Simulating the Effects of Test Score Reliability and Test Dimensionality on Teacher Value-Added Scores and Inferences

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Danielle N. Dupuis

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Michael C. Rodriguez, Adviser

December, 2015

Acknowledgements

First, I would like to first thank my adviser, Michael Rodriguez, for his thoughtful mentorship and unwavering support of my professional development and success as a scholar. Thanks also go to Michael Harwell, Asha Jitendra, and Thomas Post, members of my dissertation committee, for their valuable feedback throughout the course of my dissertation work, but also for their own unwavering commitment to my success as a scholar.

I'd also like to thank Robert Jorczak, for constantly pushing me to think more critically, and to question everything. I would not be where I am today without you. Finally, thanks also go to Caroline Hilk, Stacy Karl, and Michael Mensink, for your friendship and support throughout graduate school and beyond.

Dedication

To my family, with love, for always believing in me, even when I didn’t.

Abstract

A teacher evaluation system that has gained in popularity is value-added assessment (VAA). In VAA, student test scores are used in an attempt to isolate the effect of individual teachers on student learning. Teacher value-added scores are interpreted as measures of teacher effectiveness, and are now widely used to evaluate teachers, and in pay-for-performance programs.

Despite a body of literature on a variety of VAA issues, very little attention has been given to the psychometric properties of the student achievement tests used to derive teacher value-added scores, and the implications of violations to model assumptions on using value-added scores to make decisions about teachers. The purpose of the present study was to examine the potential bias produced in estimating teacher effectiveness (i.e., in teacher value-added scores) when there are violations to (a) the assumption of unidimensionality in the measurement model used to estimate student achievement scores, and (b) the assumption that control variables are free of measurement error in the statistical models used to estimate teacher value-added scores. An additional purpose was to examine the effect(s) of model assumption violations on the use of teacher value-added scores to make high-stakes decisions about teachers' relative effectiveness.

The results showed that teacher value-added scores were differentially biased by variations in the properties of the student achievement tests used to derive the value-added scores. Both bias in the estimation of teacher value-added scores, and the consistency of teachers' estimated relative effectiveness were most sensitive to variations in the specification of test dimensionality. The results strongly caution against the use of value-added scores to make high-stakes decisions about teachers.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

The accountability movement in U.S. public education has a long history dating to the Elementary and Secondary Education Act of 1965, which was reauthorized in 1994 as the Improving America's Schools Act. More recently, the No Child Left Behind Act (NCLB, 2002) compelled states to develop school accountability systems based on yearly assessment of student achievement. NCLB required states to set achievement standards in the domains of reading, mathematics, and science; and to create assessments of knowledge and skills defined by the achievement standards. In addition, NCLB called on educators to increase student learning by increasing the number of "highly qualified" teachers in the classroom. One reaction to concerns about teacher quality has been to develop new teacher evaluation systems.

A teacher evaluation system that has gained in popularity is value-added assessment (VAA, also knows as value-added modeling). In VAA, student test scores are used in an attempt to isolate the effect of individual teachers on student learning. Teacher value-added scores are interpreted as measures of teacher effectiveness, and are now widely used to evaluate teachers, and in pay-for-performance programs. For example, Hill, Kapitula, and Umland (2011) report that nearly one-fourth of the member districts of the Council of the Great City Schools (an organization comprised of the nation's largest public school systems) have implemented some form of a VAA-based teacher or school accountability system. Further, the 2009 Race to the Top contest required states to develop teacher value-added accountability systems to be eligible for participation. Several states had laws previously banning the use of student test score data to evaluate teachers, but those laws were subsequently overturned. Forty-six states submitted at least one Race to the Top application, suggesting unprecedented and widespread use of VAA in teacher evaluation and accountability nationwide.

Despite widespread use, growing evidence suggests that the use of student test score data to evaluate teachers suffers from many problems. Practical problems include the availability of appropriate tests (particularly in subjects other than mathematics and reading), issues associated with the attribution of teacher effects to a specific teacher (e.g., because of team teaching), and summer learning gain/loss. Methodological problems include the non-random sorting of students and teachers to classrooms, the adequacy of model specification, and the imprecision and instability of teacher value-added scores.

Despite a body of literature on a variety of VAA issues, very little attention has been given to the psychometric properties of the student achievement tests used to derive teacher value-added scores, and the implications of violations to model assumptions on using value-added scores to make decisions about teachers. For VAA to produce unbiased estimates of teacher effectiveness, the assumptions of the measurement and statistical models used to derive such estimates must be met. The purpose of the present study is to examine the potential bias produced in estimating teacher effectiveness (i.e., in teacher value-added scores) when there are violations to (a) the assumption of unidimensionality in the measurement model used to estimate student achievement scores, and (b) the assumption that control variables are free of measurement error in the statistical models used to estimate teacher value-added scores. An additional purpose is to examine the effect(s) of model assumption violations on the use of teacher value-added scores to make high-stakes decisions about teachers' relative effectiveness.

The majority of measurement models used in educational scaling applications, including VAA, assume that an examinee's performance on an achievement test is due to a single trait (i.e., unidimensional), despite agreement that multiple traits contribute to performance on most achievement tests (Birenbaum & Tatsuoka, 1982). The presence of multidimensionality in educational achievement data has lead to the development of multidimensional item response theory (IRT) models (Reckase, 1979), but the complexity of score estimation makes the application of these models impractical in many contexts, leaving the presence of test multidimensionality largely ignored. To date, no study has evaluated the effects of ignoring test multidimensionality on teacher value-added scores and inferences using real or simulated data. The lack of work examining the effects of test multidimensionality on value-added scores is particularly problematic because Martineau (2006) demonstrated mathematically that ignoring test multidimensionality can bias value-added scores and distort subsequent inferences. Furthermore, whereas states and districts have traditionally used end-of-year state achievement tests to estimate teacher effectiveness, some states are beginning to use formative assessments (e.g., curriculum-based measures) to measure teacher effectiveness. Typically, end-of-year state achievement tests show relatively high correlations (e.g., .70-.80) between subtests comprising a composite score, whereas formative assessments typically show much lower correlations between subtests comprising a composite score (e.g., .30-.50) – making test dimensionality an increased concern.

In addition to test dimensionality, the presence of measurement error in the student achievement tests used to derive value-added scores has the potential to bias value-added scores and alter subsequent inferences. Nearly all statistical models used in VAA include one or more pretest variables for the purpose of controlling for students' prior achievement. These statistical models assume that the pretest variables (and all control variables) are measured without error. However, student achievement tests always contain some measurement error. Not controlling for measurement error in the pretest variables will bias all model parameters, including the teacher value-added scores. Bias occurs because the measurement error in the pretest variables causes the pretest variables to be correlated with the model's residual terms. The reliability of the end-of-year state achievement tests used mostly frequently in VAA is often (by design) high (e.g., > .90) and considered adequate for making high-stakes decisions about students. In contrast, the reliability of formative assessments now being used by some states to evaluate K-3 teachers is usually considerably lower (e.g., .70-.80), often due to the relatively brief nature of formative assessments. As such, the potential bias produced in value-added scores by the presence of measurement error is of increasing concern.

In summary, the purpose of the present study is to examine the extent to which test multidimensionality and test measurement error bias teacher value-added scores and distort subsequent inferences about teachers' relative effectiveness. In the following chapter the literature on test dimensionality and test measurement error as it relates to value-added scores is discussed; and research questions and hypotheses are presented. In Chapter 3 a description of the method used to address the research questions is provided. In Chapter 4 the results of the study are described; and in Chapter 5 the results of the study are discussed in the context of score bias and use, and suggestions for future research are provided.

Chapter II: Literature Review

Introduction

Value-added assessment is a method of teacher evaluation that seeks to measure the effect of individual teachers on students' learning. Specifically, in value-added assessment student test score data and statistical regression models are used to estimate the portion of variance in student achievement scores attributed to classrooms, and by extension teachers. The term *value-added assessment* is used here, as opposed to the more common *value-added modeling*, to highlight that using student test score data to evaluate teachers is fundamentally an assessment process, and that teacher value-added scores should be held to the same standards as student achievement scores (American Educational Research Association, 2015). From an assessment perspective, there are a number of problems associated with using student test scores to evaluate teachers; these include: issues related to the timing of test administration, issues associated with score inflation and "teaching to the test," issues related to score scale construction and score scale properties (e.g., test dimensionality), and the effect of test measurement error on model estimates (McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

Problems related to the timing of test administration arise because most tests used in value-added assessment (VAA) are administered relatively infrequently (usually once in spring each year). As such, estimates of students' grade-to-grade growth often capture learning in two grades as well as learning gains/losses associated with the summer break, making it difficult to isolate the effect of an individual teacher on student learning. Score inflation is of particular concern when student test scores are used to evaluate teachers, as teachers feel pressure to "teacher to the test;" the result is an increase in scores and a narrowing of the implemented curriculum to match the test.

An issue of scale construction relevant to VAA involves the indeterminacy of scales connected to tests of student achievement. McCaffrey et al. (2003) argue that even though the choice of scale for measuring student achievement is arbitrary (i.e., indeterminate), the use of achievements scales in VAA assumes an interval scale of measurement. Additionally, the authors argue that "the inferences rest not only on the assumption that the scale of student performance is interval but also that this maps linearly to the latent scale of effectiveness, so that the scale of teacher effectiveness is also interval" (p. 101).

The problems associated with the indeterminacy of achievement scales are compounded when linking tests that differ in content and difficulty to create a vertical scale. Briggs and Weeks (2009) examined the influence of decisions made in creating a vertical score scale on estimates of grade-to-grade growth and school value-added scores using data from the Colorado State Assessment Program's (CSAP) test of reading. The authors collected four years of data from two cohorts of students in grades 3-7 throughout the state of Colorado. The operational vertical score scale for the CSAP reading test (grades 3-10) was created using a common-item, non-equivalent groups design and separate calibration. Scores were scaled using the 3-parameter logistic item response theory (IRT) model and maximum likelihood estimation was used to estimate person scores. Parallel tests are created each year and equated to the original vertical score scale.

To examine the influence of vertical scaling decisions on school value-added scores, Briggs and Weeks (2009) created eight vertical scales (by varying the IRT model, calibration method, and person-score estimation method), and then fit data from each vertical score scale to the layered value-added model. To examine the effect of these factors on value-added scores and inferences the authors computed correlations between the value-added score estimates from each of the eight vertical scales created. The correlations ranged from .79 to .99 with a mean of .95, which suggests that the eight scales produced similar estimates of value-added school effectiveness. The authors also classified each school as above average, average, or below average using the school value-added score estimates and then compared the classification consistency across the eight scales. The results indicated that nearly 95% of schools were consistently classified as above or below average. These results suggest that estimates of value-added school effectiveness are somewhat insensitive to the choice of vertical scale.

The Briggs and Weeks (2009) result is an important first step in understanding the effects of scaling decisions on value-added scores and inferences, but the study suffers from a number of important limitations. First, the authors only examined reading achievement; it is unclear if these results generalize to math and science achievement, particularly because math and science content tends to align more closely with curricula than reading, which is likely to produce different estimates of student growth. Second, whether the results were sensitive to the choice of scaling design and the characteristics of the common items used in constructing the operational vertical scale, as well as to the value-added model fitted is unknown. Third, the authors examined *school* value-added scores and whether these results would generalize to *teacher* value-added scores is unclear. Finally, the authors assumed that the construct of reading achievement maintains a unidimensional structure over time and across grades. This assumption is defensible and even useful for many applications of vertical scaling, however, it may not be defensible when applied to VAA. Several researchers have noted that the nature of the construct being measured, as well as the extent to which it aligns with the specific content and skills taught in the classroom may affect value-added scores (Lockwood, McCaffrey, Hamilton, Stecher, & Martinez, 2007; McCaffrey et al., 2003; Martineau, 2006; Papay, 2011; Sass, 2008). The alignment between the content taught and the content assessed is particularly important when growth is defined grade-to-grade such that only grade appropriate content is assessed, as differences in the relative weight given to specific content and skills on a test versus the emphasis placed on those content and skills in the classroom will affect the extent to which a test accurately captures instruction. These issues are exaggerated in schools that track students based on their ability level; and also when the dimensional structure of the construct of interest shifts across grade-level tests (Martineau, 2006).

Definition of Test Dimensionality

There are two definitions or perspectives from which to consider test dimensionality: psychological dimensionality and statistical dimensionality (Reckase, 1990). The psychological definition of dimensionality refers to the number of psychological constructs (e.g., math ability) necessary to respond to a set of items. In contrast, the statistical definition of dimensionality refers to the minimum number of dimensions needed to explain the variance in an item response matrix. Under the statistical definition, dimensionality is not viewed as a property of a test (or persons), but as a property of the item response matrix that results from the interaction between examinees and items (Reckase, 2009). Multidimensionality can occur when different items measure different abilities (i.e., simple structure) or when each item measures multiple abilities to varying degrees (i.e., complex structure). Items that measure multiple abilities to the same degree will appear unidimensional (Ackerman, 1992).

The definition of statistical dimensionality is strongly related to the concept of local item independence. Local item independence, also called conditional item independence, is an assumption of many measurement models including classical test theory, factor analytic, and IRT models. In classical test theory, item errors are assumed to be uncorrelated, and in IRT, items are assumed to be uncorrelated conditional on the latent trait being measured. Local item independence can be expressed as:

$$P(\{X_i = x_i\}|\theta) = \prod_{i=I}^{n} P(X_i = x_i|\theta),$$

where X_i is the score on item i. In the expression above, local item independence, as defined, is called *strong* local independence because conditional on an examinee's ability all items are assumed to be unrelated to all other items. Under the assumption of *weak* local independence (McDonald, 1979), only pair-wise correlations among items, conditional on ability, are assumed to be zero. Weak local item independence is expressed as:

$$P\left(X_i = x_i, X_j = x_j \middle| \theta\right) = P(X_i = x_i|\theta)P\left(X_j = x_j\middle|\theta\right).$$

Even when fitting multidimensional models the assumption of local item independence applies, and the presence of local item dependence would indicate that one or more important dimensions are not being modeled. McDonald (1982) linked the concept of local item independence to dimensionality by defining the dimensionality of a test based on the number of dimensions necessary to create weak local item independence.

The Sensitivity of Value-Added Scores to Different Outcome Measures

To date, no studies have explicitly examined the effects of test dimensionality on value-added scores and inferences using real or simulated data. However, three studies have examined the effects of different outcome measures on value-added scores and inferences. First, Lockwood et al. (2007) examined the effects of using different subtests as the outcome measure on teacher value-added scores, using data that included approximately 3,300 students and their teachers across four years of schooling. The outcome measures were students' Problem-Solving and Procedures subtest scores from the Stanford Achievement Test of mathematics. In addition, Lockwood et al. examined the effects of different value-added models and specifications of student control variables on teachers' value-added scores. Lockwood et al. fit four different value-added models to each outcome measure, including: the gain score model, the covariate adjustment model, and two versions of the persistence model (one where teacher effects are assumed to persist undiminished over time and one where the persistence of teacher effects is estimated from the data). In addition, the authors examined five different specifications of student control variables including: no control variables, student-level demographic variables only, student-level prior achievement scores only, both student-level demographic variables and prior performance scores, and three classroom-level aggregates of student information (variables reflecting ethnicity, socio-economic status and students' prior performance).

The results indicated that teacher value-added scores were more sensitive to the choice of outcome measure than to the value-added model fit or the specification of student control variables. The variation within teachers across outcomes was greater than the variation between teachers, which suggests that teachers are not equally effective at teaching all content and that teacher value-added scores are more sensitive to the choice of outcome measure than to various model specifications. Specifically, correlations between teacher value-added scores derived from the two outcome measures ranged from .01 to .46 across value-added models and specifications of student control variables. In contrast, average correlations between teacher value-added scores from various value-added models and specifications of study control variables ranged from .82 to .98 across the two outcome measures. The authors further examined the data by creating student composite scores under six different weighting schemes of the Problem-Solving and Procedures subtests. The results indicated that teacher value-added scores were affected by the weighting scheme for 38% of the year 2 teachers and over 60% of the year 3 teachers. Together these results suggest that teacher value-added scores are sensitive to the choice of outcome measure, with the two outcome measures producing very different rankings of teachers.

As part of a larger study examining the stability of value-added scores, Sass (2008) examined the effects of using a criterion-referenced versus norm-referenced test on value-added scores and inferences. Using data from a single county in Florida, Sass compared the consistency of elementary grade teachers' value-added scores across two measures of mathematics achievement administered to all students in grades 3-10. The measures were a criterion-referenced test named the Sunshine State Standards test, and the Stanford Achievement Test, a norm-referenced test. Sass categorized teachers into quintiles on each measure based on their value-added scores and then examined the consistency of classification across the two outcome measures. The results indicated poor consistency with only 43% of teachers classified in the top or bottom quintiles on both measures. Classification consistency was even lower for the middle three quintiles, averaging around 20%. Of even greater concern, 5% of teachers classified in the top quintile on the criterion-referenced test were classified in the bottom quintile on the norm-referenced; and 4% of teachers classified in the top quintile on the norm-referenced test were classified in the bottom quintile on the criterion-referenced test. The cross-exam correlation for the teacher value-added scores was .48. As in Lockwood et al. (2007), the two tests produced very different rankings of teachers.

Similarly, Papay (2011) examined the effects of using different tests of reading achievement as outcome measures on value-added scores and inferences. Papay collected six years of data including 55,000 students and their teachers from a large urban district in the northeastern U.S. The reading tests used as outcome measures included, (a) a state-developed achievement test, administered in the spring of each year, (b) the Stanford Achievement Test in reading, administered in the fall of each year, and (c) the Scholastic Reading Inventory, administered in the fall and spring of each year. Additionally, Papay replicated the work of Lockwood et al. (2007) by examining the consistency of teacher value-added scores when the Procedures and Problem-Solving subtests on the Stanford Achievement Test in mathematics are used as outcome measures.

As in Lockwood et al. (2007), the results indicated that teacher value-added scores were less consistent across choice of outcome measure than across value-added model specifications. Correlations between various model specifications across outcome measures were all greater than .77, and correlations between outcome measures across various model specifications were all less than .65. Regarding reading achievement, the correlations between teacher value-added scores across outcome measures ranged from .15 to .58, indicating a different rank ordering of teachers depending on the choice of reading achievement measure. To understand the consistency of teachers' rankings, Papay (2011) classified teachers into quartiles and then examined the consistency of the classifications across reading achievement measures. The results indicated that 53% and 52% of teachers were consistently classified in the top and bottom quartiles, respectively.

Of concern, 8% of teachers moved from the top to bottom quartile, and 7% of teachers moved from the bottom to top quartile across reading achievement measures.

In Lockwood et al. (2007), Sass (2008), and Papay (2011), estimates of teacher value-added scores were not consistently estimated when using different outcome measures, producing very different rankings of teachers. This finding held true in mathematics and reading and across a wide variety of value-added model specifications. Of interest though, are slight differences in the results of Lockwood et al. when compared to those of Sass and Papay. In Lockwood et al. the teacher value-added scores were more sensitive to the choice of outcome measure than in Sass and Papay, which may be due to the fact that the outcome measure examined in Lockwood et al. (i.e., mathematical procedures and problem-solving) represent narrower and more specific content than the outcome measures examined in Sass and Papay (i.e., reading achievement). Because teachers likely vary in the relative emphasis they place on mathematical procedures and problem solving during instruction, the choice of one measure over the other will affect the extent to which teacher value-added scores actually capture instruction, and by extension teachers' effectiveness. In contrast, more general measures that are intended to assess overall achievement likely capture more instruction, and as such produce more consistent rankings of teachers. However, the use of general measures of achievement fails to recognize that teachers are likely also more effective at teaching some types of content than others, as is suggested by the results of Lockwood et al., where there was more variability within teachers across measures than there was between teachers.

