ABSTRACT

Title of Dissertation:	MULTIVARIATE MULTILEVEL VALUE-ADDED MODELING: CONSTRUCTING A TEACHER EFFECTIVENESS COMPOSITE
	Anna Lissitz, Doctor of Philosophy, 2020
Dissertation directed by:	Dr. Laura Stapleton, Research, Innovation and Partnerships

This simulation study presents a justification for evaluating teacher effectiveness with a multivariate multilevel model. It was hypothesized that the multivariate model leads to more precise effectiveness estimates when compared to separate univariate multilevel models. Then, this study investigated combining the multiple effectiveness estimates that are produced by the multivariate multilevel model and produced by separate univariate multilevel models. Given that the models could produce significantly different effectiveness estimates, it was hypothesized that the composites formed from the results of the multivariate multilevel model differ from the composites formed from the results of the separate univariate models in terms of bias. The correlations between the composites from the different models were very high, providing no evidence that the model choice was impactful. Also, the differences in bias and fit were slight. While the findings do not really support a claim for the use of the more complex multivariate model over the univariate models, the increased theoretical validity from adding outcomes to the VAM does.

MULTIVARIATE MULTILEVEL VALUE-ADDED MODELING: CONSTRUCTING A TEACHER EFFECTIVENESS COMPOSITE

by

Anna Lissitz

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Laura Stapleton, Chair
Professor Paul Hanges
Professor Jeffrey R. Harring
Associate Professor Hong Jiao
Professor Greg Hancock

Table of Contents

Chapter 1: Introduction

Value added models (VAM) are a category of statistical models implemented in the educational context typically to measure the influence of a teacher, school, or district on a student’s achievement (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). This influence is often labeled as *effectiveness*. Evaluation systems are using such effectiveness estimates to compare and rank teachers, schools, and/or districts and even states. In many cases, high stakes decisions, such as tenure or dismissal and funding, are based on these evaluation systems, which has fueled the research in this field (Goldhaber, 2010; Montes, 2012; Rothstein, 2016). Evaluation systems are also using teacher effectiveness estimates to rate teachers with the purpose of reward and development. Many current evaluation systems are implemented to incentivize better performance; but further investigation into what makes teachers and schools effective is necessary before relying heavily on these systems. The American Statistical Association (ASA) published a statement on using value-added models in 2014. The ASA noted two issues that are particularly relevant to this dissertation. First, it declares that VAMs are generally based on standardized test scores, and do not directly measure teacher influences on other student outcomes. Second, the VAM scores’ rankings can change substantially when a different model or measure is used. The ASA statement declares that the ranking of teachers by VAM scores can have unintended consequences. It concludes that, with caution, VAMs could be used to distinguish characteristics of quality (ASA, 2014). It is possible that the practices of teachers determined to be effective by such models can be identified and shared with educators to improve the teaching and learning experience (Goe, 2007). These evaluation systems have garnered much criticism with respect to the methods of calculating teacher effectiveness estimates, specifically value-added modeling.

Many teacher evaluation systems are using students-within-teacher random effects multilevel models to estimate teacher effectiveness effects where student achievement, or some outcome of choice, is modeled as the VAM outcome. There can even be multiple outcomes of interest. It is common in the primary school-level for a teacher to be responsible for teaching all of the academic subjects. Students typically have one primary teacher who has the potential to impact a range of outcomes. The teacher effectiveness should reflect all of the outcomes that the teacher has influence over. Modeling multiple outcomes with a multivariate model is a more complex analysis than a univariate model, but the added information in the model could produce a more useful estimate of effectiveness. At the older grade levels, students are more likely to have a different teacher for each subject. This makes it more difficult to attribute the growth of a student to just one teacher. There are complex statistical models used to estimate effectiveness that take into account the multiple teachers, such as the cross-classification model and multiple membership model (Fielding & Goldstein, 2006).

The teacher effects are calculated post-estimation as best linear unbiased predictions (BLUPs) of the random intercept effects. This means that, prior to the calculations, teachers are assumed to be at the average level of effectiveness, until the student data evidence otherwise. When specifying a multilevel random effects model, each specified outcome has a direct influence on the resulting teacher effectiveness estimate(s). Therefore, the selection of the outcome(s) is a critical decision in VAM (Lockwood et al., 2007). Of interest to the current study is that multiple outcomes result in multiple effectiveness estimates for each teacher which can, in turn, be aggregated into one teacher effectiveness composite. Evaluation systems can choose to fit separate multilevel models to each outcome when multiple outcomes are of interest; alternatively, multiple outcomes can be modeled jointly with a multivariate multilevel model. Fitting separate multilevel models implies that the teacher effects are independent of each other, which may not be the case. In the multilevel model each outcome has an associated set of teach-level residuals, just as each separate univariate multilevel model produces a set of teacher-level residuals associated with the respective outcome. Ideally, teacher effectiveness is calculated as a composite of these teacher-level residuals across the multiple outcomes. Previous research has shown that the estimate of the population covariance structure of group-level residuals from separate univariate multilevel models was biased when compared to the estimate of the population covariance structure of the group-level residuals resulting from the multivariate model (Leckie, 2018). This finding has implications for the constructed composite of residuals, which is interpreted as the teacher effectiveness estimate. This dissertation study examined the hypothesis that the composites built from the residuals resulting from the multivariate model differ from the composites built from the univariate models. It was hypothesized that the univariate model-based composites would be biased when compared to the multivariate model-based composites.

Statement of Problem

Each outcome in the multivariate multilevel model has an associated residual at the group- (teacher) level. It is this residual that is interpreted as the effectiveness estimate. For models with multiple outcomes, multiple sets of effectiveness estimates (residuals) are produced. This means that each teacher has multiple effectiveness estimates, one for each outcome. The problem with this is that a teacher could be rated at different levels of effectiveness depending on which outcome is examined. One of the only examples of where an effectiveness composite is constructed is the state of Ohio’s EVAAS® value added reporting (SAS, 2016). Ohio’s EVAAS® implements a multivariate multilevel VAM to estimate teacher effectiveness across multiple grades, years, and academic subjects. An effectiveness composite is calculated for each teacher with the number of students used in a particular measure as the weight. The evaluation system in Ohio only focuses on grades four through eight in math and reading. There are limitations with the model where students have multiple teachers throughout the day. Typically, the elementary school level is a better fit for VAM given that students have one primary teacher throughout the day. In the upper grades, students have multiple teachers and it is more difficult to attribute growth and outcomes to one teacher. The EVAAS® model is discussed in more detail throughout this study. Alternatively, there are several empirical studies in multivariate multilevel VAM that conclude with comparing the separate multiple estimates from the multiple outcomes across models and conditions without calculating an overall combined effectiveness estimate (De Maeyer, van den Bergh, Rymenans, Van Petegem, & Rijlaarsdam, 2010; Grilli, Pennoni, Rampichini, & Romeo, 2015; Lockwood et al., 2007; Ma, 2001). Multiple outcomes and multiple estimates potentially, and likely, results in different effectiveness ratings for each teacher. An alternative to having multiple ratings for each teacher is to combine the estimates into one aggregated effectiveness estimate; however, methods for doing this have yet to be investigated. There is no VAM-related research for guidance on how to rank or classify teachers based on multiple effectiveness estimates. There are examples of other contexts where an aggregation of variables is desired. For instance, the Organization for Economic Co-operation and Development (OECD) compares and ranks countries based on a set of economic characteristics. These characteristics or ‘variables’ are aggregated into a composite that is used to represent the economic strength of the country. There are a variety of methodological options for creating a composite variable using weights and general aggregation methods (OECD, 2008; Oosterhof, 1997). This study investigated potential methods for constructing a single teacher effectiveness estimate for a teacher when multiple outcomes are modeled. Currently, in other contexts, it is most common to construct a composite based on the compensability of outcomes. As this study presents, the weighted importance of the outcomes is a theory-based decision that has a direct impact on the resulting teacher effectiveness estimates. This study employed an alternative to a theoretical weighting of the outcome-based estimates, examining the use of the covariance matrix of the residuals to inform the aggregation of the effectiveness estimates to construct an effectiveness composite. The analysis includes an examination of the consequences, in terms of bias of estimates of the residuals' covariance structure, of fitting separate univariate multilevel models as opposed to a multivariate multilevel model, thereby adding to the significant evidence that calls for a standardization of value-added models if reliable estimates of teacher effectiveness are desired.

Purpose and Significance of Study

First, this study presents a justification for evaluating teacher effectiveness with a multivariate multilevel model. A multivariate model would be most applicable in the primary school-level, where one teacher is the primary educator and is responsible for multiple subjects. It was hypothesized that the multivariate model leads to more precise effectiveness estimates when compared to separate univariate multilevel models. Then, this study investigated combining the multiple effectiveness estimates that are produced by the multivariate multilevel model and produced by separate univariate multilevel models. Given that the models could produce significantly different effectiveness estimates, it was hypothesized that the composites formed from the results of the multivariate multilevel model differ from the composites formed from the results of the separate univariate models in terms of bias. Additionally, there is no previous research that provides guidance on the aggregation of multiple sets of VAM effectiveness estimates. In fact, there is no research on best practices in combining residuals from multilevel models in any context. Multiple potential methods for weighting and aggregating the effectiveness estimates exist. Estimates can be summed in a simple equal-weight linear method or through more complex combinations of weights and aggregation methods. This study investigated viable methods for weighting and aggregating estimates with a focus on using the covariance structure of the teacher-level residuals.

In sum, as noted in the 2015 statement from the American Educational Research Association (AERA), educator evaluation systems need to heed caution when incorporating VAM due to the scientific and technical limitations (AERA, 2015). As such, AERA calls for a substantial investment in more research on VAM. This current study adds to the psychometric research on VAM, with a specific focus on multivariate multilevel models, their assumptions, and the manipulation of the resulting effectiveness estimates.

Chapter 2: Literature Review

Following the legislation of the No Child Left Behind Act (2002), federal officials demanded accountability for student achievement from schools, districts, administrators, states, and other stakeholders. Research began in order to identify methods that could attribute student learning to particular influences, focusing primarily on the effect that a particular school or teacher could have on achievement. The characteristics of teachers and schools have been investigated through statistical models to evaluate their influence on student achievement. Influences like teacher and school characteristics became the subject of value-added modeling and effectiveness research studies (Cawthorn, 2004). One focus of these studies is estimating the effect that a teacher has on student learning in order to determine how much value a teacher adds to the students' education. This is the primary focus of value-added models (VAMs).

This literature review presents a brief background of value-added modeling. This includes a review of the most common statistical models. A focus on multilevel VAM highlights support for the use of the multivariate multilevel model when multiple outcomes are of interest. After justification for the use of the multivariate model is presented, the construction of the composite and the various aggregation methods is discussed. Of particular focus is the dependency of the residual covariance structure-based aggregation method on the type of multilevel model, univariate or multivariate. This literature review highlights the complexity of issues and the lack of current research within the VAM context.

Accountability

Recent expansion of the interest in value-added models can be attributed to policymakers and legislators that have become attracted to the idea of teacher and school accountability based on student outcomes (Everson, 2017). One of the original drivers was Race to the Top (RTTT), a federal government legislation that centralized the once somewhat localized education policy of accountability (USDE, 2009). Baker et al. (2010) cites the Obama administration as encouraging states to make greater use of standardized testing results as an indicator in high-stakes teacher evaluation systems. The new act was titled Every Student Succeeds Act, ESSA, and was signed into law in 2015. The law included provisions that there will be accountability and action to effect positive change in schools and each state must have an accountability plan. Each state is required to hold schools accountable for student success. The accountability plan must include measures for academic achievement, academic progress, English language proficiency and high school graduation rates. The final plan requirement is a way to measure school quality or student success.

A 2019 check-in on state progress for implementing ESSA shows that the majority of states have methods for flagging the lowest performing schools and implementing improvement plans (Klein, 2019). When the new act was signed into law, many thought it would lead to increased innovation in education. But, according to a report from Bellwether Education Partners, this has yet to happen (Aldeman, Hyslop, Marchitello, O'Neil Schiess, & Pennington, 2017). The Bellwether review did reveal that the majority of states included a measure of year to year student growth. For example, Minnesota's plan includes a growth model that awards points to schools based on students making progress in math and English achievement levels. Others include added outcomes beyond just reading and math, including science, attendance, college readiness and even school climate. The review notes an interesting shift in the accountability systems toward norm-referenced systems rather than criterion-referenced ones. This shift is important in that it means rather than being held to predetermined criteria, schools are compared to each other. This methodology means that the system is ignoring whether or not students are on track to succeed and only looking at how they compare to one another. This phenomenon is also seen in VAM where teachers are given an effectiveness score and ranked as in the Tennessee evaluation system. The effectiveness score is the amount the teacher contributes above or below the average effect of a teacher on student outcomes. This means that there will always be a set of teachers at the bottom of the ranking, so rather than looking at what teachers can do or not do, the systems are looking at how teachers compare to each other. The implications of this are discussed in more detail.

Background on Value-added Modeling

As teachers and administrators face the reality of being evaluated based on the achievement of their students, they are asking questions about how their effort is being measured and whether the process is fair. The federal government, education associations and organizations, state and federal policy groups, and other stakeholders involved all want to know if evaluating the effectiveness of teachers is even possible and, if so, how well the methods work. Teachers are feeling the effects of the emphasis on accountability as many states have instituted teacher evaluation systems, with VAM being a significant part of those systems (Feng, Figlio, & Sass, 2010; Goldhaber & Hannaway, 2004; Jiang, Sporte, & Luppescu, 2015; Montes, 2012). There are a number of statistical approaches to VAM in use across the evaluation systems. These various models have been shown to produce varying results, adding more complexity to the teacher evaluation controversy (ASA, 2014).

State educator evaluation systems. Many teacher evaluation systems incorporate the VAM results as just one component of the teacher's overall evaluation. In May 2010, the state of Louisiana declared that school districts must make teacher effectiveness estimates account for

50% of the teacher’s evaluation score. Michigan began a pilot teacher evaluation program where the school districts were allowed some flexibility in how they meet the basic framework of the program. The majority of the districts, about 400, chose to have student achievement growth account for 21-30% of the teacher evaluation. Another 200 districts chose to have student achievement growth account for more than 31% (Keesler & Howe, 2012). The remaining percentage of the teacher’s evaluation rating in these systems was informed by administrator observation and evaluation based on a framework for teaching, such as Charlotte Danielson’s Enhancing Professional Practice for Performance of Teaching (2007). This background is relevant to the aggregation of effectiveness estimates as it gives some insight into how much weight evaluation systems give to the academic achievement component of the teacher’s evaluation. This could inform the weight for a set of residuals yielded from the academic achievement outcome(s) in VAM when modeling both achievement and non-cognitive outcomes.

A review of ten evaluation systems (Doyle & Han, 2012) identified several methods for aggregating the VAM rating with other indices of teacher effectiveness for an overall teacher evaluation. Several systems use a matrix method that examines the VAM score with other measures, such as an observation rating, to assign a qualitative label. The rationale or justification for how much the VAM score should account for in the teacher’s overall rating appears to be theoretical and assigned after a process of gathering stakeholder opinions. Clearly, the stakeholder opinions have great impact on the overall teacher’s evaluation rating. The weighting and aggregating of multiple outcomes within VAM within a teacher evaluation system where there are multiple inputs compounds this impact. A slight tweak of any of the weights in the system could produce a different outcome for a teacher. In a system where there are monetary awards or hiring decisions depending on these outcomes, the weights should be selected very carefully.

Statistical Models

Rather than providing an exhaustive review of VAM, the following discussion is intended to exhibit a sample from the wide variety of models and discuss some of the issues associated with them. This review focuses on the more common statistical approaches to VAM.

Regression models. At the K-12 education level, state systems select the VAM for evaluating teacher effectiveness. The state of Texas uses a regression model that O'Malley, Murphy, McClarty, Murphy, and McBride (2011) noted is the only one of its kind that is fully transparent and reproducible. In the typical education example, students' achievement is the outcome of interest. This multiple regression model allows for the prediction of student achievement while controlling for student characteristics including prior student achievement. The outcome from one year is predicted by the outcome from the previous year and additional covariates. For a group of students and teachers, the difference between the predicted outcome and the actual outcome is calculated for each student in the classroom, and the average of these differences is taken as the measure of the teacher effectiveness. Covariates, X_{ij} and Z_j, are used to control for factors among the students and teachers that are theorized to be above and beyond the effect of the teacher. Common student covariates include socio-economic status and prior achievement. Identifying potential teacher or group-level covariates is more difficult. Separating teacher or class characteristics from what makes a teacher effective is not straightforward. Researchers need to be careful not to control for teacher or class characteristics that contribute to teacher effectiveness. One possible example of what to control for is class size, where it is theorized that class size has an impact on a teacher's ability to be effective above and beyond the teacher's effectiveness. This equation specifies the model with student and teacher or class covariates:

$$Y_{ij(2)} = \beta_0 + \beta_1 Y_{ij(1)} + \beta_2 X_{ij} + \beta_3 Z_j + \varepsilon_{ij} , \tag{1}$$

Where $Y_{ij(2)}$ is the current test score for student *i* taught by the teacher *j*, $Y_{ij(1)}$ is the prior year test score of student *i* within teacher *j* (the current teacher), β_0 is the intercept, $\beta_1 \ldots \beta_3$ are the regression slopes, X_{ij} is the student covariate, Z_j is the teacher or class covariate, and ε_{ij} is the residual (error term), which is assumed to be normally distributed and independent of the covariates. The average of the residuals for teacher j, $\bar{\varepsilon}_{ij}$, is interpreted as teacher effectiveness.

A slight differentiation of the model above that has the current year test score predicted by the prior year and covariates is to specify the gain score as the dependent variable, where the difference between the current year score and the prior year is the gain score (McCaffrey et al., 2004). Equation 2 is the gain score model:

$$Y_{ij(2)} - Y_{ij(1)} = \beta_0 + \beta_1 X_{ij} + \beta_2 Z_j + \varepsilon_{ij}, \tag{2}$$

where the gain score, $Y_{ij(2)} - Y_{ij(1)}$ is the current year score for student *i* within teacher *j* minus the student's prior year score. Note the prior year score was within a different teacher; however, that is not specified in the model. The subscript *j* in the prior year term $Y_{ij(1)}$ in Equation 2 refers to the student's current year teacher.

Researchers posit that the inclusion of the student's prior achievement and characteristics adds validity to the teacher effectiveness estimate (McCaffrey et al., 2004). The prior year achievement and student characteristics can control for otherwise potentially extraneous variables to produce a more valid estimate of teacher effectiveness. By accounting for these extraneous effects in the model, they are removed from the effectiveness estimate. In the model where the current year test score is the dependent variable (Equation 1), the predicted achievement measure can be on a different scale than the prior achievement measure. This would be the case where there is not a vertically-scaled standardized test available.

One disadvantage of the regression model using prior year's achievement is the potential for missing data (Sanders, 2006). Students without the prior year's test scores (or current year's) would be excluded from the model, reducing the reliability of the effectiveness estimate by having fewer students in the sample. Research also suggests that this lack of data is not random, thereby introducing bias (McCaffrey et al., 2003). McCaffrey and associates (2003) discussed the significant correlation between student mobility (causing missing prior or current year data) and student achievement. Another potential disadvantage of this model is that it does not account for the common hierarchical nature of education data (Clark, Crawford, Steele, & Vignoles, 2010). This issue is explored in more depth below.

