Using Validity Criteria to Enable Model Selection:

An Exploratory Analysis

Mark Chin

Heather C. Hill

Dan McGinn

Harvard Graduate School of Education

Doug Staiger

Dartmouth College

Katie Buckley

Harvard Graduate School of Education

Paper presented at the Association for Public Policy Analysis and Management Annual Conference, Washington DC, November 7, 2013. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C090023 to the President and Fellows of Harvard College to support the National Center for Teacher Effectiveness. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Preliminary working paper. Please do not cite without authors' permission

Abstract

In this paper, the authors propose that an important determinant of value-added model choice should be alignment with alternative indicators of teacher and teaching quality. Such alignment makes sense from a theoretical perspective because better alignment is thought to indicate more valid systems. To provide initial evidence on this issue, they first calculated value-added scores for all fourth and fifth grade teachers within four districts, then extracted scores for 160 intensively studied teachers. Initial analyses using a subset of alternative indicators suggest that alignment between value-added scores and alternative indicators differ by model, though not significantly.

Using Validity Criteria to Enable Model Selection:

An Exploratory Analysis

In response to recently instituted federal incentives such as Race to the Top grants and waivers to the No Child Left Behind laws, many state and district school systems across the United States have adopted test-based accountability metric (TBAMs) models as part of their teacher evaluation programs. These models are intended to distinguish between effective and ineffective teachers by identifying their impacts on student learning, as measured by the achievement of their students on standardized tests.

Despite the growing political importance of TBAM models, little consensus exists around how to best specify these models. Research organizations and school districts utilize a range of options, from simple value-added models (VAMs) or student growth percentile (SGPs) models that control for just prior test scores, to VAMs that control for student demographics either at the individual or aggregate level, to VAMs that compare teachers to their colleagues within the same school via school fixed effects. That each model asserts to be measuring teacher effectiveness is problematic, as different specifications can result in different assessments of quality for the same teacher (Goldhaber & Theobald, 2012). Research on the appropriate model choice is important, as teacher evaluations based on such assessments inform key decisions about teacher professional development, compensation, and job security.

In this paper, we propose that an important determinant of model choice should be alignment of the model's assessment of teacher effectiveness to alternative indicators of teacher and teaching quality. From a theoretical standpoint, examining the degree of alignment between TBAMs and alternative indicators of quality is important, because better alignment indicates more valid systems. Further, because most school systems are also expanding the role of classroom observations and other non-test based indicators in response to public and academic criticism of TBAMs, alignment between the two sets of indicators is vital; ideally, multiple measures should converge and identify the same sets of teachers as belonging to specific actionable categories such as remediation and reward. The same argument holds true for less consequential decisions such as professional development or coaching; to the extent possible, different metrics should return the same diagnoses and prescriptions for teacher learning and improvement.

Though some research has explored the validity of value-added scores vis-à-vis other indicators of instruction (see for example, Hill, Kapitula & Umland, 2011; Milanowski, 2011), there are few studies that have looked at how the validity of these scores, as judged by their relationship to alternative indicators, changes depending on model specification. We address this issue by investigating alignment of TBAMs from four different models, each generated from both one and three years of student data, to a set of alternative indicators of teacher and teaching quality. Prior work has suggested these alternative indicators are either related to student outcomes or represent valued aspects of teachers and classrooms.

In our study, we find, that in line with expectations from the literature, scores from three-year models are generally more strongly related to alternative indicators than one-year models. Relationships between single-year value-added estimates and alternative measures of teacher effectiveness fluctuate across years, both in terms of magnitude and significance. Among the three-year models, we find that those with more school and classroom-level controls are typically less related to these alternative composites. Finally, we explore adjusting the alternative indicator for classroom composition or school composition, similar to adjustments made in certain value-added models. From this analysis we find that, though TBAMs from more controlled models more strongly relate to the adjusted alternative indicators than without adjustment, TBAMs with fewer controls continue to correlate just as well or better in the majority of cases.

In what follows, we first discuss the historic and current use of TBAMs and alternative indicators in teacher evaluation programs as measures of teacher effectiveness, and consider the implications of different model specifications in generating these scores. In the third section, we discuss the data and methodology used to investigate how altering specifications might result in differing relationships between TBAMs and other indicators. The fourth section, which communicates our findings to our initial questions, is followed by the description of results from a set of sensitivity analyses utilized to test the robustness of our main findings. The final section concludes by considering how our results might impact policy surrounding the use of TBAMs and alternative indicators in teacher evaluation programs, in addition to suggestions for further research.