The results of Lockwood et al. (2007), Sass (2008), and Papay (2011) provide important insight into the effects of different outcome measures on value-added scores and inferences. However, none of the studies directly addressed the effects of test dimensionality on value-added scores and inferences. In all cases the outcomes were assumed to be unidimensional both within and across grades, which left the effects of various aspects of dimensionality (e.g., the number of dimensions) ignored. As noted previously, the assumption that tests of student achievement are unidimensional is defensible for many applications of scaling, but may not be defensible when applied to VAA because of potential misalignment between the relative weight given to specific content on a test and the emphasis placed on that content by teachers in the classroom.

The Effects of Test Dimensionality on Teacher VAA

Research suggests that the effects of ignoring multidimensionality on unidimensional student score estimates are often minimal (Ansley & Forsyth, 1985; Reckase, 1979; Yen, 1984), particularly when the constructs (dimensions) are correlated .40 or greater (Drasgow & Parsons, 1983). Similarly, the effects of ignoring multidimensionality on common scaling applications are often small. For example, Camilli, Wang, and Fesq (1995) examined the effects of ignoring multidimensionality on the IRT true-score equating of six forms of the Law School Admissions Test using real data, and found an insignificant effect on bias in unidimensional person score estimates and equating tables. However, Camilli et al. also contend that, "arguments concerning dimensionality need to be validated in the context of a test's use" (p. 82).

To date, no studies have examined the effects of ignoring multidimensionality on value-added scores and inferences using real or simulated data. However, Martineau (2006) demonstrated mathematically that violations to the unidimensionality assumption bias value-added scores and distort subsequent inferences. Specifically, Martineau addressed two questions: does construct shift distort growth-based value-added scores, and to what extent are distortions to value-added scores affected by varying inter-construct correlations. To begin, Martineau (2006) provided mathematical definitions of value-added (layered effects model) purely unidimensional true scores, empirically unidimensional true scores, and multidimensional true scores. Martineau defined a value-added purely unidimensional true score as:

$$t_{hi}^{lu} = t_i + \sum_{m=0}^{h} \left(g_{mi} + a_{j_{mi}}\right),$$

where *m* represents grade level, with 0 being the lowest grade and *h* the highest grade; *i* is students; j_{mi} is the unit (e.g., teacher) that student *i* attended in grade *m*; t_{hi}^{lu} is student *i*'s layered effects purely unidimensional (lu) true score at the end of grade *h*; t_i is the true score for student *i* before entering the lowest grade 0; g_{mi} is the gain of student *i* during grade *m*; and $a_{j_{mi}}$ is the value-added to student gains (i.e., g_{mi}) by unit j_{mi}. However, as noted previously, a purely unidimensional achievement test is unrealistic, which leads to the expression for a value-added empirically unidimensional (linear combinations of constructs) true score:

$$t_{hi}^{le} = \sum_{c=1}^{C} P_c \left[t_{ci} + \sum_{m=0}^{h} \left(g_{cmi} + a_{cj_{mi}}\right) \right] \text{under the constraint} \sum_{c=1}^{C} P_c \; = \; 1,$$

where c equals construct; C equals the number of constructs that combine to make up the single true score; t_{hi}^{le} is student i's layered effects empirically unidimensional (le) true score at the end of grade h; P_c is the proportion of the combined true score that is accounted for by construct c; t_{ci} is the true score for student i on construct c before entering grade 0; g_{cmi} is the gain of student i on construct c during grade m; and $a_{cj_{mi}}$ is the value-added to student gains on construct c (i.e., g_{mci}) by unit j_{mi}.

Although it is possible to construct an empirically unidimensional score scale, its utility to VAA depends on, among other things, that "the proportional construct representations on the score scale match the importance of the various constructs in the curriculum" (Martineau, 2006, p. 46). Martineau defined a value-added empirically multidimensional true score as:

$$t_{hi}^{lm} = \sum_{c=1}^{C} P_{ch} \left[t_{ci} + \sum_{m=0}^{h} \left(g_{cmi} + a_{cj_{mi}} \right) \right]$$

with the constraints

$$\sum_{c=1}^{C} P_{ch} = 1, \sum_{c=1}^{C} \left(P_{ch} - P_{c(h-1)} \right) = \sum_{c=1}^{C} d_{ch} = 0, \text{and } d_{c0} \equiv 0,$$

where t_{hi}^{lm} is student i's layered effects empirically multidimensional (lm) combined true score at the end of grade h; P_{ch} is the grade h proportion of the combined true score that is accounted for by construct c; d_{ch} is the change in the proportional representation of construct c in true scores from the end of grade h-1 to the end of grade h.

Then, using these true score definitions and the layered effects value-added model, Martineau (2006) defined value-added scores for each of the three specifications of dimensionality described above. The expression for the value-added by unit j_{hi} to a purely unidimensional true score is: $a_{j_{hi}}$, which can be interpreted as, “the effects of a unit on its students’ gains on a single construct” (p. 45). For empirically unidimensional true scores, the expression for the value-added by unit j_{hi} is:

$$\sum_{c=1}^{C} P_c a_{cj_{hi}},$$

which can be interpreted as “the weighted combination of a unit’s effectiveness on the various constructs that combine to create the score scale, where weights reflect the constructs’ unchanging proportional representation in the single score scale” (p. 46). According to Martineau, the utility of an empirically unidimensional true score to VAA depends on a number of assumptions: (a) that the proportional construct representations on the test match those in the curriculum, (b) that the proportional construct representations in the curriculum do not change over the course of the VAA study, (c) that the proportional construct representations on the test match the developmental level of the examinees, and (d) that the score scale used is empirically unidimensional.

Finally, the value-added to empirically multidimensional true scores by unit j_{hi} is:

$$\sum_{c=1}^{C} P_{ch} a_{cj_{hi}} + \frac{1}{n_i} \sum_{c=1}^{C} \left(d_{ch} \sum_{j'=1}^{n_i'} n_{ij'} a_{cj'} \right), d_{c0} \equiv 0.$$

Here, the expression includes two terms – the first represents the value-added effect “…on the various constructs that combine to create the score scale where weights reflect the constructs’ grade-specific representation in the single score scale” (p. 46), and the second term “the weighted combination of the accumulation of all preceding units’ value-added on the various constructs that combine to create the score scale where weights reflect the constructs' grade-specific change in representation in the single score scale from the previous grade to the current grade, averaged across students in the unit" (p. 47). The first term differs from the value-added expression for empirically unidimensional true scores in that it represents a *grade-specific* construct mix interpretation. According to Martineau (2006), the utility of this term depends on the degree to which the grade-specific proportional construct representation on the test matches that in the curriculum and the degree to which the grade-specific proportional construct representation on the test matches the developmental level of the examinees.

In contrast, the latter term in the above expression, $\frac{1}{n_i}\sum_{c=1}^{C}\left(d_{ch}\sum_{j'=1}^{n_i'} n_{ij'}a_{cj'}\right)$, is never of interest because it distorts the value-added scores of a single unit by contaminating it with the effects of previous units. Specifically, current teachers are benefited by previous teachers of high value-added on constructs whose proportional representation increased from the previous grade, and by previous teachers of low value-added on constructs whose proportional representation decreased from the previous grade, because the weight (d_{ch}), which represents the change in proportional construct representation, is multiplied by the accumulation of the value-added by previous teachers. Similarly, current teachers are penalized by previous teachers of low value-added on constructs whose proportional representation increased from the previous grade, and from teachers of high value-added on constructs whose proportional representation decreased from the previous grade. The latter term resolves to zero when, (a) the changes in the proportional construct representation and the accumulation of previous teachers' value-added cancel each other out (unobservable), and (b) when the mean value-added by previous teachers is the same for all constructs (unlikely). Martineau (2006) concludes that, “the only incontestable utility of the VAM estimates using empirically multidimensional scales is for units teaching the lowest grade in the analysis.” (p 47). In addition, Martineau shows that larger changes in proportional construct representation lead to lower reliability in value-added scores (too low for high-stakes use), and that only when the correlations between the value-added scores from different dimensions are very high (near unity) do the distorting effects of multidimensionality disappear. Together with the work described earlier by Lockwood et al. (2007), Sass (2008), and Papay (2011), Martineau’s findings suggest that unidimensional scaling of student achievement tests is likely inappropriate for VAA.

Definition of Test score reliability & Measurement Error

Test score reliability is defined as the ratio of variance in true scores to variance in observed scores:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2},$$

and represents the consistency of observed scores across testing conditions, when those conditions are modeled in the testing procedure. In classical test theory, an examinee’s (*i*) observed score is thought to consist of two components: the examinee’s true score (the expected value over test replications) and measurement error:

$$X_i = T_i + E_i.$$

Across examinees, it is assumed that measurement errors are randomly distributed with a mean equal to zero and variability defined by the standard error of measurement (SEM), where the SEM is defined as:

$$SEM_X = SD_X\sqrt{1 - r_{XX}},$$

where SD_X is the standard deviation of the observed scores, and r_{XX} is the test score reliability.

In multiple regression, the presence of measurement error in the outcome variable results in lessened precision, which in turn leads to less statistical power. As a result, R-squared values are smaller and regression coefficients are attenuated toward zero. The presence of measurement error in one or more predictor variable biases estimation of all regression coefficients. Specifically, the regression coefficient associated with the predictor variable containing measurement error will be biased toward zero, and it will be under adjusted for in the estimation of other regression coefficients.

Most statistical models used in VAA include one or more predictor variables for the purpose of controlling for students' prior achievement. Because the statistical models used in VAA are regression models they assume that the predictor variables are measured without error (i.e., that observed scores equal true scores). However, student achievement tests always contain measurement error, and as noted above, not accounting for test measurement error in the predictor variables will bias all model estimates, including the value-added score estimates. The bias occurs because the measurement error in the predictor variables causes them to be correlated with the model's residual terms.

The Effects of Test Measurement Error on Teacher VAA

To date, only one published study has examined the effects of test measurement error (TME) on value-added scores and inferences (an additional unpublished study exists but is unavailable). Koedel, Leatherman, and Parsons (2012) examined the effects of TME on teacher value-added scores and inferences using real and simulated data. In the simulation study, the authors generated gain scores by summing three components: a student component, a measurement error component, and a teacher effect component. Both the teacher and student components were based on a random normal distribution with a mean equal to zero. The measurement error component was generated by using conditional gain score SEMs reported for the operational test used in the real data study (described below). The variance in gain scores equaled the sum of the component variances, where teachers accounted for 9% of the variance in gain scores, measurement error accounted for 50% of the variance in gain scores, and students accounted for the remaining 41% of the variance in gain scores. Once each of the components was generated, the student component and the measurement error component were summed such that students with scores in the tails were associated with greater measurement error. Finally, students were randomly assigned to teachers and the teacher components were added accordingly.

Once the data was generated, Koedel et al. (2012) then fitted the observed scores to a value-added model where teachers were treated as fixed effects. The fixed teacher effects were also weighted such that students with more precise scores contributed more to the teacher's estimated effect. Across a range of class size specifications, the correlation between teachers' true value-added scores and teachers' estimated value-added scores ranged from .42 to .93 for the unweighted estimates, and .50 to .95 for the weighted estimates. For both the unweighted and weighted estimates the magnitude of the correlations increased as the within-teacher sample size increased. Further, the authors found that if the portion of variance in gain scores attributed to measurement error is increased from 50% to 80% the magnitude of the correlations decreases such that they range from .37 to .94 across weighting schemes.

The study conducted by Koedel et al. (2012) using real data included over 35,000 grade 5 students and 1,600 teachers from over 650 schools in the state of Missouri. Because the authors used the gain score model, data for all participants was available for two consecutive academic years. Over 40% of students were eligible for free or reduced priced lunch, 13% were receiving special education services, and 17% of students were minority students. Using data from Missouri's state level mathematics test, the authors implement the same weighting schemes as in the simulation study, such that students with more precise gain scores (as defined by conditional SEMs) were weighted more in the estimation of teacher effects. The results showed that the average correlation between teachers' effects in Year 1 and their effects in Year 2 was .45 without weighting, and .48 with weighting. As in the simulation study, the results of the real data study suggest that teacher value-added scores are sensitive to the effects of test measurement error.

Research Questions

The primary purpose of the present study is to examine the extent to which test dimensionality and test score reliability bias the estimation of teacher value-added scores. In addition, to mimic the use of teacher value-added scores by states and school districts, a secondary purpose of the present study is to examine the effects of test dimensionality and test score reliability on the consistency of teachers' estimated relative effectiveness. Accordingly, the following research questions guide this study:

1. Does test dimensionality bias the estimation of teacher value-added scores?
    1.1. To what extent, if any, does the number of test dimensions bias the estimation of teacher value-added scores?
    1.2. To what extent, if any, does the strength of the association among test dimensions bias the estimation of teacher value-added scores?
    1.3. Does the number of test dimensions interact with the strength of the association among test dimensions to bias the estimation of teacher value-added scores?
2. Does test dimensionality affect the consistency of teachers' estimated relative effectiveness?
    2.1. To what extent, if any, does the number of test dimensions affect the consistency of teachers' estimated relative effectiveness?
    2.2. To what extent, if any, does the strength of the association among test dimensions affect the consistency of teachers' estimated relative effectiveness?

2.3. Does the number of test dimensions interact with the strength of the association among test dimensions to affect the consistency of teachers' estimated relative effectiveness?

3. To what extent, if any, does variability in test score reliability bias the estimation of teacher value-added scores?
4. To what extent, if any, does variability in test score reliability affect the consistency of teachers' estimated relative effectiveness?

In addition, the extent to which test dimensionality and test score reliability might bias the estimation of unidimensional person (student) estimates was examined, as it is the bias in student scores that is hypothesized to lead to bias in teacher value-added scores.

Hypotheses

Regarding research questions 1.1 and 2.1, bias in the estimation of teachers' value-added scores was hypothesized to decrease as the number of test dimensions increased (an inverse relation) for the multidimensional cases. Prior research shows that unidimensional student score estimates derived from multidimensional data show less bias as the number of dimensions increases (Harrison, 1986). As such, when the data shows evidence of multidimensionality, a greater number of test dimensions was hypothesized to lead to less bias in the estimation of teacher value-added scores. Similarly, it was hypothesized that teachers' estimated relative effectiveness would increase in consistency as the number of test dimensions increased.

Regarding research questions 1.2 and 2.2, it was hypothesized that value-added scores derived from data sets where the strength of the association among test dimensions was greater would be less biased, with bias increasing as the strength of the association among dimensions decreases. More specifically, it was hypothesized that value-added scores would show little bias when the correlation among test dimensions was .50 or greater, as unidimensional student score estimates are often robust to violations of the unidimensionality assumption when the correlations among test dimensions are moderate to large (Ackerman, 1989; Drasgow & Parsons, 1983; Harrison, 1986). In contrast, when the correlations among test dimensions are small, and data show clear evidence of multidimensionality, unidimensional student score estimates are frequently biased, which was hypothesized to lead to bias in the estimation of teacher value-added scores. Similarly, it was hypothesized that teachers' estimated relative effectiveness would be more consistent when correlations among dimensions were moderate to large (i.e., equal to or greater than .50), and less consistent when correlations among dimensions were small.

Regarding research questions 3 and 4, bias in the estimation of teachers' value-added scores was hypothesized to decrease as the level of test score reliability increased (and test measurement error decreased). As noted previously, all model parameter estimates will be biased to some degree by the presence of measurement error in any of the predictor (control) variables. Because a student's prior achievement is almost always used as a control variable in VAA, any measurement error present in the test used to measure prior achievement (most often a student's end-of-year state test score from the previous grade) will bias the estimation of teacher value-added scores. It was also hypothesized that consistency in teachers' estimated relative effectiveness would increase as the level of test score reliability increased.

Chapter III: Method

Research Design

Independent variables. To address the research questions, three variables of the student achievement test used to estimate teachers' value-added scores were manipulated: (a) the number of test dimensions (i.e., subscales), (b) the strength of the association among test dimensions, and (c) the level of test score reliability. As will be described below, the present study's focus was on the effect of teachers on students' fifth-grade mathematics achievement, controlling for students' fourth-grade mathematics achievement. As such, scores for two tests were simulated for each student, a fourth-grade test score and a fifth-grade test score, and all independent variables were manipulated for both sets of test scores.

Number of test dimensions. Data was simulated using one, two, and four test dimensions. One dimension was chosen to serve as a baseline to which the multidimensional levels of two and four dimensions could be compared. Two dimensions was chosen for two reasons, (a) several well-known and frequently used standardized achievement tests (e.g., Stanford Achievement Test) comprise two subtests/dimensions, and (b) many studies that have examined the effects of violating the unidimensionality assumption on person parameter (student) estimates simulated data using two dimensions, which allowed for easier comparison between the present study's results and the results of previous studies. Modeling four dimensions was chosen because the operational end-of-year state mathematics achievement test used by the district, which served as the basis for the present study, comprises four subtests (as do many formative assessments).

Correlation among test dimensions. To manipulate the strength of the association among test dimensions, the magnitude of the correlations among test dimensions was specified at five different levels: .05, .30, .50, .80, and .95, reflecting varying degrees of multidimensionality in latent structure. The correlation of .95 was chosen because it reflects data that is essentially unidimensional in its latent structure, whereas .05 was chosen because it reflects data that is essentially fully multidimensional in its latent structure. The correlations of .30, .50, and .80 were chosen to reflect the types of correlations among test dimensions often found in real data, and represent what are typically viewed as weak, moderate, and strong associations among test dimensions, respectively. Further, the correlations of .30 and .50 were chosen to reflect the (relatively new) use of formative assessments in value-added assessment (VAA), where the correlations among subtests making up a composite score are often weak to moderate in strength. The correlation of .80 reflects the use of more traditional end-of-year state assessments in VAA, where the correlations among subtests are often high by design.

Test score reliability. Finally, students' achievement scores were simulated under two levels of test score reliability, .95 and .75, which represent the levels of test score reliability common to end-of-year state achievement tests and formative assessments, respectively.

The experimental conditions described above, result in a 3 (number of test dimensions) x 5 (correlation among test dimensions) x 2 (test score reliability) between-subjects, design matrix containing 22 partially-crossed cells, each corresponding to a different simulated data set (see Table 1). The design matrix is partially-crossed because the independent variable representing the correlation among test dimensions is only crossed with two levels of the independent variable representing the number of test dimensions (levels two and four). In other words, the correlation among test dimensions is not considered for the one-dimensional case because there are not multiple dimensions to correlate.

Table 1: Research Design Matrix

Number of dimensions	Correlation among dimensions	Test score reliability	
		.75	.95
One	NA	*	*
Two	.05	*	*
	.30	*	*
	.50	*	*
	.80	*	*
	.95	*	*
Four	.05	*	*
	.30	*	*
	.50	*	*
	.80	*	*
	.95	*	*

Dependent variables. For each of the 22 data sets, two dependent variables were considered: bias in the estimation of teachers' value-added scores, and consistency in teachers' estimated relative effectiveness.

