0.93*	-0.89*
				Large	0.09	0.78*	-1.92*	-2.03*
			Small	Moderate	0.05	0.14*	-0.77*	-0.72*
				Large	0.03	0.75*	-1.76*	-1.87*
			Moderate	Large	-0.02	0.60*	-0.99*	-1.15*
		Expanded	None	Small	0.03	0.02	-0.16	-0.12
				Moderate	0.08	0.17*	-0.96*	-0.88*
				Large	0.08	0.64*	-1.91*	-1.95*
			Small	Moderate	0.05	0.15*	-0.79*	-0.75*
				Large	0.05	0.61*	-1.74*	-1.83*
			Moderate	Large	0.00	0.46*	-0.95*	-1.07*
	Pairwise	Narrow	None	Small	0.06	0.00	-0.17	-0.19
				Moderate	0.09	-0.01	-0.94*	-0.96*
				Large	0.01	-0.15*	-1.86*	-1.86*
			Small	Moderate	0.03	-0.01	-0.76*	-0.77*
				Large	-0.05	-0.15*	-1.69*	-1.68*
			Moderate	Large	-0.08	-0.14	-0.93*	-0.91*
		Expanded	None	Small	0.03	0.01	-0.17	-0.20
				Moderate	0.06	-0.03	-0.95*	-0.96*
				Large	0.00	-0.16*	-1.88*	-1.87*
			Small	Moderate	0.03	-0.05	-0.78*	-0.76*
				Large	-0.04	-0.18*	-1.71*	-1.67*
			Moderate	Large	-0.06	-0.13	-0.93*	-0.91*
3-Dim	Separate	Narrow	None	Small	0.02	0.02	-0.06	-0.06
				Moderate	0.04	0.07	-0.19	-0.18
				Large	0.06	0.29*	-0.42*	-0.47*
			Small	Moderate	0.02	0.05	-0.14	-0.12
				Large	0.04	0.27*	-0.36*	-0.41*
			Moderate	Large	0.02	0.22*	-0.23	-0.29
		Expanded	None	Small	0.00	0.01	-0.07	-0.04
				Moderate	0.04	0.12	-0.22	-0.21
				Large	0.07	0.27*	-0.42*	-0.45*
			Small	Moderate	0.04	0.11	-0.15	-0.17
				Large	0.07	0.26*	-0.36*	-0.42*
			Moderate	Large	0.03	0.15*	-0.21	-0.24
	Pairwise	Narrow	None	Small	0.01	-0.01	-0.05	-0.05
				Moderate	0.03	-0.01	-0.19	-0.19
				Large	0.03	-0.02	-0.41*	-0.37*
			Small	Moderate	0.01	0.00	-0.14	-0.14
				Large	0.02	-0.01	-0.35*	-0.32*
			Moderate	Large	0.01	-0.01	-0.21	-0.18
		Expanded	None	Small	-0.01	0.00	-0.07	-0.04
				Moderate	0.02	-0.01	-0.21	-0.17
				Large	0.03	-0.01	-0.41*	-0.37*
			Small	Moderate	0.03	-0.01	-0.14	-0.13
				Large	0.04	-0.01	-0.35*	-0.32*
			Moderate	Large	0.01	0.00	-0.21	-0.20

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C17. Grade 8 Comparisons of Vertical Scaling Set

Ability Generation	Common Item Set	Format Effect	Scaling Method		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Narrow		Separate	Pairwise	-0.02	-0.51*	0.10	0.56*
	Expanded		Separate	Pairwise	-0.02	-0.51*	0.11	0.54*
2-Dim	Narrow	None	Separate	Pairwise	-0.01	-0.52*	0.09	0.46*
		Small	Separate	Pairwise	-0.02	-0.56*	0.07	0.43*
		Moderate	Separate	Pairwise	-0.04	-0.71*	0.08	0.39*
		Large	Separate	Pairwise	-0.10	-1.45*	0.14	0.62*
	Expanded	None	Separate	Pairwise	-0.01	-0.54*	0.08	0.48*
		Small	Separate	Pairwise	-0.01	-0.55*	0.07	0.41*
		Moderate	Separate	Pairwise	-0.03	-0.74*	0.09	0.40*
		Large	Separate	Pairwise	-0.10	-1.34*	0.11	0.57*
3-Dim	Narrow	None	Separate	Pairwise	-0.02	-0.32*	0.06	0.25*
		Small	Separate	Pairwise	-0.03	-0.36*	0.06	0.26*
		Moderate	Separate	Pairwise	-0.03	-0.41*	0.06	0.24*
		Large	Separate	Pairwise	-0.05	-0.64*	0.08	0.36*
	Expanded	None	Separate	Pairwise	-0.02	-0.33*	0.06	0.26*
		Small	Separate	Pairwise	-0.02	-0.35*	0.06	0.25*
		Moderate	Separate	Pairwise	-0.03	-0.46*	0.07	0.30*
		Large	Separate	Pairwise	-0.05	-0.61*	0.07	0.34*