Literature Review

Below, we review the history of TBAMs' development and use in teacher evaluation. We consider in particular the decisions policy-makers face regarding the choice of models, including whether and how much to control for student, classroom and school characteristics and whether to base teacher scores on one or more years of student outcome data.

Test-based Accountability Metrics and Teacher Evaluation

Though the framework for differentiating teachers based on their contributions to student gains on tests has existed since the early seventies (Hanushek, 1971), the use of test-based accountability metrics (TBAMs) in teacher evaluation programs did not begin in earnest until the late 90s (Sanders & Horn, 1998), when yearly student testing became common in many states and districts. For almost a decade, districts and states used TBAM scores mainly in performance incentive systems; in Dallas, Houston, Minneapolis, Florida and other locations, for instance, teachers could earn bonuses based on their students' performance. More recent federal polices such as Race to the Top and waivers to the No Child Left Behind act, which both typically require teacher TBAMs be included in performance evaluations, have helped popularize the use of TBAMs at scale. Political enthusiasm for the use of TBAMs in performance evaluations has often outstripped academic enthusiasm for the idea, with cautious proponents (Glazerman, Loeb, Goldhaber, Staiger, Raudenbush, & Whitehurst, 2010; Gordon, Kane, & Staiger, 2006) often appearing outnumbered by opponents (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012) as well as the sheer volume of research critical of these scores (e.g., Amrein-Beardsley, 2008; Kupermintz, 2003; Rothstein, 2009).

Yet despite the widespread prevalence and interest in TBAM models and scores, little consensus exists in both political and academic circles on how best to specify these models. One class of TBAMs takes a "value-added" modeling approach, or VAM. All VAMs, though differing in controls included, follow the same statistical framework: for a given school year, the performance on some measure of growth in student cognitive ability is associated with the student's teacher. The teacher's ultimate value-added estimate is the aggregation of the under- or over-performance, beyond what is expected given the controls, of the teacher's students on the measure, usually a standardized test.

More recently, up to 25 states have begun using estimates of teacher effectiveness derived from Student Growth Percentile (SGPs) models as a component of their teacher evaluation programs (Castellano & Ho, 2013). SGPs were initially developed to provide a student-growth-based measure for school accountability systems, particularly for systems in which the assessments are not vertically aligned across grades (Betenbenner, 2009; Betenbenner, 2010). Unlike VAM measures, SGPs are not intended to provide a causal estimate of a teacher's contribution to student achievement, but instead a descriptive measure of a student's relative growth. To do so, SGPs define how "normal" a student's current achievement is, in percentile rankings, when compared with the current achievement of all students who had similar scores in the prior year(s) (for a more formal description of the SGP model, see Castellano & Ho, 2013). SGPs can be aggregated easily to the teacher or school level, where the summary measure is the median growth percentile (MGP), or the median of every student's growth percentile in each classroom or school.

Comparing TBAMs from Different Models

Decisions about how to specify TBAM models have non-negligible impacts on how teachers are categorized and evaluated. Though TBAMs generated from different models for a single teacher are generally highly correlated across models (Goldhaber & Theobald, 2012), research has shown that the rank of a teacher's effectiveness based on TBAM models is not perfectly preserved from specification to specification. For example, Goldhaber and Theobald (2012) demonstrate that 11.4% of teachers in the bottom quintile for VA when the model controls for just student background and prior achievement move out of the bottom quintile when the model controls for school fixed effects. Similarly, 11.2% of the top VA quintile teachers from the less controlled model move out of the quintile when adding a school fixed effect control. Conflicting categorizations, particularly in these quintiles, where rewards and sanctions are often targeted, challenge the validity and utility of the conclusions reached from TBAMs. Model decisions can be grouped into two categories: decisions about whether and how to control for student, peer, and school characteristics; and decisions about whether to use a single-year or multi-year TBAM model. We review each in brief.

Controlling for Student, Peer, and School Characteristics. The logic behind controlling for student, peer and school effects relates to the comparison groups states and districts may wish to construct for teachers. By using only prior student test scores in the creation of TBAMs, as SGPs and simple VAM models do, the implicit comparison group is the entire district (or state). Teachers are compared to all others in a similar grade level of subject within their unit, with no controls for the composition of students in their classroom or for the unique characteristics of their school. This is also the approach favored by the Tennessee Value Added Assessment System, one of the original teacher accountability systems in the U.S. (Ballou, Sanders, & Wright, 2004).