Bias in score estimation. Bias in the estimation of teachers' value-added scores (and students' achievement scores) was assessed using two measures: mean bias and root mean squared error (RMSE). Mean bias (MB) measures the direction of bias in the estimation of teachers' value-added scores, and RMSE measures the amount of bias in the estimation of teachers' value-added scores. In both cases bias was assessed by comparing the true population parameters (γ) used to generate the simulated data to estimates of the same parameters ($\hat{\gamma}$) obtained after fitting the simulated data to a value-added model (estimated random effects for teachers and fitted values for students). Because the effects of interest were random effects, the generating parameters differed for each replication. As such, bias was calculated for each person within a replication, and then averaged across replications (R). Mean bias for each person was defined as:

$$MB_j = \frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_j - \gamma_j),$$

and RMSE for each person was defined as:

$$RMSE_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\gamma}_j - \gamma_j)^2}.$$

Consistency in estimated relative effectiveness. Consistency in teachers' estimated relative effectiveness was examined by comparing teachers' estimated value-added scores across study conditions. For each teacher within a design cell, the estimated value-added score across replications was computed as:

$$VA\ score_j = \frac{1}{R}\sum_{r=1}^{R}\hat{\gamma}_j,$$

and used to examine consistency of estimated value-added scores across study conditions. Large correlations between teachers' estimated value added scores across study conditions would indicate that teachers' estimated relative effectiveness is unaffected by the test dimensionality and/or test score reliability. In addition, within a design cell teachers were classified into quartiles (representing varying levels of effectiveness) based on their estimated value-added score. Contingency tables and chi-squared values were used to assess the consistency of teachers' effectiveness classification across study conditions. Large chi-squared values would indicate that value-added teacher effectiveness estimates are unaffected by test dimensionality and/or test score reliability.

Internal validity. The experimental nature of the present simulation study significantly reduces potential threats to internal validity, and allows for causal inferences about the effects of interest on the study outcomes. In addition, the models used to generate the simulated data and those fitted to the simulated data were purposely chosen to be similar, so as to reduce the effect(s) of model variations on the study's results. Consequently, variations in the latent structure of the test of mathematics achievement (i.e., dimensionality) and the level of test score reliability are the only differences between the two sets of models, thereby eliminating additional potential threats to the internal validity of the study's results.

External validity. To increase the generalizability of the results several aspects of the study (value-added model specification, sample size, and parameter generating values) were designed to mimic the use of VAA in Minneapolis Public Schools. Minneapolis Public Schools (MPS) is a large, urban school district with over 3,700 teachers and approximately 34,000 students. The distribution of ethnicities among students in the district is 36% Black, 33% White, 19% Hispanic, 8% Asian American, and 4% American Indian. Further, 19% of students are receiving special education services, 21% of students are English language learners, and 66% of students are eligible to receive free or reduced price lunch.

Model specification details and fixed effect parameter estimates for the model (found by fitting the model to real MPS data from the 2011-2012 and 2012-2013 academic years) were obtained from technical documentation provided to the public by MPS (see Value-Added Research Center, 2013). The value-added model used by MPS is a univariate, posttest-on-pretest, multi-level regression model, where students' composite scores on the Minnesota Comprehensive Assessment–III (MCA-III) in mathematics for their current grade are treated as the outcome variable, and their composite scores on the MCA-III (in mathematics and reading) for their previous grade serve as control variables. The MCA-III in mathematics assesses students' mathematics achievement in the domains of Number and Operation, Algebra, Geometry and Measurement, and Data Analysis. The value-added model used by MPS is also an error-in-variables model, which accounts for test measurement error in the estimation of teachers' value-added scores. However, because the effect of test measurement error was of primary interest to the present study the value-added model used in the present study is not an error-in-variables model. In addition, variables representing student demographic characteristics (free or reduced price lunch status, English language learner status, special education status, race, sex, and mobility status) are included as control variables in the model used by MPS, but were not included in the present study's model because they were not of interest and were not expected to have an effect on the study's results.

Simulation Procedures

The simulation comprised two primary stages, (a) generating simulated data under the study conditions described above, and (b) fitting a value-added model to the simulated data to obtain teachers' value-added scores, which were then used to evaluate the effects of the independent variables on the dependent variables. The two stages and their corresponding steps were replicated 500 times. Within a replication, data for n = 2,500 students and j = 100 teachers was simulated for each of the 22 design cells, which resulted in a total N = 55,000 students and J = 2,200 teachers per replication. As described above, values of interest were aggregated across replications within cell. A single seed value (8279) was used when generating all data. The simulation was implemented using RStudio version 0.99.473 (2015); in addition to the base package, the 'mvtnorm' and 'psych' R packages were used. The code used to generate and model the simulated data is presented in Appendix C.

Generating simulated data. For both the fourth- and fifth-grade tests, generation of the simulated data took place in three steps. The first step involved generating students' true subscale scores following the dimensionality specifications detailed above, and then computing students' true composite scores. The second step involved creating students' observed composite scores by adding measurement error related to the two levels of test score reliability to their true composite scores. The third step involved simulating teachers' true value-added scores and adding those scores to each teacher's respective students' observed composite scores. Steps 1-3 were repeated for students' fourth- and fifth-grade test scores. Then, after students' fourth- and fifth-grade test scores were generated a correlation was imposed between the fourth- and fifth-grade tests scores.

Step 1. As described above, two dimensionality variables were considered, the number of dimensions (1, 2, 4) and the correlation among dimensions (.05, .30, .50, .80, .95). The multidimensional cases are described first, followed by the one-dimensional case.

To simulate true subscale scores, 10 sets of multivariate, random normal variables were generated under different specifications of multidimensionality. For the two-dimensional case, 5 sets of two random variables were generated; and for the four-dimensional case, 5 sets of four random variables were generated. For both cases, the correlation among dimensions for one set of variables equaled .05, .30, .50, .80, or .95. The mean of all variables equaled 0, and the standard deviation equaled 1. Then, true composite scores were created by calculating the mean of each student's subscale scores. Data for the one-dimensional case were simulated by generating a univariate, random normal variable with a mean equal to 0 and a standard deviation equal to 1. The above described steps resulted in 11 true score variables generated under various specifications of test dimensionality.

Step 2. To simulate students' observed scores, a measurement error component related to test score reliability was added to the 11 true score variables generated in Step 1. First, measurement errors were simulated by generating two sets of random normal deviates for each of the 11 true score variables. The mean of the random normal deviates equaled 0 and the standard deviation equaled the standard error of measurement (SEM) related to the two levels of test score reliability under study (.75 and .95). In classical test theory, the SEM of a test is defined as:

$$SD_x\sqrt{1 - r_{xx}}.$$

where SD_x is the standard deviation of the observed scores, and r_{xx} is test score reliability. Because the standard deviations of the tests simulated here equaled 1, the expression reduces to:

$$\sqrt{1 - r_{xx}}$$

in the present study. Accordingly, the SEM equaled .5 when test score reliability equaled .75, and .2236 when test score reliability equaled .95. Each set of random normal deviates (measurement errors) was then added to the 11 true score variables, resulting in 22 observed score variables each associated with one of the 22 design cells.

Step 3. To simulate teachers' true value-added scores, 22 random normal variables were generated, with means equal to 0 and standard deviations equal to 0.5. The standard deviation of 0.5 was chosen so that differences in teachers would be associated with 20% of the variance in students' test scores (i.e., intra-class correlation equal to .20).

To assign teachers' value-added scores to students' observed scores, students and teachers were randomly assigned to classrooms (clusters), where the within-teacher sample size (i.e., class size) equaled 25. Teachers' value-added scores were then added to the observed scores of the students in their classroom. The simulation was designed so that students stayed clustered in the same classroom in fourth and fifth grade, but the teacher associated with that classroom differed in each grade.

In summary, students' final observed scores were defined as the sum of three components: a true score component (Step 1), a measurement error component (Step 2), and a teacher value-added score component (Step 3). Steps 1-3 were repeated for students' fourth- and fifth-grade tests independently.

Once the data for both grades were generated the final step was to impose a correlation between students' fourth- and fifth-grade scores. A correlation of .70 (and the corresponding slope of 0.96) was chosen because it is the correlation between students' fourth- and fifth-grade scores on the MCA-III in mathematics in MPS. To impose the correlation of .70 on the relationship between students' fourth- and fifth-grade scores the following adjustment was applied to students' fifth-grade scores:

$$X_{5th}^{*} = X_{5th} + .96(X_{4th}),$$

where,

$$var(X_{5th}^{*}) = var(X_{5th}) + var(X_{4th}) + 2cov(X_{5th}, X_{4th}),$$

which reduces in the present study to:

$$var(X_{5th}^{*}) = var(X_{5th}) + var(X_{4th}),$$

because students' fourth- and fifth-grade scores were generated independently and were therefore uncorrelated. After imposing the correlation of .70 on the relationship between students' fourth- and fifth-grade scores, students' test scores were fitted to a value-added model.

Modeling simulated data. To obtain estimated teacher value-added scores under the various study conditions, students' simulated test scores generated in the steps described above were fitted to a value-added model. The value-added model used in the present study is a univariate posttest-on-pretest, multi-level model, where students' fifth-grade observed composite scores were the dependent variable, and students' fourth-grade observed composite scores were a predictor (control) variable. The following model was fitted to each of the 22 sets of student test scores generated in Steps 1-3:

$$5th\ grade\ score_{ij} = \gamma_{00} + \gamma_{01} 4th\ grade\ score_{ij} + u_{0j} + r_{ij},$$

where i indexes students; j indexes teachers; γ_{00} = the model intercept; γ_{01} = the fixed effect associated with fourth-grade achievement; u_{0j} = the random effect associated with teacher j; and r_{ij} = the random effect (residual) for student i associated with teacher j. The random teacher effects (u_{0j}) are teachers' estimated value-added scores and were used to calculate the dependent variables (i.e., MB and RMSE).

Data Analysis

Bias in value-added score estimation. The first research question asked whether test dimensionality and test score reliability bias the estimation of teacher value-added scores. Two bias estimates were considered, MB and RMSE. First, a series of plots were generated to examine the effects, if any, of the independent variables on MB and RMSE, including: histograms of overall average bias and marginal mean bias, scatterplots between generating parameters and estimated values, and interaction plots of cell mean bias. Descriptive statistics of overall values, main effects, and interactions were also examined.

Mean bias. Based on the results of the descriptive analyses (detailed in Chapter 4), inferential analysis of MB in score estimation was limited to one independent variable, number of test dimensions, for teachers' value-added scores only. The descriptive analyses showed no evidence of differences in MB between the levels of correlation among test dimensions and test score reliability for teachers' value-added scores. Similarly, there was no evidence of differences between the levels of any of the independent variables in MB in the estimation of students' scores. The effect of the number of test dimensions on MB in the estimation of teachers' value-added scores was examined by performing an ANOVA. Preliminary analyses showed no variability between design cells in MB, so a single-level analysis was performed.

Root-mean-squared-error. Regarding RMSE, preliminary analyses indicated sufficient clustering between design cells (ICC = .94 for teachers' value-added scores, and .96 for students' scores) to warrant fitting multi-level models. Because the correlation among test dimensions is not fully crossed with the number of test dimensions, two models were fitted. In both models, Level 1 was an unconditional model with teachers (j = 100) nested within design cells (k = 22). In Model 1, all independent variables were included at Level 2, as well as the interaction between the number of test dimensions and the correlation among test dimensions. However, because the correlation among test dimensions was included, data for the one-dimensional case was ignored, and so only two and four test dimensions were considered. In Model 2, the correlation among test dimensions was not included in the analysis, which allowed for examination of the one-dimensional case. As such, Model 2 included the number of test dimensions and test score reliability at Level 2. For Model 1, the number of test dimensions was coded so that four dimensions = 1 and two dimensions = 0. For Model 2, the number of test dimensions was dummy coded such that one dimension was the reference group. In both models, the correlation among test dimensions was dummy coded such that the correlation of .95 was the reference group, and test score reliability was coded so that a reliability of .95 = 1 and .75 = 0.

The mixed form of Model 1 can be written as:

$$\begin{aligned} RMSE_{jk} = \gamma_{00} &+ \left(\gamma_{01} numb.dim_{jk}\right) + \left(\gamma_{02} cor.dim = .05_{jk}\right) + \left(\gamma_{03} cor.dim = .30_{jk}\right) \\ &+ \left(\gamma_{04} cor.dim = .50_{jk}\right) + \left(\gamma_{05} cor.dim = .80_{jk}\right) + \left(\gamma_{06} reliab_{jk}\right) \\ &+ \left(\gamma_{07} numb.dim_{jk} cor.dim = .05_{jk}\right) \\ &+ \left(\gamma_{08} numb.dim_{jk} cor.dim = .30_{jk}\right) \\ &+ \left(\gamma_{09} numb.dim_{jk} cor.dim = .50_{jk}\right) \\ &+ \left(\gamma_{010} numb.dim_{jk} cor.dim = .80_{jk}\right) + u_{0k} + r_{jk}. \end{aligned}$$

Where *j* indexes teachers; *k* indexes design cells; γ_{00} = the model intercept; γ_{01} = the fixed effect associated with the number of test dimensions; γ_{02} = the fixed effect associated with the dummy code when the correlation among test dimensions = .05; γ_{03} = the fixed effect associated with the dummy code when the correlation among test dimensions = .30; γ_{04} = the fixed effect associated with the dummy code when the correlation among test dimensions = .50; γ_{05} = the fixed effect associated with the dummy code when the correlation among test dimensions = .80; γ_{06} = the fixed effect associated with test score reliability; γ_{06} = the fixed effect associated with test score reliability; γ_{07} = the fixed effect associated with the interaction between the number of test dimensions and dummy code when the correlation among dimensions = .05; γ_{08} = the fixed effect associated with the interaction between the number of test dimensions and dummy code when the correlation among dimensions = .30; γ_{09} = the fixed effect associated with the interaction between the number of test dimensions and dummy code when the correlation among dimensions = .50; γ_{010} = the fixed effect associated with the interaction between the number of test dimensions and dummy code when the correlation among dimensions = .80; u_{0k} = the random effect associated with design cell *k*; and r_{jk} = the random effect (residual) for teacher *j* associated with design cell *k*.

The mixed form of Model 2 can be written as:

$$RMSE_{jk} = \gamma_{00} + \left(\gamma_{01} numb.dim = 2_{jk}\right) + \left(\gamma_{02} numb.dim = 4_{jk}\right) + \left(\gamma_{03} reliab_{jk}\right)$$
$$+ u_{0k} + r_{jk}.$$

Where j indexes teachers; k indexes design cells; γ_{00} = the model intercept; γ_{01} = the fixed effect associated with the dummy code when number of test dimensions = 2; γ_{02} = the fixed effect associated with the dummy code when number of test dimensions = 4; γ_{03} = the fixed effect associated with the level of test score reliability; u_{0k} = the random effect associated with design cell *k*; and r_{jk} = the random effect (residual) for teacher *j* associated with design cell *k*.

Consistency in teachers’ estimated effectiveness. The second research question addressed the consistency of teachers’ estimated relative effectiveness across levels of the independent variables. Teachers’ mean estimated value-added scores across replications were used to examine consistency in relative effectiveness rankings. First, Pearson product-moment correlations between teachers’ value-added scores across study conditions were examined. High correlations would indicate that teachers’ relative effectiveness rankings were relatively similar across levels of the independent variables, and therefore unaffected by test dimensionality and/or test score reliability. Next, within each of the 22 design cells teachers’ value-added scores were used to classify teachers into one of four equal-sized effectiveness groups. Then, contingency tables were examined to evaluate the consistency of teachers’ effectiveness classification across levels of the independent variables. Because the effectiveness classification is an ordinal variable, the Kruskal-Wallis chi-square test for ordered contingency tables was used to examine the statistical significance of the consistency of the teachers’ effectiveness classification.

Data analyses were performed using RStudio version 0.99.473 (2015); in addition to the base package, the ‘lme4’, ‘stats’, and ‘psych’ R packages were used. The results of these analyses are described in the next chapter.

Chapter IV: Results

Bias in Value-Added Score Estimation

The first research question asked, *Does test dimensionality bias the estimation of teacher value-added scores?* The third research question asked, *Does test score reliability bias the estimation of teacher value-added scores?* Bias in the estimation of teacher value-added scores was assessed using two measures of bias, mean bias (MB) and root mean squared-error (RMSE). Bias in the estimation of student scores was also assessed.

Descriptive analyses. The overall average MB in the estimation of teachers' value-added scores is -0.000014, showing that on average teachers' value-added scores are slightly underestimated in the present study. In contrast, the overall average MB in the estimation of students' scores is 2.24e-18, which shows that on average students' scores are not systematically over- or underestimated in the present study. Histograms of overall MB for teachers' value-added scores and students' scores are presented in Figure 1. The overall average RMSE is 0.1796 for the estimation of teachers' value-added scores, and 0.9095 for the estimation of students' scores, which reveals that there is considerably more bias in the estimation of students' scores than teachers' value-added scores. Histograms of overall RMSE for teachers' value-added scores and students' scores are presented in Figure 2.

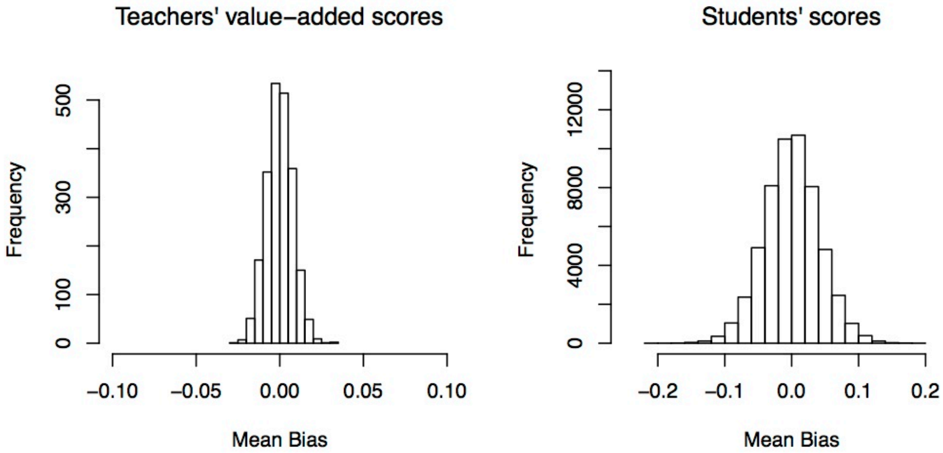

Figure 1: Overall average MB in the estimation of teachers' value-added scores and students' scores.