* significant at *p* < 0.05 and Cohen's *d* > 0.20

Table C18. Grade 8 Comparisons of Common Item Set

Ability Generation	Scaling Method	Format Effect	Common Item Set		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Separate		Narrow	Expanded	-0.01	-0.02	-0.02	0.02
	Pairwise		Narrow	Expanded	-0.01	-0.02	-0.01	0.00
2-Dim	Separate	None	Narrow	Expanded	0.01	0.03	0.02	-0.03
		Small	Narrow	Expanded	-0.02	0.02	0.01	0.01
		Moderate	Narrow	Expanded	-0.02	0.02	-0.01	-0.02
		Large	Narrow	Expanded	0.00	-0.11	0.03	0.05
	Pairwise	None	Narrow	Expanded	0.01	0.01	0.01	0.00
		Small	Narrow	Expanded	-0.02	0.03	0.01	-0.01
		Moderate	Narrow	Expanded	-0.01	-0.01	0.00	0.00
		Large	Narrow	Expanded	0.00	0.00	0.00	-0.01
3-Dim	Separate	None	Narrow	Expanded	0.00	0.00	0.01	-0.01
		Small	Narrow	Expanded	-0.02	-0.01	0.00	0.01
		Moderate	Narrow	Expanded	0.00	0.05	-0.01	-0.04
		Large	Narrow	Expanded	0.00	-0.02	0.00	0.01
	Pairwise	None	Narrow	Expanded	0.01	-0.01	0.01	0.00
		Small	Narrow	Expanded	-0.02	0.00	0.00	0.00
		Moderate	Narrow	Expanded	0.00	-0.01	0.00	0.02
		Large	Narrow	Expanded	0.00	0.00	0.00	0.00

* significant at *p* < 0.05 and Cohen's *d* > 0.20

Table C19. Grade 8 Comparisons of Grade Level Separation

Ability Generation	Scaling Method	Common Item set	Format Effect	Grade Level Separation		Mean BIAS Mean *diff*	Mean RMSE Mean *diff*
Uni	Separate	Narrow		Small	Large	0.46*	-0.29*
		Expanded		Small	Large	0.49*	-0.25*
	Pairwise	Narrow		Small	Large	-0.03	0.07
		Expanded		Small	Large	0.00	0.08
2-Dim	Separate	Narrow	None	Small	Large	0.48*	-0.27*
			Small	Small	Large	0.45*	-0.28*
			Moderate	Small	Large	0.55*	-0.23*
			Large	Small	Large	1.17*	-0.38*
		Expanded	None	Small	Large	0.50*	-0.32*
			Small	Small	Large	0.49*	-0.28*
			Moderate	Small	Large	0.60*	-0.24*
			Large	Small	Large	1.06*	-0.36*
	Pairwise	Narrow	None	Small	Large	-0.03	0.10
			Small	Small	Large	-0.09	0.08
			Moderate	Small	Large	-0.12	0.08
			Large	Small	Large	-0.18	0.10
		Expanded	None	Small	Large	-0.02	0.08
			Small	Small	Large	-0.04	0.06
			Moderate	Small	Large	-0.12	0.08
			Large	Small	Large	-0.18	0.10
3-Dim	Separate	Narrow	None	Small	Large	0.30*	-0.08
			Small	Small	Large	0.30*	-0.08
			Moderate	Small	Large	0.33*	-0.07
			Large	Small	Large	0.53*	-0.13
		Expanded	None	Small	Large	0.30*	-0.10
			Small	Small	Large	0.32*	-0.07
			Moderate	Small	Large	0.38*	-0.09
			Large	Small	Large	0.50*	-0.13
	Pairwise	Narrow	None	Small	Large	0.00	0.11
			Small	Small	Large	-0.03	0.12
			Moderate	Small	Large	-0.04	0.12
			Large	Small	Large	-0.06	0.15
		Expanded	None	Small	Large	-0.02	0.10
			Small	Small	Large	-0.01	0.12
			Moderate	Small	Large	-0.05	0.14
			Large	Small	Large	-0.06	0.15

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C20. Grade 8 Comparisons of Data Generation Models

Scaling Method	Common Item Set	Format Effect	Ability Generation		Mean BIAS Small Separation	Mean BIAS Large Separation	Mean RMSE Small Separation	Mean RMSE Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Separate	Narrow	None	Uni	2-Dim	0.01	0.03	-0.04	-0.01
				3-Dim	0.00	-0.17*	0.42*	0.63*
		2-Dim	2-Dim	3-Dim	-0.01	-0.19*	0.46*	0.64*
		Small	2-Dim	3-Dim	-0.06	-0.21*	0.56*	0.75*
		Moderate	2-Dim	3-Dim	-0.08	-0.30*	1.19*	1.35*
		Large	2-Dim	3-Dim	-0.04	-0.68*	1.96*	2.20*
	Expanded	None	Uni	2-Dim	0.03	0.04	0.00	-0.06
				3-Dim	0.01	-0.19*	0.45*	0.60*
		2-Dim	2-Dim	3-Dim	-0.02	-0.23*	0.45*	0.67*
		Small	2-Dim	3-Dim	-0.06	-0.23*	0.54*	0.75*
		Moderate	2-Dim	3-Dim	-0.06	-0.28*	1.19*	1.33*
		Large	2-Dim	3-Dim	-0.04	-0.59*	1.93*	2.16*
Pairwise	Narrow	None	Uni	2-Dim	0.02	0.02	-0.05	-0.02
				3-Dim	0.00	0.02	0.39*	0.42*
		2-Dim	2-Dim	3-Dim	-0.02	0.00	0.43*	0.44*
		Small	2-Dim	3-Dim	-0.06	-0.01	0.55*	0.58*
		Moderate	2-Dim	3-Dim	-0.08	0.00	1.17*	1.21*
		Large	2-Dim	3-Dim	0.01	0.13	1.89*	1.94*
	Expanded	None	Uni	2-Dim	0.03	0.01	-0.03	-0.02
				3-Dim	0.01	-0.01	0.40*	0.42*
		2-Dim	2-Dim	3-Dim	-0.02	-0.02	0.42*	0.44*
		Small	2-Dim	3-Dim	-0.06	-0.03	0.53*	0.59*
		Moderate	2-Dim	3-Dim	-0.06	0.00	1.17*	1.23*
		Large	2-Dim	3-Dim	0.01	0.13	1.89*	1.94*

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C21. Grade 9 Comparisons of Item Format Effects