Others argue that it is fairer to teachers to enter one or more controls into TBAMs. SGPs are not amenable to such adjustments, but controls for student demographics and classroom composition (peer effects) can be easily added to most VAMs. Doing so creates comparison groups for each teacher that correspond to the set of teachers situated in similarly-comprised classrooms. Many view these comparisons as more fair, as the composition of students in classrooms is thought to influence outcomes above and beyond teacher effects themselves (Hanushek, Kain, Markman & Rivkin, 2003; Uribe, Murnane & Willett, 2005). Models that control for student background characteristics and/or peer effects are widely used (see Goldhaber & Theobald, 2012). Another possibility is to control for school fixed effects, which compare teachers only to other teachers in his or her school (McCaffrey, Lockwood, Koretz, & Hamilton, 2003) by including a dummy variable that captures any school-specific influence on scores. School fixed effects models do not appear to be used in practice, but are common in the research literature (e.g., Chetty, Friedman, Hilger, Saez, & Yagan, 2011; Nye, Konstantopolous & Hedges, 2004).

As Goldhaber and Theobald (2012) note, there is no way to clearly theoretically justify more- or less-controlled models. One important issue is that including student demographics in a model implies that expectations for performance on tests differ from subgroup to subgroup; correspondingly, expectations for student growth are lower for teachers instructing more students of low-performing subgroups (Tekwe et al., 2004). This may not be acceptable to many, and has been a significant point of contention around the construction of VAMs. Second, objective measures of teacher quality (e.g., preparation, experience, mathematical knowledge) appear to correlate with student demographics, with more poorly prepared teachers more often in higher-poverty classrooms (Clotfelter, Ladd, & Vigdor, 2006; Hill, Rowan, & Ball, 2005; Loeb & Reininger, 2004). In this case, constructing comparison groups based on student demographics may inadvertently conceal poorer teacher quality, because, for instance, low-knowledge teachers are only compared to other lower-knowledge teachers in the model. In fact, it is difficult to disentangle the source of variation in student test scores attributable to differences between teachers, classrooms, or schools; by adding adjustments in TBAM models for the composition of a teacher's classroom and school, some of the actual teacher variation in effectiveness may be misassigned (McCaffrey et al., 2003).

Single- and Multi-Year TBAM Models. More recently, research on TBAMs has also explored how inclusion of multiple years of student-teacher links into the models might impact assessments of teacher effectiveness. Decisions surrounding teacher compensation and job security made from single-year evaluations are more problematic if TBAM scores vary substantially from year to year. The literature provides a basis for concern, as correlations between TBAM scores across subsequent years for a teacher are moderate, averaging around 0.50 (Aaronson, Barrow, & Sander, 2007; Ballou, 2005; Goldhaber & Hansen, 2012; Koedel & Betts, 2011; Schochet & Chiang, 2010). McCaffrey, Sass, Lockwood, and Mihaly (2009) find that 11 to 16 percent of elementary math teachers who scored in the bottom quintile one year in terms of VA placed into the top quintile the following year. Similarly, 11 to 15 percent of elementary math teachers who scored in the top quintile one year placed into the bottom quintile the next. Whether variability in TBAM scores reflects true variability in teaching quality and effects is an open question; to date, there are few studies examining the stability of observational metrics although one recent study (Polikoff, 2013) found roughly the same variability as exists in VA scores.

Because of the unstable nature of TBAM evaluations, school districts, such as the New York City Department of Education transitioned from single-year to multi-year TBAM scores, often with years nested within teachers estimated as a separate variance component. Multi-year models may be less susceptible to noise and more ably identify persistent teacher effects (Goldhaber & Hansen, 2012). Furthermore, Koedel and Betts (2011) argue that multi-year VA models can mitigate the sorting bias identified by Rothstein (2009). However, multi-year models lack feasibility for key targets of accountability policies, namely new teachers.

Alternative Indicators

Though interest in and use of TBAMs have increased substantially due to recent policy changes, these metrics represent just one type of teacher effectiveness measure. Researchers have also investigated measuring effective teachers and classrooms using non-test-based metrics, and up to 41 states have already adopted such alternative measures to supplement their evaluation systems (Hull, 2013).