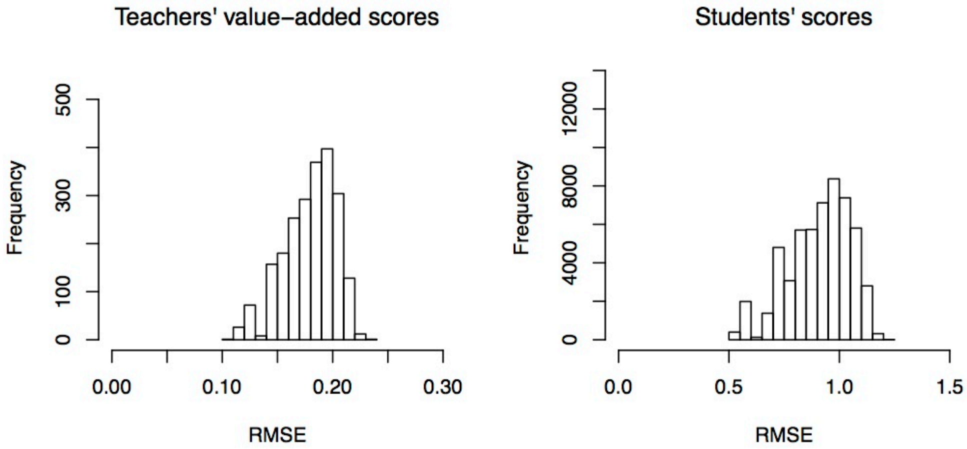

Figure 2: Overall average RMSE in the estimation of teachers' value-added scores and students' scores.

To further examine the estimation bias in teachers' value-added scores, the correlation and scatterplot (Figure 3) representing the association between the generating parameters and estimates for teachers' overall value-added scores was examined. The overall correlation between the generating parameters and estimates for teachers' value-added scores is .94, which shows that many, but not all teachers' value-added scores are estimated accurately across study conditions. An examination of the scatterplot in Figure 3 supports this finding, and reveals a strong, linear association between generating parameters and estimates for teachers' value-added scores. The average MB and RMSE by levels of the independent variables, for teachers' value-added scores and students' scores, are presented in Table 2 and discussed next.

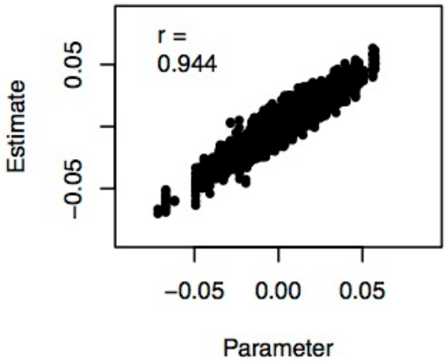

Figure 3: Scatterplot between generating parameters and estimates for teachers' value-added scores across independent variables.

Number of test dimensions. Regarding the number of dimensions, the means presented in Table 2 for MB show that teachers' value-added scores are overestimated for the one- and two-dimensional cases (i.e., MB > 0), but underestimated for the four-dimensional case (MB < 0). Students' scores are overestimated for all three cases, but that overestimation is negligible. Histograms of MB in the estimation of teachers' value-added scores by number of test dimensions are presented in Figure 4. Histograms of MB in the estimation of students' scores by number of test dimensions are presented in Figure 5 (Appendix B).

Table 2: Marginal Means of Bias Measures for Levels of the Independent Variables

	Mean Bias		RMSE	
Independent Variables	Teachers	Students	Teachers	Students
Number of dimensions				
One	0.0015	5.17e-18	0.2041	1.0520
Two	0.0002	3.31e-18	0.1833	0.9313
Four	-0.0006	0.59e-18	0.1710	0.8593
Correlation among dimensions				
.05	-0.0002	7.72e-18	0.1491	0.7242
.30	-0.0002	-0.91e-18	0.1661	0.8249
.50	-0.0002	4.42e-18	0.1773	0.8958
.80	-0.0002	-1.61e-18	0.1934	0.9933
.95	-0.0002	0.12e-18	0.2001	1.0383
Test score reliability				
.75	-0.00001	4.69e-18	0.1884	0.9640
.95	-0.00001	-0.22e-18	0.1708	0.8551

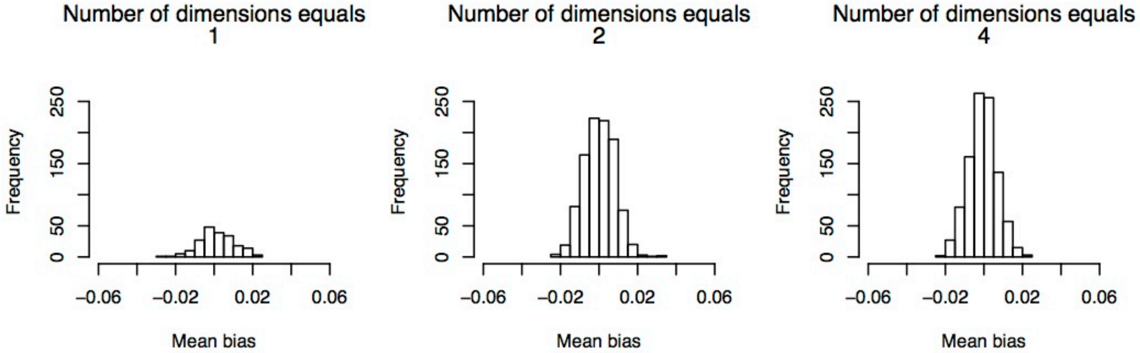

Figure 4: Histograms of MB in the estimation of teachers' value-added scores by number of test dimensions.

For both teachers' value-added scores and students' scores, the amount of estimation bias (i.e., RMSE) decreases as the number of test dimensions increases. The bias in the estimation of teachers' value-added scores is considerably less than the bias in the estimation of students' scores. Histograms of RMSE in the estimation of teachers' value-added scores by number of test dimensions are presented in Figure 6. Histograms of RMSE in the estimation of students' scores by number of test dimensions are presented in Figure 7 (Appendix B).

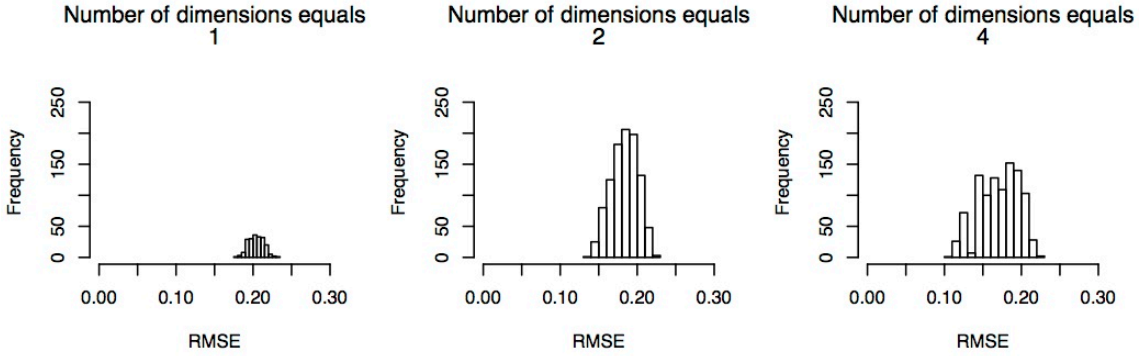

Figure 6: Histograms of RMSE in the estimation of teachers' value-added scores by number of test dimensions.

Correlations between generating parameters and estimates for teachers' value-added scores are .90, .94, and .96 for one, two, and four dimensions, respectively. As the number of dimensions increases teachers' value-added scores are estimated more accurately. Scatterplots associated with these correlations for one, two, and four dimensions are presented in Figure 8. The scatterplots show a strong, linear association between generating parameters and estimates for teachers' value-added scores regardless of the number of test dimensions.

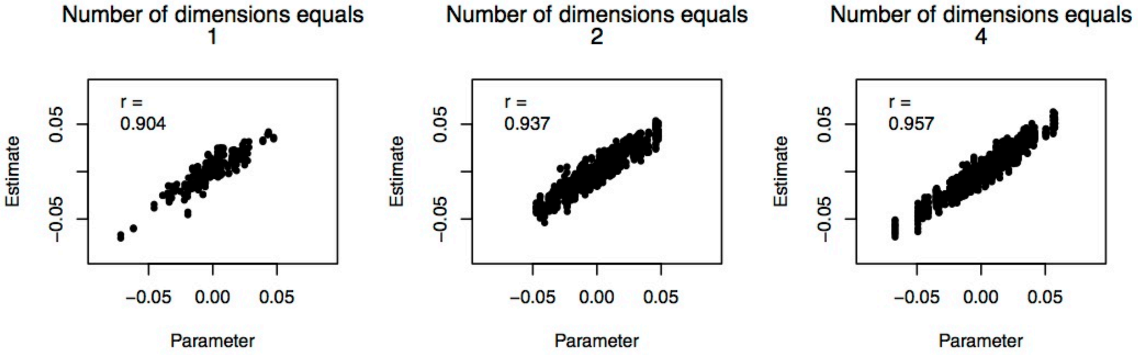

Figure 8: Scatterplots between generation parameters and estimates for teachers' value-added scores by number of test dimensions.

Correlation among test dimensions. Regarding the correlation among test dimensions, the results presented in Table 2 for MB show that teachers' value-added scores are consistently underestimated across the different levels of correlation. Although it is not evident in Table 2, the means for teachers' MB do differ across the different levels of correlation, however, those differences cannot be seen until the 16th place value and are therefore unimportant. The MB in the estimation of students' scores is very small, and there is no clear pattern across the different levels of correlation regarding systematic over- or underestimation. Histograms of MB in the estimation of teachers' value-added scores by the correlation among test dimensions are presented in Figure 9. Histograms of MB in the estimation of students' scores by the correlation among dimensions are presented in Figure 10 (Appendix B).

For both teachers' value-added scores and students' scores, the amount of estimation bias (i.e., RMSE) increases as the correlation among dimensions increases. As with the number of dimensions, bias in the estimation of teachers' value-added scores is considerably less than bias in the estimation of students' scores. Histograms of RMSE in the estimation of teachers' value-added scores by the correlation among test dimensions are presented in Figure 11. Histograms of RMSE in the estimation of students' scores by the correlation among test dimensions are presented in Figure 12 (Appendix B).

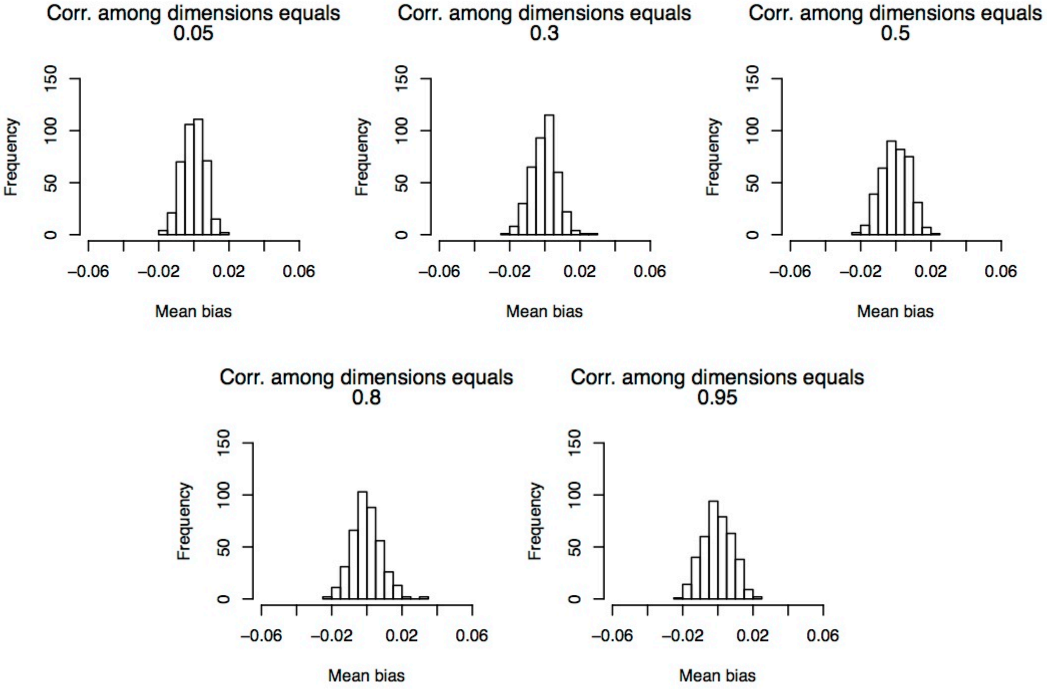

Figure 9: Histograms of mean bias in the estimation of teachers' value-added scores by the correlation among test dimensions.

Correlations between generating parameters and estimates for teachers' value-added scores are .97, .95, 0.94, 0.94, and 0.94 for correlations of .05, .30, .50, .80, and .95, respectively. As the correlations among dimensions increases teachers' value-added scores are estimated less accurately, with no differences in accuracy when the correlation among dimensions is .50 or greater. Scatterplots associated with these correlations for correlations of .05, .30, .50, .80, and .95 are presented in Figure 13. The scatterplots show a strong, linear association between generating parameters and estimates for teachers' value-added scores.

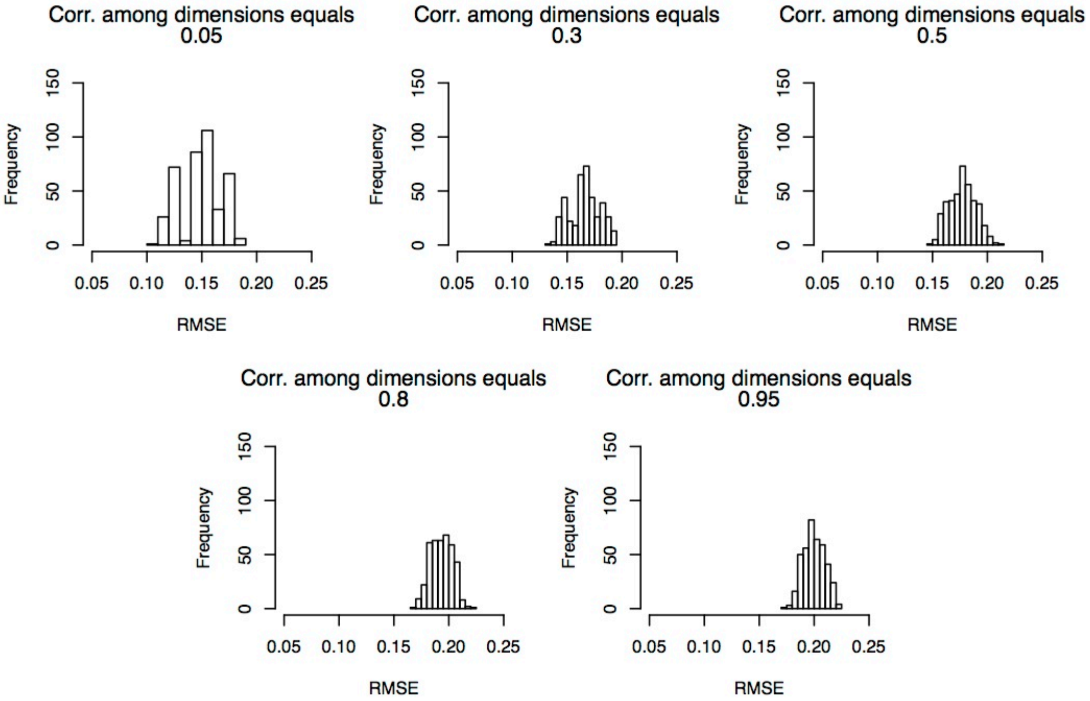

Figure 11: Histograms of RMSE in the estimation of teachers' value-added scores by the correlation among test dimensions.

Test score reliability. The results presented in Table 2 related to test score reliability show that teachers' value-added scores are underestimated for both levels of test score reliability (i.e., MB < 0). Again, although it can't be seen in the table, the means for the two levels of test score reliability do differ, but not until the 12th place value, so the difference is unimportant. Students' scores are overestimated when test score reliability equals .75 and underestimated when test score reliability equals .95, however, in both cases the amount of over/underestimation is very small. Histograms of

MB in the estimation of teachers' value-added scores by test score reliability are presented in Figure 14. Histograms of MB in the estimation of students' scores by test score reliability are presented in Figure 15 (Appendix B).

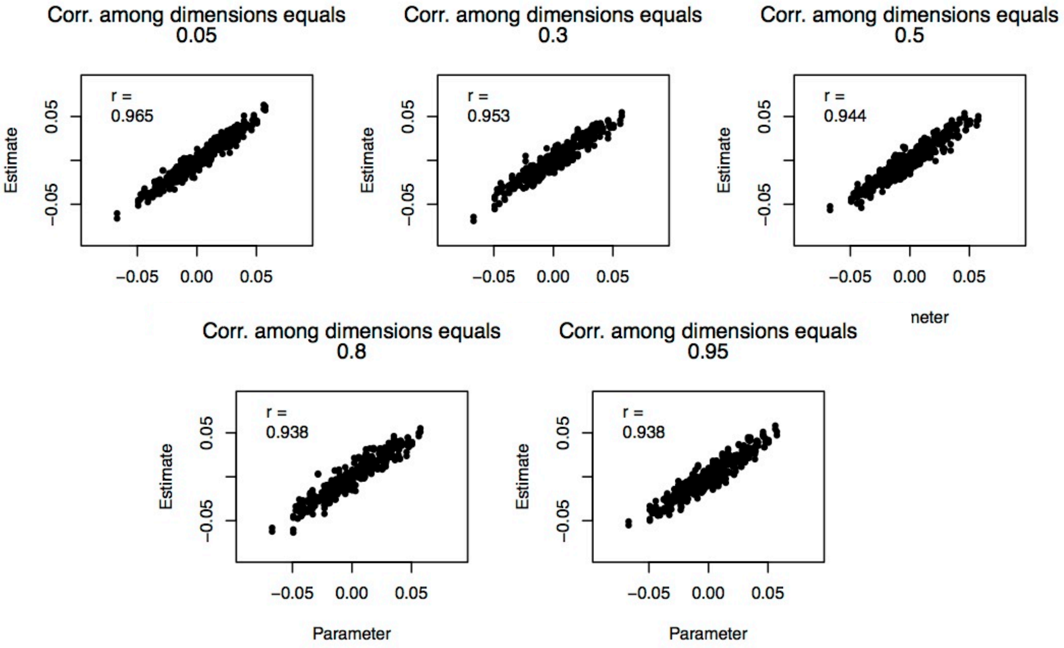

Figure 13: Scatterplots between generating parameters and estimations for teachers' value-added scores by correlation among test dimensions.

The correlation between generating parameters and estimates for teachers value-added scores was .94 when test score reliability equaled .75, and .95 when test score reliability equaled .95, providing further evidence that teachers' value-added scores are more accurately estimated as test score reliability increases. Scatterplots between generating parameters and estimates for teachers' value-added scores are presented in Figure 18 for test reliabilities of .75 and .95.

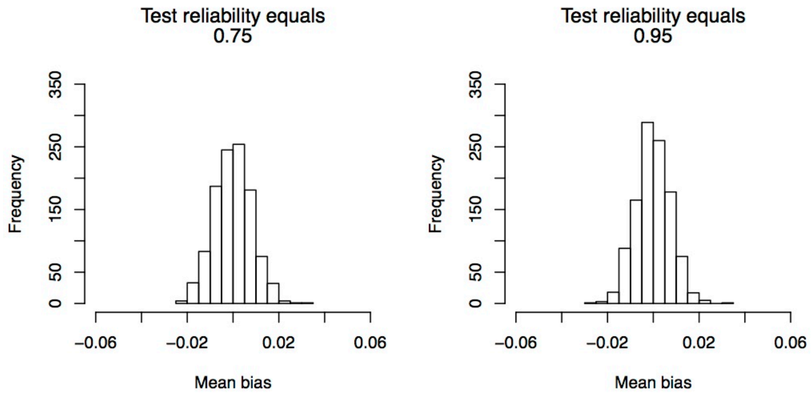

Figure 14: Histograms of MB in estimation of teachers' value-added scores by test score reliability.