Ability Generation	Scaling Method	Common Item Set	Format Effect		Mean BIAS Small Separation Mean *diff*	Mean BIAS Large Separation Mean *diff*	Mean RMSE Small Separation Mean *diff*	Mean RMSE Large Separation Mean *diff*
2-Dim	Separate	Narrow	None	Small	0.04	0.10	-0.15	-0.15
				Moderate	0.12	0.58*	-1.05*	-0.85*
				Large	0.15	2.15*	-2.03*	-2.46*
			Small	Moderate	0.09	0.48*	-0.90*	-0.70*
				Large	0.11	2.04*	-1.88*	-2.31*
			Moderate	Large	0.03	1.57*	-0.98	-1.61*
		Expanded	None	Small	0.03	0.14	-0.13	-0.18
				Moderate	0.13	0.68*	-0.93*	-1.00*
				Large	0.13	2.03*	-1.99*	-2.35*
			Small	Moderate	0.10	0.54*	-0.80*	-0.82*
				Large	0.09	1.89*	-1.86*	-2.17*
			Moderate	Large	0.00	1.35*	-1.06*	-1.35*
	Pairwise	Narrow	None	Small	0.04	-0.14	-0.15	-0.07
				Moderate	0.17	1.26*	-0.93*	-0.88*
				Large	0.30	2.82*	-1.93*	-2.36*
			Small	Moderate	0.13	1.39*	-0.77*	-0.81*
				Large	0.25	2.96*	-1.77*	-2.29*
			Moderate	Large	0.12	1.57*	-1.00	-1.48*
		Expanded	None	Small	0.04	-0.43	-0.11	-0.13
				Moderate	0.18	0.41	-0.89*	-0.82*
				Large	0.27	2.77*	-1.93*	-2.47*
			Small	Moderate	0.13	0.84*	-0.78	-0.68*
				Large	0.23	3.20*	-1.82*	-2.34*
			Moderate	Large	0.09	2.36*	-1.04*	-1.66*
3-Dim	Separate	Narrow	None	Small	0.03	0.08	-0.07	-0.03
				Moderate	0.07	0.29	-0.25	-0.19
				Large	0.11	0.77*	-0.47	-0.59*
			Small	Moderate	0.04	0.21	-0.18	-0.16
				Large	0.08	0.69*	-0.40	-0.56*
			Moderate	Large	0.05	0.48	-0.22	-0.40
		Expanded	None	Small	0.01	0.06	-0.02	-0.04
				Moderate	0.06	0.32	-0.21	-0.23
				Large	0.10	0.73*	-0.44	-0.59*
			Small	Moderate	0.05	0.26	-0.18	-0.19
				Large	0.09	0.66*	-0.41	-0.54*
			Moderate	Large	0.03	0.41	-0.23	-0.35
	Pairwise	Narrow	None	Small	0.03	-0.01	-0.06	-0.10
				Moderate	0.07	0.46	-0.23	-0.43
				Large	0.12	0.70*	-0.42	-0.63*
			Small	Moderate	0.04	0.47	-0.17	-0.33
				Large	0.08	0.71*	-0.37	-0.52*
			Moderate	Large	0.04	0.24	-0.20	-0.20
		Expanded	None	Small	0.02	0.04	-0.07	-0.09
				Moderate	0.07	0.09	-0.23	-0.34
				Large	0.13	0.25	-0.51	-0.63*
			Small	Moderate	0.06	0.05	-0.16	-0.25
				Large	0.11	0.21	-0.44	-0.54*
			Moderate	Large	0.05	0.16	-0.28	-0.29

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C22. Grade 9 Comparisons of Vertical Scaling Set

Ability Generation	Common Item Set	Format Effect	Scaling Method		Mean BIAS Small Separation Mean *diff*	Mean BIAS Large Separation Mean *diff*	Mean RMSE Small Separation Mean *diff*	Mean RMSE Large Separation Mean *diff*
Uni	Narrow		Separate	Pairwise	0.00	-0.41	-0.03	0.06
	Expanded		Separate	Pairwise	0.00	-0.24	-0.01	-0.04
2-Dim	Narrow	None	Separate	Pairwise	0.00	-0.64	-0.02	-0.06
		Small	Separate	Pairwise	0.01	-0.84*	-0.02	0.02
		Moderate	Separate	Pairwise	0.06	0.03	0.10	-0.09
		Large	Separate	Pairwise	0.15	0.03	0.08	0.04
	Expanded	None	Separate	Pairwise	0.00	-0.47	-0.03	-0.02
		Small	Separate	Pairwise	0.01	-1.04*	-0.01	0.02
		Moderate	Separate	Pairwise	0.05	-0.74	0.01	0.16
		Large	Separate	Pairwise	0.14	0.27	0.03	-0.15
3-Dim	Narrow	None	Separate	Pairwise	-0.01	-0.55	0.02	-0.01
		Small	Separate	Pairwise	0.00	-0.63	0.04	-0.08
		Moderate	Separate	Pairwise	0.00	-0.38	0.05	-0.25
		Large	Separate	Pairwise	-0.01	-0.62	0.07	-0.04
	Expanded	None	Separate	Pairwise	-0.02	-0.27	0.08	0.10
		Small	Separate	Pairwise	-0.01	-0.29	0.03	0.06
		Moderate	Separate	Pairwise	-0.01	-0.50	0.05	0.00
		Large	Separate	Pairwise	0.01	-0.76	0.00	0.06

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C23. Grade 9 Comparisons of Common Item Set