One key alternative indicator is observations of practice. Observations of practice have historically been a major method for measuring teacher quality (see Peterson, 2000), and most states and districts already mandate that school principals observe teachers multiple times a year for various purposes. Though in the past, school systems developed their own evaluation systems for these observations, the federal push for Race to the Top has led many states and districts to a common set of instruments, for example, the Danielson Group's Framework for Teaching (FFT), and the National Institute for Excellence in Teaching's TAP System, whose standards for observation share many similarities to certain instructional standards of FFT (Daley & Kim, 2010). Other observational instruments have been popular in research, such as the Classroom Assessment Scoring System (CLASS) (Pianta, LaParo, & Hamre, 2007) as well as content-specific observational instruments such as the Mathematical Quality of Instruction (MQI) (Hill, Blunk, Charalambous, Lewis, Phelps, Sleep, & Ball, 2008). In general, observation instruments feature domains that include topics like classroom management and organization, the clarity and accuracy of the lesson, the depth of the material covered, and student inquiry-oriented practices.

Another often-used alternative indicator consists of measures of teacher knowledge. Such measures are primarily used for policy purposes at the time of teacher certification, with some evidence that better performance on these assessments predicts better student performance (Clotfelter et al., 2006). Researchers have also taken a strong interest in these metrics as well. For example, several groups have created measures of teachers' mathematics- specific knowledge and used scores from those instruments to predict student outcomes (Baumert et al., 2010; Hill et al., 2005; Hill et al., 2011). Others have focused on measures of what teachers know about students, for instance how students learn difficult topics in mathematics or common students misconceptions of science facts, and hypothesize that such knowledge might allow teachers to produce better student outcomes (Baumert et al., 2010; Hill, Ball & Schilling, 2008; Sadler, Sonnert, Coyle, Cook-Smith, & Miller, 2013).

Finally, the last alternative metric currently popular in the research and policy realms is students' reports of teacher and classroom traits. Ron Ferguson's TRIPOD (2012) is one such survey that aims to address this question. For TRIPOD, students score their teacher's quality of instruction on seven categories, called the "Seven C's": Care, Control, Clarify, Challenge, Captivate, Confer, and Consolidate. Together, responses to these surveys capture, in a mathematics classroom, students' perceptions of the environment teachers establish for instruction of mathematics and how the mathematics is delivered.

These alternative measures represent desirable normative qualities for effective teachers and instruction, such as teachers who are more attuned to the needs of their students, who are able to create classroom environments conducive for learning, and who can deliver rich, error free instruction. Further, some of these indicators, such as teacher knowledge and instruction, can be directly targeted for improvement through additional training or professional development.

Because researchers and policymakers are both interested in investigating and measuring teacher and teaching effectiveness through TBAMs and these alternative indicators, and because there is no clear consensus on how best to specify TBAM models, exploring how these alternative measures relate to different TBAMs may inform model choice. In theory, TBAMs and alternative indicators of quality should be aligned, in that they are both measuring underlying traits of teacher quality. Therefore a higher degree of alignment between the two sets of measures should provide a stronger measure of a teacher's overall ability (Goe & Holdheide, 2011). Further, because schools and districts are using a variety of metrics to classify teacher effectiveness, alignment is vital to allow states and districts to leverage these measures in their accountability systems to impact teacher and teaching effectiveness.

Methods

Sampling

To answer our research questions, we use a subsample of data from a larger study investigating ways to measure effective mathematics teachers and teaching in fourth- and fifth-grade classrooms across three large East coast public school districts. Our final sample included 150 teachers, for which we estimated TBAMs and alternative indicator composite scores.

Data-Analysis - TBAMs

After retrieving district administrative data, including student demographics and test scores on state mathematics and reading exams, we estimated TBAMs for all fourth- and fifth-grade teachers using student-teacher links from three years of roster data. Students were only included in the model when they met the following restrictions: the student was reliably linked to a single primary mathematics teacher in the given school year; the student was not missing any prior mathematics achievement or demographic data; the student's tested mathematics grade at time *t* followed sequence with regards to his or her tested grade at time *t-1*. Classroom-year combinations were included in the model if fewer than 50% of the class *j* in year *t* were marked with special education status; fewer than 50% of their students were missing mathematics achievement scores from time *t-1*; and, after all these restrictions, there were at least 5 students in the class *j* during year *t*.[1] Because a focus of this paper is one-year vs. three-year models, teachers in our larger dataset who had less than three classroom-year combinations ($n = 100$) were not included in the data on which this paper is based.

[1] After imposing these restrictions, 80,709 of the 96,331 original student-teacher links remained for TBAM calculation. On average, teachers in our sample taught 62 students over three years.

Four models corresponding to the empirical debates above were created using this data. Each model predicted student mathematics achievement and estimated teacher scores with a different set of methods and controls.