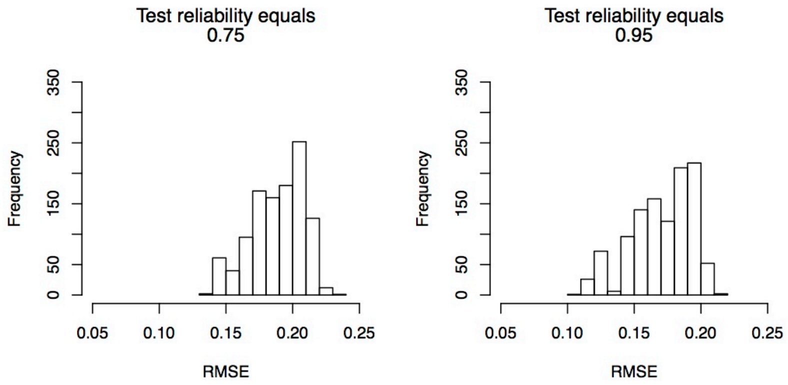

Figure 16: Histograms of RMSE in the estimation of teachers' value-added score by test score reliability.

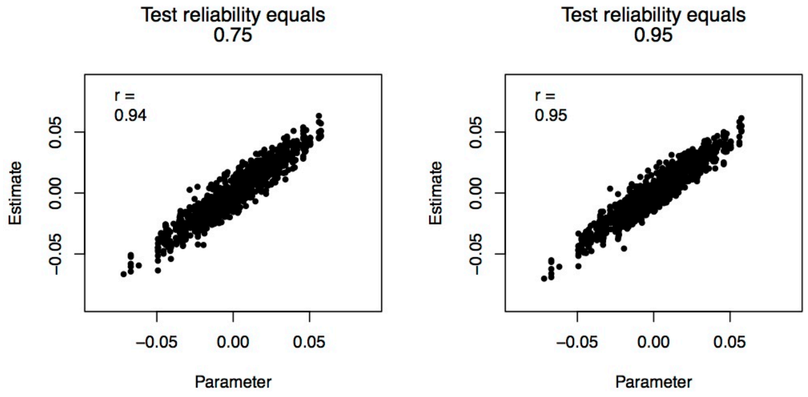

Figure 18: Scatterplots of generating parameters and estimates for teachers' value-added scores by test score reliability.

Interaction between number of dimensions and correlation among dimensions. The next step in the descriptive analyses was to examine the interaction between the number of test dimensions and the correlation among test dimensions. The interactions between test score reliability and the dimensionality variables were also examined, and no evidence of an interaction was found. Because no evidence of an interaction was found, and no interaction was hypothesized, the interactions between test score reliability and the dimensionality variables are not considered further. An examination of the interaction between the number of test dimensions and the correlation among test dimensions for MB revealed little information, because the values of MB are essentially the same for all levels of correlation among test dimensions. As such, the only interaction considered further is the interaction between the number of test dimensions and the correlation among test dimensions in regard to RMSE in the estimation of teachers' value-added scores and students' scores.

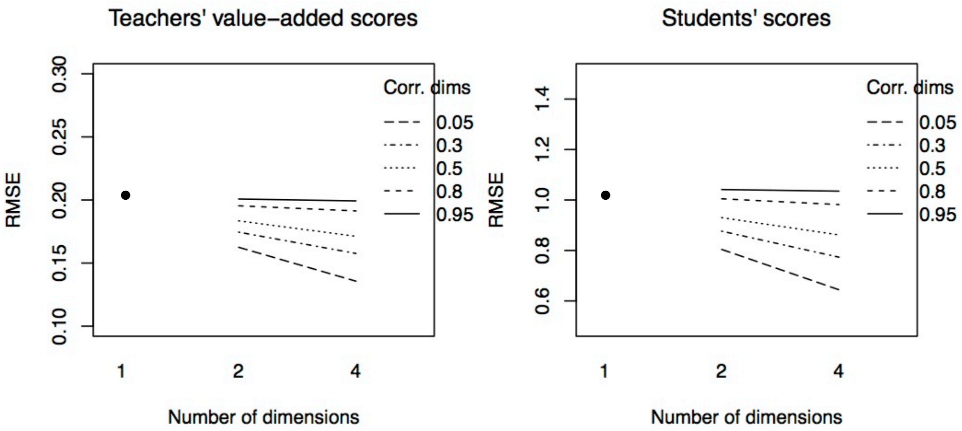

Figure 19: Interaction between the number of test dimensions and the correlation among dimensions in RMSE.

To examine the interaction between the number of test dimensions (two and four) and the correlation among test dimensions in regard to RMSE, plots comparing cells means for the crossed levels of the dimensionality variables were generated (see Figure 19). The plots reveal evidence of an interaction between the number of test dimensions and the correlation among test dimensions for both teachers' value-added scores and students' scores. When the correlation among test dimensions is high (.80 and .95), the amount of estimation bias is the same for two and four dimensions, however, when the correlation among test dimensions is moderate to low (.05, .30, or .50), there is less estimation bias when the number of test dimensions equals four relative to when the number of test dimensions equals two. Further, for the moderate to low levels of correlation among dimensions, the magnitude of the difference in the amount of bias for two and four dimensions increases as the correlation between dimensions decreases.

Inferential analyses. As noted in Chapter 3, inferential analysis of MB was performed with only one independent variable, number of test dimensions, and only for teachers' value-added scores. Descriptive results showed no evidence of differences between levels of the other independent variables in MB in the estimation of teachers' value-added scores. Further, descriptive results showed no differences between levels of any of the independent variables in MB in the estimation of students' scores. Preliminary analyses showed no evidence of clustering between design cells so a single-level analysis was performed. Results of an ANOVA show evidence of statistically significant differences between the number of test dimensions in terms of MB in the estimation of teachers' value-added scores, $F(2, 2197) = 6.88$, $p = .001$. Beta coefficients where the one-dimensional case is the reference group are -0.0013 for the two-dimensional case and -0.0021 for the four-dimensional case, showing a greater difference between one and four dimensions, than between one and two dimensions.

As described in Chapter 3, two models were fitted to RMSE. Model 1 included all variables and the interaction between the number of test dimensions and the correlation among test dimensions, but ignored data for the one-dimensional case. Model 2 excluded the correlation among test dimensions from the analysis, which allowed for examination of the one-dimensional case. Results of Model 1 for teachers' value-added scores are presented in Table 3, and the results of Model 2 for teachers' value-added scores are presented in Table 5.

Table 3: Results of Multilevel Analysis of RMSE in the Estimation of Teachers' Value-Added Scores: Model 1

Fixed Effects	Estimate	*SE*	*t*-value
Intercept (γ_{00})	0.2098	0.001	168.01
Number of dimensions (ND; γ_{01})	-0.0015	0.002	-0.88
Correlation among dimensions (CD)			
.05 (γ_{02})	-0.0383	0.002	-22.73
.30 (γ_{03})	-0.0263	0.002	-15.61
.50 (γ_{04})	-0.0173	0.002	-10.28
.80 (γ_{05})	-0.0054	0.002	-3.20
Test score reliability (γ_{06})	-0.0179	0.001	-23.90
Interaction between ND and CD			
ND * CD = .05 (γ_{07})	-0.0254	0.002	-10.65
ND * CD = .30 (γ_{08})	-0.0154	0.002	-6.46
ND * CD = .50 (γ_{09})	-0.0108	0.002	-4.52
ND * CD = .80 (γ_{10})	-0.0026	0.002	-1.09
Random Effects	Variance	*SD*	
Design cells, Level 2	0.000003	0.002	
Teachers, Level 1	0.000029	0.005	

The results presented in Table 3, show no statistically significant difference between two and four dimensions in RMSE in the estimation of teachers' value-added scores. The results related to the dummy variables representing the correlation among dimensions show that all four levels of correlation represented in the dummy variables are statistically significantly different than the correlation of .95 in regard to RMSE in the estimation of teachers' value-added scores. The pattern of significance is such that the largest difference with the correlation of .95 is for the correlation of .05, with the magnitude of the difference decreasing as the size of the correlation being compared to .95 increases. Regarding the dummy variables representing the interaction between the number of test dimensions and the correlation among test dimensions, the results show that three of the four dummy variables are significant. When the correlation among dimensions is .50 or less, there is a significant interaction between the number of test dimensions and the correlation among test dimensions. The results related to test score reliability show a statistically significant difference in RMSE in the estimation of teachers' value-added scores between test reliabilities of .75 and .95. The results of the same model fitted to students' scores revealed an identical pattern of findings, and are presented in Table 4 (Appendix A).

The results presented in Table 5 show a statistically significant difference in RMSE in the estimation of teachers' value-added scores between one and four dimensions, but not between one and two dimensions. The results also show evidence of a statistically significant difference between the two levels of test score reliability in RMSE in the estimation of teachers' value-added scores. As with Model 1, Model 2 was also fitted to students' scores and the results were the same as for teachers, and are presented in Table 6 (Appendix A).

Table 5: Results of Multilevel Analysis of RMSE in the Estimation of Teachers' Value-Added Scores: Model 2

Fixed Effects	Estimate	*SE*	*t*-value
Intercept (γ_{00})	0.2129	0.013	15.86
Number of dimensions			
Two (γ_{02})	-0.0207	0.014	-1.47
Four (γ_{03})	-0.0330	0.014	-2.35
Test score reliability (γ_{04})	-0.0176	0.008	-2.28

Random Effects	Variance	*SD*
Design cells, Level 2	0.00033	0.018
Teachers, Level 1	0.00003	0.006

Consistency in Estimated Relative Effectiveness

The second research question asked, *Does test dimensionality affect the consistency of teachers' estimated relative effectiveness?* The fourth research question asked, *Does test score reliability affect the consistency of teachers' estimated relative effectiveness?* To address these questions correlations between teachers' estimated effectiveness scores (i.e., value-added scores) were compared across levels of the independent variables. In addition, teachers were classified into one of four effectiveness levels based on their estimated effectiveness, and consistency of classification was examined.

Number of test dimensions. Data for the one-dimensional case was considered by using data for the two- and four-dimensional cases where level of the correlation among test dimensions equaled .95. The correlation between teachers' value-added scores when the number of test dimensions equaled one versus two was -.01, and when the number of test dimensions equaled one versus four was -.02. These correlations indicate no (to a slightly negative) relationship between teachers' value-added scores when the test is unidimensional versus multidimensional. The correlation between teachers' value-added scores when the number of test dimensions equaled two versus four was .22. This small correlation indicates a small relationship between teachers' value-added scores when the test comprises two versus four dimensions. The correlations comparing the unidimensional case to the multidimensional cases are likely artificially smaller than the correlation comparing the two multidimensional cases because the correlation between dimensions is not held constant when comparing the unidimensional case to the multidimensional case, but is held constant when comparing the two multidimensional cases. Nonetheless, these findings are consistent with the results regarding bias and the number of test dimensions. Recall that when the number of test dimensions equaled one or two, teachers' value-added scores were overestimated on average, and when the number of test dimensions equaled four, teachers' value-added scores were underestimated on average. Scatterplots between teachers' estimated value-added scores when the number of test dimensions equaled one versus two, one versus four, and two versus four are presented in Figure 20.

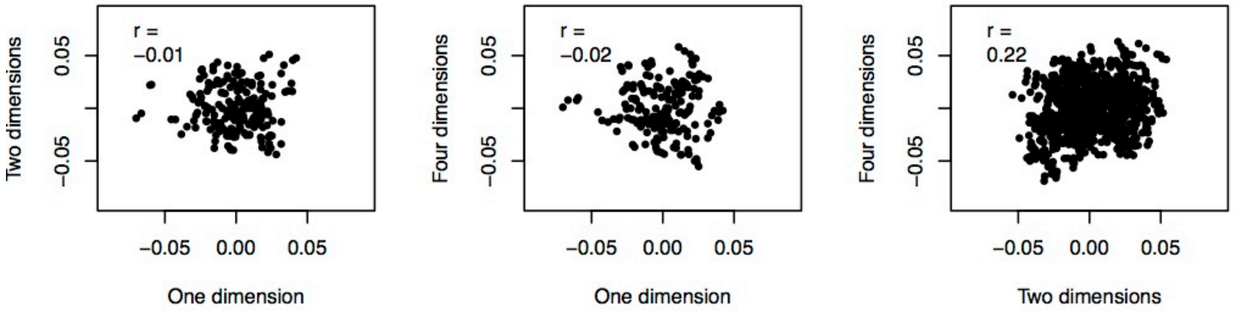

Figure 20: Scatterplots between teachers' estimated value-added scores across number of test dimensions.

Next, teachers' were classified into four effectiveness groups based on their value-added scores, and the consistency of those classifications across the number of test dimensions was examined. The percentages of teachers classified in the same effectiveness group was 26% when comparing one and two dimensions, 30% when comparing one and four dimensions, and 31% when comparing two and four dimensions. The percentages of teachers moving up or down one classification level were 39%, 32%, and 39%, and two classification levels were 20%, 26%, and 19%, when comparing one and two, one and four, and two and four dimensions, respectively. Of most concern, 15%, 12%, and 11% of teachers moved from the highest effectiveness level to the lowest effectiveness level (or vice versa), when comparing one and two, one and four, and two and four dimensions, respectively.

The results of Kruskal-Wallis tests of ordered contingency tables are, $\chi^2_{df=3}$ = 0.61, p = .895 for one versus two dimensions, $\chi^2_{df=3}$ = 9.07, p = .028 for one versus four dimensions, and $\chi^2_{df=3}$ = 34.85, p < .001 for two versus four dimensions. The larger chi-squared value for the comparison of the two multidimensional cases versus the chi-squared values comparing the unidimensional case to the two multidimensional cases suggests teachers are classified more consistently across the multidimensional cases than when comparing the unidimensional case to the multidimensional cases.

Correlation among test dimensions. Results comparing teachers estimated relative effectiveness (i.e., value-added scores) across levels of the correlation among test dimensions are presented in Table 7. The results indicate that across all possible pairwise comparisons of the correlation among test dimensions, teachers' value-added scores are estimated fairly consistently, particularly relative to the results regarding the number of test dimensions. The correlations between teachers' value-added scores across various levels of correlation among test dimensions range from .89 to .95. Similarly, across various levels of correlation among test dimensions, the percentage of teachers classified in the same effectiveness level ranges from 62% to 74%, the percentage of teachers moving up or down one effectiveness level ranges from 26% to 36%, and the percentage of teachers moving up or down two effectiveness levels is small and ranges from 1% to 2%. No teachers moved from the highest to lowest effectiveness level (or vice versa) when comparing the correlation among test dimensions. Scatterplots between teachers' value-added scores across various correlations among test dimensions are presented in Figure 21.

Table 7: Results Comparing Teachers' Estimated Relative Effectiveness for Levels of the Correlation Among Test Dimensions

		Percentage of teachers by number of classification levels moved					
Correlation among test dimensions	r	0	1	2	3	$\chi^2_{df=3}$	p-value
.05 vs. .30	.95	74%	26%	0%	0%	319.50	<.001
.05 vs. .50	.92	65%	34%	1%	0%	291.00	<.001
.05 vs. .80	.93	70%	29%	1%	0%	304.76	<.001
.05 vs. .95	.92	65%	34%	1%	0%	291.99	<.001
.30 vs. .50	.91	67%	32%	1%	0%	292.28	<.001
.30 vs. .80	.93	69%	29%	2%	0%	298.79	<.001
.30 vs. .95	.91	66%	33%	1%	0%	297.93	<.001
.50 vs. .80	.92	67%	32%	1%	0%	295.44	<.001
.50 vs. .95	.91	64%	34%	2%	0%	290.87	<.001
.80 vs. .95	.89	62%	36%	2%	0%	278.26	<.001

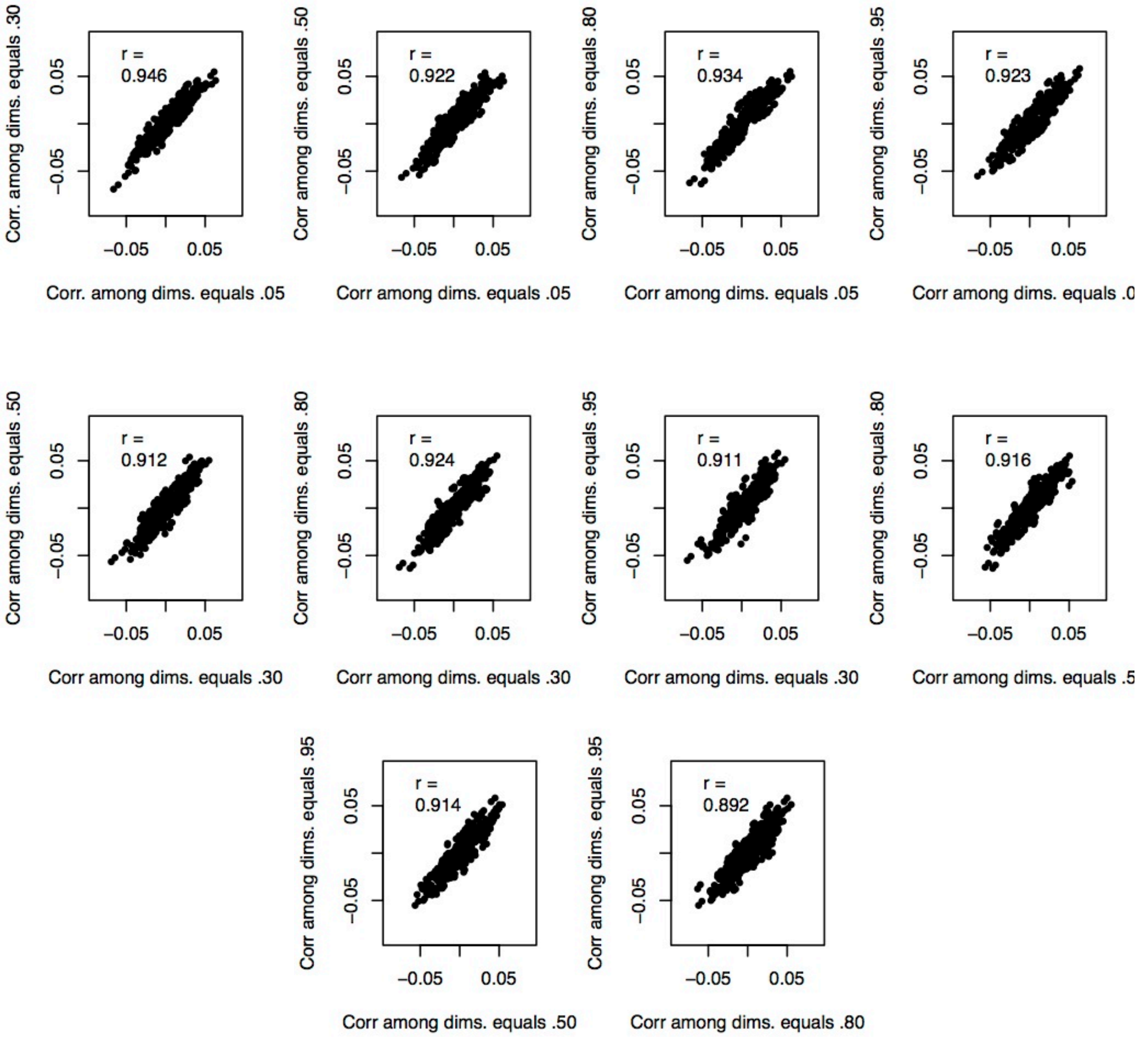

Figure 21: Scatterplots between teachers' value-added scores by correlation among test dimensions.