Ability Generation	Scaling Method	Format Effect	Common Item Set		Mean BIAS Small Separation Mean *diff*	Mean BIAS Large Separation Mean *diff*	Mean RMSE Small Separation Mean *diff*	Mean RMSE Large Separation Mean *diff*
Uni	Separate		Narrow	Expanded	-0.01	-0.02	0.02	0.03
	Pairwise		Narrow	Expanded	-0.01	0.14	0.00	0.05
2-Dim	Separate	None	Narrow	Expanded	0.02	-0.02	-0.02	0.04
		Small	Narrow	Expanded	0.01	0.01	0.00	0.01
		Moderate	Narrow	Expanded	0.02	0.07	0.10	-0.11
		Large	Narrow	Expanded	-0.01	-0.14	0.02	0.16
	Pairwise	None	Narrow	Expanded	0.02	0.15	-0.03	0.08
		Small	Narrow	Expanded	0.01	-0.19	0.02	0.02
		Moderate	Narrow	Expanded	0.02	-0.70	0.01	0.14
		Large	Narrow	Expanded	-0.01	0.09	-0.03	-0.03
3-Dim	Separate	None	Narrow	Expanded	0.02	0.00	-0.03	0.00
		Small	Narrow	Expanded	0.00	-0.01	0.02	-0.01
		Moderate	Narrow	Expanded	0.01	0.04	0.02	-0.04
		Large	Narrow	Expanded	0.00	-0.03	0.00	0.00
	Pairwise	None	Narrow	Expanded	0.01	0.28	0.03	0.11
		Small	Narrow	Expanded	-0.01	0.33	0.01	0.13
		Moderate	Narrow	Expanded	0.01	-0.09	0.02	0.20
		Large	Narrow	Expanded	0.02	-0.18	-0.06	0.11

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C24. Grade 9 Comparisons of Grade Level Separation

Ability Generation	Scaling Method	Common Item set	Format Effect	Grade Level Separation		Mean BIAS Mean *diff*	Mean RMSE Mean *diff*
Uni	Separate	Narrow		Small	Large	1.38*	-0.68*
		Expanded		Small	Large	1.37*	-0.63*
	Pairwise	Narrow		Small	Large	0.98*	-0.71*
		Expanded		Small	Large	0.92*	-0.67*
2-Dim	Separate	Narrow	None	Small	Large	1.40*	-0.69*
			Small	Small	Large	1.47*	-0.68*
			Moderate	Small	Large	1.86*	-0.49*
			Large	Small	Large	3.40*	-1.12*
		Expanded	None	Small	Large	1.36*	-0.62*
			Small	Small	Large	1.47*	-0.68*
			Moderate	Small	Large	1.91*	-0.69*
			Large	Small	Large	3.27*	-0.98*
	Pairwise	Narrow	None	Small	Large	0.76	-0.73*
			Small	Small	Large	0.62	-0.64*
			Moderate	Small	Large	1.84*	-0.68*
			Large	Small	Large	3.29*	-1.16*
		Expanded	None	Small	Large	0.91*	-0.62*
			Small	Small	Large	0.42	-0.64*
			Moderate	Small	Large	1.13*	-0.55*
			Large	Small	Large	3.39*	-1.16*
3-Dim	Separate	Narrow	None	Small	Large	0.81*	-0.21
			Small	Small	Large	0.86*	-0.17
			Moderate	Small	Large	1.03*	-0.15
			Large	Small	Large	1.46*	-0.33
		Expanded	None	Small	Large	0.80*	-0.18
			Small	Small	Large	0.85*	-0.20
			Moderate	Small	Large	1.06*	-0.21
			Large	Small	Large	1.43*	-0.33
	Pairwise	Narrow	None	Small	Large	0.27	-0.24
			Small	Small	Large	0.23	-0.29
			Moderate	Small	Large	0.66	-0.44
			Large	Small	Large	0.86*	-0.44
		Expanded	None	Small	Large	0.54	-0.16
			Small	Small	Large	0.57	-0.18
			Moderate	Small	Large	0.56	-0.26
			Large	Small	Large	0.66	-0.27

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C25. Grade 9 Comparisons of Data Generation Models

Scaling Method	Common Item Set	Format Effect	Ability Generation		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Separate	Narrow	None	Uni	2-Dim	0.00	0.01	0.01	0.01
		None		3-Dim	0.08	-0.50	0.56*	1.04*
			2-Dim	3-Dim	0.08	-0.51	0.55*	1.03*
		Small	2-Dim	3-Dim	0.08	-0.54	0.64*	1.15*
		Moderate	2-Dim	3-Dim	0.03	-0.80*	1.35*	1.69*
		Large	2-Dim	3-Dim	0.05	-1.89*	2.11*	2.90*
	Expanded	None	Uni	2-Dim	0.02	0.01	0.01	0.02
		None		3-Dim	0.10	-0.47	0.55*	1.00*
			2-Dim	3-Dim	0.09	-0.48	0.55*	0.99*
		Small	2-Dim	3-Dim	0.06	-0.55	0.65*	1.13*
		Moderate	2-Dim	3-Dim	0.02	-0.84*	1.27*	1.76*
		Large	2-Dim	3-Dim	0.06	-1.78*	2.10*	2.75*
Pairwise	Narrow	None	Uni	2-Dim	0.00	-0.23	0.02	0.00
		None		3-Dim	0.07	-0.64	0.62*	1.09*
			2-Dim	3-Dim	0.07	-0.42	0.60*	1.09*
		Small	2-Dim	3-Dim	0.07	-0.33	0.70*	1.05*
		Moderate	2-Dim	3-Dim	-0.03	-1.21*	1.30*	1.53*
		Large	2-Dim	3-Dim	-0.11	-2.54*	2.10*	2.82*
	Expanded	None	Uni	2-Dim	0.02	-0.22	-0.02	0.04
		None		3-Dim	0.08	-0.50	0.64*	1.15*
			2-Dim	3-Dim	0.07	-0.28	0.65*	1.11*
		Small	2-Dim	3-Dim	0.04	0.19	0.69*	1.16*
		Moderate	2-Dim	3-Dim	-0.04	-1.31*	1.31*	1.59*
		Large	2-Dim	3-Dim	-0.08	-2.81*	2.07*	2.96*