Multilevel models. The multilevel model is built on the regression model used in Texas by adding a level to account for the fact that students are grouped within schools, within classrooms, or within some other category, and such grouping creates a dependency among the data. The two-level random intercept multilevel model can be specified as:

$$Y_{ij(2)} \; = \beta_0 + \beta_1 Y_{ij(1)} + \beta_2 X_{ij} + \beta_3 Z_j + u_j + \varepsilon_{ij} \; , \qquad (3)$$

where $Y_{ij(2)}$ is the current test score for student *i* taught by teacher *j*, $Y_{ij(1)}$ is the prior year test score of student *i* within teacher *j* (same current year teacher *j*), β_0 is the intercept, $\beta_1 \ldots \beta_3$ are the regression slopes, X_{ij} is the student covariate, Z_j is the teacher covariate, u_j is the effect of teacher *j* on student achievement, and ε_{ij} is the residual at the student level (error term).

Teacher effects, u_j, and student residuals, ε_{ij}, are assumed normally distributed and to have zero means and constant variances. The multilevel model addresses the issue of missing data due to mobility that is a disadvantage of the regression model discussed above. If a case is missing either the prior or current year data in the regression model, it must be excluded from the analyzed data. Multilevel models can incorporate missing data and mobility into the analysis using full information maximum likelihood estimation (Peugh & Enders, 2004). There is no need to remove cases with incomplete data unless the missing data is in the covariates. Parameters are estimated using all the available data and are informed by the relations among the variables – even when these variables have missing data. (Note, there are certain restrictions based on the type of missing data; see Peugh and Enders, 2004, for further information.)

The three-level random intercept model adds a level of nesting and therefore another subscript to represent this level, m, and a new residual at level three, ω_m .

$$Y_{ijm(2)} = \beta_0 + \beta_1 Y_{ijm(1)} + \beta_2 X_{ijm} + \beta_3 Z_{jm} + u_{jm} + \varepsilon_{ijm} + \omega_m \tag{4}$$

A third level could be, for example, a school. Students are nested within teachers and teachers are nested within a school. There could be a commonality among teachers of the school that should be accounted for in the model. Covariates could also be added at the third level to control for characteristics related to this level (not depicted in Equation 4). Using the example of the school as the level-three grouping variable, a relevant covariate could be the percent of students receiving free or reduced school lunch.

A potential disadvantage of the multilevel model is the inherent complexity, particularly when covariates at multiple levels are included. Some stakeholders will have difficulty tracking the multiple levels and/or variables, which could lead to questions among a population that may already be predisposed to doubt the fairness of teacher effectiveness estimates (O'Malley et al., 2011). The multilevel model is made even more complex with the addition of multiple outcomes. It is the hypothesis of the current study that additional outcomes, modeled jointly, leads to a more precise teacher effectiveness estimate because more information (data) going into the calculation of the estimate, i.e. increased sample size, results in decreased standard error. A multilevel model with multiple outcomes is termed a multivariate multilevel model.

Multivariate multilevel models. The multivariate class of multilevel VAM specifies multiple dependent variables, also referred to as outcomes in this study, as a function of the explanatory variables, also referred to as covariates. In value-added modeling, if the theory is that teacher effectiveness influences more than just one outcome, then these additional outcomes can be specified to provide information in the model for estimating effectiveness. Despite the prevalence of VAM with student achievement outcomes, teachers have been shown to affect non-cognitive outcomes as well (Chetty, Friedman, & Rockoff, 2011; Jackson, 2012). There are a limited number of studies that have shown how multiple outcomes could be modeled in this context (De Maeyer, van den Bergh, Rymenans, Van Petegem, & Rijlaarsdam, 2010; Lockwood et al., 2007; Ma, 2001; Papay, 2010). The researchers used achievement-based outcomes, but the model could be applied to non-achievement outcomes and a combination of the two types of outcomes as well. This study investigated modeling additional outcomes and manipulating the resulting multiple sets of residuals.

Multivariate multilevel model equations. The multivariate multilevel model builds on the hierarchical linear model by using level 1 to account for the multiple outcomes. The individual student is now modeled at level 2. In the model examined in this study, the underlying assumption is that the outcome measures (level 1) are nested within students (level 2), which are nested within teachers (level 3). The model notation follows the multivariate multilevel notation format of Hox (2002).

The equation for level 1 is:

$$Y_{hij} = \psi_{1ij}A_{1ij} + \psi_{2ij}A_{2ij} + \cdots + \psi_{Dij}A_{Dij} \tag{5}$$

$$A_{dij} = 1 \text{ when } h = d$$

$$A_{dij} = 0 \text{ when } h \neq d,$$

where A_{dij} is a dummy variable used to distinguish the outcomes, Y_{hij} is the outcome *h* of student *i* within teacher *j* (*d* = 1, …, *D*; *i* = 1,…, N_i ; *j* = 1,…, *n*), *D* is the number of outcomes, (when *h* = *d*, the dummy variable is ‘on,’ equal to 1; when *h* ≠ *d*, the dummy variable is ‘off,’ equal to 0), N_i is the number of students for teacher *j*, and *n* is the number of teachers. The level 1 model of the multivariate model excludes the usual intercept found in the univariate model. There is also no error term for level 1. The number of outcomes in the model determines the number of $\psi_{dij}A_{dij}$ terms in the level 1 model.

The level 2 model of the multivariate model defines the coefficients, ψ_{dij}, from level 1, where ψ_{dij} is the score for student *i* within teacher *j* on outcome *d*:

$$\psi_{dij} = \beta_{d0j} + \sum_{q=1}^{Q} \beta_{qdj}X_{qdij} + \varepsilon_{dij} \tag{6}$$

The intercept term, β_{d0j}, is a measure of the average outcome *d* performance for teacher *j,* adjusted for any level 2 covariates, X_{qdij}, where Q is the number of covariates. Without grand mean centering of the predictors, the intercept, β_{d0j}, is interpreted as the expected value on the *d*th outcome for student *i* within teacher *j* who has a value of zero on the covariates. Grand mean centering of the predictors allows for the interpretation of the intercept as the expected value on the *d*th outcome for student *i* within teacher *j* who has the overall sample mean value on the covariates. The alternate centering method, group mean centering, was not applied because the research interest is in the level 3 effect. Group mean centering would remove the between-group variation for the covariates and would thus remove a valuable aspect of the model and yield misleading results given the theoretical basis for the research. The vector of student-level residuals is $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{1ij} \dots \varepsilon_{dij})$ and is assumed distributed as multivariate normal with a mean of zero and a variance-covariance matrix of $\boldsymbol{\Sigma}'$, a *D*×*D* matrix for the *D* outcomes with variance components σ_{dd} (on the diagonal) and covariance components $\sigma_{dd'}$ (off the diagonal).

The level 3 model defines the β_{qdj} coefficients from level 2.

$$\beta_{d0j} = \gamma_{d00} + \gamma_{0j} Z_{d0j} + u_{d0j} \tag{7a}$$

$$\beta_{1dj} = \gamma_{1d0} + \sum_{m=1}^{M} \gamma_{md} Z_{mdj} \tag{7b}$$

$$\vdots$$

$$\beta_{Qdj} = \gamma_{Qd0} + \sum_{m=1}^{M} \gamma_{md} Z_{mdj}$$

For each teacher *j*, there is a vector of the residuals, $\boldsymbol{u_j} = (u_{1j}, \dots, u_{dj})$, that is distributed as multivariate normal with a mean of zero and a variance-covariance matrix, **T**. Each element, $(u_{1j}, \dots, u_{dj})$, has a mean of zero and a variance of τ_{dd} and each pair of elements has a covariance of $\tau_{dd'}$. The intercepts, γ_{d00}, are the grand means of the outcomes, *d*, each adjusted for any level three covariates, Z_{mdj}, where M is the number of level three covariates. Equation 7b presents only fixed effects of level two covariates as indicated by the lack of a random error term. Level three covariates should be selected carefully. Because the level three grouping variable is the teacher, including covariates about the teacher could remove the unique characteristics of the teacher that leads to the value that the teacher adds to the outcome(s) (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). For example, if the number of years of experience was added as a covariate at level three, the model essentially removes the variation in the outcome score explained by the covariate. But, it could be the years of experience that leads to the teacher having more knowledge and instructional techniques that results in higher student outcome scores. The higher outcome score relates to a high effectiveness score, but this is suppressed if the covariate is specified.

Substituting the level two and three equations into level one yields a combined model:

$$\begin{aligned} Y_{hij} = \sum_{d=1}^{D} \gamma_{d00} A_{hdij} + \sum_{d=1}^{D}\sum_{q=1}^{Q} \gamma_{qd0} X_{qdij} A_{hdij} + \sum_{d=1}^{D}\sum_{m=1}^{M} \gamma_{md0} Z_{mdij} A_{hdij} \\ + \sum_{d=1}^{D} \varepsilon_{dij} A_{hdij} + \sum_{d=1}^{D} u_{dj} A_{hdij} \end{aligned} \qquad (8)$$

Note that the model is essentially a three-level model with repeated outcomes (level one) nested within students (level two) nested within teachers (level three). Thus, there is a set of level three residuals across teachers – one for each outcome *d*. The multivariate multilevel model is expected to follow a set of assumptions for these parameters. These assumptions as well as guidance for sufficient sample sizes are discussed below, following an example of a multivariate multilevel model, EVAAS.

EVAAS, a multivariate multilevel model. The state of Ohio implements an extensive evaluation system, including state and district-level analyses in addition to teacher-level evaluation. The system applies a multivariate response model (MRM) for tests given in consecutive grades (e.g., reading, math). The MRM is a gain-based model that measures growth in student achievement between two points in time for a group of students. The system applies a univariate response model (URM) when the test is given in non-consecutive grades (e.g., science). The URM measures the difference between the students' predicted scores for a particular subject/year with their observed scores. Both models include multiple years of test data to minimize the influence of measurement error and accommodate students with missing test scores. EVAAS Technical documentation claims that adjusting for student characteristics is not necessary due to the inclusion of multiple years of student testing data. The teacher evaluation model includes the percentage of instructional responsibility that the teacher had for the student. The technical documentation provides the teacher model equation (SAS, 2016).

$$Y_{iskl} = \gamma_{skl} + \left(\sum_{k* \leq k} \sum_{j=1}^{\mathrm{T}_{isk^*l^*}} w_{isk^*l^*j} \times u_{isk^*l^*j}\right) + \varepsilon_{iskl}, \qquad (9)$$

where Y_{iskl} is the test score for the *i*th student in the *s*th subject in the *k*th grade in the *l*th year, for the current year; $u_{isk*l*j}$ is the teacher effect of the *j*th teacher on the *i*th student in the *s*th subject in the grade k* in year *l**, where k* and *l** are previous grades and years; and $w_{isk^*l^*j}$ is the fraction of the *i*th student's instructional time claimed by teacher *j*. The inner summation is over all the teachers of the *i*th student in a particular subject/grade/year. The outer summation accumulates teacher effects for the current and previous grades in the same subject. This is referred to as a *layered* model. The Tennessee value-added assessment system also implements a layered model, termed the TVAAS (McCaffrey et al., 2004).

The estimated teacher effects are treated as random effects in the EVAAS and TVAAS models. These estimates are obtained by shrinkage estimation known as best linear unbiased prediction (BLUP). From Equation 9, the teacher-level residuals, $u_{isk^*l^*j}$, are assumed to be a linear function of the outcome, Y_{iskl}, and is said to be unbiased so that the mean of the difference between the estimates and the ‘true’ value is zero and the variance of that difference is no larger than the variance of the difference between any other linear and unbiased predictor and the true value. The shrinkage estimation considers the information available about the specific teacher and the information about all teachers (Tate, 2004). The combination of information includes a weight that depends, in part, on the amount of information available for the individual teacher. The equation for the BLUP, using the notation from Equation 9, is as follows:

$$u_{isk^*l^*j} = \frac{\sigma^2_{u_{isk^*l^*j}}}{\sigma^2_{u_{isk^*l^*j}} + \sigma^2_{\varepsilon_{iskl}}} \left(Y_{\varepsilon_{iskl}} - \bar{Y}_{..}\right), \tag{10}$$

where the estimate is a product of the individual score and overall mean score.

This method protects teachers from being incorrectly classified due to random measurement error in the test scores when there is an insufficient number of students. A greater number of students results in a greater weight of the individual teacher’s information in the combination, such that the resulting shrinkage estimate would not be much different from the observed mean. On the other hand, if the class size were very small, the estimate is weighted more heavily on the information of all the teachers. The shrinkage estimate is therefore closer to the overall mean of all teachers.

Multivariate models, like the EVAAS, result in teacher effectiveness estimates for each subject/grade/year. With separate teacher effectiveness estimates, the evaluation system is not likely be able to provide one ranking for the teacher. Across the different subjects, grades and years, the teacher ranking could vary and evaluation systems that depend on the ranking to, for example, make decisions, give bonuses, or evaluate effective characteristics, would lack the ability to do this. It is possible that systems would give a set of rankings and address each subject area separately, but there could also be the need for just one. For example, the New York State Annual Professional Performance Review Guidelines require a composite rating of highly effective, effective, developing or ineffective (New York State Education Department, 2019). The rating is based on growth scores on state tests and on local assessments. Teachers with an ineffective rating are required to develop an improvement plan. Having separate subject-level scores in addition to the composite score could be helpful in identifying specific areas of weakness for the improvement plan. In Ohio's system, a composite is constructed from the estimates across subjects. The composite weights each subject based on the number of students used in the measure within a year. For example, the composite for a teacher with two effects, math and reading, would be calculated as shown in Equation 11:

$$C_j = \frac{n_r}{n_r+n_m}\left(u_{rj}\right) + \frac{n_m}{n_r+n_m}\left(u_{mj}\right), \tag{11}$$

where C_j is the composite, n_r is the number of students in the reading sample, n_m is the number of students in the math sample, and u_{mj} is the teacher effectiveness estimate for teacher *j* based on the math outcome, and u_{rj} is the teacher effectiveness estimate for teacher *j* based on the reading outcome.

The EVAAS example is of interest to this study because it provides the precedence of implementing a multivariate multilevel model to estimate teacher effectiveness. EVAAS also provides an example of how an effectiveness composite can be constructed. While the basic method of using the sample size is logical, alternative methods should be explored. Methods based on statistical properties of the measures, such as reliability and covariance structure, could lead to more accurate estimations of effectiveness and are worthy of exploration. There may be theoretical reasons for wanting a different structure of the composite. An outcome may have more perceived importance than another, so the aggregation method should allow for this to be reflected in the composite construction. This is proposed in more detail in the aggregation methods section. Aggregation methods is yet another example of the many choices to be made in value-added modeling that can lead to very different effectiveness estimates.

Univariate versus multivariate modeling. The decision of separately or jointly modeling related outcomes is of great relevance to this study. First, the justification for implementing a more complex multivariate model over separate univariate models is warranted, particularly when the consumers of the VAM results are not typically statisticians. Univariate models are complex enough to the lay person and adding multiple outcomes and correlation structures is likely to reduce transparency even more. However, because modeling the outcomes separately is often inferred to mean that they are independent, the complexity of the multivariate model is warranted given that the outcomes are seldom independent of each other.

Griffiths, Brown, and Smith (2003) examined the application of univariate and multivariate multilevel models for repeated measures, with the purpose of comparing the two modeling approaches. The researchers first applied a univariate three-level logistic regression model to empirical pregnancy-related data derived from mothers across time with multiple pregnancies within primary sampling units. The model assumed that the probability of the use of antenatal care over successive pregnancies (limited to three pregnancies per mother) within primary sampling units were independent Bernoulli trials. So, the three levels of the model were pregnancy (the use of antenatal care or not) within mother within primary sampling unit. The model specified three outcomes, the use of antenatal care or not for each of the three pregnancies. Pregnancies of one mother, and more specifically the use of antenatal care, are likely not independent events, so the researchers hypothesized that applying a univariate approach would violate the independence assumption. To assess violation of the independence assumption, the model was run where the variation at the pregnancy-level, the binomial standard deviation, was constrained to one (the required binomial variation at the pregnancy-level) and then unconstrained (free to be estimated). If the estimated variance was significantly greater than one or less than one, the independence assumption was violated. If the estimated variance is greater than one, this implies overdispersion of the data at the pregnancy-level. If it is less than one, this implies underdispersion. The underdispersion suggests a strong correlation between the outcomes. The estimated variation indicated a problem with underdispersion in the model. The results displayed a severe violation of the independence assumption and found that the mother-level residuals were significantly increased when the variance was not constrained at the pregnancy-level. The results displayed a relation between the outcome variable and the grouping variable, the mother, that was not explained through covariates or the inclusion of the mother-level random effect. Next, they fit a multivariate multilevel model to the same data where the multiple outcomes were the use or not of antenatal care for each pregnancy. Equation 12 is the multivariate model.

$$Y_{ijk} = \pi_{ijk} + \varepsilon_{ijk} z_{ijk}, \qquad (12)$$

where Y_{ijk} is the use of antenatal care or not with pregnancy *i* where *i* = 1, 2, 3. The probability of using antenatal care, π_{ijk}, depends on characteristics of the pregnancy, the mother, *j*, and the primary sampling unit, *k*. z_{ijk} is the binomial standard deviation and ε_{ijk} is the pregnancy-level residuals.

The pregnancy-level residuals have a covariance matrix of:

$$\begin{pmatrix} \sigma^2_{\varepsilon_1} & & \\ \sigma_{\varepsilon_{21}} & \sigma^2_{\varepsilon_2} & \\ \sigma_{\varepsilon_{31}} & \sigma_{\varepsilon_{32}} & \sigma^2_{\varepsilon_3} \end{pmatrix}, \qquad (13)$$

where $\sigma_{\varepsilon_{21}} = \sigma_{\varepsilon_{32}}$ and $\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \sigma^2_{\varepsilon_3}$. There were incomplete cases where not all mothers had three outcomes, (three pregnancies). To address this issue, the error covariance for outcomes one and two was set equal to the error covariance for outcomes two and three.

As in the multivariate VAM model where there is a dependency among the multiple outcomes, this model allowed for the specification of the correlation between the pregnancies to mothers. The multivariate multilevel model corrected the violation of the independence assumption among pregnancies and the researchers saw no significant difference between the random parameter estimates, pregnancy-level variance and mother-level residuals' variances, resulting from the models with constrained variance and non-constrained variance at the pregnancy-level. The multivariate multilevel model also allowed researchers to examine the relation between successive outcomes in a more flexible way than the univariate multilevel model, i.e. the relations between the use of antenatal care and the mother-level and pregnancy-level covariates. This is analogous to the relation between the student outcomes in a VAM with the student-level and teacher-level covariates. This study exemplifies the impact of applying univariate models to multivariate data and the advantage of employing the multivariate model.