Test score reliability. The correlation between teachers' value-added scores derived from tests where the reliability equaled .75 versus tests where the reliability equaled .95 was .98, indicating that test score reliability has a small effect on the consistency of teachers' estimated effectiveness. However, when comparing the consistency of teachers' classification into effectiveness levels, the results show that 15% of teachers are inconsistently classified by one effectiveness level across test reliabilities. The remaining 85% of teachers are classified the same across test reliabilities. The large chi-squared value from the Kruskal-Wallis test of ordered contingency tables confirmed this results, $\chi^2_{df=3}$ = 969.32, p < .001. A scatterplot comparing teachers' value-added scores are test reliabilities is presented in Figure 22.

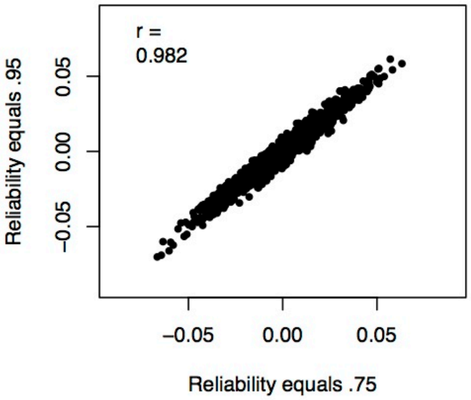

Figure 22: Scatterplot between teachers' value-added by test score reliability.

Classification accuracy. The final step in the analysis was to examine whether the accuracy of teachers' effectiveness classification varied by levels of the independent variables. Results comparing teachers' effectiveness classification based on their true value-added scores to their classification based on their estimated value-added scores, by levels of the independent variables, are presented in Table 8. Regarding the number of test dimensions, the results presented in Table 8 show that the accuracy of teacher's effectiveness classification increases as the number of test dimensions increases. This result is consistent with the results presented in Figure 8, where the correlations between teachers' true value-added scores and their estimated value-added scores increased in magnitude as the number of test dimensions increased. Together, the results presented in Table 8 and Figure 8 show that even when the correlation between teachers' true value-added scores and estimated value-added scores is large (> .90), up to 30% of teachers can be misclassified by at least one effectiveness level.

Table 8: Classification Accuracy of Teacher Effectiveness by Levels of the Independent Variables

	Percentage of teachers by number of classification levels moved					
	0	1	2	3	$\chi^2_{df=3}$	p-value
Number of test dimensions						
One	66%	31%	3%	0%	139.11	<.001
Two	73%	27%	<1%	0%	785.04	<.001
Four	76%	24%	<1%	0%	807.34	<.001
Correlation among test dimensions						
.05	91%	8%	<1%	0%	331.89	<.001
.30	91%	9%	<1%	0%	328.22	<.001
.50	88%	12%	0%	0%	311.81	<.001
.80	90%	10%	<1%	0%	315.99	<.001
.95	88%	12%	<1%	0%	303.48	<.001
Test score reliability						
.75	69%	30%	1%	0%	853.11	<.001
.95	72%	28%	<1%	0%	878.30	<.001

The results presented in Table 8 for the correlation among test dimensions show that classification accuracy decreases as the correlation among test dimensions increases for small to moderate correlations, while classification accuracy is similar for moderate to large correlations, which is consistent with the results presented in Figure 13. Classification accuracy was greater for the correlation among test dimensions than for the number of test dimensions. However, nearly 10% of teachers are misclassified by one effectiveness level when considering the correlation among test dimensions. The accuracy of teachers' effectiveness classification was greater when test score reliability equaled .95 than when test score reliability equaled .75, although the difference is small (2-3%).

In the next chapter the results detailed above are discussed, as are limitations of the present study, and suggestions for future research.

Chapter V: Discussion

The present study examined the effects of the psychometric properties of the tests used to measure student achievement on teachers' value-added scores and inferences. Specifically, the present study examined the effects of test dimensionality (number of test dimensions and correlation among test dimensions) and test score reliability on bias in the estimation of teachers' value-added scores, and consistency in teachers' estimated relative effectiveness.

Consistency of Teachers' Estimated Relative Effectiveness

Teachers' estimated value-added scores derived from the unidimensional test were unrelated to teachers' estimated value-added scores derived from either of the multidimensional tests (*r*'s = 0). Further, the relation between teachers' value-added scores estimated from the two multidimensional tests was small (*r* = .22). As a result, over 60% of teachers were classified into different effectiveness levels depending on the number of test dimensions. Most importantly, approximately 10% of teachers moved from the highest effectiveness level to the lowest effectiveness level (or vice versa) depending on the number of test dimensions. These results can be explained by the inconsistency in the direction of estimation bias (i.e., mean bias) across the number of test dimensions. Across most levels of the independent variables, teachers' value-added scores were underestimated in the present study. The exceptions were for one and two dimensions, where across levels of the other independent variables, teachers' value-added scores were overestimated. Because teachers' value-added scores were overestimated for one and two dimensions, but underestimated for four dimensions, there was very little consistency in teachers' estimated relative effectiveness rankings across the numbers of test dimensions.

Correlations between teachers' value-added scores estimated from tests with varying correlations among test dimensions were large (all $r > .89$). However, despite large correlations between teachers' estimated value-added scores across correlations among test dimensions, over 25% of teachers were classified differently depending on the correlation among test dimensions. There were no differences between the levels of correlation among test dimensions in the direction of estimation bias in teachers' value-added scores, which explains the large correlations between teachers' value-added scores estimated from tests with different correlations among test dimensions. For all levels of correlation among test dimensions, teachers' value-added scores were slightly underestimated. Teachers' estimated relative effectiveness was considerably more consistent across the correlations among test dimensions, than across numbers of test dimensions.

The correlation between teachers' value-added scores derived from a test with reliability equal to .75 and a test with reliability equal to .95 was .98, indicating high consistency across levels of test score reliability in teachers' estimated value-added scores. However, as with the correlation among test dimensions, despite a large correlation between value-added scores estimated from tests with differing reliabilities, 15% of teachers were classified differently based on the two tests. There were no differences between test reliabilities in the direction of bias in teachers' estimated value-added scores, which explains the large correlation between value-added scores derived from tests with different reliability. For both levels of test score reliability, teachers' value-added scores were slightly underestimated.

The results described above suggest that even when the overall MB in teachers' estimated value-added scores is small, and similar across levels of the independent variables, variations in the properties of the tests used to measure student achievement have a considerable detrimental effect on the use of value-added scores to make high-stakes decisions about teachers, as 15% to 60% of teachers are classified into different effectiveness levels based on variations in test dimensionality and reliability.

Estimation Accuracy

The amount of bias in the estimation of teachers' value-added scores (as measured by the root-mean-squared-error [RMSE]) was statistically significantly different for the two levels of test score reliability (.75 and .95), although descriptively the difference is small (0.02), so the statistical significance can likely be attributed to the large sample size

(J = 2,000). Nonetheless, the amount of bias in the estimation of teachers' value-added scores for the test with reliability equal to .95 (RMSE = 0.17) was less than the test with reliability equal to .75 (RMSE = 0.19). This difference can be explained by the fact that tests with lower reliability by definition contain more measurement error, which introduces more random variability into students' observed scores than tests that contain less measurement error. Greater variability in students' observed scores in turn introduces greater variability into teachers' value-added scores, which leads to greater bias in the estimation of teachers' value-added scores. As the reliability of a test increases, teachers' value-added scores are estimated more accurately.

Descriptively, as the number of test dimensions increased, teachers' value-added scores were estimated more accurately, which is consistent with previous work examining the effect of the number of test dimensions on bias in the estimation of students' scores (e.g., Harrison, 1986). However, the only statistically significant difference in the amount of estimation bias was between one and four dimensions, and statistical significance can again likely be attributed to the large sample size. In addition, it is unlikely the non-statistically significant findings associated with the difference between one and two dimensions, and two and four dimensions, are associated with a lack of power given the large sample size in the present study. As such, the present study found no important differences between the numbers of test dimensions in the amount of estimation bias in teachers' value-added scores. It is important to note that even though the present study found no differences in the *amount* of estimation bias based on the number of test dimensions, teachers' value-added scores were estimated least consistently across the different numbers of test dimensions, because there were significant differences between the numbers of test dimensions in the *direction* of estimation bias.

As the correlation among test dimensions decreased, the amount of bias in the estimation of teachers' value-added scores also decreased, the opposite of what was hypothesized (see research questions 1.2 and 2.2). It was hypothesized that the amount of estimation bias would decrease as the correlation among test dimensions *increased*, because it was thought that treating multidimensional data as unidimensional (by computing a single composite score) would bias students' score estimates and consequently teachers' value-added score estimates. However, estimation bias in teachers' value-added scores increased as the correlation among test dimensions increased because composite scores associated with larger correlations among test dimensions had greater variability than composite scores associated with smaller correlations among test dimensions. The variance of a linear combination, such as a composite score, is equal to the sum of its component variances and their covariance. When the correlation among dimensions is greater, the covariance is also greater, so the resulting composite score has greater variability, which leads to more bias in the estimation of students' scores and consequently teachers' value-added scores. Further, when the correlation among dimensions was large (i.e., .80 and .95), the amount of bias in the estimation of teachers' value-added scores was the same for two and four dimensions. In contrast, when the correlation among dimensions was moderate to small (i.e., .05, .30, and .50), the amount of estimation bias was significantly greater for two dimensions than for four dimensions. The dimensionality specification that led to the least amount of bias in the estimation of teachers' value-added scores was for the test comprised of four dimensions, where the correlation among dimensions equaled .05. Nevertheless, estimation bias was present for all levels of correlation among test dimensions, which confirms the work of Martineau (2006). Martineau proved mathematically that teachers' value-added scores are biased by the presence of (ignored) multidimensionality, even when the correlation among dimensions is large.

Limitations and Future Research

The present study has several important limitations. First, because the value-added model used to estimate teachers' value-added scores was constant in the present study, it is not known if the results are sensitive to the choice of value-added model. The choice of value-added model is probably most relevant to issues of test score reliability, because how growth is specified in a value-added model will affect how measurement error is distributed throughout the model. Value-added models that use gain scores may be associated with more bias in the estimation of teachers' value-added scores, especially if the measurement errors from the two scores used to calculate the gain scores are unrelated. Future research should examine the interaction between test score reliability and how growth is measured in a value-added model on bias in the estimation of teachers' value-added scores and consistency in teachers' estimated relative effectiveness.

It is also not known how sensitive the results regarding teachers' estimated relative effectiveness are to the decision to classify teachers into four effectiveness groups, and to the decision to make the groups of equal size. The choice to use four groups was based on how states and districts use value-added scores to make decisions about teachers, but there is no evidence to suggest that four is the ideal number of groups, or that classifying teachers into effectiveness groups is useful at all. The choice to make the groups of equal size also reflects the use of value-added scores by states and districts, but may not be ideal, particularly given the normative nature of value-added data.

Another limitation of the present study is that measurement error is assumed to be constant across the score scale. However, this assumption is unrealistic because there are often fewer items providing information in the tails of the score scale distributions, which leads to greater measurement error in the tails of score scale distributions. Future research should consider the effects of conditional standard errors of measurement on teachers' value-added scores, particularly in light of the limitation discussed next.

Students and teachers in the present study were randomly assigned to classrooms, which is unrealistic given the presence of tracking in many schools. The precise nature of tracking in schools is unknown, because data related to tracking is rarely available publically, but it is often assumed that teachers and students are not assigned to classrooms randomly. Instead, one of two scenarios likely takes place (implicitly or explicitly): a) experienced teachers are systematically assigned to classrooms containing a higher percentage of lower performing students, and higher performing students are taught by less experienced teachers, or b) experienced teachers are systematically assigned to classrooms with a higher percentage of high performing students, and lower performing students are taught by less experienced teachers. Whereas the first scenario may be more desirable, the increased use of student test score data to evaluate teachers, likely increases the frequency with which the second scenario occurs, as experienced teachers with seniority may opt into classrooms with higher performing students in an attempt to increase their evaluation.

Issues related to student tracking are particularly problematic for estimation bias in teachers' value-added scores, because measurement error is often distributed differently across the score scale. As such, teachers' with more students in the tails of the score distribution will have value-added scores comprised of more measurement error than teachers with students in the middle of the score distribution. In addition to examining the effects of conditional standard errors of measurement on bias in the estimation of teachers' value-added scores, future research should examine how measurement error and student tracking combine to effect the amount of bias in value-added score estimates.

Finally, the present study is a simulation study only, and the results would have been strengthened by an examination of real data. Unfortunately, real value-added data was not available. Future research should seek to confirm the findings of the present study using real data.

Conclusions

The results of the present study show that teachers' value-added scores are affected by the psychometric properties of the student achievement tests used to derive such scores. In particular, even when the amount and direction of estimation bias is small, teachers' estimated relative effectiveness can be greatly affected, with up to 60% of teachers misclassified into different effectiveness levels based on variations in test properties. These findings are consistent with previous methodological work on the use of student test score data to evaluate teachers, and caution against the use of student test score data to make high-stakes decisions about teachers. Given the results of the present study regarding number of test dimensions, the use of student test score data to evaluate teachers is likely invalid even for low-stakes decisions. The primary principle of all of measurement is that the validity of a test's use depends on the intended purpose of the test. Given that none of the tests currently used to estimate teachers' value-added scores are primarily designed to evaluate teachers, their use for such a purpose is inappropriate.

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113 – 127. doi: 10.1177/014662168901300201

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67 – 91.

American Educational Research Association (2015). AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. *Educational Researcher, 44*, 448-452. doi: 10.3102/0013189X15618385

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37–48. doi: 10.1177/014662168500900104

Birenbaum, M., & Tatsuoka, K. K. (1982). On the dimensionality of achievement test data. *Journal of Educational Measurement, 19*, 259 – 266.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Educational Measurement: Issues and Practice, 28*, 3 – 14.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admissions Test. *Journal of Educational Measurement, 32*, 79 – 96.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189 – 199. doi: 10.1177/014662168300700207

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

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. American Educational Research Journal, 48, 794 - 831. doi: 0.3102/0002831210387916

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44,* 47-67.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics 31,* 35-62.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND Corporation.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48,* 163-193. doi: 10.3102/0002831210362589

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

Reckase, M. D. (1990, April). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. Paper presented at the annual meeting of the American Educational Research Association: Boston, MA.

Reckase, M. D. (2009). *Multidimensional Item Response Theory: Statistics for Social and Behavioral Sciences.* New York, NY: Springer.

Sass, T. R. (2008). *The stability of value added measure of teacher quality and implications for teacher compensation policy.* Briefing paper 4. Washington, DC: Urban Institute, Center for Analysis of Longitudinal Data in Education Research.

Value-Added Research Center (2013) Technical Report: Minneapolis Value-Added Model. Author: Madison, WI.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Appendix A

This appendix contains two tables related to bias in the estimation of students' scores. Table 4 contains the results of Model 1 for students' scores. Table 6 contains the results of Model 2 for students' scores.

Table 4.

Results of Multilevel Analysis of RMSE in the Estimation of Students' Scores: Model 1

Fixed Effects	Estimate	*SE*	*t*-value
Intercept (γ_{00})	1.0969	0.006	176.95
Number of dimensions (ND; γ_{01})			
Correlation among dimensions (CD)	-0.0065	0.008	-0.78
.05 (γ_{02})	-0.2375	0.008	-28.41
.30 (γ_{03})	-0.1652	0.008	-19.77
.50 (γ_{04})	-0.1117	0.008	-13.36
.80 (γ_{05})	-0.0368	0.008	-4.40
Test score reliability (γ_{06})	-0.1107	0.004	-29.61
Interaction between ND and CD			
ND * CD = .05 (γ_{07})	-0.1533	0.012	-12.97
ND * CD = .30 (γ_{08})	-0.0965	0.012	-8.16
ND * CD = .50 (γ_{09})	-0.0617	0.012	-5.22
ND * CD = .80 (γ_{10})	-0.0165	0.012	-1.39
Random Effects	Variance	*SD*	
Design cells, Level 2	0.00007	0.008	
Students, Level 1	0.00082	0.029	

Table 6.

Results of Multilevel Analysis of RMSE in the Estimation of Students' Scores: Model 2

Fixed Effects	Estimate	*SE*	*t*-value
Intercept (γ_{00})	1.1065	0.082	13.42
Number of dimensions			
Two (γ_{02})	-0.1207	0.087	-1.40
Four (γ_{03})	-0.1928	0.087	-2.23
Test score reliability (γ_{04})	-0.1090	0.048	-2.29
Random Effects	Variance	*SD*	
Design cells, Level 2	0.0125	0.112	
Students, Level 1	0.0008	0.029	

Appendix B

This appendix contains 6 figures related to the estimation of students' scores. Figures 5 and 7 contain histograms related to bias in the estimation of students' scores and the number of test dimensions. Figures 10 and 12 contain histograms related to bias in the estimation of students' scores and the correlation among test dimensions. Figures 15 and 17 contain histograms related to bias in the estimation of students' scores and test score reliability.

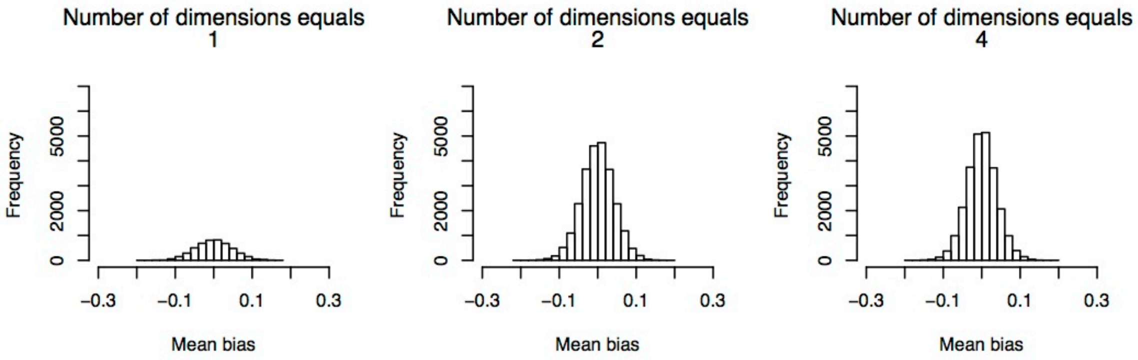

Figure 5: Histograms of mean bias in the estimation of students' scores by number of test dimensions.