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C26. Grade 10 Comparisons of Item Format Effects

Ability Generation	Scaling Method	Common Item Set	Format Effect		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
2-Dim	Separate	Narrow	None	Small	-0.04	0.14	-0.13	-0.21
				Moderate	-0.05	0.61	-0.86*	-0.92*
				Large	-0.31	2.02*	-1.76*	-2.54*
			Small	Moderate	-0.01	0.46	-0.73*	-0.72*
				Large	-0.27	1.88*	-1.63*	-2.33*
			Moderate	Large	-0.26	1.41*	-0.90*	-1.62*
		Expanded	None	Small	0.01	0.11	-0.05	-0.15
				Moderate	-0.02	0.79	-0.76*	-1.09*
				Large	-0.28	1.91*	-1.71*	-2.39*
			Small	Moderate	-0.03	0.68	-0.71*	-0.94*
				Large	-0.29	1.80*	-1.65*	-2.24*
			Moderate	Large	-0.27	1.12*	-0.95*	-1.30*
	Pairwise	Narrow	None	Small	-0.03	-0.27	-0.15	-0.10
				Moderate	-0.05	1.17*	-0.86*	-0.82*
				Large	-0.37	2.26*	-1.76*	-2.04*
			Small	Moderate	-0.02	1.44*	-0.72*	-0.72*
				Large	-0.34	2.52*	-1.62*	-1.93*
			Moderate	Large	-0.32	1.08*	-0.90*	-1.22*
		Expanded	None	Small	0.01	-0.70	-0.10	-0.14
				Moderate	-0.03	0.12	-0.83*	-0.80*
				Large	-0.37	2.16*	-1.77*	-2.14*
			Small	Moderate	-0.05	0.83	-0.73*	-0.66*
				Large	-0.38	2.87*	-1.67*	-2.00*
			Moderate	Large	-0.33	2.04*	-0.94*	-1.34*
3-Dim	Separate	Narrow	None	Small	0.02	0.05	-0.03	-0.03
				Moderate	0.05	0.39	-0.22	-0.24
				Large	0.08	1.04*	-0.47*	-0.74*
			Small	Moderate	0.03	0.34	-0.19	-0.21
				Large	0.07	0.99	-0.43	-0.71*
			Moderate	Large	0.04	0.65	-0.25	-0.50*
		Expanded	None	Small	-0.02	0.09	-0.01	-0.07
				Moderate	0.00	0.46	-0.20	-0.33
				Large	0.04	0.98	-0.39	-0.70*
			Small	Moderate	0.02	0.37	-0.19	-0.26
				Large	0.06	0.89	-0.38	-0.64*
			Moderate	Large	0.05	0.52	-0.19	-0.38
	Pairwise	Narrow	None	Small	0.02	-0.04	-0.05	-0.11
				Moderate	0.03	-0.51	-0.20	-0.40
				Large	0.03	0.70	-0.41	-0.55*
			Small	Moderate	0.01	0.55	-0.15	-0.30
				Large	0.01	0.74	-0.36	-0.45*
			Moderate	Large	0.01	0.19	-0.21	-0.15
		Expanded	None	Small	0.02	0.36	-0.06	-0.07
				Moderate	0.02	-0.38	-0.22	-0.31
				Large	0.06	0.47	-0.44	-0.53*
			Small	Moderate	0.01	0.02	-0.16	-0.23
				Large	0.04	0.11	-0.39	-0.45*
			Moderate	Large	0.04	0.09	-0.23	-0.22

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C27. Grade 10 Comparisons of Vertical Scaling Set

Ability Generation	Common Item Set	Format Effect	Scaling Method		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Narrow		Separate	Pairwise	0.04	-1.03*	0.21	0.48*
	Expanded		Separate	Pairwise	0.04	-1.01*	0.22	0.48*
2-Dim	Narrow	None	Separate	Pairwise	0.04	-1.32*	0.19	0.43*
		Small	Separate	Pairwise	0.05	-1.73*	0.18	0.53*
		Moderate	Separate	Pairwise	0.03	-0.76*	0.19	0.53*
		Large	Separate	Pairwise	-0.03	-1.09*	0.19	0.93*
	Expanded	None	Separate	Pairwise	0.03	-1.10*	0.22	0.50*
		Small	Separate	Pairwise	0.03	-1.91*	0.17	0.51*
		Moderate	Separate	Pairwise	0.01	-1.77*	0.16	0.79*
		Large	Separate	Pairwise	-0.05	-0.85*	0.16	0.75*
3-Dim	Narrow	None	Separate	Pairwise	-0.02	-0.96*	0.17	0.27
		Small	Separate	Pairwise	-0.02	-1.05*	0.15	0.19
		Moderate	Separate	Pairwise	-0.04	-0.84*	0.18	0.10
		Large	Separate	Pairwise	-0.07	-0.96*	0.22	0.46*
	Expanded	None	Separate	Pairwise	-0.06	-0.92*	0.21	0.36
		Small	Separate	Pairwise	-0.02	-0.65	0.16	0.35
		Moderate	Separate	Pairwise	-0.03	-1.00*	0.19	0.38
		Large	Separate	Pairwise	-0.04	-1.43*	0.15	0.54*

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C28. Grade 10 Comparisons of Common Item Set