Baldwin, Imel, Braithwaite, and Atkins (2014) also investigated the application of multivariate multilevel models in comparison to univariate multilevel models; however, their purpose was not to compare the results for bias or error, but rather to bring awareness to the mismatch between study design and study analysis. The researchers noted that multivariate models have not been widely adopted in the psychotherapy research community; yet it is rare to find studies in the field that involve only one outcome. In a meta-analysis, they found only one study out of 60 where the multivariate design was examined with a multivariate model. The researchers claimed this model misspecification creates a disadvantage with regard to testing important theoretical questions that are best examined in the multivariate context. They argued for the use of multivariate models to examine hypotheses about the relations among the multiple outcomes, the correlations among the residuals and treatment effects across the outcomes. This is beneficial in VAM research as well where the relation between the outcomes and the covariates could be examined. The researchers illustrated their arguments with simulated longitudinal treatment data as they investigated hypotheses regarding fixed and random effects. To compare univariate and multivariate models, the researchers examined comparative model fit, specifically the deviance statistic associated with the models. The deviance of each separate univariate model can be summed and compared to the deviance of the multivariate model. A likelihood ratio test compares the difference between the deviances to a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the total of the univariate models and the multivariate model. This method for comparing the univariate and multivariate models was applied in this dissertation study.

The second consideration with how the model is specified concerns the covariance structure of the residuals. There is research, however limited, to show that the covariance calculated among residuals that result from separate univariate multilevel models will differ substantially from the covariance calculated from the residuals of the multivariate model (Leckie, 2018). In a study with empirical data, Leckie (2018) found that the set of school effects (residuals) for each school estimated from the multivariate model were shrunken toward the overall mean of the school effects. The covariance among the residuals associated with the independently modeled outcomes was smaller than the covariance among the residuals estimated with the multivariate model. This is of interest to this study because the covariance structure of the residuals informed the weights of the residuals in one of the effectiveness composites used in this study. If there is bias in the estimated covariance between residuals from the univariate models, this study provides further evidence and a rationale for implementing the more complex multivariate model over separate univariate models when an effectiveness composite is desired.

Multivariate multilevel model assumptions. Assumptions for the multivariate multilevel model consist of the assumptions for the multilevel univariate model, regarding the distribution of the residuals as well as the relation between covariates and residuals, as described in detail in the section below.

First, the multivariate multilevel model includes the assumption that the residuals of the continuous variables have multivariate normal distributions. The linearity assumption specifies that the outcome variables are a linear function of the covariates in the model. If the relations among the variables are truly not linear, a linear model will underestimate the strength of the relation or fail to find the existence of a relation.

Second, the residual at the student-level is expected to have no covariance with the covariates at the student-level, $(COV(\varepsilon_{ij}, X_{ij}) = 0)$. And, like the residuals at the student-level, the residuals at the teacher-level are assumed to be unrelated to the covariates at the teacher-level, $(COV(u_j, Z_j) = 0)$.

Third, the residual homoscedasticity assumption at the group-level of the model (level three) implies that the level two residual variance and the variance-covariance for the level three residuals is held constant across all groups at level three (Snijders & Berkhof, 2008). Violating the homoscedasticity assumption for level three residuals could result in incorrect hypothesis tests for the level three covariates and biased standard errors (Raudenbush & Bryk, 2002). That said, research has shown that the influence of heteroscedasticity on the level one variance and the standard error is minimal (Snijders & Bosker, 1992). Given the inconclusive research, it is strongly advocated to evaluate the homoscedasticity assumption and model it if heteroscedasticity is found (Korendijk, Maas, Moerbeek, & Van der Heijden, 2008).

A guideline for the multivariate multilevel model involves the multiple outcomes selected for the model. The specified outcomes of the model are expected to be moderately correlated, not too high and not too low (Finch & French, 2013; Maxwell, 2001). If the correlation among the outcomes is too high, the model could be specified with only one of the outcomes, as the additional outcome brings little additional information. It would be statistically redundant to include too highly correlated variables (Thum, 1997). If there is no correlation among the outcomes, then the joint modeling is not necessary as the purpose of jointly modeling the outcomes is to account for the relation and its joint influence on the effectiveness estimates. Exactly what is too high and what is too low is difficult to define and existing research does not specify exact correlation values. Maxwell (2001) offered a rule of thumb, greater than 0.3 and less than 0.7, but he offered no significant research-based evidence for the rule. Finch and French (2013) cited the dependent variable correlations of small = 0.2 and large = 0.8, and stated that these are consistent with prior research.

To evaluate the adequacy of the application of the univariate and multivariate models, there are several comparative fit measures (Heck & Thomas, 2000; McCoach, 2010). Often, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are employed to evaluate the comparative fit of non-nested models (Heck & Thomas, 2000). In this study, the set of univariate models is nested within the multivariate model, which jointly models the outcomes. In this case, the deviance was examined as in the Baldwin et al. (2014) study, described above.

Achievement outcomes. The correlation between reading and math scores has been found to be quite high among standardized assessments as well as curriculum-based assessments. The Programme for International Assessment (PISA) consists of reading and math assessments given to a population of 15-year-old students from over 44 countries; the correlation between the reading and math scores was found to be .85 (PISA, 2012). In 2011, the International Association for the Evaluation of Educational Achievement combined PIRLS (reading) and TIMSS (math) databases, including fourth grade students responding to both instruments. Grilli et al. (2015) examined the correlation between the 2011 PIRLS reading and TIMSS math tests for Italy and found the correlation to be slightly smaller at .76. Additionally, Larwin (2010) examined the reading and math assessments from the Educational Longitudinal Study (ELS, 2004) database for 10th graders and found the correlation to be .75. Larwin’s study provided evidence to conclude that a student’s reading score was a significant predictor of the student’s math score. The Stanford Achievement Test claims a correlation of about .70 between its reading and math tests (Pearson, 2014). However, the findings from Villa (2008) support a smaller correlation, .49, based on scores he examined from Stanford’s math problem solving subtest and reading skills for a small sample (n=58) of students between sixth and eleventh grade. To summarize, researchers have found evidence that there is a substantial correlation between students’ standardized assessment-based reading and math scores. The standardized assessments’ correlations range from .49 to .85. There is no shortage of evidence to conclude that math performance and reading performance are strongly related. When examined in the VAM context, the relation of the effectiveness estimates based on these different subjects is the correlation of interest.

Some researchers (Fox, 2016; Goldhaber, Cowan, & Walch, 2012; Lefgren & Sims, 2012) have theorized that elementary school teachers, who teach all subjects, could be more capable of raising achievement in one subject over another. These researchers suggested that by incorporating multiple subjects into the VAM, inferences made from the effectiveness estimates are more valid. Fox (2016) presented research that concluded that teachers who are good at teaching in one subject are generally as good at teaching another. Fox estimated the teacher value-added scores for reading and math separately and found a correlation of .70 between the results. Koedel and Betts (2007) also estimated the value-added scores for teachers across two subjects using the Stanford 9 reading and math assessments. The researchers presented a lower and upper bound of correlations between the two sets of value-added scores from .35 to .64, resulting in the same conclusion as Fox (2016) that the ability to be an effective teacher is not strongly subject-specific. Lefgren and Sims (2012) posed the hypothesis that it is unlikely that the value-added score from each specific subject is equally informative about overall teacher effectiveness. They examined a gain score OLS regression model to determine if including both reading and math in the equation improved the predictive power of the value-added model. They modeled the teacher’s value-added score for math as the outcome conditioned on the gain score for math and the gain score for reading. Contrary to the previous findings, the results showed that while math was a stronger predictor of future overall teaching ability than reading, the two subjects together were even stronger by over 25%. They found that incorporating both subjects in the model increased the precision of teacher value-added across a composite average of the subjects. Goldhaber and associates (2012) examined the same dataset from North Carolina as Lefgren and Sims (2012) but found contradictory results. Goldhaber et al. (2012) corrected the estimates for sampling error and produced correlations between value-added estimates based on math and reading scores within the same year of 0.8 to 0.9. They used the effectiveness estimates to place teachers within a quintile and found that 75% of the teachers were in the same or adjacent quintile across subjects. They concluded that teachers who were determined to be effective in one subject were as effective in the other. However, the results also suggest that 25% of the teachers are inconsistently classified, depending on which subject is used as the basis for the value-added score. With some evaluation systems placing high stakes decisions on the value-added scores, potentially negatively impacting 25% of the teachers is significant and justifies further examination into the topic.

Findings from Fox (2016), Goldhaber et al. (2012), and Koedel and Betts (2007) could suggest that a single achievement measure is sufficient; however, the variability among value-added scores across subjects in other studies suggests otherwise (Lefgren & Sims, 2012; Lockwood et al., 2007; Papay, 2010; Rose et al., 2012). There is enough conflicting evidence to justify further investigation into the influence of subject on VAM. One focus of this study is on jointly modeling two achievement outcomes and a non-achievement outcome. The correlation between the two achievement outcomes, reading and math, and between the achievement and non-achievement outcomes, was simulated to examine the influence of these relations on the effectiveness estimates. Previous research suggests that the achievement outcomes strongly correlate and the resulting effectiveness estimates strongly correlate. This study contributes new evidence to the study of the influence of choice and modeling of outcomes in the value-added context.

Non-cognitive outcomes. The discussion above was about achievement outcomes, but non-cognitive outcomes can be used in VAM as well. Other outcomes of teacher effectiveness could include, for example, student motivation, satisfaction and engagement. Chickering and Gamson (1997) present seven principles of effective education models, one of which is engagement. If their effective education model is valid, then the addition of an engagement indicator in the value-added model as an outcome could increase the utility of the modeling for student improvement. As an outcome, student engagement provides an indicator of student success much like student achievement. Researchers propose that engagement and achievement are positively correlated and both desired outcomes of an effective education model (Korobova & Starobin, 2015). Research also highlights a positive correlation of motivation to academic performance. It is theorized that a teacher could influence a student's motivation, but increased motivation is not enough to guarantee increased academic achievement; however, student motivation has been linked to successful outcomes in the long-run (Jackson, 2012). Teachers should be recognized for their ability to positively influence student motivation. This makes motivation a good example in the discussion of the use of a non-cognitive outcome in VAM. Adding outcome variables that correlate to the effective teaching principles can add validity to the value-added model because then the teacher is not just rated on one measure, it is a compilation of measures that should all correlate to theoretical teacher effectiveness. This simulation employed motivation as an example of a non-cognitive outcome.

Collins, Hanges, and Locke (2004) conducted a meta-analysis of 28 studies examining the relation between student motivation and academic performance. They coded the studies as *known group* studies when the researchers' unit of analysis was a group of students and as *individual* when the researchers' unit of analysis was the individual student. The mean correlation between student motivation and academic performance was .46 among known group studies (*n* = 20) and .18 among individual studies (*n*=8); the difference between the mean correlations from these two types of studies was statistically significant. This indicates that the correlation between motivation and academic performance could depend on how the data are modeled, either as aggregate data or as independent data. This dissertation study models the correlation between academic performance and motivation at the student level. The study by Collins, Hanges and Locke (2004) provides guidance on the values to use in the simulation. As previously stated, the specific non-cognitive and cognitive outcomes selected in this study are not a focus. The focus is on the ability to model multiple outcomes with varying relations to produce an aggregated effectiveness estimate constructed from either theory-based or statistical-based weighted components, that is, the teacher-level residuals from each of the outcomes.

Aggregated Effectiveness Estimates

As discussed, the multivariate model produces multiple sets of effectiveness estimates, one for each outcome. Previous studies evaluated the effectiveness of teachers separately for each set of estimates, without attempting to combine sets into one indicator of effectiveness (Lockwood et al., 2007; Ma, 2001). However, when multiple sets of estimates create contrasting rankings or groupings of teachers, it is logical to develop a composite of the estimates to produce one effectiveness measure on which to rank or group teachers. This study examined possible methods of aggregating the estimates.

OECD (2008) suggests considering the following criteria when determining the aggregation method:

- Intensity of preference for indicators
- Weighting method
- Desired amount of compensability
- Relation of each indicator to all other indicators
- Relation of each grouping unit to all other grouping units

These criteria suggest that the researcher should have a solid theory of the relations among and between the teacher effectiveness estimates derived from various outcomes and the teachers and the ranking thereof. Many state educator evaluation systems have formed evaluation models of weighted components, such as effectiveness rankings and observation ratings, based on a theory of relative importance.

The OECD (2008) handbook on constructing composite indicators stated that the most common method of aggregating indicators is equal weighting, where all the variables are given the same weight. In the case of aggregating multiple effectiveness estimates into one indicator of teacher quality, it is theoretically justified to assume that the informative value of estimates yielded from one outcome would be higher than estimates from another. For example, suppose the model relies on two outcomes, student achievement and student satisfaction. The researcher could justifiably assume that achievement is more objective than student satisfaction, which was derived from end of course evaluations with questionable reliability, and should be a larger component of the teacher overall effectiveness indicator (Hendrickson, Patterson, & Ewing, 2010; Kane & Case, 2003). This idea of using the reliability of the indicators to inform the composition of an aggregated variable is a potential method for combining the multiple sets of residuals produced from the multivariate multilevel value-added model (Cunningham, Fina, Adams, & Welch, 2011; OECD, 2008; Rudner, 2001). Rudner (2001) presented a model where the composite weights are a function of the reliability and validity of the composite. He went on to show that the validity increases with more reliable indicators receiving a higher weight, up to a point, and then the validity begins to decrease. Kane and Case (2003) supported this claim and cautioned that weighting the more reliable indicators in a composite too highly can harm the validity of the composite. Researchers need to examine the composite consistency after weighting based on reliability to ensure the composition is realistic and an indicator of what it was intended to measure. At some point, if the more reliable variable(s) are weighted too heavily, the less reliable variables have little meaning in the composite, thus defeating the purpose of aggregating all of them into one composite. This study applied a simple reliability weighting, where the standardized value of the teacher effectiveness estimate is weighted by the reliability of the instrument used to measure the respective estimate. After each teacher effectiveness estimate for each outcome is weighted, the weighted estimates are added to form a composite effectiveness estimate for each teacher.

$$CO_j = \sum \rho_d u_{dj}, \tag{14}$$

where CO_j is the effectiveness composite for teacher j and ρ_d is the reliability coefficient of the instrument used to measure the outcome, d.

Burns and Clemen (1993) presented covariance structure models as a multivariate procedure that involves observable and unobservable variables, where the unobserved variables are measured by the observed variables. This is analogous to the effectiveness composite concept where an unobserved effectiveness composite is constructed from the observed effectiveness estimates. The researchers proposed the use of the covariance structure of the observed variables to create and analyze linear combinations to serve as composite indices of the corresponding unobserved variable of interest. Burns and Clemen (1993) illustrated the method with empirical data based on the risk associated with transporting hazardous materials. The general composite equation is that the unobserved variable is equal to the sum of observed variables each weighted by its respective contribution to the estimated variation of the unobserved variable. The researchers used the observed variables' regression coefficients calculated in the multivariate model as weights in the composite construction. This example is not directly related to value-added modeling or teacher effectiveness models where the residuals are the outcome of interest, but it is a good example of how the covariance structure of observed variables is used to construct a composite variable. Struppeck (2014) presented an example from actuarial science where multiple loss estimates were combined to obtain one estimate of loss. The procedure weighted the estimates by multiplying the estimates by the variance-covariance matrix. The weighted estimates were then summed to obtain the composite. The composite construction in this study considered the relative precision of the observed variables and the extent to which they were correlated among themselves. These considerations could be extended to the multivariate multilevel random effects model where the covariance structure of the random effects (residuals) informs the construction of the composite.

Applying this model to the teacher effectiveness context yields Equation 15, where $\boldsymbol{u_j}$ is a 1×*D* vector of the teacher effectiveness estimates for teacher *j* for a given set of outcome variables and **T** is the *D*×*D* variance-covariance matrix for the teacher-level residuals and the vector **1** is *D*×1.

$$CO_j = (\boldsymbol{u_j} \times \mathbf{T})\,(\mathbf{1}) \tag{15}$$

The effectiveness composite, CO_j, is the sum of the elements in the resulting 1×D product matrix of $\boldsymbol{u_j} \times \mathbf{T}$. This is the basis for one of the aggregation methods employed in the dissertation study. Another logical method for weighting indicators of a composite is based on the indicator's theoretical importance relative to the other indicators. Rothstein (2000) debated the process of determining weights based on importance in composites of school effectiveness. Rothstein was part of the effort of the Educator Preparation Institute (EPI) to assess teacher preparation institutions based on a composite of how each institution met the three goals of EPI (EPI, 2014). The three goals focused on an assessment of the teachers from the institutions in these areas: content knowledge, satisfaction and perception survey data, and the value-added rating scores. The weighting of each of the metrics for the goals into one institution effectiveness composite was based on an expert-determined importance of each goal. Generally, the proportion of the metric in the composite was highest for goal one (the teachers' knowledge assessment), less for goal two (satisfaction and perception data) and lowest for goal three (value-added rating score). The general equation for a composite weighted by varying degrees of importance is

$$CO_j = \sum_{d=1}^{D}(w_d u_{dj}), \tag{16}$$

where the weight, w_d, is given a value that corresponds to the relative importance of the associated outcome.

Rothstein (2000) stressed the importance of clearly articulating the weights of the indicators and suggested stakeholders could use the indicators from EPI to apply their own weights based on their own theory of relative importance to determine composites. This is an example of a participatory aggregation method to construct a composite. For demonstration purposes, this method was included in this study. Using the example from above, this method assumes that a lack of effectiveness based on the satisfaction rating can be compensated for by higher effectiveness based on the achievement outcome.

This study not only examined four methods for aggregating variables into a composite, but also examined the difference in results when the residuals are a product of the multivariate multilevel model and separate univariate models. To summarize, the four aggregation methods are simple linear summation with equal weights, weights based on reliability, weights based on the covariance structure of the residuals, and weights based on the relative importance of the outcomes that are aligned to the residuals. This dissertation hypothesized that calculating the composite from residuals produced from the multivariate multilevel model, independent of the aggregation method, differ from calculating the composite from residuals resulting from separate univariate multilevel models. As previous empirical research by Leckie (2018) showed, the relation among the multivariate model residuals estimates were more highly correlated than the estimates from the independent outcomes models, this study expected the same. The estimated residuals from the separate univariate models were expected to have more bias than the estimated residuals from the multivariate model because the univariate models ignored the dependency among the data.

It seems logical to explore the two categories of methods for aggregating effectiveness estimates, participatory and statistical methods, to highlight the consequences of selecting one method over the other. The public opinion/ relative importance method appears to be a justified choice based on the applied methods discussed above for constructing an overall teacher rating in the state evaluation systems. This study examined the reliability of the measures for the outcomes to provide component weights for each resulting set of effectiveness estimates. The underlying covariance structure of the residuals was used to provide weights for the estimates in the composite as well. This examination also provides evidence of the bias when modeling related outcomes in separate univariate multilevel models as opposed to jointly modeling them with a multivariate multilevel model. These methods appear as viable options for combining sets of teacher effectiveness estimates. In addition to the impact of the aggregation method, sample size likely influences VAM results.