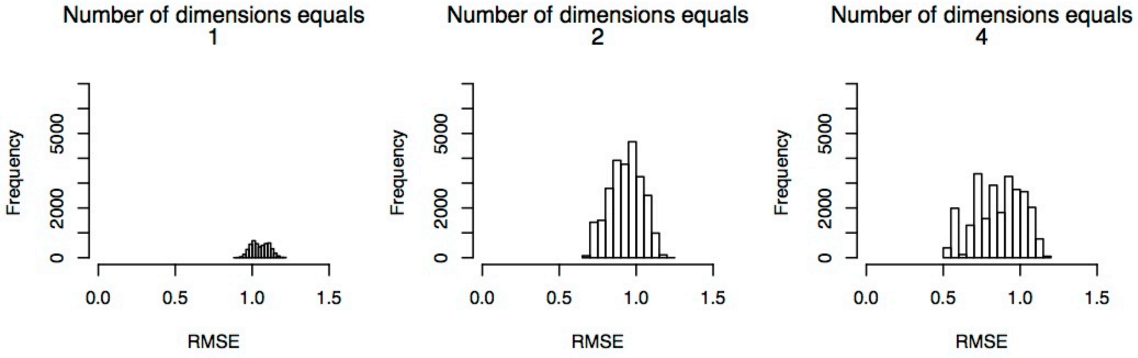

Figure 7: Histograms of RMSE in the estimation of students' scores by number of test dimensions.

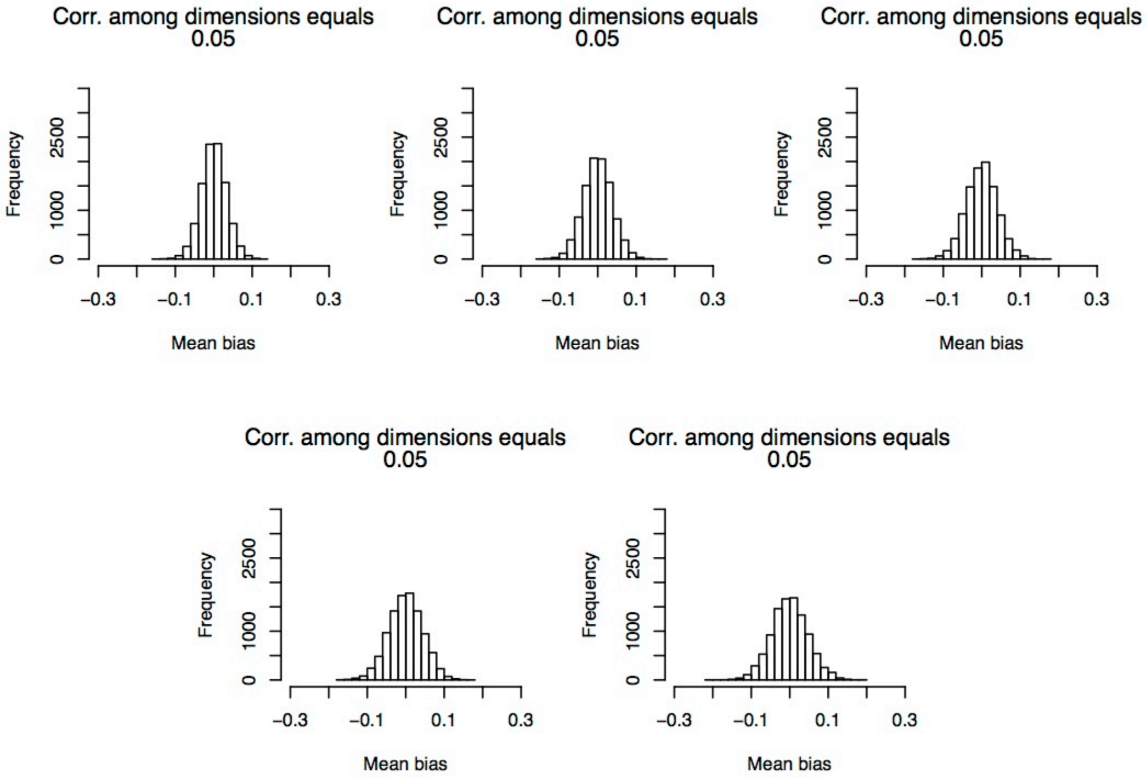

Figure 10: Histograms of mean bias in the estimation of students' scores by the correlation among test dimensions.

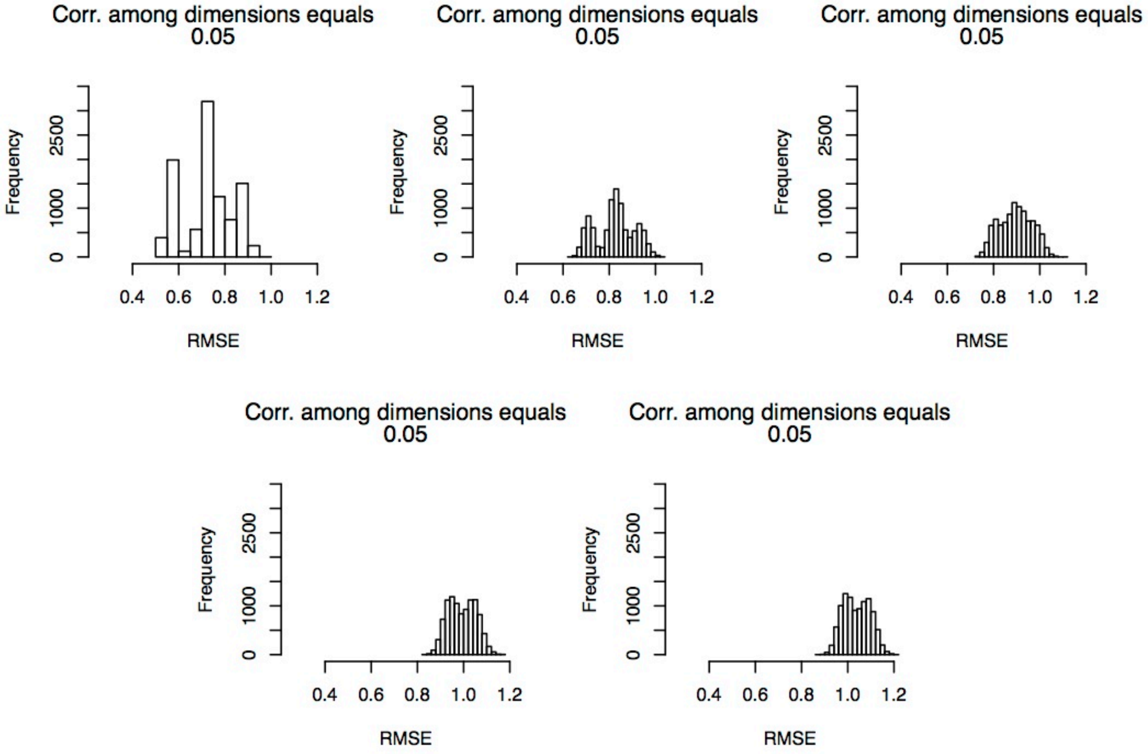

Figure 12: Histograms of RMSE in estimation of students' scores by correlation among test dimensions.

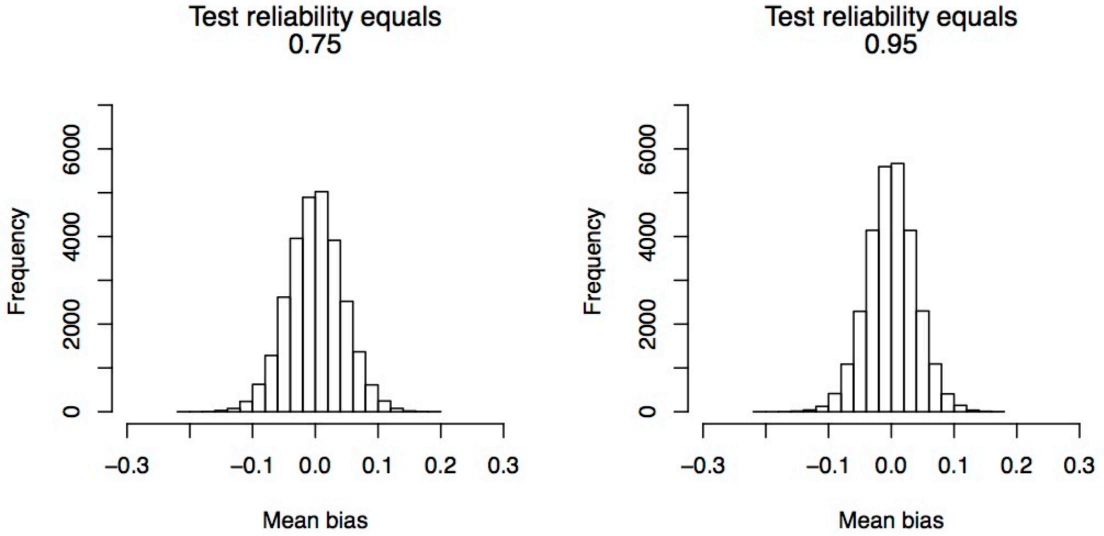

Figure 15: Histograms of MB in the estimation of students' scores by test score reliability.

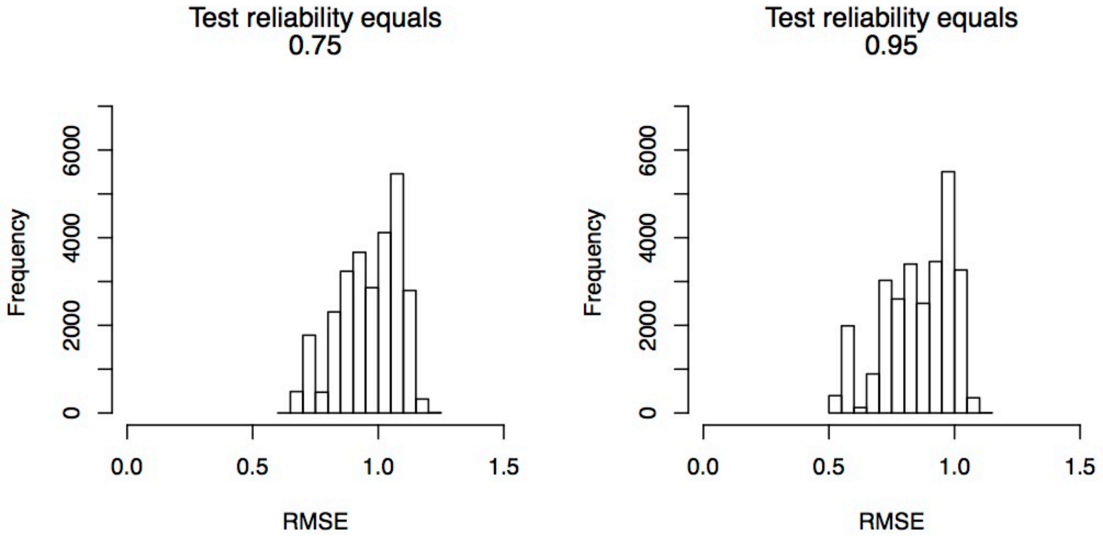

Figure 17: Histograms of RMSE in estimation of students' scores by test score reliability.