Ability Generation	Scaling Method	Format Effect	Common Item Set		Mean BIAS		Mean RMSE	
					Small Separation	Large Separation	Small Separation	Large Separation
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Uni	Separate		Narrow	Expanded	0.00	-0.09	-0.01	0.07
	Pairwise		Narrow	Expanded	0.00	0.17	0.00	0.07
2-Dim	Separate	None	Narrow	Expanded	-0.03	-0.02	-0.05	-0.01
		Small	Narrow	Expanded	0.02	-0.06	0.03	0.05
		Moderate	Narrow	Expanded	0.01	0.16	0.05	-0.18
		Large	Narrow	Expanded	0.00	-0.13	0.00	0.14
	Pairwise	None	Narrow	Expanded	-0.03	0.20	-0.03	0.06
		Small	Narrow	Expanded	0.01	-0.24	0.02	0.02
		Moderate	Narrow	Expanded	-0.01	-0.85	0.01	0.08
		Large	Narrow	Expanded	-0.02	0.11	-0.03	-0.04
3-Dim	Separate	None	Narrow	Expanded	0.04	0.00	-0.03	0.02
		Small	Narrow	Expanded	0.01	0.04	0.00	-0.02
		Moderate	Narrow	Expanded	0.00	0.07	-0.01	-0.07
		Large	Narrow	Expanded	0.01	-0.06	0.05	0.06
	Pairwise	None	Narrow	Expanded	0.01	0.04	0.02	0.11
		Small	Narrow	Expanded	0.00	0.44	0.01	0.14
		Moderate	Narrow	Expanded	0.00	-0.09	0.00	0.21
		Large	Narrow	Expanded	0.03	-0.19	-0.02	0.14

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C29. Grade 10 Comparisons of Grade Level Separation

Ability Generation	Scaling Method	Common Item set	Format Effect	Grade Level Separation		Mean BIAS	Mean RMSE
						Mean *diff*	Mean *diff*
Uni	Separate	Narrow		Small	Large	2.37*	-0.82*
		Expanded		Small	Large	2.28*	-0.74*
	Pairwise	Narrow		Small	Large	1.30*	-0.55*
		Expanded		Small	Large	1.47*	-0.48*
2-Dim	Separate	Narrow	None	Small	Large	2.28*	-0.82*
			Small	Small	Large	2.47*	-0.89*
			Moderate	Small	Large	2.94*	-0.88*
			Large	Small	Large	4.62*	-1.60*
		Expanded	None	Small	Large	2.29*	-0.78*
			Small	Small	Large	2.39*	-0.87*
			Moderate	Small	Large	3.10*	-1.11*
			Large	Small	Large	4.48*	-1.46*
	Pairwise	Narrow	None	Small	Large	0.93	-0.58*
			Small	Small	Large	0.70	-0.54*
			Moderate	Small	Large	2.16*	-0.54*
			Large	Small	Large	3.56*	-0.86*
		Expanded	None	Small	Large	1.16*	-0.49*
			Small	Small	Large	0.44	-0.54*
			Moderate	Small	Large	1.32*	-0.47*
			Large	Small	Large	3.69*	-0.87*
3-Dim	Separate	Narrow	None	Small	Large	1.22*	-0.06
			Small	Small	Large	1.25*	-0.06
			Moderate	Small	Large	1.56*	-0.08
			Large	Small	Large	2.17*	-0.34
		Expanded	None	Small	Large	1.17*	-0.02
			Small	Small	Large	1.28*	-0.08
			Moderate	Small	Large	1.64*	-0.15
			Large	Small	Large	2.11*	-0.34
	Pairwise	Narrow	None	Small	Large	0.28	0.04
			Small	Small	Large	0.22	-0.02
			Moderate	Small	Large	0.76	-0.16
			Large	Small	Large	0.94	-0.11
		Expanded	None	Small	Large	0.31	0.13
			Small	Small	Large	0.66	0.12
			Moderate	Small	Large	0.67	0.04
			Large	Small	Large	0.72	0.05

* significant at $p < 0.05$ and Cohen's $d > 0.20$

Table C30. Grade 10 Comparisons of Data Generation Models

| Scaling Method | Common Item Set | Format Effect | Ability Generation | | Mean BIAS | | Mean RMSE | |
| | | | | | Small Separation | Large Separation | Small Separation | Large Separation |
					Mean *diff*	Mean *diff*	Mean *diff*	Mean *diff*
Separate	Narrow	None	Uni	2-Dim	0.04	-0.04	0.04	0.05
				3-Dim	0.23	-0.92	0.63*	1.39*
			2-Dim	3-Dim	0.19	-0.88	0.58*	1.34*
		Small	2-Dim	3-Dim	0.25	-0.97	0.69*	1.52*
		Moderate	2-Dim	3-Dim	0.29	-1.10*	1.23*	2.02*
		Large	2-Dim	3-Dim	0.58	-1.86*	1.88*	3.14*
	Expanded	None	Uni	2-Dim	0.02	0.03	0.00	-0.03
				3-Dim	0.28	-0.83	0.61*	1.33*
			2-Dim	3-Dim	0.26	-0.86	0.61*	1.37*
		Small	2-Dim	3-Dim	0.23	-0.88	0.65*	1.45*
		Moderate	2-Dim	3-Dim	0.28	-1.19*	1.17*	2.13*
		Large	2-Dim	3-Dim	0.59	-1.79*	1.93*	3.05*
Pairwise	Narrow	None	Uni	2-Dim	0.04	-0.33	0.02	-0.01
				3-Dim	0.18	-0.85	0.58*	1.17*
			2-Dim	3-Dim	0.14	-0.52	0.56*	1.18*
		Small	2-Dim	3-Dim	0.19	-0.29	0.66*	1.18*
		Moderate	2-Dim	3-Dim	0.22	-1.18*	1.22*	1.60*
		Large	2-Dim	3-Dim	0.54	-2.08*	1.91*	2.66*
	Expanded	None	Uni	2-Dim	0.01	-0.30	0.00	-0.02
				3-Dim	0.18	-0.74	0.60*	1.21*
			2-Dim	3-Dim	0.17	-0.68	0.60*	1.23*
		Small	2-Dim	3-Dim	0.18	0.39	0.64*	1.29*
		Moderate	2-Dim	3-Dim	0.23	-0.42	1.21*	1.72*
		Large	2-Dim	3-Dim	0.60	-2.37*	1.92*	2.84*

* significant at $p < 0.05$ and Cohen's $d > 0.20$

REFERENCES

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-24.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87-96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17,* 239-251.