Sample size for VAM. The impact that class size has on teacher effectiveness has been an area of educational research for decades. Class size research with specific reference to VAM can be found in the K-12 sector with mixed findings (Lipscomb, Teh, Gill, Chiang, & Owens, 2010; McCaffrey et al., 2003; Wright, Horn, & Sanders, 1997). Wright and associates (1997) did not find evidence to support the claim that class size has an effect on the teacher effectiveness estimate. Lefgren and Sims (2012) included class size in the weighting of multiple subjects in a gain-score OLS regression model. They found that including weights for the class size for each teacher had no significant effect on the resulting value-added scores. Incorporating class size in the model did not improve the predictive power, mostly because there was low variance for class size in the dataset. McCaffrey et al. (2003) concluded that class size does influence the variability in the effectiveness estimate for an individual teacher due to sampling error. They found that the teachers with the largest classes (20 – 32 students) had less variability in their teacher effectiveness estimates than teachers with the smaller classes (10 – 19 students). Among the K-12 research literature is a study on the growth model used by Pittsburgh Public Schools to estimate teacher effectiveness. In the growth model, Lipscomb et al. (2010) included student characteristics such as race, gender and up to three years of achievement data. The researchers ran variations of the VAMs to compare the effectiveness estimates. Class size was one of the control variables that the researchers examined. They found that the teachers with smaller classes, and therefore less precise estimates, will be overrepresented at the high and low ends of the estimated performance distribution. Overall, the research is inconclusive on what the standards for class size are explicitly. There are two related questions with regard to class size: (1) What affect does class size have on a teacher's ability to impact student achievement (or other outcomes of interest)? and (2) What is the smallest class size that can be used to make effectiveness estimates? The sample size must be large enough to provide stable and accurate effectiveness estimates, especially when used to make personnel decisions or judgments of teacher quality.

The latter question is examined below, followed by a discussion of the sample size guidelines with respect to multivariate multilevel models. Some of the state accountability models have specified a minimum number of students per teacher required before a value-added estimate can be calculated. It varies, with no concrete evidence of an optimal minimum number, from five (Harris & Sass, 2009) to 20 students (Koedel & Betts, 2010).

Multivariate and multilevel modeling research provides guidance on the question of how many teachers and students the study must include overall. The effect from the violation of the independence assumption is likely impacted by the sample size. Previous research on assumptions and sample size suggest that the estimates will be more robust, less biased, under assumption violations when the sample size is relatively larger (Maas & Hox, 2005). An important examination in this study attempted to determine the relation of class size and bias in estimates of teacher effectiveness and how it differs across the univariate models and the multivariate model.

According to Maas and Hox (2005), for multilevel models, it is evident that estimates and standard errors are more accurate as the sample sizes at all levels are increased. Maas and Hox cited the ‘30/30 rule’ offered by Kreft (1996) which suggests a sample of at least 30 groups with at least 30 subjects per group. Maas and Hox went on to suggest that this rule is best when the interest is in the fixed parameters. When the interest is in the residuals, Maas and Hox (2005) suggest greatly increasing the number of groups and subjects per group to a ‘100/10 rule.’ Maas and Hox (2005) further provided evidence to support this suggestion in a school effectiveness simulation study using a two-level model. The researchers showed that only a small sample size at level two, of 50 or less, resulted in biased estimates of the level two standard errors of the residuals (Maas & Hox, 2005). Simulations of larger sample sizes at level two resulted in no significant effects on the estimates of standard errors, regression coefficients, or variance components. Maas and Hox (2005) also varied the intraclass correlation of the groups and found no significant effects from the manipulation. Snijders (2005) suggested that if the focus of study is on the effect of the level one variable, then the level one sample size is of greatest importance. In this study, the level of interest is the grouping-level: level three. Therefore, a focus of this study was on the sample size of the teachers.

Summary

In summary, VAM presents opportunities and challenges for improving education. Hopefully, by assessing students and identifying effective teachers, best instructional practices can be determined and shared. Incorporating multiple outcomes in the VAM adds more information on which to base the effectiveness estimate when compared to single outcome models; therefore, this study focused on the multivariate multilevel model to investigate the use of multiple outcomes and to address the hierarchical nature of the data. There is a challenge with the multivariate multilevel model's resulting multiple effectiveness estimates across teachers. Theoretically, teachers could be ranked differently depending on which outcome's residuals are examined. This study proposed a composite of the effectiveness estimates for each teacher. However, the literature points out that there are a variety of potential methods for aggregating estimates and the selected method likely has a significant impact on the teacher rating. The background research for this study has resulted in four research questions.

Research Questions

The research questions examined in this study are:

1) *Does the use of univariate or multivariate models result in different levels of bias in estimated group- (teacher-) level residuals?*

2) *Does the use of univariate or multivariate models result in different levels of model fit?*

3) *In terms of teacher ranking, what is the influence of constructing teacher effectiveness composites with equal weight, weight by theory, weight by reliability, and weight by residual covariance structure aggregation methods?*

4) *What is the influence of the combinations of small, medium and large-sized teacher groups with small and average-sized student groups on teacher effectiveness estimates?*

The next chapter of this study presents the methods for a Monte Carlo simulation that evaluated these research questions.

Chapter 3: Methods

This chapter discusses the methods used to conduct the Monte Carlo simulation and to evaluate the results in order to draw conclusions about the research questions posed above. The first section of this chapter presents the simulation design including the empirical research upon which simulation parameter values were based. The second section discusses the conditions and manipulated factors. The third section presents the sequence of simulation steps including data generation and the procedures for checking the quality of the data generation. The fourth section presents the statistical analyses applied to examine and compare the estimates, including model fit, absolute bias of teacher-level residuals and the ranking of teachers based on the composite effectiveness estimates.

Simulation Design

The multivariate multilevel model was used to generate the outcomes data. As is common in educational research, the multilevel model was used due to the nested nature of the data. In this study, the outcomes are nested within students, which are nested within teachers. The multivariate model allows for the joint specification of multiple outcomes. Each of the outcomes produces a set of residuals for each teacher. These residuals are the teacher effectiveness estimates. In this study, the three-level multivariate multilevel model does not include covariates at either the student- or teacher-level. The influence of covariates on the estimation of residuals is outside the scope of this study.

This section presents the values for generating the data used in the simulation. For demonstration, three outcomes were modeled, two achievement-related outcomes and one non-cognitive outcome. As is common in value-added modeling research, the two achievement outcomes are assumed measured with a standardized math test and a standardized reading test

(Fox, 2016; Goldhaber, Cowan, & Walch, 2012; Lefgren & Sims, 2012; Lockwood et al., 2007; Papay, 2010; Rose, Henry, & Lauen, 2012). These two subject areas are the most commonly studied subjects in the K-12 research because standardized testing in these areas is required across the U.S., per RTTT (USDE, 2009). This study explored the reliability of each for use in constructing the effectiveness composite. Lockwood et al. (2007) incorporated the Stanford 9 standardized achievement test in the VAM study of multiple math measures. Papay's (2010) follow-up study incorporated the Stanford 9 reading sub-tests. This standardized achievement test has been in use for decades and Pearson provides reliability estimates which can be used in the weighting of the components in the effectiveness composite. This study used reliability indices and parameters from the Stanford 10 math and reading achievement tests. Statistics Solutions (2016) cites the reliability for the Stanford 10 math section between .80 and .87 (an average of .84 was employed in this simulation) and the reliability for the reading section .87. The correlation between the Stanford reading and math achievement tests was found to be about .70 (Pearson, 2014). These values were used in the baseline data generation model.

As discussed in the literature review, Collins et al. (2004) calculated a correlation between student motivation and academic performance of .18 among individual studies (*n*=8). They found a larger correlation between motivation and performance among known group studies (.48, *n*=20). Because motivation was assumed to be measured for each student in this study, the correlation from the individual studies (.18) was used to inform the correlation between the cognitive and non-cognitive outcomes, academic achievement and motivation, at the student-level in the baseline data generation model. The reliability of the motivation measure was also needed to inform the weighting of the effectiveness estimates in the teacher effectiveness composite. As discussed in the literature review, Fredricks and McColskey (2012) evaluated 11 self-report measures and proposed the guideline that reliabilities above .70 are acceptable for motivation measures. Following this review, this study assumed .70 as the reliability for the student motivation measure for use in the weighting of the effectiveness estimates in the composite.

The specification of relevant covariates influences the independence between the residuals of the outcomes and the covariates. Omitting a relevant covariate can falsely attribute effects to the variables that are present in the model. Based on a review of VAM research, common student-level covariates are SES and prior achievement. Lockwood et al. (2007) incorporated these as well as gender, age, race, English proficiency and special education status in an empirical study to examine the influence of covariates on teacher effectiveness estimates. The study included four VAM, the gain score model, the covariate adjustment model, the complete persistence model and the variable persistence model and applied the models [separately] with two different math-related outcomes. The researchers applied five different covariate configurations in all four VAM, for each of the two outcomes. For all four models, the average correlation for the effectiveness estimates ranged from .92 to .98 across both outcomes. The results did show a slightly greater sensitivity to the inclusion of covariates at the teacher level compared to the student level, but the correlations were still very high. The researchers found the effectiveness estimates to be robust to the exclusion of these variables in the model. This suggests the added model complexity of including covariates does not influence the effectiveness estimates, thus was not justified. No teacher-level covariates were modeled as it is theorized that teacher characteristics are what makes the teacher effective and controlling for these would result in an underestimated effectiveness estimate. The data generation correlations between the student-level outcomes are presented in Table 1.

Table 1

Correlations Between Outcomes

Variable	Math	Reading	Motivation
Math	1	.7	.18
Reading		1	.18
Motivation			1

The multivariate multilevel equation (in its combined form) used to simulate the outcomes data was:

$$Y_{hij} = \sum_{d=1}^{D} \gamma_{d00} A_{hdij} + \sum_{d=1}^{D} \varepsilon_{dij} A_{hdij} + \sum_{d=1}^{D} u_{dj} A_{hdij}, \tag{17}$$

where A_{hdij} is a dummy variable used to distinguish the outcomes, Y_{hij} is the outcome *h* of student *i* within teacher *j* ($d = 1, \ldots, D$; $i = 1,\ldots, N_i$; $j = 1,\ldots, n$), *D* is the number of outcomes, (when $h = d$, the dummy variable is 'on,' equal to 1; when $h \neq d$, the dummy variable is 'off,' equal to 0), N_i is the number of students for teacher *j*, and *n* is the number of teachers. For each teacher *j*, there is a vector of the residuals, $\boldsymbol{u_j} = (u_{1j}, \ldots, u_{dj})$. The vector of student-level residuals is $\boldsymbol{\varepsilon}_{ij} = (\varepsilon_{1ij} \ldots \varepsilon_{dij})$. The intercepts, γ_{d00} , are the grand means of the outcomes. The population covariance at the teacher-level was:

Table 2

Covariance of Teacher-level Residuals

Variable	Math	Reading	Motivation
Math	.25	.175	.0225
Reading		.25	.0225
Motivation			.0625

The population covariance at the student-level was:

Table 3

Covariance of Student-level Residuals

Variable	Math	Reading	Motivation
Math	1	.2	.0225
Reading		1	.0225
Motivation			1

Simulation conditions. This simulation is a within-cell and between-cell factor design where the four aggregation methods were applied to the results of the univariate and multivariate models for each of the six sample size combinations. Table 4 summarizes the factors that were involved in this design of six between-cell conditions and eight within-cell methods of obtaining teacher effectiveness composites.

Table 4

Study Design and Factors

		Sample Size					
		Students per teacher (N)					
		Small			Average		
		10			22		
		Teachers (N)					
	Composite Method	15	220	500	15	220	500
Univariate model	Equal weights (Ew)	Ew: $CO_j = (1 \times u_{1j}) + (1 \times u_{2j}) + (1 \times u_{3j})$					
	Reliability-based (Rb)	Rb: $CO_j = (.87 \times u_{1j}) + (.84 \times u_{2j}) + (.70 \times u_{3j})$					
	Covariance-based (RC)	Cb: $CO_j = (\boldsymbol{u}_j \times \mathbf{T})\,(\mathbf{1})$					
	Theory-based (Tb)	Tb: $CO_j = (1 \times u_{1j}) + (1 \times u_{2j}) + (.5 \times u_{3j})$					
Multivariate model	Equal weights						
	Theory-based						
	Covariance-based						
	Reliability-based						

First, these categories are described, followed by a detailed account of the values for the factors and levels within each category. There are two samples to consider in the study: teachers and students per teacher. Multilevel modeling literature provides general guidance on the number of groups and the number of observations per group for accurate estimation. The second category of factors involves the aggregation method used to combine the teacher effectiveness estimates for each of the outcomes. The approaches include participatory (theory-based, equal weighting) and statistical (reliability-based, covariance structure-based).

Sample size. Investigating the advice of multilevel researchers (Goldstein, 1999; Heck & Thomas, 2000; Hox, 2005; Snijders, 2005) and VAM researchers (McCaffrey et al., 2004), this study examined the influence of sample size with conditions based on combinations of the number of teachers and ratio of students per teacher. The baseline data were generated on the assumption that the average class size is 22 students. This was contrasted with a small class which was defined as 10 students. Ten is the minimum number of students required for the TVAAS to provide teacher reporting (SAS, 2016). Following multilevel recommended guidelines of group to unit of analysis ratio of 100/10, the initial teacher sample size was 220. To contrast this, the small sample of teachers, 15, is less than the recommendation by Maas and Hox (2005) of at least 50 and less than the recommended ratio, 100/10. A third teacher sample size of 500 was also employed to examine the use of a large sample, more than twice the recommended ratio. The number of teachers in the VAM sample had three levels, small = 15 teachers, medium = 220 teachers, and large = 500 teachers.

Aggregation method. The VAM and multivariate research is limited regarding methods for combining residuals. In other fields of study, there are common methods for weighting and aggregating variables to form composites. These most common methods were simulated in this study. These include equal weights, weights based on reliability, weights based on the residual covariance structure, and weights based on a theory of perceived outcome importance. The equal weights method assigned equal weights to each teacher effectiveness estimate and then summed them to form the overall teacher effectiveness composite. The second method gave more weight to the indicators from more reliable sources, using the reliability coefficient as the weight. Previous empirical research was examined to inform the reliability of the math and reading assessments and the motivation measure. The reliability of the Stanford 10 reading test is .87 and the math test is .84 (Statistics Solutions, 2016). Based on a review of the literature, the reliability of the motivation measure was set at .70 (Fredricks & McColskey, 2012). The covariance structure-based method refers to the structure of teacher-level residuals, also known as the teacher effects. This method was of particular interest when comparing the separate multilevel model results to the multivariate multilevel model results. The final method is referred to as participatory and is based on perceived importance of type of outcome. Based on a theory, a higher proportion was applied to each of the sets of estimates from the achievement outcomes. The two cognitive and one non-cognitive outcomes were given weights of 1, 1, and .5 to place more importance on the cognitive outcomes.

Replications. The number of replications used in this study was determined by examining two parameters, the mean bias of the MM reading residual and the mean bias of UM reading residual. In general, smaller sample size usually leads to less stable parameter estimates. Therefore, the small teacher sample size and small student sample size data set was used for this analysis. Figure 1 below shows that the mean values appear to stabilize around 300 replications. Around 900 replications, the values are fairly constant. This study ran all analyses for 1000 replications.

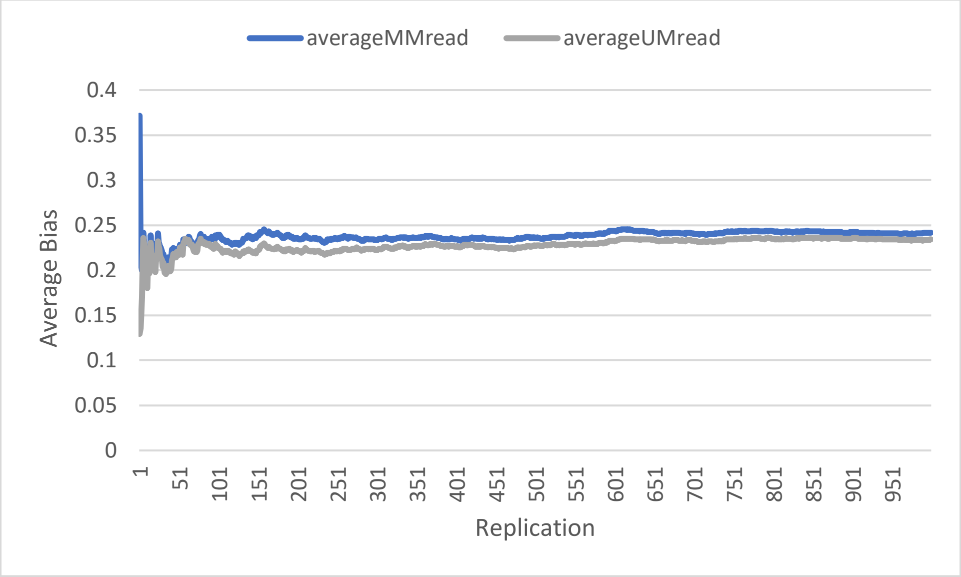

Figure 1. Average MM reading residual bias and average UM reading bias by replication for the sample where n_j=10, J=15.

Simulation Procedures

The sequence of simulation procedures is as follows:

A. Dataset Generation

a. To simulate the residuals associated with each outcome for the student-level and the teacher-level, two sets of three variables were generated. It was assumed the mean of each residual variable was zero. The variances of teacher residuals associated with the math and reading outcomes were set to be the same. As in several previous related simulation studies, the variance of the student-level residuals, σ^2, was fixed at 1 (Coleman, Hoffer & Kilgore, 1982; Donoghue & Jenkins, 1992).

b. The grand mean of each outcome was specified. The values were standardized to account for the difference in the scales used for the achievement outcomes and the non-cognitive outcome, motivation. The grand means were 50, 50 and 30, respectively.

c. The outcome scores, Y_{hij}, were generated from the grand mean, γ_{oo}, plus the student-level residual, ε_{dij}, and the teacher-level residual, u_{dj}.

B. Data Generation Check

a. Descriptive statistics were calculated for the outcomes, teacher residuals and student residuals of the largest generated dataset. These statistics were compared to the known values used in the data generation code.

b. The distributions of the residuals were examined for normality.

c. The correlation matrix between the teacher-level residuals was calculated and compared to the correlation matrix used in the data generation code.

C. Estimating effectiveness from models

a. SAS was used to analyze the data. Code for one sample size condition is in the Appendix.

i. The PROC MIXED command was used to apply the univariate and multivariate models to the datasets. Refer to Equations 3 and 8.

ii. The student-level and teacher-level residuals were saved.

iii. Restricted Maximum Likelihood (REML) estimation method was employed. REML estimation was selected over Maximum likelihood (ML) estimation because REML optimizes the likelihood of the full residuals as opposed to the observations directly. REML partials out the fixed effects and maximizes the portion which is free of fixed effects (Harville, 1977). ML estimates are unbiased for the variance components of the fixed effects, but biased for the variance components of the random effects, whereas REML is the opposite, biased for the variance components of the fixed effects and unbiased for the variance components of the random effects (Swallow & Monahan, 1984; Wu, Gumpertz & Boos, 2001). The difference between estimation methods is often more pronounced in small samples. Since the random effects (residuals) are the focus of this study, REML was selected.

iv. In the cases where the analysis yielded a non-positive definite G matrix, the cases were removed from the dataset. This resulted in smaller datasets, < 1000 replications, for the two samples with the small teacher sample size condition. This design decision has potential implications for the results as discussed later in the paper.