Appendix C

This appendix contains the R code used to generate the simulated data, and to fit the simulated data to a value-added model to obtained teachers' estimated value-added scores. The code represents one replication.

```
#Install the following packages then load libraries.
library(psych)
library(lme4)

#Clear environment.
rm(list=ls())

#Set seed.
set.seed(8279)

#Specify the number of students and teachers.
j <-100
nsubj <-25
n <-nsubj*j

##### Start replications loop for 1 dimension ####
file.i <-vector()
file.j <-vector()
for (d in 1:500){
print(d)
  #### Generate X (4th grade math achievement) ####
  #Create true scores.
  x.1 <-data.frame(rnorm(n, 0, 1))
  names(x.1) <-c("x.dim1.1")

  #Create observed scores for two levels of reliability.
  x.1$x.msmt.75.1 <-rnorm(n, 0, .5000) #Reliability = .75
  x.1$x.msmt.95.1 <-rnorm(n, 0, .2236) #Reliability = .95
  x.1$x.comp.75.1 <-x.1$x.dim1.1 + x.1$x.msmt.75.1
  x.1$x.comp.95.1 <-x.1$x.dim1.1 + x.1$x.msmt.95.1

  #Restructure data and add id vars.
  x.1 <-stack(x.1[,4:5])
  x.1$ind <-rep(c(".75",".95"), each=2500)
  names(x.1) <-c("X","reliability")
  x.1$no.dim <-"1"
  x.1$cor.dim <-NA
  x.1$N.id <-rep(1:n, times=2)
  x.1$J.id <-rep(rep(1:j, each=nsubj), times=2)

  #### Generate Y (5th grade math achievement) ####
  #Create true scores.
  y.1 <-data.frame(rnorm(n, 0, 1))
  names(y.1) <-c("y.dim1.1")

  #Create observed scores for two levels of reliability.
  y.1$y.msmt.75.1 <-rnorm(n, 0, .5000) #Reliability = .75
  y.1$y.msmt.95.1 <-rnorm(n, 0, .2236) #Reliability = .95
```

```
y.1$y.comp.75.1 <-y.1$y.dim1.1 + y.1$y.msmt.75.1
y.1$y.comp.95.1 <-y.1$y.dim1.1 + y.1$y.msmt.95.1

#Restructure data and add id vars.
y.1 <-stack(y.1[,4:5])
y.1$ind <-rep(c(".75",".95"), each=2500)
names(y.1) <-c("Y","reliability")
y.1$no.dim <-"1"
y.1$cor.dim <-NA
y.1$N.id <-rep(1:n, times=2)
y.1$J.id <-rep(rep(1:j, each=nsubj), times=2)

#### Merge X and Y and restructure ####
data.i <-cbind(x.1, y.1)
data.i <-data.i[,c(5:6,2:4,1,7)]

#### Add teacher effects to X and Y ####
#Generate random teacher effects.
U.x <-rep(rnorm(j, 0, .5), times=2) #Grade 4
U.y <-rep(rnorm(j, 0, .5), times=2) #Grade 5
J.id <-rep(seq(1:j), times=2)
data.j <-data.frame(cbind(J.id, U.x, U.y))
names(data.j) <-c("J.id","U.x.parm","U.y.parm")
data.j$reliability <-rep(c(".75",".95"), each=100)
data.j$no.dim <-1
data.j$cor.dim <-NA

#Merge teacher effects with student data.
temp <-merge(data.i, data.j, by=c("J.id", "reliability"),
             sort=FALSE)
temp <-temp[order(temp$reliability, temp$J.id, temp$N.id),]
data.i <-temp

#Add random teacher effects to X and Y.
data.i$X <-data.i$X + data.i$U.x.parm
data.i$Y <-data.i$Y + data.i$U.y.parm

#### Adjust Y to reflect correlation with X (.70) ####
data.i$Y <-data.i$Y + (data.i$X*.96)

#### Rename data.i variables to reflect parameters ####
data.i <-data.i[,c(1,3,2,4:7)]
names(data.i) <-
  c("J.id","N.id","reliability","no.dim","cor.dim",
                "X.parm","Y.parm")

#### Restructure data.j and add id variables ####
data.j <-data.j[,c(1,4:6,2:3)]
```

```
  #### Fit VAA model to simulated data ####
  out.i <-vector()
  out.j <-vector()
  for (r in c(".75",".95")){
    data.temp <-data.i[data.i$reliability==r,]
    lmer.temp <-lmer(Y.parm ~ 1 + X.parm + (1|J.id),
                     data=data.temp)
    fit.temp <-as.data.frame(unlist(fitted(lmer.temp)))
    vam.temp <-as.data.frame(unlist(ranef(lmer.temp)))
    out.i <-rbind(out.i, fit.temp)
    out.j <-rbind(out.j, vam.temp)
  }

  #### Merge student estimates with student data ####
  out.i <-data.frame(out.i, row.names=NULL)
  names(out.i) <-c("Y.estm")
  temp.i <-cbind(data.i, Y.estm=out.i)

  #### Compute bias estimates ####
  temp.i$Y.bias <-temp.i$Y.estm - temp.i$Y.parm  #Bias
  temp.i$Y.rmse <-temp.i$Y.bias^2 #RMSE

  #### Merge teacher estimates with teacher data ####
  out.j <-data.frame(out.j, row.names=NULL)
  names(out.j) <-c("U.estm")
  temp.j <-cbind(data.j, U.y.estm=out.j$U.estm)

  #### Compute bias estimates ####
  temp.j$U.bias <-temp.j$U.y.estm - temp.j$U.y.parm  #Bias
  temp.j$U.rmse <-temp.j$U.bias^2 #RMSE
  #### Close replication loop ####
  file.i <-rbind(file.i, temp.i)
  file.j <-rbind(file.j, temp.j)
}
file.i <-data.frame(file.i)
file.j <-data.frame(file.j)

#### Aggregate values across 500 replications ####
#Create variables to aggregate on.
file.i$rep.id <-rep(1:5000, times=500)
file.j$rep.id <-rep(1:200, times=500)

#Aggregate data across 500 replications.
aggr.i <-aggregate(file.i[,6:11], by=list(file.i$rep.id), mean)
aggr.j <-aggregate(file.j[,5:10], by=list(file.j$rep.id), mean)

#Merge aggregated data with id vars.
```

```
merg.i <-cbind(data.i[,1:5], aggr.i[,2:6])
merg.j <-cbind(data.j[,1:4], aggr.j[,2:6])

#### Finish computing RMSE ####
merg.i$Y.rmse <-sqrt(merg.i$Y.rmse)
merg.j$U.rmse <-sqrt(merg.j$U.rmse)

#### Write files ####
write.csv(file.i, file="data.i.long.1.csv")
write.csv(file.j, file="data.j.long.1.csv")
write.csv(merg.i, file="data.i.1.csv")
write.csv(merg.j, file="data.j.1.csv")

##### Start replications loop for 2 dimensions ####
file.i <-vector()
file.j <-vector()
for (d in 1:500){
print(d)
  #### Generate X (4th grade math achievement) ####
  #Define means and SDs.
  x.means.2 <-c(0,0)
  x.sigma.1.2 <-matrix(c(1, .05, .05, 1), nrow=2, byrow=TRUE)
  x.sigma.2.2 <-matrix(c(1, .30, .30, 1), nrow=2, byrow=TRUE)
  x.sigma.3.2 <-matrix(c(1, .50, .50, 1), nrow=2, byrow=TRUE)
  x.sigma.4.2 <-matrix(c(1, .80, .80, 1), nrow=2, byrow=TRUE)
  x.sigma.5.2 <-matrix(c(1, .95, .95, 1), nrow=2, byrow=TRUE)
  x.sigma.2 <-list(x.sigma.1.2, x.sigma.2.2, x.sigma.3.2,
                   x.sigma.4.2, x.sigma.5.2)

  #Create subscale scores.
  x.2 <-data.frame()
  for (s in 1:5){
    temp.1.2 <-matrix(unlist(x.sigma.2[s]), nrow=2, byrow=TRUE)
    temp.2.2 <-data.frame(rmvnorm(n = n, mean = x.means.2, sigma
                          = temp.1.2))
    names(temp.2.2) <-c("x.dim1.2", "x.dim2.2")
    x.2 <-rbind(x.2, temp.2.2)
  }
  x.2 <-data.frame(x.2)

  #Create true composite score.
  x.2$X.comp.2 <-(x.2$x.dim1.2 + x.2$x.dim2.2)/2

  #Create observed scores for two levels of reliability.
  x.2$x.msmt.75.2 <-rep(rnorm(n, 0, .5000), 5) #Reliability = .75
  x.2$x.msmt.95.2 <-rep(rnorm(n, 0, .2236), 5) #Reliability = .95
  x.2$X.comp.75.2 <-x.2$X.comp.2 + x.2$x.msmt.75.2
  x.2$X.comp.95.2 <-x.2$X.comp.2 + x.2$x.msmt.95.2
```

```
#Restructure data and add id vars.
x.2 <-stack(x.2[,6:7])
x.2$ind <-rep(c(".75",".95"), each=12500)
names(x.2) <-c("X","reliability")
x.2$no.dim <-"2"
x.2$cor.dim <-rep(c(".05",".30",".50",".80",".95"), each=2500)
x.2$N.id <-rep(1:n, times=10)
x.2$J.id <-rep(rep(1:j, each=nsubj), times=10)

#### Generate Y (5th grade math achievement) ####
#Define means and SDs.
y.means.2 <-c(0,0)
y.sigma.1.2 <-matrix(c(1, .05, .05, 1), nrow=2, byrow=TRUE)
y.sigma.2.2 <-matrix(c(1, .30, .30, 1), nrow=2, byrow=TRUE)
y.sigma.3.2 <-matrix(c(1, .50, .50, 1), nrow=2, byrow=TRUE)
y.sigma.4.2 <-matrix(c(1, .80, .80, 1), nrow=2, byrow=TRUE)
y.sigma.5.2 <-matrix(c(1, .95, .95, 1), nrow=2, byrow=TRUE)
y.sigma.2 <-list(y.sigma.1.2, y.sigma.2.2, y.sigma.3.2,
                 y.sigma.4.2, y.sigma.5.2)

#Create subscale scores.
y.2 <-data.frame()
for (s in 1:5){
  temp.1.2 <-matrix(unlist(y.sigma.2[s]), nrow=2, byrow=TRUE)
  temp.2.2 <-data.frame(rmvnorm(n = n, mean = y.means.2, sigma
                      = temp.1.2))
  names(temp.2.2) <-c("y.dim1.2", "y.dim2.2")
  y.2 <-rbind(y.2, temp.2.2)
}
y.2 <-data.frame(y.2)

#Create true composite score.
y.2$Y.comp.2 <-(y.2$y.dim1.2 + y.2$y.dim2.2)/2

#Create observed scores for two levels of reliability.
y.2$y.msmt.75.2 <-rep(rnorm(n, 0, .5000), 5) #Reliability = .75
y.2$y.msmt.95.2 <-rep(rnorm(n, 0, .2236), 5) #Reliability = .95
y.2$Y.comp.75.2 <-y.2$Y.comp.2 + y.2$y.msmt.75.2
y.2$Y.comp.95.2 <-y.2$Y.comp.2 + y.2$y.msmt.95.2

#Restructure data and add id vars.
y.2 <-stack(y.2[,6:7])
y.2$ind <-rep(c(".75",".95"), each=12500)
names(y.2) <-c("Y","reliability")
y.2$no.dim <-"2"
y.2$cor.dim <-rep(c(".05",".30",".50",".80",".95"), each=2500)
y.2$N.id <-rep(1:n, times=10)
y.2$J.id <-rep(rep(1:j, each=nsubj), times=10)
```

```
#### Merge X and Y and add ID vars ####
data.i <-cbind(x.2, y.2)
data.i <-data.i[,c(5:6,2:4,1,7)]

#### Add teacher effects to X and Y ####
#Generate random teacher effects.
U.x <-rnorm(j, 0, 1) #Grade 4
U.y <-rnorm(j, 0, 1) #Grade 5
J.id <-seq(1:j)
data.j <-data.frame(cbind(J.id, U.x, U.y))
names(data.j) <-c("J.id","U.x.parm","U.y.parm")
#Merge teacher effects with student data.
temp <-merge(data.i, data.j, by="J.id", sort=FALSE)
temp <-temp[order(temp$reliability, temp$cor.dim, temp$J.id,
                  temp$N.id),]
data.i <-temp

#Add random teacher effects to X and Y.
data.i$X <-data.i$X + data.i$U.x.parm
data.i$Y <-data.i$Y + data.i$U.y.parm

#### Adjust Y to reflect correlation with X (.70) ####
data.i$Y <-data.i$Y + (data.i$X*.96)

#### Rename data.i variables to reflect parameters ####
data.i <-data.i[,c(-8,-9)]
names(data.i) <-
  c("J.id","N.id","reliability","no.dim","cor.dim",
                  "X.parm","Y.parm")

#### Restructure data.j and add id variables ####
data.j <-rbind(data.j, data.j, data.j, data.j, data.j,
               data.j, data.j, data.j, data.j, data.j)
data.j$reliability <-rep(c(.75, .95), each=500)
data.j$no.dim <-2
data.j$cor.dim <-rep(rep(c(.05,.30,.50,.80,.95), each=100),
                     times=2)
data.j <-data.j[,c(1,4:6,2:3)]

#### Fit VAA model to simulated data for 2 dimensions ####
out.i <-vector()
out.j <-vector()
for (r in c(".75",".95")){
  data.1 <-data.i[data.i$reliability==r,]
  out.1 <-vector()
  out.2 <-vector()
  for (c in c(".05",".30",".50",".80",".95")){
    data.temp <-data.1[data.1$cor.dim==c,]
```

```
      lmer.temp <-lmer(Y.parm ~ 1 + X.parm + (1|J.id),
                       data=data.temp)
      fit.temp <-as.data.frame(unlist(fitted(lmer.temp)))
      vam.temp <-as.data.frame(unlist(ranef(lmer.temp)))
      out.1 <-rbind(out.1, fit.temp)
      out.2 <-rbind(out.2, vam.temp)
    }
    out.i <-rbind(out.i, out.1)
    out.j <-rbind(out.j, out.2)
  }

  #### Merge student estimates with student data ####
  out.i <-data.frame(out.i, row.names=NULL)
  names(out.i) <-c("Y.estm")
  temp.i <-cbind(data.i, Y.estm=out.i)

  #### Compute bias estimates ####
  temp.i$Y.bias <-temp.i$Y.estm - temp.i$Y.parm  #Bias
  temp.i$Y.rmse <-temp.i$Y.bias^2 #RMSE

  #### Merge teacher estimates with teacher data ####
  out.j <-data.frame(out.j, row.names=NULL)
  names(out.j) <-c("U.estm")
  temp.j <-cbind(data.j, U.y.estm=out.j$U.estm)

  #### Compute bias estimates ####
  temp.j$U.bias <-temp.j$U.y.estm - temp.j$U.y.parm  #Bias
  temp.j$U.rmse <-temp.j$U.bias^2 #RMSE

  #### Close replication loop ####
  file.i <-rbind(file.i, temp.i)
  file.j <-rbind(file.j, temp.j)
}
file.i <-data.frame(file.i)
file.j <-data.frame(file.j)

#### Aggregate values across 500 replications ####
#Create variables to aggregate on.
file.i$rep.id <-rep(1:25000, times=500)
file.j$rep.id <-rep(1:1000, times=500)

#Aggregate data across 500 replications.
aggr.i <-aggregate(file.i[,6:11], by=list(file.i$rep.id), mean)
aggr.j <-aggregate(file.j[,5:10], by=list(file.j$rep.id), mean)

#Merge aggregated data with id vars.
merg.i <-cbind(data.i[,1:5], aggr.i[,2:6])
merg.j <-cbind(data.j[,1:4], aggr.j[,2:6])
```

```
#### Finish computing RMSE ####
merg.i$Y.rmse <-sqrt(merg.i$Y.rmse)
merg.j$U.rmse <-sqrt(merg.j$U.rmse)

#### Write files ####
write.csv(file.i, file="data.i.long.2.csv")
write.csv(file.j, file="data.j.long.2.csv")
write.csv(merg.i, file="data.i.2.csv")
write.csv(merg.j, file="data.j.2.csv")

##### Start replications loop for 4 dimensions ####
file.i <-vector()
file.j <-vector()
for (d in 1:500){
print(d)
  #### Generate X (4th grade math achievement) ####
  #Define means and SDs.
  x.means.4 <-c(0,0,0,0)
  x.sigma.1.4 <-matrix(c(1.00, 0.05, 0.05, 0.05,
                         0.05, 1.00, 0.05, 0.05,
                         0.05, 0.05, 1.00, 0.05,
                         0.05, 0.05, 0.05, 1.00), nrow=4,
               byrow=TRUE)
  x.sigma.2.4 <-matrix(c(1.00, 0.30, 0.30, 0.30,
                         0.30, 1.00, 0.30, 0.30,
                         0.30, 0.30, 1.00, 0.30,
                         0.30, 0.30, 0.30, 1.00), nrow=4,
               byrow=TRUE)
  x.sigma.3.4 <-matrix(c(1.00, 0.50, 0.50, 0.50,
                         0.50, 1.00, 0.50, 0.50,
                         0.50, 0.50, 1.00, 0.50,
                         0.50, 0.50, 0.50, 1.00), nrow=4,
               byrow=TRUE)
  x.sigma.4.4 <-matrix(c(1.00, 0.80, 0.80, 0.80,
                         0.80, 1.00, 0.80, 0.80,
                         0.80, 0.80, 1.00, 0.80,
                         0.80, 0.80, 0.80, 1.00), nrow=4,
               byrow=TRUE)
  x.sigma.5.4 <-matrix(c(1.00, 0.95, 0.95, 0.95,
                         0.95, 1.00, 0.95, 0.95,
                         0.95, 0.95, 1.00, 0.95,
                         0.95, 0.95, 0.95, 1.00), nrow=4,
               byrow=TRUE)
  x.sigma.4 <-list(x.sigma.1.4, x.sigma.2.4, x.sigma.3.4,
                   x.sigma.4.4, x.sigma.5.4)

  #Create subscale scores.
  x.4 <-data.frame()
```

```
for (s in 1:5){
  temp.1.4 <-matrix(unlist(x.sigma.4[s]), nrow=4, byrow=TRUE)
  temp.2.4 <-data.frame(rmvnorm(n=n, mean=x.means.4,
                       sigma=temp.1.4))
  names(temp.2.4) <-
   c("x.dim1.4","x.dim2.4","x.dim3.4","x.dim4.4")
  x.4 <-rbind(x.4, temp.2.4)
}
x.4 <-data.frame(x.4)

#Create true composite score.
x.4$X.comp.4 <-(x.4$x.dim1.4 + x.4$x.dim2.4 + x.4$x.dim3.4 +
                x.4$x.dim4.4)/4

#Create observed scores for two levels of reliability.
x.4$x.msmt.75.4 <-rep(rnorm(n, 0, .5000), times=5)
x.4$x.msmt.95.4 <-rep(rnorm(n, 0, .2236), times=5)
x.4$X.comp.75.4 <-x.4$X.comp.4 + x.4$x.msmt.75.4
x.4$X.comp.95.4 <-x.4$X.comp.4 + x.4$x.msmt.95.4

#Restructure data and add id vars.
x.4 <-stack(x.4[,8:9])
x.4$ind <-rep(c(".75",".95"), each=12500)
names(x.4) <-c("X","reliability")
x.4$no.dim <-"4"
x.4$cor.dim <-rep(rep(c(".05",".30",".50",".80",".95"),
                 each=2500), times=2)
x.4$N.id <-rep(1:n, times=10)
x.4$J.id <-rep(rep(1:j, each=nsubj), times=10)

#### Generate Y (5th grade math achievement) ####
#Define means and SDs.
y.means.4 <-c(0,0,0,0)
y.sigma.1.4 <-matrix(c(1.00, 0.05, 0.05, 0.05,
                       0.05, 1.00, 0.05, 0.05,
                       0.05, 0.05, 1.00, 0.05,
                       0.05, 0.05, 0.05, 1.00), nrow=4,
               byrow=TRUE)
y.sigma.2.4 <-matrix(c(1.00, 0.30, 0.30, 0.30,
                       0.30, 1.00, 0.30, 0.30,
                       0.30, 0.30, 1.00, 0.30,
                       0.30, 0.30, 0.30, 1.00), nrow=4,
               byrow=TRUE)
y.sigma.3.4 <-matrix(c(1.00, 0.50, 0.50, 0.50,
                       0.50, 1.00, 0.50, 0.50,
                       0.50, 0.50, 1.00, 0.50,
                       0.50, 0.50, 0.50, 1.00), nrow=4,
               byrow=TRUE)
```

```
y.sigma.4.4 <-matrix(c(1.00, 0.80, 0.80, 0.80,
                       0.80, 1.00, 0.80, 0.80,
                       0.80, 0.80, 1.00, 0.80,
                       0.80, 0.80, 0.80, 1.00), nrow=4,
               byrow=TRUE)
y.sigma.5.4 <-matrix(c(1.00, 0.95, 0.95, 0.95,
                       0.95, 1.00, 0.95, 0.95,
                       0.95, 0.95, 1.00, 0.95,
                       0.95, 0.95, 0.95, 1.00), nrow=4,
               byrow=TRUE)
y.sigma.4 <-list(y.sigma.1.4, y.sigma.2.4, y.sigma.3.4,
                 y.sigma.4.4, y.sigma.5.4)

#Create subscale scores.
y.4 <-data.frame()
for (s in 1:5){
  temp.1.4 <-matrix(unlist(y.sigma.4[s]), nrow=4, byrow=TRUE)
  temp.2.4 <-data.frame(rmvnorm(n=n, mean=y.means.4,
                        sigma=temp.1.4))
  names(temp.2.4) <-
  c("y.dim1.4","y.dim2.4","y.dim3.4","y.dim4.4")
  y.4 <-rbind(y.4, temp.2.4)
}
y.4 <-data.frame(y.4)
#Create true composite score.
y.4$Y.comp.4 <-(y.4$y.dim1.4 + y.4$y.dim2.4 + y.4$y.dim3.4 +
                y.4$y.dim4.4)/4

#Create observed scores for two levels of reliability.
y.4$y.msmt.75.4 <-rep(rnorm(n, 0, .5000), times=5)
y.4$y.msmt.95.4 <-rep(rnorm(n, 0, .2236), times=5)
y.4$Y.comp.75.4 <-y.4$Y.comp.4 + y.4$y.msmt.75.4
y.4$Y.comp.95.4 <-y.4$Y.comp.4 + y.4$y.msmt.95.4

#Restructure data and add id vars.
y.4 <-stack(y.4[,8:9])
y.4$ind <-rep(c(".75",".95"), each=12500)
names(y.4) <-c("Y","reliability")
y.4$no.dim <-"4"
y.4$cor.dim <-rep(rep(c(".05",".30",".50",".80",".95"),
                  each=2500), times=2)
y.4$N.id <-rep(1:n, times=10)
y.4$J.id <-rep(rep(1:j, each=nsubj), times=10)

#### Merge X and Y and add ID vars ####
data.i <-cbind(x.4, y.4)
data.i <-data.i[,c(5:6,2:4,1,7)]
```

```
#### Add teacher effects to X and Y ####
#Generate random teacher effects.
U.x <-rep(rnorm(j, 0, .5), times=10) #Grade 4
U.y <-rep(rnorm(j, 0, .5), times=10) #Grade 5
J.id <-rep(seq(1:j), times=10)
data.j <-data.frame(cbind(J.id, U.x, U.y))
names(data.j) <-c("J.id","U.x.parm","U.y.parm")
data.j$reliability <-rep(c(".75",".95"), each=500)
data.j$no.dim <-4
data.j$cor.dim <-rep(rep(c(".05",".30",".50",".80",".95"),
                     each=100), times=2)

#Merge teacher effects with student data.
temp <-merge(data.i, data.j, by=c("J.id", "reliability",
                                 "cor.dim"),
            sort=FALSE)
temp <-temp[order(temp$reliability, temp$cor.dim, temp$J.id,
            temp$N.id),]
data.i <-temp

#Add random teacher effects to X and Y.
data.i$X <-data.i$X + data.i$U.x.parm
data.i$Y <-data.i$Y + data.i$U.y.parm

#### Adjust Y to reflect correlation with X (.70) ####
data.i$Y <-data.i$Y + (data.i$X*.96)

#### Rename data.i variables to reflect parameters ####
data.i <-data.i[,c(1,4,2,5,3,6:7)]
names(data.i) <-
  c("J.id","N.id","reliability","no.dim","cor.dim",
                  "X.parm","Y.parm")

#### Restructure data.j and add id variables ####
data.j <-data.j[,c(1,4:6,2:3)]

#### Fit VAA model to simulated data for 2 dimensions ####
out.i <-vector()
out.j <-vector()
for (r in c(".75",".95")){
  data.1 <-data.i[data.i$reliability==r,]
  out.1 <-vector()
  out.2 <-vector()
  for (c in c(".05",".30",".50",".80",".95")){
    data.temp <-data.1[data.1$cor.dim==c,]
    lmer.temp <-lmer(Y.parm ~ 1 + X.parm + (1|J.id),
                    data=data.temp)
    fit.temp <-as.data.frame(unlist(fitted(lmer.temp)))
```

```
      vam.temp <-as.data.frame(unlist(ranef(lmer.temp)))
      out.1 <-rbind(out.1, fit.temp)
      out.2 <-rbind(out.2, vam.temp)
    }
    out.i <-rbind(out.i, out.1)
    out.j <-rbind(out.j, out.2)
  }

  #### Merge student estimates with student data ####
  out.i <-data.frame(out.i, row.names=NULL)
  names(out.i) <-c("Y.estm")
  temp.i <-cbind(data.i, Y.estm=out.i)

  #### Compute bias estimates ####
  temp.i$Y.bias <-temp.i$Y.estm - temp.i$Y.parm  #Bias
  temp.i$Y.rmse <-temp.i$Y.bias^2 #RMSE

  #### Merge teacher estimates with teacher data ####
  out.j <-data.frame(out.j, row.names=NULL)
  names(out.j) <-c("U.estm")
  temp.j <-cbind(data.j, U.y.estm=out.j$U.estm)

  #### Compute bias estimates ####
  temp.j$U.bias <-temp.j$U.y.estm - temp.j$U.y.parm  #Bias
  temp.j$U.rmse <-temp.j$U.bias^2 #RMSE

  #### Close replication loop ####
  file.i <-rbind(file.i, temp.i)
  file.j <-rbind(file.j, temp.j)
}
file.i <-data.frame(file.i)
file.j <-data.frame(file.j)

#### Aggregate values across 500 replications ####
#Create variables to aggregate on.
file.i$rep.id <-rep(1:25000, times=500)
file.j$rep.id <-rep(1:1000, times=500)

#Aggregate data across 500 replications.
aggr.i <-aggregate(file.i[,6:11], by=list(file.i$rep.id), mean)
aggr.j <-aggregate(file.j[,5:10], by=list(file.j$rep.id), mean)

#Merge aggregated data with id vars.
merg.i <-cbind(data.i[,1:5], aggr.i[,2:6])
merg.j <-cbind(data.j[,1:4], aggr.j[,2:6])

#### Finish computing RMSE ####
merg.i$Y.rmse <-sqrt(merg.i$Y.rmse)
merg.j$U.rmse <-sqrt(merg.j$U.rmse)

#### Write files ####
write.csv(file.i, file="data.i.long.4.csv")
write.csv(file.j, file="data.j.long.4.csv")
write.csv(merg.i, file="data.i.4.csv")
write.csv(merg.j, file="data.j.4.csv")
```