Baker, F. B. (1997). Empirical sampling distributions of equating coefficients for graded and response instruments. *Applied Psychological Measurement, 21*, 157-172.

Baker, F. B. & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Becker, D. F., & Forsyth, R. A. (1992). An Empirical Investigation of Thurstone and IRT Methods of Scaling Achievement Tests. *Journal of Educational Measurement, 29*(4), 341-354.

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000). Effect of multidimensionality on separate and concurrent estimation in IRT equating. Downloaded from http://www.b-a-h.com/papers/paper0002.html.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14*, 151-162.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77-92.

Briggs, D. C. (2011, December). Making inferences about growth and value-added: Design issues for the PARCC consortium. [White paper]. Retrieved from http://www.parccon line.org/sites/parcc/files/BriggsPARCCGrowthFINAL022412.pdf.

Briggs, D. C. & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3-14.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO 2.1: Flexible, multidimensional, multiple categorical IRT modeling. [Computer software]. Chicago, IL: Scientific Software International.

Camilli, G., Yamamoto, K., & Meng, M-m. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379-388.

Cao, Y. (2008). Mixed-Item Format Equating: Effects of Test Dimensionality and Common-Item Sets. Dissertation University of Maryland.

Chen, F.F., Hayes, A., Carver, C. S., Laurenceau, J.P., & Zhang, Z. (2012). Modeling General and Specific Variance in Multifaceted Constructs: A Comparison of the Bifactor Model to Other Approaches. Journal of Personality, 80(1), 219-251.

Chin, T-Y., Kim, W., & Nering, M. L. (2006, April). Five statistical factors that influence IRT vertical scaling. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Cohen, A. S. & Kim, S. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement, 22*, 116-130.

Cohen, J. (1998). *Statistical Power Analysis for the Behavioral Sciences* (revised ed.). Hillsdale, NJ: Erlbaum.

Dorans, N. J. & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*(4), 249-262.

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*(2), 189-199.

Dunbar, Koretz, & Hoover, (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289-304.

Eastwood, M. (2014). The effects of construct shift and model-data misfit on estimates of growth using vertical scales. (Unpublished Doctoral Dissertation). University of Connecticut, Storres, CT.

Ercikan, K., Schwarz, R., Julian, M. W., Burket, G. R., Weber, M. W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response test item types. *Journal of Educational Measurement, 35*, 137-154.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement 33*(3), 291-314.

Fukuhara, H., & Kamata, A. (2012) A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement, 35*, 604-612.

Gibbons, R. D. & Hedeker, D. R. (1992). Full-information item bi-factor analysis. P*sychometrika, 57*(3), 423-436.

Gibbons, R. D., Bock, R. D., Hedecker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4-19.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144-149.

Hagee, S. L. & Kolen, M. J. (2011). Equating mixed-format tests with format representative and non-representative common items. In M. J. Kolen, & W.-C. Lee (Eds.), *Mixed-format tests: Psychometric properties with a primary focus on equating Volume 1* (pp. 95-123). University of Iowa, Iowa City, IA: Center for Advanced Studies in Measurement and Assessment (CASMA).

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Luwer-Nijhoff Publishing.

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

Hanson, B. A. & Béguin, A. A. (1999). *Separate versus concurrent calibration estimation of IRT parameters in the common item equating design*. ACT Research Report 99-8. Iowa City, IA: ACT, Inc.

Harris, D. J. (2007). Practical issues in vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233-251). New York, NY: Springer Science + Business Media.

Harris, D. J., & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement In Education, 6*(3), 195-240.

Harris, D. J., & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*(2), 151-159.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2), 91-115.

Hendrickson, Cao, Chin, and Lee, (2006, April). *Effect of base year on IRT vertical scaling from the common-item design*. Pater presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Holland, P. W. (2002). Two measures of change in the gaps between CDFs of test-score distributions. *Journal of Educational Behavioral Statistics, 27*(1), 3-18.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19*(2), 139-147.

Ito, K., Sykes, R. C., & Lao, Y. (2008). Concurrent and separate grade-groups linking procedures for vertical scaling. *Applied Measurement in Education, 21,* 187-206.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71*(3), 229-250.

Kamata, A. & Tate, R. (2005). The performance of a method for the long-term equating of mixed-format assessment. *Journal of Educational Measurement, 42*, 193-213.

Karkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003, April). Separate versus concurrent calibration methods in vertical scaling. Paper presented at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*(1), 25-41.

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143.

Kim, S. & Kolen, M. J. (2004). STUIRT: A computer program for scale transformation under unidimensional item response theory models. Iowa City, IA: Iowa Testing Programs, The University of Iowa (Available from: http://www.education.uiowa.edu/casma.

Kim, S. & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19,* 357-381.

Kim, S. & Lee, W. (2004). IRT scale linking methods for mixed format tests. *ACT Research Paper Series 2004-5*, Iowa City, IA: ACT, Inc.