D. Comparing the results

a. To assess research question one, I calculated the absolute bias, $Bias(u_{dj})$, for the teacher effectiveness estimates from the univariate and multivariate models within each set of conditions. This evaluated the question of the impact of applying univariate models to multivariate data. Let u_{dj} be the value of the teacher effectiveness estimate for outcome d for teacher j. The absolute bias measures the difference between the 'true' value (the data generation values) and the estimated value, $\hat{u}_{dj}$. Absolute bias was employed rather than relative and actual bias because the estimated effectiveness value can be less than or greater than the ‘true’ value. The focus is on how much the estimate differs from truth, not the direction of the difference.

$$Bias(u_{dj}) = \frac{\sum\left|u_{dj} - \hat{u}_{dj}\right|}{N} \tag{18}$$

The difference between the true effectiveness and the estimated effectiveness was summed for all teachers for each outcome. The sum was then divided by the number of teachers in the sample, *N*. I compared the absolute bias for the teacher effectiveness estimates for each outcome, within each sample size combination. An average absolute bias value was calculated across the replications, R, where r = 1, ..., 1000.

$$AVERAGE\ of\ Bias(u_{dj}) = \frac{\sum_{r=1}^{R} Bias(u_{dj})_r}{1000} \tag{19}$$

b. To assess research question two, the likelihood ratio test (LRT) compared the difference between the deviances (De) of the models to a chi-square distribution with a degrees of freedom equal to the difference in the number of parameters between models. Deviance describes the difference between the specified model and the best possible, i.e. saturated, model. In general, models with lower deviance indicate better fit than models with higher deviance. The univariate models were nested within the multivariate model because the univariate models use the same data and are constrained versions of the multivariate model (the correlation between outcomes is assumed to be zero). When models are nested, as is the case with independent univariate models and the associated multivariate model, the deviance measures can be compared using the chi-square difference test, i.e. likelihood ratio test (Hox, 2002). The combined deviance (De_{um}) of the univariate multilevel models was compared to the deviance of the multivariate model De_{mm} (Baldwin et al., 2014).

$$\chi^2(df_{um} - df_{mm}) = De_{um} - De_{mm} \tag{20}$$

A significant likelihood ratio test indicates if the multivariate model is a better fit to the data than the separate univariate models. The degrees of freedom for the chi-square test is equal to the difference in the number of parameters between models (where the set of independent univariate models is considered one model and the multivariate model is the other). The difference in the degrees of freedom was six, the critical value was 12.592 and an alpha level of .05 was applied.

E. Aggregating the effectiveness estimates

a. I constructed an overall teacher effectiveness estimate from the three sets of residuals for each teacher for each aggregation method, from both the multivariate model and separate univariate models.

i. Equal weights and summation, based on Equation 11:

$$CO_j = (1 \times u_{1j}) + (1 \times u_{2j}) + (1 \times u_{3j}) \tag{21}$$

ii. Reliability weights and summation, based on Equation 12:

$$CO_j = (.87 \times u_{1j}) + (.84 \times u_{2j}) + (.70 \times u_{3j}) \tag{22}$$

iii. Covariance structure-based, refer to Equation 13, where the composite is the sum of the elements in the product matrix of the vector of teacher effectiveness estimates, $\boldsymbol{u}_j$, and the covariance matrix of residuals, **T**, multiplied by the vector **1**:

$$CO_j = (\boldsymbol{u_j} \times \ \mathbf{T})\ (\mathbf{1}) \tag{23}$$

iv. Participatory method with importance on achievement outcomes, based on equation 11:

$$CO_j = \left(1 \times u_{1j}\right) + \left(1 \times u_{2j}\right) + \left(.5 \times u_{3j}\right) \tag{24}$$

For this study, I assumed that the weight of the importance of the achievement outcomes was twice that of the weight of the non-cognitive outcome, motivation (i.e., weights were 1, 1, and .5).

F. Comparison of rankings

a. For each multivariate model dataset, I rank ordered the teachers based on each aggregated effectiveness estimate, CO_j.

b. For research question three, Spearman and Pearson correlations were employed to compare the teacher effectiveness estimates (Newton, Darling-Hammond, Haertel, & Thomas, 2010; Wei, Hembry, Murphy, & McBride, 2012). Additionally, like other VAM studies, I compared the change in rankings through an analysis of quintiles (Koedel & Betts, 2007; McCaffrey et al., 2008; Sass, 2008).

i. I calculated a Spearman rank correlation for each ordered, paired set of aggregated effectiveness estimates:

$$r_s = 1 - \frac{6\sum D^2}{N^3 - N}, \tag{25}$$

where D is the difference between the ranks of the corresponding values $CO_{1j}, \ldots CO_{8j}$, and N is the number of values in each data set. Based on statistical significance tests with an alpha level of .05 and degrees of freedom of N - 2, where N is the number of pairwise cases,

ii. I calculated a Pearson correlation, r_p, for each paired set of aggregated effectiveness estimates.

$$r_p = \frac{Cov(CO_{jc}, CO_{jt})}{\sigma_{CO_{jc}} \times \sigma_{CO_{jt}}}, \tag{26}$$

where t is not equal to c, c = 1, …, 8. I summarized the results from the Pearson correlation in the same manner as the Spearman rank correlation above. The Pearson correlation was applied to evaluate the linear relation between two sets of effectiveness estimates.

iii. For each composite, across the models, I placed the teacher effectiveness estimate composite into a quintile. Each teacher had eight effectiveness quintiles, four from the composites that resulted from the univariate models and four from the composites that resulted from the multivariate model. Then, the difference in quintiles for each teacher across the models was examined. The frequency of quintile changes is reported. Also, the difference in quintiles from composite to composite within model was examined.

c. For research question four, the variation in the outcomes of interest across the replications was analyzed by sample size. First, the ANOVA assumptions were evaluated; normality, independence of cases, and homogeneity of variance. The cases are assumed independent given the experimental design. PROC Univariate in SAS was used to produce the statistics and visuals to check the normality and homogeneity. Histograms showed that the estimated bias of the teacher-level residuals were not normally distributed. The correlations between the residuals and composites were not normally distributed either. The P-P and Q-Q plots confirm this as well. For each of the outcomes, examining box plots of the groups revealed that there is homogeneity of variance across sample size conditions. Given that the sample sizes were all greater than 10,000 and ANOVA has been shown to be robust to the violation of the normality assumption (Blanca, Alarcon, Arnau, Bono & Bendayan, 2017; Kahn & Rayner, 2003) the ANOVA were conducted despite the violation of the normality assumption. When the ANOVA resulted in significant variation, the effect size was calculated. The effect size, partial eta squared, was calculated as the sum of squares for the factor of interest divided by the sum of squares for that factor plus its associated error sum of squares.

i. ANOVA was conducted for bias.

ii. Repeated ANOVA was conducted for the Spearman correlations between the univariate residuals and the multivariate residuals.

iii. Repeated ANOVA was conducted for the Spearman correlations between the composites from the different aggregation methods.

iv. Repeated ANOVA was conducted for the Spearman correlations between the composites from the different models.

Summary

In summary, this study presents a Monte Carlo simulation to assess four research questions.

1) *Does the use of univariate or multivariate models result in different levels of bias in estimated group- (teacher-) level residuals?*

2) *Does the use of univariate or multivariate models result in different levels of model fit?*

3) *In terms of teacher ranking, what is the influence of constructing teacher effectiveness composites with equal weight, weight by theory, weight by reliability, and weight by residual covariance structure aggregation methods?*

4) *What is the influence of the combinations of small, medium and large-sized teacher groups with small and average-sized student groups on teacher effectiveness estimates?*

The Monte Carlo simulation generated datasets based on known means and variance-covariance values for the outcomes, the student-level residuals and the teacher-level residuals. Datasets were generated for the sample size combinations, resulting in six datasets. To analyze the datasets and inform the conclusions to the research questions, univariate and multivariate multilevel models were applied to the datasets. As described in the simulation procedures, to address research question one, the observed effectiveness estimates were compared to the generating effectiveness estimate values with the absolute bias statistic. Research question two was addressed with an examination of the model fit statistics for each model within each sample size combination. Research question three was addressed with ranking the teachers based on the composite effectiveness estimates and evaluating pairwise correlations as well as the number of teachers that jump ranked quintiles across paired models and paired composites. Research question four was addressed with analysis of variance. The effect size, partial eta squared, within-subject was calculated when the analysis of variance resulted in statistical significance. The evidence collected to assess the research questions in this study informs the VAM research literature as well as multivariate multilevel modeling literature.

Chapter 4: Results

Data Generation Check

The process for generating the simulation data was presented in Chapter III. In order to confirm the accuracy of the generated data in representing the desired population distributions, descriptive statistics for the variables in a large sample with 500 teachers (*J*) and 22 students per teacher (*n*j) for one replication are presented in Table 5.

Table 5

Descriptive statistics of the variables in the large sample (nj=22, J=500), one replication

	N	Mean	SD	Skewness	Kurtosis
$\boldsymbol{Y_{1ij}}$ math	11,000	50.00	1.12	0.01	-0.03
$\boldsymbol{Y_{2ij}}$ read	11,000	50.00	1.12	0.04	0.00
$\boldsymbol{Y_{3ij}}$ mot	11,000	30.01	1.03	-0.00	-0.02
$\boldsymbol{u_{1j}}$ math	500	-0.00	0.51	-0.17	-0.18
$\boldsymbol{u_{2j}}$ read	500	0.01	0.51	-0.02	-0.11
$\boldsymbol{u_{3j}}$ mot	500	0.00	0.24	0.02	0.12
$\boldsymbol{\varepsilon_{1ij}}$ math	11,000	-0.00	0.99	-0.00	0.00
$\boldsymbol{\varepsilon_{2ij}}$ read	11,000	-0.00	0.99	0.00	-0.00
$\boldsymbol{\varepsilon_{3ij}}$ mot	11,000	0.00	1.00	0.00	0.00

The data that were generated in the largest sample reflects the values used in the data generation code. The means and standard deviations for the outcomes, teacher residuals and student residuals are as expected. There are no issues with the skewness or kurtosis. More results for this and the other samples are discussed below.

The resulting correlations between the teacher-level residuals are within one standard error of the generating values. The data generation correlation between the math and reading teacher-level residuals was .70. The data generation correlation between the math and motivation teacher-level residuals was .18. The data generation correlation between the reading and motivation teacher-level residuals was .18. The resulting teacher-level residual correlations for one replication are presented in Table 6.

Table 6

Correlations between teacher-level residuals in the large sample (n_j=22, J=500), one replication

	Math	Read	Mot
Math	1		
Read	.699	1	
Mot	.180	.178	1

Multivariate and univariate multilevel models were applied to the generated data. Across all 1000 replications for each sample size combination, the multivariate multilevel model and the three univariate multilevel models converged. However, there were cases in the smallest sample, (n_j=10, *J*=15), where the analysis resulted in a non-positive definite G matrix. The G matrix is the estimated covariance matrix for the subject-specific effects. The multivariate model had 53% of replications result in a non-positive definite G matrix. The math, reading and motivation univariate models had .6%, .2% and 15.5%, respectively, resulting in non-positive definite G matrices under this small sample scenario. The values for the residuals resulting from cases with the non-positive definite G matrix were identified as missing in the dataset. In these cases, the analyses are based on the number of non-missing records.

The residuals from the univariate models and the multivariate models were analyzed for variation across the teacher sample size and student sample size as well as the interaction between the teacher and student samples sizes. Due to the dependency between the univariate and multivariate measures, a repeated measures ANOVA was applied to the data. The results are in Table 7.

Table 7

Repeated measures ANOVA, within-subjects effects, for the residuals from the univariate models and the multivariate model across all replications for teacher and student sample sizes

	DF	Sum of squares	F value	P value
Math	1	0.0000	0.00	.944
TSamp	2	0.0002	0.02	.981
SSamp	1	0.0007	0.18	.669
TSamp * SSamp	2	0.0062	0.79	.456
Read	1	0.0014	0.36	.548
TSamp	2	0.0002	0.02	.980
SSamp	1	0.0008	0.19	.659
TSamp * SSamp	2	0.0144	1.81	.163
Mot	1	0.0000	0.01	.925
TSamp	2	0.0018	0.22	.807
SSamp	1	0.0000	0.07	.787
TSamp * SSamp	2	0.0014	1.75	.173

Notes. *n*=147,000 TSamp = teacher sample size, SSamp = student sample size

The results of the analysis of variance, within-subjects effects, show that there is no evidence that the univariate residuals differ significantly from the multivariate residuals on average. There were no differences found in the means of the residuals across the univariate and multivariate models, the teacher sample sizes or the student sample sizes. There were no differences found in the means of the residuals across the interaction of teacher and student sample sizes.

The following results are organized by research question one, two and three. The results for research question four, regarding the impact of sample sizes, are presented along with each of the other research questions since each was examined for differences across samples.

Research Question One

Does the use of univariate or multivariate models result in different levels of bias in estimated group- (teacher-) level residuals?

To assess research question one, the absolute bias for the teacher effectiveness estimates from the univariate and multivariate models within each set of conditions was calculated. Generally, the absolute bias is greater for univariate model effectiveness estimates over the multivariate model effectiveness estimates, for each outcome, for each sample size combination. The small samples where both the teacher sample size and student sample size are small are exceptions to this. For these two samples, the absolute bias for the multivariate model is slightly greater. Table 8 presents the average absolute bias for the teacher effectiveness estimates for each outcome for each model.

Table 8

Average absolute teacher effectiveness estimate bias across replications by sample

	Average Bias					
	Multivariate model			Univariate model		
Sample (n_j, J)	Math	Read	Mot	Math	Read	Mot
1: 10, 15	.241	.239	.188	.238	.237	.170
2: 10, 220	.201	.201	.158	.215	.215	.158
3: 10, 500	.206	.206	.157	.214	.214	.157
4: 22, 15	.185	.184	.148	.185	.184	.141
5: 22, 220	.154	.154	.130	.159	.158	.130
6: 22, 500	.153	.153	.129	.157	.157	.130

Figure 2 presents the distributions of absolute bias of the teacher effectiveness estimates by each sample. The box plots below show the average and the range of absolute bias values.

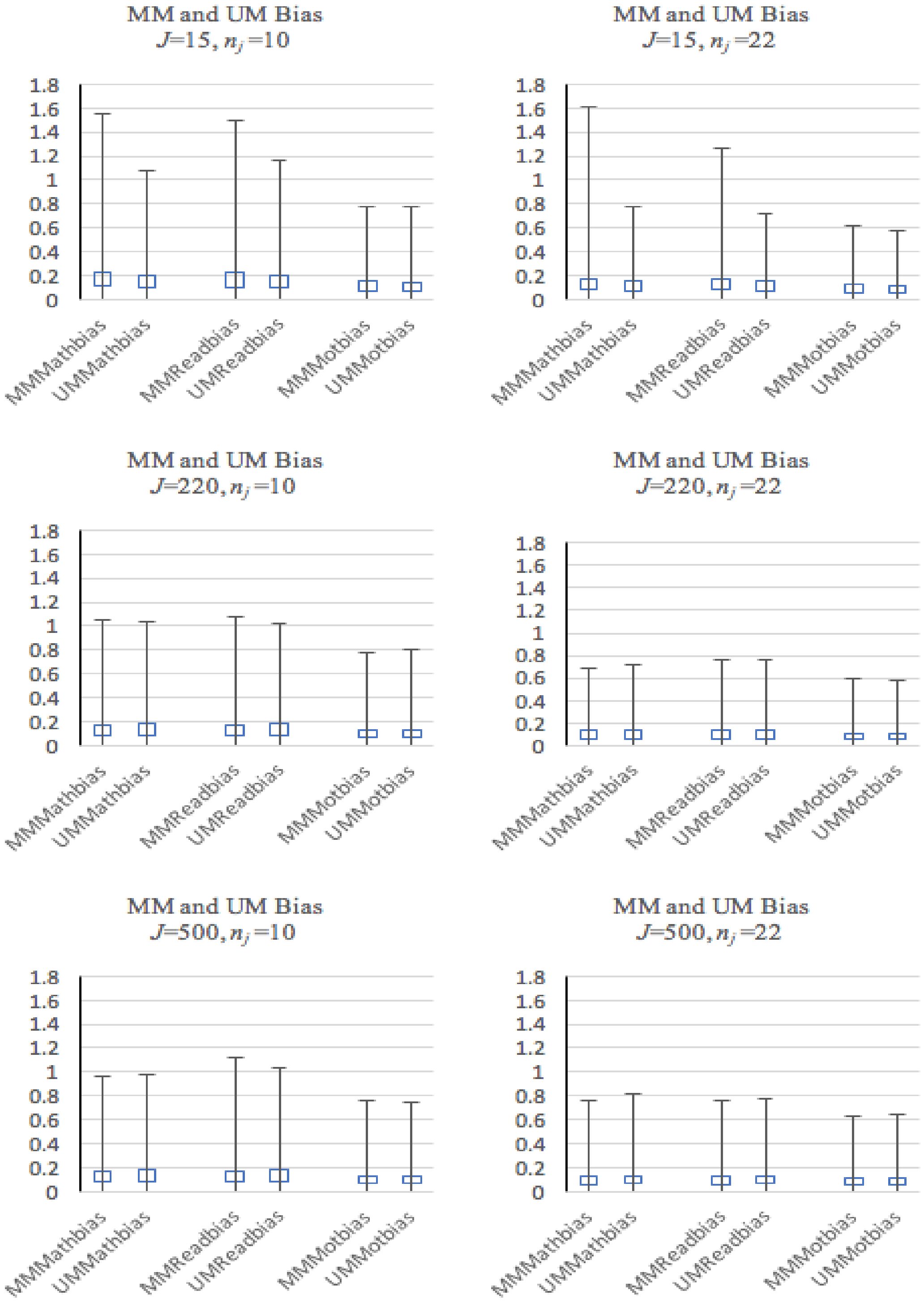

Figure 2. Distribution of bias of the teacher effectiveness estimates by sample.

An analysis of variance of bias across all replications by teacher sample size, student sample size and teacher by student sample size interaction showed a significant difference for the means of the absolute bias variables. The effect size, partial eta squared (ηp^2), is presented in Table 9. Conventionally, values of .01, .06, and .14 represent small, moderate, and large effects, respectively (Green, Salkind, & Akey, 2000). The ANOVA results indicate that the mean absolute bias for each outcome for each model varies significantly across the teacher sample sizes and the student sample sizes. The effect sizes for the teacher sample size variable are small across all bias variables. The effect sizes for the student sample sizes are between small and medium. The effect sizes are not calculated for the factors where the effect is not significant. For the interaction effect, only the differences in mean bias for multivariate motivation, univariate math and univariate reading were significant. The effect sizes for these were all small, <0.0001.