Kim, J., Lee, W., & Kim, D. (2008 March). *The effect of choosing a base grade on the vertical scale using various IRT calibration methods.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York City, NY.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among designs for equating mixed-format tests in large-scale assessments. *Journal of Educational Measurement, 47*(1), 36-53.

Koepfler, J. R. (2012). Examining the bifactor IRT model for vertical scaling in K-12 assessment. (Unpublished Doctoral Dissertation). James Madison University, Harrisonburg, VA.

Kolen, M. J. (2011, September). Issues associated with vertical scales for PARCC assessments. [White paper]. Retrieved from http://www.parcc_online_.org/sites/par_cc/files/PAR_CC VertScal289-12-201129.pdf.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp.155-186). Westport, CT: American Council on Education and Praeger Publishers.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking* (2nd ed.). New York, NY: Springer Science and Business Media, Inc.

Kolen, M. J. & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice, 29*(3), 8-14.

Lane, S., & Stone, C. A. (2006). Performance Assessment. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp.387-431). Westport, CT: American Council on Education and Praeger Publishers.

Li, T. (2006). The effect of dimensionality on vertical scaling. (Unpublished Doctoral Dissertation). Michigan State University, East Lansing, MI.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21.

Li, Y. H., Lissitz, R. W., & Yang, N. (1999, April). Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items. Paper presented at the Annual Meeting of the National Council on Measurement, Montreal, Quebec, Canada.

Li, Y., & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement, 36*(1), 3-20.

Lin, P., Wei, H., & Lissitz, R. W. (2007, April). Equivalent test structure across grades: A multi-group confirmatory factor analysis approach. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 179-193.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*(3), 234-250.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

McKinley, R. L., & Reckase, M. D. (1983). An extension of the two-parameter logistic model to the multidimensional item space. (Research Rep. ONR 83-2). Iowa City, IA: The American College Testing Program. Retrieved from ERIC database. (ED241581).

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-format tests in vertical scaling.* (Unpublished Doctoral Dissertation). The University of Iowa, Iowa City, IA.

Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*(1), 73-90.

Muraki, E., Hombo, C. M., & Lee, Y.-W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325-337.

Ostini, R. & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage Publishing.

Partnership for Assessment of Readiness for College and Careers (PARCC). (2013, September). PARCC frequently asked questions. Retrieved from: http://ww w.parc conline .org/sites/ parcc/files/PARCCFAQ_9-18-2013.pdf.

Patz, R. J., & Yao, L. (2007). Methods and models for vertical scaling. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp.253-272). New York, NY: Springer Science + Business Media.

Perkhounkova, Y., & Dunbar, S. B. (1999, April). *Influences of item content and format on the dimensionality of tests combining multiple-choice and open-response items: An application of the Poly-DIMTEST procedure.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Peterson, N. S. (2010). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp.59-72). New York, NY: Springer Science + Business Media.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New York, NY: Springer Science + Business Media.

Reckase, M. D. (1979). Unifactor latent trait model applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230.

Reise, S. P. (2012). The rediscovery of the bifactor measurement model. *Multivariate Behavioral Research, 47*, 667-696.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.

Rijmen, F. (2010). Formal realtions and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.

Rindskopf, D. & Rose, T. (1988). Some theory and applications of confirmatory second order factor analysis. *Multivariate Behavioral Research, 23*(1), 51-67.

Rosenthal, R. & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw Hill.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Skaggs, G. & Lissitz, R. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*(4), 495-529.

Slinde, J. A. & Linn, R. L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement 16*(3), 159-165).

Smarter Balanced Assessment Consortium (SMARTER). (2012, February). *Master Work Plan Narrative – Summative.* Retrieved from: http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/03/Summative-Assessment-Master-Work-Plan-Narrative.pdf

Smith, Z. R., Finkelman, M., Nering, M. L., & Kim, W. (2008, March). *Vertical scaling: A comparison of linking methods with unidimensional and multidimensional data.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002, April). *Multidimensionality and the equating of a mixed-format math examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Sykes, R. C. & Yen, W. M. (2000). The scaling of mixed format tests with the one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 37*(3), 221-244.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed-response items. *Journal of Educational Measurement, 37*(4), 329-346.

Tate, R. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36*(4), 336-346.

Thissen, D. (1991). MULTILOG 6.3. [Computer software]. Chicago, IL: Scientific Software International.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional then multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*(2), 113-123.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*(7), 433-451.

Tong, Y., & Kolen, M. J. (2010). Scaling: An ITEMS module. *Educational Measurement: Issues and Practice, 29*(4), 39-48.

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education, 20*(2), 227-253.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp. 29-44). Hillsdale, NJ: Erlbaum.

U.S. Department of Education. (2009, November). *Race to the Top Program Executive Summary*. Retrieved from: http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

U.S. Department of Education. (2010, April). *Race to the Top Assessment Program Executive Summary*. Retrieved from: http://www2.ed.gov/programs/racetothetop assessment/executive-summary-042010.pdf.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Williams, V. S. L., Pommerich, M., & Thissen, D. (1998). A Comparison of Developmental Scales Based on Thurstone Methods and Item Response Theory. *Journal of Educational Measurement, 35*(2), 93-107.

Yao, C. (2008). *Mixed-format test equating: effects of test dimension and common-item sets.* Unpublished Doctoral Dissertation, University of Maryland, College Park.

Yao, C. & Mao, (2004, April). Unidimensional and multidimensional estimation of vertical scaled tests with complex structure. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Yao, L, & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*(4), 299-325.

Yen, W. M., & Burket, G. R. (1997). Comparison of item response theory and Thurstone methods of vertical scaling. *Journal of Educational Measurement, 34*(4), 293-313.

Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling*. Unpublished Doctoral Dissertation, Michigan State University, Lansing.