Table 9

Results of the ANOVA for Bias across all replications by sample size

	DF	Sum of squares	F value	P value	Effect size ηp^2
MM math bias	5	1153.93	11917.90	<.0001	0.04
TSamp	2	136.98	3535.96	<.0001	0.00
SSamp	1	1018.13	52576.70	<.0001	0.04
TSamp * SSamp	2	0	0	1	--
MM read bias	5	1131.98	11740.60	<.0001	0.04
TSamp	2	127.56	3307.66	<.0001	0.00
SSamp	1	1005.46	52142.00	<.0001	0.03
TSamp * SSamp	2	0	0	1	--
MM mot bias	5	304.3	5109.56	<.0001	0.02
TSamp	2	20.49	859.92	<.0001	0.00
SSamp	1	282.91	23751.50	<.0001	0.02
TSamp * SSamp	2	0.91	38.21	<.0001	0.00
UM math bias	5	1195.94	11744.80	<.0001	0.04
TSamp	2	22.37	549.16	<.0001	0.00
SSamp	1	1173.33	57613.90	<.0001	0.04
TSamp * SSamp	2	0.24	5.84	0.003	0.00
UM read bias	5	1175.67	11598.20	<.0001	0.04
TSamp	2	17.8	439.01	<.0001	0.00
SSamp	1	1157.62	57100.40	<.0001	0.04
TSamp * SSamp	2	0.25	6.27	0.0020	0.00
UM mot bias	5	284.95	4773.36	<.0001	0.02
TSamp	2	5.08	212.68	<.0001	0.00
SSamp	1	279.81	23436.30	<.0001	0.02
TSamp * SSamp	2	0.06	2.55	0.0800	--

Notes. *n*=147,000 TSamp = teacher sample size, SSamp = student sample size

To better understand the relation between the teacher effectiveness estimates from the univariate and multivariate models, the correlations were calculated. The average correlations between the effectiveness estimates from the univariate models and the multivariate models are significantly high, for each outcome, for each sample size combination. The correlations are presented in Table 10.

Table 10

Ave. Pearson correlation between the effectiveness estimates from the UM and MM across replications

	Correlations		
Sample *(*n$_j$, J*)*	MMMath, UMMath	MMRead, UMRead	MMMot, UMMot
1: 10, 15	.965	.963	.849
2: 10, 220	.984	.984	.975
3: 10, 500	.985	.985	.982
4: 22, 15	.999	.989	.949
5: 22, 220	.995	.995	.993
6: 22, 500	.995	.995	.994

Notes. UM = Univariate multilevel model, MM= Multivariate multilevel model

A repeated measures analysis of variance across all replications across teacher sample size and student sample size shows there is not enough evidence to conclude there is a significant difference between the means of the correlations of the teacher effectiveness estimates from the univariate and multivariate models. The results of the ANOVA are presented in Table 11.

Table 11

Repeated measures ANOVA, within-subjects effects for correlations between UM and MM estimates across replications by samples

	DF	Sum of squares	F value	P value
Math	1	.00	0.00	0.94
TSamp	2	.00	0.02	0.98
SSamp	1	.00	0.18	0.67
TSamp * SSamp	2	.01	0.79	0.46
Read	1	.00	0.36	0.55
TSamp	2	.00	0.02	0.98
SSamp	1	.00	0.19	0.66
TSamp * SSamp	2	.01	1.81	0.16
Mot	1	.00	0.01	0.93
TSamp	2	.00	0.22	0.81
SSamp	1	.00	0.07	0.79
TSamp * SSamp	2	.00	1.75	0.17

Notes. *n*=147,000 TSamp = teacher sample size, SSamp = student sample size

Research Question Two

2) *Does the use of univariate or multivariate models result in different levels of model fit?* The deviance statistic from the multivariate multilevel model was compared to the sum of the deviance statistics from the univariate multilevel models, for each data set. The LRT with six degrees of freedom was employed to compare the difference in the deviances to a chi-square distribution to determine significance. The use of univariate or multivariate models results in different levels of model fit (p <.05). The percent of the replications where the multivariate multilevel model had a greater model fit over the univariate multilevel models is summarized in Table 12. Only the two smallest samples had less than 100% of the replications resulting in a non-significant LRT, indicating that the multivariate multilevel model was not a significantly better fit in a very few instances.

Table 12

LRT and average deviance statistic for the MM and UM models across replications

Sample (n_j, J)	N	MM	Sum of UM	UM math	UM reading	UM motivation	% LRT Sig.
1: 10, 15	932	1299.96	1317.91	442.83	443.26	431.82	75.8%
2: 10, 220	1000	19220.97	19393.30	6522.75	6518.87	6351.68	100%
3: 10, 500	1000	43683.79	44069.22	14819.48	14816.48	14433.27	100%
4: 22, 15	994	2849.36	2875.39	963.44	961.77	950.18	96.8%
5: 22, 220	1000	41921.76	42232.77	14145.63	14151.19	13935.95	100%
6: 22, 500	1000	95260.71	95963.37	32158.61	32148.36	31656.41	100%

The model fit analysis provides evidence that the models produce different levels of model fit, with the multivariate model outperforming the univariate models. This finding, alone, cannot lead to a conclusion that the multivariate model should be used over the univariate models. There are other issues to consider than model fit. This is examined in more detail in the discussion.

Research Question Three

3) *In terms of teacher ranking, what is the influence of constructing teacher effectiveness composites with equal weight, weight by theory, weight by reliability, and weight by residual covariance structure aggregation methods?* The descriptive statistics for the aggregated teacher effectiveness estimates composites are presented in Table 13. Consistent across all samples, the standard deviation for the composite constructed from weights from the residual covariance structure was less than the standard deviation for the other composites. The means and standard deviations appear to be comparable for the equal weight, theory weight, and reliability weight composites. While all of the composite equations are structurally similar, the residual covariance-based composite differs in one key way. In the other three equations, the weights are relatively of the same magnitude, ranging from .5 to 1. The equation for the covariance structure-

based composite contains a set of weights based on the covariance matrix, which has values that are much smaller and closer together than the weights used in the other aggregation methods, ranging from .022 to .2.

Table 13

Descriptive statistics for the composites, mean and standard deviation

			Composite			
Sample *(*nj, J*)*			Equal Wt.	Theory Wt.	Reliable Wt.	Residual Cov.
1: 10, 15	MM	Mean	0.0017	0.0015	0.0014	0.0003
		SD	(0.772)	(0.725)	(0.644)	(0.153)
	UM	Mean	0	0	0	0
		SD	(0.762)	(0.735)	(0.642)	(0.135)
2: 10, 220	MM	Mean	-0.0001	-0.0001	-0.0001	-0.0000
		SD	(0.836)	(0.802)	(0.704)	(0.173)
	UM	Mean	0	0	0	0
		SD	(0.774)	0.754)	(0.655)	(0.124)
3: 10, 500	MM	Mean	-0.0001	-0.0001	-0.0001	-0.0000
		SD	(0.860)	(0.827)	(0.725)	(0.201)
	UM	Mean	0	0	0	0
		SD	(0.777)	(0.756)	(0.657)	(0.136)
4: 22, 15	MM	Mean	-0.0018	-0.0017	-0.0015	-0.0000
		SD	(0.855)	(0.808)	(0.715)	(0.06)
	UM	Mean	0	0	0	0
		SD	(0.864)	(0.834)	(0.729)	(0.081)
5: 22, 220	MM	Mean	0.0000	0.0001	0.0000	0.0000
		SD	(0.919)	(0.881)	(0.773)	(0.269)
	UM	Mean	0	0	0	0
		SD	(0.874)	(0.845)	(0.737)	(0.227)
6: 22, 500	MM	Mean	0.0000	0.0000	0.0000	0.0000
		SD	(0.934)	(0.897)	(0.786)	(0.243)
	UM	Mean	0	0	0	0
		SD	(0.878)	(0.848)	(0.740)	(0.205)

Notes. UM = Univariate multilevel model, MM= Multivariate multilevel model

The two tables below present the correlations, Pearson and Spearman Rho, between the composites constructed from the four different aggregation methods, across conditions of the number of students and the number of teachers. As evidenced by the analysis of variance results, presented in Table 16, the mean correlations are consistent across the teacher and student sample size conditions and across the different composites, except for the sample with the small teacher sample size and the average student sample size. These mean correlations are lower than the other pairs across the teacher and student sample conditions for the paired correlations that include the residual covariance structure-based composite. The smallest sample with the small teacher sample size and the small student sample size, also has slightly lower correlations for the pairs that involve the residual covariance structure-based composites, but not as low as the correlations from the sample with the small teacher sample size and average student sample size. There were no noticeable differences between the Pearson correlations and the Spearman Rho correlations.

Table 14

Pearson correlations between composites from the aggregation methods

		Aggregation Methods Combinations					
Sample (n_j, J)		EW&T	EW&R	EW&RC	T&R	T&RC	R&RC
1: 10, 15	MM	.994	.999	.972	.998	.967	.970
	UM	.994	.999	.892	.998	.879	.885
2: 10, 220	MM	.996	.999	.983	.999	.969	.978
	UM	.995	.999	.950	.998	.933	.943
3: 10, 500	MM	.996	.999	.979	.999	.967	.974
	UM	.995	.999	.936	.998	.917	.927
4: 22, 15	MM	.995	.999	.568	.998	.504	.542
	UM	.994	.999	.740	.998	.672	.714
5: 22, 220	MM	.995	.999	.971	.998	.957	.966
	UM	.994	.999	.950	.998	.934	.943
6: 22, 500	MM	.995	.999	,965	.998	.948	.958
	UM	.994	.999	.937	.998	.918	.929

Notes. Aggregation methods: EW: equal weight, T: theory, R: reliability, RC: residual covariance structure

Table 15

Spearman Rho correlations between rankings for aggregation methods

		Aggregation Methods Combinations					
Sample (n_j, J)		EW&T	EW&R	EW&RC	T&R	T&RC	R&RC
1: 10, 15	MM	.994	.999	.973	.997	.967	.971
	UM	.993	.999	.882	.997	.867	.875
2: 10, 220	MM	.996	.999	.981	.998	.966	.975
	UM	.995	.999	.946	.998	.927	.937
3: 10, 500	MM	.996	.999	.977	.998	.963	.972
	UM	.995	.999	.930	.998	.910	.921
4: 22, 15	MM	.995	.999	.547	.998	.484	.522
	UM	.994	.999	.733	.997	.667	.708
5: 22, 220	MM	.995	.999	.968	.998	.953	.963
	UM	.994	.999	.946	.997	.928	.938
6: 22, 500	MM	.995	.999	.961	.998	.944	.954
	UM	.994	.999	.932	.997	.910	.923

Notes. Aggregation methods: EW: equal weight, T: theory, R: reliability, RC: residual covariance structure

The results of the analysis of variance indicate significant differences between the means of the correlations for paired composites for both the univariate models and the multivariate model across the aggregation method and teacher and student sample size conditions. The aggregation method effect is large for both the univariate and multivariate models. This is evident by looking at the correlations for the small teacher/average student sample size condition in Table 15. The pairs that include the residual covariance-based composites are much lower than the others. The teacher sample size effects are large for both the multivariate and univariate models. The student sample size effects are lower, with the effects being between small and medium for the multivariate model and between medium and large for the univariate model. The interactions of teacher sample size and student sample size with aggregation method are also significant effects. The effects for the interaction between the aggregation method and the teacher sample size is large for both the multivariate and univariate models. The interaction effect between the aggregation method and the student sample size is between small and medium for the multivariate model and between medium and large for the univariate model.

Table 16

ANOVA and effect size for correlations of paired aggregation methods across replications by samples

		DF	Sum of squares	F value	P value	Effect size ηp^2
AM	MM	5	673.59	3733.87	<.0001	.34
	UM	5	104.18	5416.06	<.0001	.47
TSamp	MM	2	361.65	5011.78	<.0001	.22
	UM	2	37.43	4864.63	<.0001	.24
SSamp	MM	1	51.08	1415.70	<.0001	.04
	UM	1	12.21	3174.27	<.0001	.09
TSamp* SSsamp	MM	2	816.48	11314.80	<.0001	.39
	UM	2	6.89	894.99	<.0001	.06
AM * TSamp	MM	10	362.54	1004.83	<.0001	.22
	UM	10	36.45	947.40	<.0001	.24
AM*SSamp	MM	5	51.14	283.49	<.0001	.04
	UM	5	12.54	652.08	<.0001	.10

Notes. n_{mm}=36,000 n_{um}=30,648 AM= aggregation method, TSamp = teacher sample size, SSamp = student sample size

Although most of the composite correlations don't appear to differ much, the differences from the residual covariance-based composite and the small teacher sample size condition appear to impact the results enough to create significant effects across the conditions.

The correlations between the composite from the univariate models and the corresponding composite from the multivariate model are high for each aggregation method and each sample. The correlations are presented in Table 17.

Table 17

Average correlation between UM and MM composites

	Aggregation Methods			
Sample (n_j, J)	EW	R	T	RC
1: 10, 15	.989	.991	.993	.955
2: 10, 220	.999	.999	.999	.988
3: 10, 500	.999	.999	.999	.988
4: 22, 15	.998	.998	.999	.981
5: 22, 220	.999	.999	.999	.995
6: 22, 500	.999	.999	.999	.996

Notes. Aggregation methods: EW: equal weight, T: theory, R: reliability, RC: residual covariance structure

Analysis of variance showed that across the replications, the difference between the means of the correlations between the composite from the multivariate model and the corresponding composite from the univariate models did differ significantly across the teacher samples sizes and the student sample sizes conditions. The interaction effect between teacher and student sample sizes was significant for all the aggregation methods except for the equal weight method, with the covariance structure-based composite having a very large effect size. The teacher sample effect was large for all four aggregation methods as shown in Table 18.

Table 18

ANOVA for correlations of UM and MM composites across replications by samples

Aggregation method	DF	Sum of squares	F value	P value	Effect size ηp^2
Equal Wt	5	56.36	1064.16	<.0001	0.18
TSamp	2	56.04	2644.99	<.0001	0.18
Ssamp	1	0.27	25.11	<.001	0.00
TSamp*SSamp	2	0.06	2.85	.058	--
Theory Wt	5	50.81	928.50	<.0001	0.16
TSamp	2	50.20	2293.37	<.0001	0.16
Ssamp	1	0.00	0.07	.799	--
TSamp*SSamp	2	0.61	27.85	<.0001	0.00
Reliable Wt	5	53.37	1011.11	<.0001	0.17
TSamp	2	53.07	2513.47	<.0001	0.17
Ssamp	1	0.11	10.76	0.001	0.00
TSamp*SSamp	2	0.19	8.93	0.0001	0.00
Residual Cov.	5	3065.02	15081.80	<.0001	0.76
TSamp	2	930.78	11450.00	<.0001	0.49
Ssamp	1	244.19	6007.93	<.0001	0.20
TSamp*SSamp	2	1890.05	23250.60	<.0001	0.66

Notes. *n*=23,999 TSamp = teacher sample size, SSamp = student sample size

When examining the findings for the correlations between the composites, across the conditions, there are a few patterns that emerge. The small teacher sample size condition appears to influence the correlations between the compositions. The residual covariance-based composite also stands out when looking at the standard deviations across sample size conditions, correlations between composites across the aggregation methods and correlations between composites across the models.

In addition to correlations, the teacher effectiveness composites were ranked and placed in quintiles and analyzed for differences across the composites derived from the univariate models and the multivariate model. For each aggregation method, within each sample, the frequency of quintile changes was examined. Across all aggregation methods and samples, most composites did not change quintiles when comparing the univariate-based composite to the multivariate-based composite except for two samples. The two samples with the small number of teachers (*J*=15) experienced proportionally more quintile changes than the other samples. The frequency of zero quintile changes for the equal weight, theory-based weight, and reliability weight composites ranges from 58% to 98%. The frequency of quintile changes for the covariance-based composite ranges from 36% to 74%. Table 19 summarizes the changes when going from the univariate model-based composite to the multivariate model-based composite.

Table 19

Frequency of Quintile Changes from Univariate to Multivariate Model-based Composites

Aggregation method	Sample	<-2	-2	-1	0	1	2	>2
EqualWt	$J=15$, $n_j=10$	2%	4%	16%	58%	15%	4%	2%
	$J=220$, $n_j=10$	0	1%	7%	83%	7%	1%	0
	$J=500$, $n_j=10$	0	1%	3%	90%	3%	1%	0
	$J=15$, $n_j=22$	1%	3%	9%	71%	9%	3%	2%
	$J=220$, $n_j=22$	0	1%	3%	94%	2%	1%	0
	$J=500$, $n_j=22$	0	0	1%	97%	1%	0	0
TheoryWt	$J=15$, $n_j=10$	2%	3%	15%	59%	15%	3%	2%
	$J=220$, $n_j=10$	0	1%	7%	83%	7%	1%	0
	$J=500$, $n_j=10$	0	1%	4%	90%	3%	1%	0
	$J=15$, $n_j=22$	2%	3%	9%	73%	9%	3%	2%
	$J=220$, $n_j=22$	0	1%	2%	94%	2%	1%	0
	$J=500$, $n_j=22$	0	0	1%	98%	1%	0	0
ReliableWt	$J=15$, $n_j=10$	2%	3%	15%	58%	15%	3%	2%
	$J=220$, $n_j=10$	0	1%	7%	83%	7%	1%	0
	$J=500$, $n_j=10$	0	1%	3%	90%	3%	1%	0
	$J=15$, $n_j=22$	2%	3%	9%	72%	9%	3%	2%
	$J=220$, $n_j=22$	0	1%	2%	94%	2%	1%	0
	$J=500$, $n_j=22$	0	0	1%	98%	1%	0	0
CovWt	$J=15$, $n_j=10$	3%	6%	20%	42%	21%	6%	3%
	$J=220$, $n_j=10$	0	1%	16%	64%	17%	1%	0
	$J=500$, $n_j=10$	1%	2%	17%	61%	17%	1%	0
	$J=15$, $n_j=22$	1%	6%	24%	36%	25%	6%	1%
	$J=220$, $n_j=22$	0	0	12%	74%	12%	0	0
	$J=500$, $n_j=22$	0	0	12%	74%	13%	0	0

It was also hypothesized that there would be changes in the quintiles when comparing across aggregation methods. The composites were placed into quintiles and the frequency of changes between quintiles was examined. Table 20 presents the quintile changes for the composites from the multivariate model. Like the comparison between the composites from the univariate models and the multivariate model, the comparisons that include the residual covariance-based composite exhibit more quintile changes than the other comparisons. Each of the comparisons with the covariance-based composites had changes of more than 2 quintiles for 16% of the sample where n_j=22, J=15. The other composite comparisons, without the covariance-based composite, had zero changes of more than 2 quintiles at this and the other sample sizes. At this sample size, n_j=22, J=15, only 18-19% of the composite comparisons with the covariance-based composite did not change quintiles. This is compared to 79%, 83% and 86% for the other composite comparisons at the same sample size. This analysis supports the other common finding where the small teacher sample size condition produced more quintile changes than the other sample conditions.

Table 20

Frequency of Quintile Changes Between Composites, Multivariate Model

Composite Pair	Sample	<-2	-2	-1	0	1	2	>2
EqualWt, TheoryWt	J=15, n_j=10	0	0	11%	77%	11%	0	0
	J=220, n_j=10	0	0	9%	81%	9%	0	0
	J=500, n_j=10	0	0	9%	82%	9%	0	0
	J=15, n_j=22	0	0	10%	79%	10%	0	0
	J=220, n_j=22	0	0	11%	78%	11%	0	0
	J=500, n_j=22	0	0	11%	78%	11%	0	0
EqualWt, ReliableWt	J=15, n_j=10	0	0	4%	92%	4%	0	0
	J=220, n_j=10	0	0	3%	93%	3%	0	0
	J=500, n_j=10	0	0	3%	93%	3%	0	0
	J=15, n_j=22	0	0	4%	93%	4%	0	0
	J=220, n_j=22	0	0	4%	92%	4%	0	0
	J=500, n_j=22	0	0	4%	92%	4%	0	0
EqualWt, CovWt	J=15, n_j=10	0	2%	19%	55%	21%	2%	0
	J=220, n_j=10	0	0	20%	57%	21%	1%	0
	J=500, n_j=10	0	2%	19%	59%	19%	2%	0
	J=15, n_j=22	16%	12%	13%	18%	13%	12%	16%
	J=220, n_j=22	0	4%	20%	50%	21%	4%	0
	J=500, n_j=22	1%	4%	29%	50%	20%	5%	1%
TheoryWt, ReliableWt	J=15, n_j=10	0	0	7%	85%	8%	0	0
	J=220, n_j=10	0	0	6%	88%	6%	0	0
	J=500, n_j=10	0	0	6%	88%	6%	0	0
	J=15, n_j=22	0	0	7%	86%	7%	0	0
	J=220, n_j=22	0	0	7%	86%	7%	0	0
	J=500, n_j=22	0	0	7%	86%	7%	0	0
TheoryWt, CovWt	J=15, n_j=10	0	2%	18%	56%	20%	3%	0
	J=220, n_j=10	0	3%	22%	50%	22%	3%	0
	J=500, n_j=10	0	4%	20%	52%	20%	4%	0
	J=15, n_j=22	16%	11%	13%	18%	13%	11%	16%
	J=220, n_j=22	1%	6%	20%	45%	20%	6%	1%
	J=500, n_j=22	1%	6%	20%	45%	20%	6%	1%
ReliableWt, CovWt	J=15, n_j=10	0	2%	19%	56%	20%	2%	0
	J=220, n_j=10	0	2%	21%	54%	21%	2%	0
	J=500, n_j=10	0	2%	20%	56%	20%	2%	0
	J=15, n_j=22	16%	11%	13%	19%	13%	11%	16%
	J=220, n_j=22	1%	5%	20%	48%	21%	5%	1%
	J=500, n_j=22	1%	5%	20%	48%	20%	5%	1%

The comparisons for the composites from the univariate models follow the same patterns as the composites from the multivariate model. The quintile changes for the composites from the univariate model are presented in Table 21. The frequencies of composite comparisons that contain the covariance-based composite that resulted in zero quintile changes range from 15% to 42% across the sample size conditions. This is compared to a range of 76% to 94% for the comparisons that do not contain the covariance-based composite that resulted in zero quintile changes. The comparisons with the covariance-based composite for the sample where n_j=22, J=15 resulted in a greater frequency of quintile changes compared to the other samples. The frequencies of composite comparisons that contain the covariance-based composite that resulted in more than two quintile changes range from 22% to 25% for the small sample size, n_j=22, J=15. The frequencies of the same comparisons, change of more than two quintiles, at the other sample sizes range from 2% to 8%. Again, the small teacher sample size condition and the comparisons that include the residual covariance-based composite exhibit more quintile changes than the others.

Table 21

Frequency of Quintile Changes Between Composites, Univariate Model

Composite Pair	Sample	<-2	-2	-1	0	1	2	>2
EqualWt, TheoryWt	J=15, n_j=10	0	0	11%	78%	11%	0	0
	J=220, n_j=10	0	0	10%	79%	10%	0	0
	J=500, n_j=10	0	0	10%	79%	10%	0	0
	J=15, n_j=22	0	0	11%	78%	11%	0	0
	J=220, n_j=22	0	0	12%	76%	12%	0	0
	J=500, n_j=22	0	0	12%	76%	12%	0	0
EqualWt, ReliableWt	J=15, n_j=10	0	0	5%	91%	4%	0	0
	J=220, n_j=10	0	0	4%	92%	4%	0	0
	J=500, n_j=10	0	0	4%	94%	4%	0	0
	J=15, n_j=22	0	0	4%	92%	4%	0	0
	J=220, n_j=22	0	0	4%	91%	4%	0	0
	J=500, n_j=22	0	0	4%	91%	4%	0	0
EqualWt, CorrWt	J=15, n_j=10	2%	8%	21%	36%	21%	8%	3%
	J=220, n_j=10	1%	5%	24%	41%	24%	5%	1%
	J=500, n_j=10	2%	7%	21%	42%	20%	7%	2%
	J=15, n_j=22	12%	14%	16%	16%	17%	14%	11%
	J=220, n_j=22	2%	7%	19%	42%	19%	7%	2%
	J=500, n_j=22	2%	7%	19%	42%	19%	7%	2%
TheoryWt, ReliableWt	J=15, n_j=10	0	0	7%	85%	7%	0	0
	J=220, n_j=10	0	0	7%	86%	7%	0	0
	J=500, n_j=10	0	0	7%	86%	7%	0	0
	J=15, n_j=22	0	0	7%	86%	7%	0	0
	J=220, n_j=22	0	0	8%	84%	8%	0	0
	J=500, n_j=22	0	0	8%	85%	8%	0	0
TheoryWt, CovWt	J=15, n_j=10	4%	8%	21%	34%	21%	9%	4%
	J=220, n_j=10	1%	7%	23%	37%	23%	7%	1
	J=500, n_j=10	2%	8%	20%	38%	20%	8%	2
	J=15, n_j=22	13%	13%	15%	15%	16%	13%	12%
	J=220, n_j=22	4%	8%	19%	39%	19%	8%	4%
	J=500, n_j=22	4%	8%	18%	38%	18%	8%	4%
ReliableWt, CovWt	J=15, n_j=10	4%	8%	21%	35%	21%	8%	4%
	J=220, n_j=10	1%	6%	23%	39%	23%	6%	1%
	J=500, n_j=10	2%	7%	20%	40%	20%	7%	2%
	J=15, n_j=22	12%	14%	16%	16%	16%	14%	12%
	J=220, n_j=22	2%	8%	19%	41%	19%	8%	2%
	J=500, n_j=22	4%	8%	19%	40%	19%	8%	4%

Summary

The use of univariate or multivariate models leads to slightly different levels of bias in estimated group- (teacher-) level residuals and different levels of model fit. Generally, the multivariate model produced estimates with slightly lower bias and the multivariate model deviance statistics indicate better model fit. Teacher sample size had small significant effects for the math, reading and motivation bias variables for both the multivariate and univariate models. Teacher sample size did not have a significant effect on the correlations between the residuals from the univariate and multivariate models. The correlations between the composites both within and across the composites from the univariate and multivariate models are very high, all above .955 for the paired composites between the models. For the paired composites within each model, the correlations were above .917 for all of the samples except where n_j=22 and J=15 for the correlations that included the covariance-based composite. This sample had correlations between composites within the model of .484 to .733 for the pairs that included the covariance-based composite. Teacher sample size had significant effects for the correlations between the composites from the multivariate and univariate models. Composites from the univariate and multivariate models changed quintiles across the aggregation methods and sample sizes. The greatest differences were seen for the comparisons that included the covariance-based composite and in the small sample size condition where n_j=22, J=15. The differences in the frequencies of zero quintile changes were up to 60% in some cases for composite comparisons both within and across the models in the small size condition where n_j=22, J=15.

Two consistent issues appeared in several of the analyses. First was the influence of the small teacher sample size. The differences in the correlations between and across composites as well as the quintile classifications were greater in the small teacher sample size condition. The differences are great enough to caution against the use of such a small teacher sample size. The second issue was the difference in the results for the residual covariance-based composite compared to the other composites, particularly when looking at the correlations of the different composites and in the quintile classifications.

Examination of the univariate and multivariate model composites provide some support for the argument to use the multivariate model over the univariate models. While the correlations between effectiveness estimates and composites did not appear to vary across the models as much as was expected, the model fit and quintile changes suggest that the multivariate model should be used over the univariate models. This is discussed in more detail in Chapter V.

Chapter 5: Discussion

Given the attention on evaluating teachers, schools, districts and states to quantify the effectiveness on student learning (Goldhaber, 2010; Montes, 2012; Rothstein, 2016), this study employed a multivariate multilevel model that yields a composite effectiveness estimate. The findings provided mixed statistical evidence toward the conclusion that the multivariate model should be used over the univariate models; however, a theoretical argument could be made for the increased validity provided by the multiple outcomes in the multivariate model (Baldwin et al., 2014). Interpretation of the findings, limitations, implications and future research are discussed below.

Interpretation of the Findings

Research Question 1. *Does the use of univariate or multivariate models result in different levels of bias in estimated group- (teacher-) level residuals?*

In the study, the use of univariate or multivariate models results in slightly different levels of bias in estimated group- (teacher-) level residuals. The absolute bias is slightly greater for the univariate model effectiveness estimates over the multivariate model effectiveness estimates, for each outcome, for each sample size combination except for the conditions where the teacher sample size was small (J=15). These findings are consistent with the previous literature that found the advantage of employing the multivariate model (Griffiths et al., 2003). However, it should be noted that the differences in absolute bias are very slight across all the samples and outcomes, ranging from .000 to .018. The evidence does support the hypothesis that the choice of model has an impact on the bias of the estimates.

There is a note to make about the bias statistic and its interpretation. The bias statistic was calculated as the difference between the estimated parameter and the ‘true value’, the generating value. The bias statistic calculated in this study could also be a measure of variability, rather than ‘bias’. It is difficult to parse the value of the difference between the estimated parameter and the true value of the parameter (Fortmann-Roe, 2012). The difference is likely made up of both bias and variance, but this could not be parsed in this study. It is important to note in this study, that the results focus on bias, but it is possible that the term ‘bias’ contains both bias and variance.

Research Question 2. *Does the use of univariate or multivariate models result in different levels of model fit?*

In the study, the use of univariate or multivariate models results in different levels of model fit. The LRT shows that the deviances from the multivariate multilevel models differ from the summed deviances from the univariate multilevel models. These findings are consistent with the previous literature that argued for the utilization of multivariate models to test hypotheses about the associations among the multiple outcomes. The results confirm prior research (Baldwin et al., 2014). Baldwin et al. (2014) examined multivariate multilevel models and univariate multilevel models. Baldwin et al. (2014) argued for the use of multivariate models to examine hypotheses about the relations among the multiple outcomes. These results are useful in the discussion of which model to use when the data are multivariate. If not modeled correctly, the inferences based on hypothesis tests about the relation between the outcomes could be inaccurate.

Research Question 3. *In terms of teacher ranking, what is the influence of constructing teacher effectiveness composites with equal weight, weight by theory, weight by reliability, and weight by residual covariance structure aggregation methods?*

In the study, teacher effectiveness composites were highly correlated across the equal weight, weight by theory, weight by reliability, and weight by residual covariance structure aggregation methods from the multivariate models and the univariate models but the analysis of variance for the means of the correlations between the composites from the univariate models and the multivariate model revealed significant effects from the aggregation method and teacher sample sizes. The residual covariance-based composite and the small teacher sample conditions produced larger effects than the other conditions. These findings are consistent with the limited previous literature that found that the relation between the residuals estimated with the multivariate model was larger than the relation between those associated with the independently modeled outcomes (Leckie, 2018). The effects found on the correlations between composites and the quintile changes suggest that the aggregation method and whether the residuals are from the univariate models or the multivariate models are impactful decisions for researchers and stakeholders such as administrators and teachers. These results support those from research questions one and two that provide evidence that the model specification does impact the teacher effectiveness estimates. These findings suggest that further research is required. As discussed in the limitations below, the simulation and study design decisions greatly impact the generalizability of the results. The modeling specifications and variety of aggregation methods could lead to very different findings.

Research Question 4. *What is the influence of the combinations of small, medium and large-sized teacher groups with small and average-sized student groups on teacher effectiveness estimates?*

There were three teacher sample sizes and two student sample sizes. The residuals, bias, model fit, and composites were evaluated across the conditions. There was significant difference in the means across the replications across conditions for bias, deviance, and residual and composite correlations. As expected, the smaller samples had greater effects. There already exist some guidelines on the sample sizes for multivariate and univariate multilevel models. As discussed, Maas and Hox (2005) suggested a ratio of 100/10 for groups to subjects. This dissertation study supports the researchers' claim. The conditions where the teacher to student ratios were only 15/10 and 15/22 exhibited greater effects in bias, deviance and composites than the other samples where the teacher samples were 220 and 500. The differences in the means for the bias, deviances and composites across samples and models were significant. This finding could be a problem for evaluations in small schools where the teacher sample size is limited. They could model multiple schools to increase the teacher sample size, but then would need to add a new grouping level for 'school'. The impact of this needs to be researched in more detail.

Limitations

This study was designed to address the specific questions regarding the composites and the methods of constructing them. This simulation, like others, required the input of specific fixed variables and values which led to limitations. The generalizability of the results is potentially impacted by each design decision in this study.

The target population was students and teachers. The findings cannot be generalized to other people or groups (Remler & Van Ryzin, 2010). One might want to generalize the results to VAM studies at the school- or district-level rather than the teacher-level. It is likely that the sample sizes would differ for those groups, therefore the results of this study may not be comparable. The teacher sample sizes used in this simulation may not accurately represent the typical values of the number of teachers being evalauted. It is not likely that there would be a sample of 220 or 500 teachers within one school. This sample size could exist if there were multiple schools, but in that case a fourth level should be added to the model to account for the school differences. Given the consistent finding that the small teacher sample size condition was notable and that the likelihood that most teacher samples will be small, the use of the multivariate model or even VAM in general is questionable and requires further research. Future research could examine the differences in sample sizes in more detail.

This study employed one set of correlations among the outcomes, where reading and math were correlated .70 and the achievement outcomes were correlated with motivation at .18. The results may be impacted with a different set of correlations. This study assumed that two of the outcomes were very closely related. It is realistic to expect that a researcher might select three outcomes that are dependent but not as highly correlated as math and reading. Generally, the results for the math effectiveness estimates were the same as the results for the reading effectiveness estimates. If they were not as highly correlated, they could differ as the results for motivation do. The motivation outcome was generated with a lower variance and a lower covariance with the other outcomes. This produced some differences in the results, such as the average absolute bias. The motivation average absolute bias was lower than math and reading across both models and the sample size conditions. Alternatively, if the outcomes were all highly correlated, the results would be more similar. As noted earlier, there could be an issue when including outcomes that are too highly correlated. The additional outcomes are adding complexity to the models without adding more information. The correlation between outcomes also greatly impacts the covariance-based composites. Since the covariance was used as a weight in the covariance-based composite, the values set for the correlations directly impact the composites which then impacts the ranking and quintile placement for the teacher.

Related to the correlation between the outcomes, is the intra-class correlation (ICC). This is a measure of how much of the variance in the outcomes is between the level two units. The higher the ICC, the more homogeneity there is within teacher groups. An ICC above zero indicates some degree of dependence in the data. The data were generated with an assumed ICC of .2, which is common in student achievement scores (Hedges & Hedberg, 2007). The influence of the ICC value was not examined in this study. If a higher ICC was selected, indicating more dependence in the data within the teacher group, there could have been more variation among the teacher groups (less within), resulting in a wider range of teacher effectiveness scores (McCoach & Adelson, 2010). A lower ICC value would have created more variability within the teacher group, creating less variation across teachers. The effectiveness scores would have had a smaller range and the differences between the effectiveness estimates would have been less. The impact of this would have been less differences across the composites and likely less differences in the estimates across the models.

Another implication of the choice of outcomes is the scale of the measure and the impact this has on the composite. For example, in this study math and reading were assumed to be on the same scale, but motivation was assumed to come from a measure with a different scale. This creates a different scale for the effectiveness estimates which are then aggregated into the composite. This could be addressed by standardizing the outcomes to be on the same scale. This was not addressed in this study. This means that the motivation outcome was weighted less in the composites given its scale, even in the equal weights method where it was assumed they were all equally weighted. Another interesting potential issue is with the reliability values for the measures. Often in psychology, a measure's reliability can vary depending on the sample size (Holland et al., 2018). This simulation assumed the reliability values were constant across sample size conditions, but in reality some measures vary in reliability across sample sizes. If the reliability values varied across sample size conditions, the composites would likely have been more variable like the covariance matrix-based composites. The weights would have been different for each sample size condition rather than constant like the equal weights and theory-based weights. It is important to examine the weights and aggregation methods with respect to the measures of the outcome variables that are aggregated.

The number of outcomes modeled is also a potential limiting factor. If there are more outcomes, there is an increased risk of including outcomes that are too highly correlated rather than introducing new information into the model or including outcomes that do not correlate at all that makes the use of the multivariate model unnecessary. On the other hand, not including outcomes that contribute to teacher effectiveness causes, in some cases, incorrect decisions leading to unfair actions, such as missed promotions, missed bonuses, probation, transfers and termination. It is difficult to know how many and which outcomes to include in VAM. It could be possible that there are a set of outcomes that measure the teacher effectiveness on some students that is not the right set of outcomes for other students (Lefgren & Sims, 2012). Teachers have a variety of impacts on students and depending on the characteristics of the students, it is possible that those impacts are not consistent. For example, one teacher may be able to bring up the test scores for part of her class but not all (Lockwood & McCaffrey, 2009; Rothstein, 2009). Lockwood and McCaffrey (2009) found modest effects from student achievement level on teacher effectiveness estimates indicating that teacher effectiveness could vary for different student achievement levels within the same class. Suppose that for students that have more trouble with the achievement tests, the teacher has a positive impact in other areas, like motivation, confidence or imagination. Many VAM only focus on achievement tests, but teachers can provide other positive influences on students that should be rewarded (Chetty, Friedman, & Rockoff, 2011; Jackson, 2012).

This study did not include covariates in the model, although the research is inconclusive on the impact of covariates (Lockwood et al., 2007). It is possible that including student and/or teacher level covariates could have impacted the results. Covariates could have controlled for some of the variance in the model decreasing the amount attributed to the teacher effectiveness. However, if the outcome had been a growth score rather than a direct score on some measure, the need for covariates is removed. A gain score or growth score is the difference between the prior score(s) and current score. This controls for extraneous variables; therefore, the use of covariates is not necessary (Rose, Henry, & Lauen, 2012). In this simulation, without covariates, it may have been more appropriate to use outcomes that were gain or growth scores to help with the interpretation of the results given the lack of covariates in the model. Given that the modeling specifications would most likely impact the results, these results cannot be generalizable to all multivariate multilevel VAM. Future studies could examine the variety of VAM specifications within the multivariate multilevel class of models.

The aggregation methods were selected to demonstrate that there are a variety of ways to combine the effectiveness estimates to form a composite. This study did not include all the possible aggregation methods. It is likely that there are aggregation methods that would significantly impact the ranking and correlation of the composites, such as the methods that weight for the number of students. This method would change the results for systems where teachers had different class sizes or even within the class where the measurements were not completed by all students. In the teacher evaluation systems where the stakes are high and the evaluation score depends significantly on the effectiveness composite, the decision of which aggregation method to use could seriously impact a teacher's career. The analyses on the composites shows that while there is a high correlation between the composites for most sample conditions, there are significant differences in the means of composite correlations across aggregation methods. The small teacher sample size condition has composite correlation values lower than the others which cautions against the reliability of teacher effectiveness estimates in small teacher samples. Future studies could examine the different categories of aggregation methods in more detail.

Another study design decision that could lead to limitations in the generalizability of the results includes the handling of missing data. Missing data is common in education data and not an issue that was modeled in this simulation. There are different types of missing data that could be expected in data of this kind. For example, the model included multiple outcomes which in this study, each student had a score for each one. A student could have reading and math assessment scores, but failed to do the survey that measured motivation. In this case, there are decisions to be made of how to deal with the missing data. The student could be excluded from the dataset, the student's motivation score could be imputed using the mean, or the teacher's effectiveness composite could consist of unequal residual sample sizes. These are just a few of the potential methods for dealing with the missing data and only one example of the type of missing data that is possible. A future study could examine the impact of missing data on univariate and multivariate multilevel model as well as the impact on a variety of aggregation methods.

During the application of the multivariate and univariate models a portion of the analyses resulted in a non-positive definite G matrix. The G matrix is the estimated covariance matrix for the residuals. The multivariate model analysis with the smallest sample size condition, where the teacher sample size was small and the student sample size was small, yielded a non-positive definite G matrix 53% of the time. These results were removed from the further analysis. This created missing data in two of the data sets. While the number of replications was still around 500 in the smaller sample, the decrease in replications could have impacted the findings. The small teacher sample size conditions resulted in notable findings and the impact of the reduced number of replications was not examined. The analysis could have been run for more replications to make up for those that resulted in the non-positive definite G matrix so that all of the datasets had 1000 replications. The impact of the number of replications could be examined in future research.

There are two common reasons for why the PROC MIXED procedure might result in a non-positive G matrix. One is that there is not enough variation in the outcomes, meaning that after controlling for the fixed effects, there isn't much variation in the residuals. The other cause is that the model is misspecified. If the number of observations is too close to the number of parameters in the model, the analysis can result in a non-positive definite matrix. In this study, the non-positive definite G matrix finding was only in the smallest sample. This is more evidence that the teacher sample size of 15 is too small to produce usable teacher effectiveness estimates.

Implications for Practice and Further Research

Although this dissertation was limited in scope, researchers can utilize the study to examine if the use of univariate or multivariate models results in different levels of bias in estimated group- (teacher-) level residuals and different levels of model fit. This study also provides a demonstration of constructing an effectiveness composite. A composite allows for the evaluation system to assign a rating based on one combined value rather than assigning separate ratings for each effectiveness score. This is useful when the separate effectiveness scores result in conflicting ratings and a decision or outcome relies on the rating. However, it is useful to have the profile of scores available as well. When the composite results in a low rating, the profile can be used to identify the areas of weakness for improvement. This is useful in the primary school setting where a teacher is responsible for multiple subject areas. A composite allows for the teacher to get an overall rating and then the profile allows for the drill down to the specific subject area or outcome of weakness. Given the usefulness of the composite, the aggregation method potentially has a great impact on the resulting rating.

The correlations between the composites from the different models were very high, providing no evidence that the model choice was impactful. Also, the differences in bias and fit were slight. While the findings do not really support a claim for the use of the more complex multivariate model over the univariate models, the increased theoretical validity from adding outcomes to the VAM does. The trade-off between validity and complexity is not a simple concept. While the more complex model may provide more accurate or precise results, the stakeholders who consume the results may not be appreciative. Tensions are already high over the use of VAM to measure teacher effectiveness for use in high stakes decisions, such as terminations. In 2017, the Houston Independent School District won a lawsuit in which a group of teachers declared the use of the EVAAS for termination is unlawful (AFT, 2017). The teachers, and the courts agreed, claimed that the EVAAS model was flawed, unfair and incomprehensible. This follows previous lawsuits in Tennessee and Florida (Sawchuk, 2014). These lawsuits are over the use of VAM in teacher evaluation systems, regardless of the model employed. The use of a seemingly incomprehensible model such as the multivariate multilevel model would not contribute to acceptance of VAM.

The study also adds to the VAM research literature. There are limited studies on the multivariate multilevel VAM. This study informs the examination of multiple measures of teacher effectiveness. This study adds to the evidence that suggests the specifications for VAM are complex and the decisions could influence the results. Making high stakes decisions based on the results of VAM should be done so with caution, if at all.

Further research should include empirical data. While the values used in this simulation were selected from the available research literature, applying the models to real data could reveal complexities and issues not seen in the simulated data. One such issue is missing data. This study did not address missing data that is commonly found in educational data sets.

Further research should include more complex value-added models. The three-level multivariate model could be expanded to include covariates at the student-level. As discussed, the research is not conclusive on the influence of covariates on teacher effectiveness estimates. More research could compare the various model complexities and the impact that has on the estimates. The model could also be expanded to include a fourth level. A four-level model could include the outcomes, students, teachers, and schools. One could even imagine a five-level model that adds the school district or even state. Testing more complex models for the influence on the teacher effectiveness estimates and the resulting composites is valuable research for those interested in increasing the validity of the teacher effectiveness models and VAM.

More research is needed around the validity of VAM. There is limited research and guidance about which outcomes should be used to estimate teacher effectiveness. As this study highlighted, the use of non-cognitive outcomes could have a valuable place in VAM; though few states are using non-cognitive outcomes in teacher evaluation. Models with multiple outcomes could allow for the inclusion of non-cognitive outcomes without losing the achievement-based outcomes.

Bibliography

Adelman, C., Hyslop, A., Marchitello, M., O'Neal Schiess, J., & Pennington, K. (2017). *An independent review of ESSA state plans.* Washington, DC: Bellwether Education Partners. Retrieved from https://bellwethereducation.org/publication/independent-review-essa-state-plans

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.

American Federation of Teachers. (2017, October 10). *Federal suit settlement: End of value-added measures for teacher termination in Houston* [Press release]. Retrieved from https://www.aft.org/press-release/federal-suit-settlement-end-value-added-measures-teacher-termination-houston

American Statistical Association. (2014). *ASA Statement on using value-added models for educational assessment.* Retrieved from the ASA website: https://www.amstat.org/asa/files/pdfs/POL-ASAVAM-Statement.pdf

Baker, E., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., & Linn, R. L. (2010). *Problems with the use of student test scores to evaluate teachers* (Briefing Paper #278). Washington, DC: Economic Policy Institute. Retrieved from http://epi.3cdn.net/b9667271ee6c154195_t9m6iij8k.pdf

Baldwin, S., Imel, Z., Braithwaite, S., & Atkins, D. (2014). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Consulting and Clinical Psychology*, *82*, 920-930.

Blanca, M.J., Alarcon, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29,* 552-557.

Burns, W., & Clemen, R. (1993). Covariance structure models and influence diagrams. *Management Science, 39*, 816-834.

Chetty, R., Friedman, J. N., & Rockoff, J. E., (2012). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (NBER Working Paper No. 17699). Cambridge, MA: National Bureau of Economic Research.

Coleman, J., Hoffer, T., & Kilgore, S. (1982). Cognitive outcomes in public and private schools. *Sociology of Education, 55*, 65-76.

Collins, C. J., Hanges, P. J., & Locke, E. A. (2004). The relationship of achievement motivation to entrepreneurial behavior: A meta-analysis. *Human Performance*, *17*, 95-117.

Cunningham, P., Fina, A., Adams, W., & Welch, C. (2011). *Impact of end-of-course tests on accountability decisions in mathematics and science*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, New Orleans, LA.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (1997). *Doing what matters most: Investing in quality teaching*. New York: The National Commission on Teaching and America's Future.

De Maeyer, R., van den Bergh, H., Rymenans, R., Van Petegem, P., & Rijlaarsdam, G. (2010). Effectiveness criteria in school effectiveness studies: Further research on the choice for a multivariate model. *Educational Research Review, 5*, 81-96.

Donoghue, J., & Jenkins, F. (1992). *A monte carlo study of the effects of model misspecification on HLM estimates*. ETS Research Report Series, i-41. doi:10.1002/j.2333-8504.1992.tb01500.x

Doyle, D., & Han, J. G. (2012). *Measuring teacher effectiveness: A look "under the hood" of teacher evaluation in 10 sites.* New York: 50CAN; New Haven, CT: ConnCAN; and Chapel Hill, NC: Public Impact Report.

Educator Preparation Institution (EPI). (2014). *Technical manual: A guide to component and overall score calculation*. Lansing, MI: Michigan Department of Education.

Everson, K. (2017). Value-added modeling and educational accountability: are we answering the real questions? *Review of Educational Research, 87*, 35-70.

Fielding, A. & Goldstein, H. (2006). *Cross classified and multiple membership structure in multilevel models: An introduction and review* (Research Report No 791). Birmingham, AL: University of Birmingham. Retrieved from https://dera.ioe.ac.uk/6469/1/RR791.pdf

Feng, L., Figlio, D., & Sass, T. (2010). *School accountability and teacher mobility* (CALDER Working Paper No. 47). Washington DC: CALDER. Retrieved from http://www.urban.org/uploadedpdf/1001396-school-accountability.pdf

Finch, H., & French, B. (2013). A monte carlo comparison of robust MANOVA test statistics. *Journal of Modern Applied Statistical Methods, 12*, 35-81.

Fortmann-Roe, S. (2012). Understanding the bias – variance tradeoff. Retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html

Fox, L. (2016). Playing to teachers' strengths: Using multiple measures of teacher effectiveness to improve teacher assignments. *Education Finance and Policy, 11,* 70-96.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763-782). New York, NY: Springer.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://www.ncctq.org/publications/LinkBetweenTQandOutcomes.pdf

Goldhaber, D. (2010). *When the stakes are high, can we rely on value-added?* Retrieved from the Center for American Progress website: https://americanprogress.org/wp-content/uploads/issues/2010/12/pdf/vam.pdf

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *American Economic Review*, *100*, 250-255.

Goldhaber, D., & Hannaway, J. (2004). Accountability with a kicker: Preliminary observations on the Florida A+ accountability plan. *Phi Delta Kappan, 85,* 598–605.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review*, *36*, 216-228.

Goldstein, H. (1995). *Multilevel statistical models*, (2nd ed.). London: Arnold.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, *8*, 369-395.

Goldstein, H. (1999). *Multilevel statistical models*. London: Institute of Education, Multilevel Models Project.

Green, S. B., Salkind, N. J., & Akey, T. M. (2000). *Using SPSS for windows: Analyzing and understanding dData*, (2nd ed.). Upper Saddle River: Prentice Hall.

Grilli, L., Pennoni, F., Rampichini, C., & Romeo, I. (2015). Exploiting TIMSS and PIRLS combined data: Multivariate multilevel modelling of student achievement. *The Annals of Applied Statistics*, *10*, 2405-2426.

Harris, D. & Sass, T. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management*, *28*, 55-80.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320-340.

Heck, R. H. & Thomas, S. L. (2000). *An introduction to multi-level modeling techniques*. Mahwah, NJ: Erlbaum.

Hendrickson, A., Patterson, B., & Ewing, M. (2010). *Developing form assembly specifications for exams with multiple choice and constructed response items: Balancing reliability and validity concerns.* Paper presented at the Annual Conference of the National Council for Measurement in Education, Denver, CO.

Holland, D., Kraha, A., Zientek, L., Nimon, K., Fulmore, J., Johnson, U., Ponce, H., Aguiilar, M., & Henson, R. (2018). Reliability generation of the motivated strategies for learning questionnaire: A meta-analytic view of reliability estimates. *Sage Open*, *8*(3).

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Jackson, K., (2012). *Non-cognitive ability, test scores and teacher quality: Evidence from 9th grade teachers in North Carolina* (NBER Working paper 18624). Cambridge, MA: National Bureau of Economic Research.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform Chicago's REACH students. *Educational Researcher*, *44*(2), 105-116.

Khan, A., & Rayner, G. (2003). Robustness to non-normality of common tests for the many-sample location problem. *Journal of Applied Mathematics and Decision Sciences, 7,* 187-206.

Kane, M., & Case, S. (2003). *The reliability and validity of weighted composite scores*. Paper presented at the Annual Conference of the National Council for Measurement in Education, Chicago, IL.

Keesler, V., & Howe, C. (2012). *Understanding educator evaluations in Michigan: Results from year 1 implementation*. Lansing, MI: Michigan Department of Education.

Klein, A. (2019). States, districts tackle the tough work on making ESSA a reality. Education Week, December 3, 2019. Retrieved from https://www.edweek.org/ew/articles/2019/04/03/states-districts-tackle-the-tough-work-of.html

Koedel, C., & Betts, J. R. (2007). *Re-Examining the role of teacher quality in the educational production function*. (Working Paper). Columbia, MO: University of Missouri, Columbia.

Koedel, C., & Betts, J. R. (2008). Proceedings from JSM, Social Statistics Section: *Test-Score ceiling effects and value-added measures of school quality*. Alexandria, VA: American Statistical Association.

Koedel, C., & Betts, J. R. (2010). Value added to what? How a ceiling in the testing instrument influences value added estimation. *Education Finance and Policy, 5,* 54-81.

Korendijk, E., Maas, C., Moerbeek, M., & Van der Heijden, P. (2008). The influence of misspecification of the heteroscedasticity on multilevel regression parameter and standard error estimates. *Methodology, 4*, 67-72.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies* (Unpublished manuscript). California State University at Los Angeles. Retrieved from http://www.calstatela.edu/faculty/ikreft/quarterly.html

Larwin, K. (2010). Reading is fundamental in predicting math achievement in 10th graders. *International Electronic Journal of Mathematics Education*, *5*, 131-145.

Leckie, G. (2018). Avoiding bias when estimating the consistency and stability of value-added school effects using multilevel models. *Journal of Educational and Behavioral Statistics, 43,* 440-468.

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value added. *Educational Evaluation and Policy Analysis*, *34,* 109–121.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*, 101-128.

Lipscomb, S., Teh, B., Gill, B., Chiang, H., & Owens, A. (2010). *Teacher and principal value added: Research findings and implementation practices* (Final Report 06815.100). Cambridge, MA: Mathematica Policy Research.

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.

Ma, X. (2001). Stability of school academic performance across subject areas. *Journal of Educational Measurement, 38*, 1-18.

Maas, C. J. M., & Hox, J. J. (2003). The influence of violation of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, *46*, 427-440.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*, 86-92.

McCaffrey, D. F. (2012). *Do value-added methods level the playing field for teachers?* Carnegie Knowledge Network.

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.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, L., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 68-101.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The inter-temporal variability of teacher effect estimates. *Education Finance and Policy, 4,* 572-606.

McCoach, D. B. (2010). Hierarchical linear modeling. In G. R. Hancock & R. O. Mueller (Eds.) *The reviewer's guide to quantitative methods in the social sciences* (pp. 123-140). New York: Routledge.

McCoach, D. B., & Adelson, J. (2010). Dealing with dependence (part 1): Understanding the effects of clustered data. *Gifted Child Quarterly, 54,* 152-155.

Montes, G. (2012). Race to the Top, value-added models and the catholic view of education. *The Catholic Social Science Review, 17,* 337-344.

National Council of Teachers of English. (1998). *NCTE position on class size and teaching workload, K-college.* Urbana, IL: NCTE.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives, 18*(23), 1-26. Retrieved from http://epaa.asu.edu/ojs/article/view/810

No Child Left Behind Act of 2001, P.L. 107-110, 29 U.S.C. § 6319 (2002).

O'Malley, K., Murphy, S., McClarty, K., Murphy, D., & McBride, Y. (2011). *Overview of student growth models* (White paper). Iowa City, IA: Pearson.

Organization for Economic Co-operation and Development. (2008). *Handbook on constructing composite indicators.* France: Organization for Economic Co-operation and Development Publications. Retrieved from http://www.oecd.org/std/42495745.pdf

Oosterhof, A. (1987). Obtaining intended weights when combining students' scores. *NCME Instructional Module: Instructional Topics in Educational Measurement, 6*, 29-36.

Papay, J. P. (2010). Different tests, different answers: The stability of teacher value added estimates across outcome measures. *American Educational Research Journal, 48*, 163-193.

Pearson. (2014). *Stanford Achievement Test Series*, (Tenth Edition). Iowa City, IA: Pearson Assessment.

Peugh, J., & Enders, C. (2004). Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*, 525-556.

Popham, J. (1997). The moth and the flame: Student learning as a criterion of instructional competence. In J. Millman (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 264-274). Thousand Oaks, CA: Corwin.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Remler, D., & Van Ryzin, G. (2010). *Research methods in practice: Strategies for description and causation*. Thousand Oaks, CA: Sage Publications.

Rose, R., Henry, G., & Lauen, D. (2012). *Comparing value-added models for estimating teacher effectiveness* (Technical Briefing to NC DPI). Raleigh, NC: Caroline Institute for Public Policy. Retrieved from http://www.ncpublicschools.org/effectiveness-model/evaas/selection/

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy, 4*, 537-571.

Rothstein, J. (2016). *Can value-added models identify teachers' impacts?* Retrieved from The Institute for Research on Labor and Employment, UC Berkeley website: https://irle.berkeley.edu/files/2016/IRLE-Can-value-added-models-identify-teachers-impacts.pdf

Rothstein, R. (2000). Toward a composite index of school performance. *The Elementary School Journal, 100,* 409-441.

Rudner, L. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice, 20*, 16-20.

Sanders, W. L. (2006). *Comparisons among various educational assessment value-added models.* Presented at The Power of Two-National Value-Added Conference, Columbus, OH. Retrieved on July 17, 2012, from http://www.sas.com/govedu/edu/services/vaconferencepaper.pdf

SAS. (2016). *Technical documentation of EVAAS analyses* (EVOHODE-547-30SEP16). Cary, NC: The SAS Institute.

Sass, T. (2008). *The stability of value*-*added measures of teacher quality and implications for teacher compensation policy* (CALDER Brief 4). Washington, DC: The Urban Institute.

Sawchuck, S. (2014). Tenn. teachers' union takes evaluation fight into the courtroom. *Education Week, 33*(27), 8-9.

Snijders, T. A. B., & Bosker, R. J. (1992). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18* (3), 237-259.

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science, volume 3 (1570-1573)*. Chicester (etc.): Wiley, 2005.

Snijders, T. A. B., & Berkhof, J. (2008). Diagnostic checks for multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 141-175). New York: Springer.

Statistics Solutions. (2016). *Stanford Achievement Test-10 (SAT-10).* Retrieved from http://www.statisticssolutions.com/stanford-achievement-test-10-sat-10/

Struppeck. T. (2014). Combining estimates. *Casualty Actuarial Society E-Forum*, *2*, 1-14. Retrieved from https://www.casact.org/pubs/forum/14sumforumv2/Struppeck.pdf

Thum, Y. M. (1997). Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics*, *22*, 77-108.

United States Department of Education. (2009). *Race to the top executive summary.* Washington, DC: U.S. Government Printing Office. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

Villa, S. M. (2008). *Correlation between reading skills and mathematics performance: An analysis of Stanford Achievement Test scores from grades 6 to 11* (Doctoral dissertation). Retrieved from ProQuest. (1453851)

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Wu, C., Gumpertz, M., & Boos, D. (2001). Comparison of Gee, MINQUE, ML and REML estimating equations for normally distributed data. *The American Statistician*, *55*, 215-130.