The first three models used the Hiearchical Linear Model (HLM) framework to account for the nesting of students within teachers. HLM provides Empirical Bayes estimates, or "shrunken" estimates, which incorporate differences in the reliability of teacher-level estimates – usually caused by differences in the number of students taught by teacher – by shrinking less reliable estimates towards the mean. Models were run within school year, district, and grade. To attain these VA scores for a single year, we took the student-weighted average of these shrunken estimates within a single year for each teacher. To attain three-year value-added scores, we similarly took the student-weighted average of these shrunken estimates across three years for each teacher.

Model 1 Simple Model: Adjusting for student prior achievement and demographics:

$$y_{ijkst} = \beta_0 + \beta_1 A_{ijkst-1} + \beta_2 X_{ijkst} + \tau_{ks} + \epsilon_{ijkst}$$

where the outcome of interest, y_{ijkst}, represents the standardized state mathematics score for student i, in classroom j, of teacher k, in school s, at time t. $A_{ijkst-1}$represents a vector of prior achievement for $ijkst$'s student, consisting of linear, quadratic, and cubic terms for student j's mathematics score at time $t-1$, an imputed reading score for student j at time $t-1$, and an indicator if the student did not take the reading exam at time $t-1$. X_{jkst} represents a vector of demographic indicators for $ijkst$'s student, including the student's modal gender and race, and free- or reduced-price lunch eligibility, special education status, and limited English proficiency at time t. From the model, we also estimate τ_{ks}, a teacher random effect, or the teacher's VA score.

Model 2 Classroom Peer Effects: Adjusting for student prior achievement, student demographics, and classroom-level aggregates of prior achievement and demographics:

$$y_{ijkst} = \beta_0 + \beta_1 A_{ijkst-1} + \beta_2 X_{ijkst} + \beta_3 C_{jkst} + \tau_{ks} + \epsilon_{ijkst}$$

where C_{jkst} is a vector of variables representing the averages of $A_{ijkst-1}$and X_{ijkst} for classroom j.

Model 3 School Fixed Effects Model: Adjusting for student prior achievement, student demographics, plus a school fixed effect:

$$y_{ijkst} = \beta_0 + \beta_1 A_{ijkst-1} + \beta_2 X_{ijkst} + \delta S_{st} + \tau_{ks} + \epsilon_{ijkst}$$

where $S_{st}$is a vector of indicator variables for a student receiving instruction in school *s*.

The fourth model used a different logic to create teacher scores:

Model 4 Student Growth Percentiles: SGPs are calculated using the open-sourced statistical software program R (Betebenner, 2010). They are estimated using quantile regression, whereby a student's growth percentile is based on his/her current test score as a function of his/her prior test scores (Betebenner, 2008). Unlike with ordinary least squares (OLS) regression, which summarizes the relationship between the outcome variable (i.e., current test score) and independent variable (i.e. a vector of prior test scores) based on the conditional mean of the outcome, quantile regression summarizes the relationship using conditional quantile lines (Koenker, 2005). A student's growth percentile based on the midpoint between the quantile lines that border the student's observed current score (Castellano & Ho, 2013). There are two additional nuances to the SGP calculation. First, SGPs use Bspline parameterization instead of linear parameterization in order to account for potential nonlinearly, heteroscedasticity, and skewness in the test score data (Betebenner, 2009). Second, the software code for calculating SGPs uncrosses the fitted quantile lines prior to estimating a student's SGP to allow for monotonicity of the fitted quantiles (see Castellano & Ho, 2013, for a thorough explanation of SGP estimation).

For our analyses, we utilized only one year of prior achievement scores. To recover a teacher-level SGP score in a given year, we took the median SGP value of all students taught by the specified teacher in a given year.

Data Analysis – Alternative Indicators

Ultimately, two alternative indicators of teacher quality were created based on seven individual metrics derived from five different instruments. We describe the five instruments and the metrics created from them, then describe the process of data reduction that resulted in the two alternative indicators.

Observation scores – CLASS & MQI. Teachers in our sample had their instruction videotaped on up to three different occasions each in two school years. Teachers selected the dates for taping, under the restrictions that they would choose lessons typical of the teacher's instruction to be taped, and that teachers would not choose to tape lessons composed mainly of student testing. Each lesson taped lasted approximately one hour, and was scored by a set of highly trained raters on the Classroom Assessment Scoring System (CLASS) and Mathematical Quality of Instruction (MQI) observation instruments.

CLASS. The CLASS is subject-matter-independent observation tool organized to capture three primary domains of interaction: emotional support, classroom organization, and instructional support. A single rater scores each code of the CLASS rubric on a scale from one to seven, for each 15-minute segment of instruction. For a more complete description of what CLASS captures, see Pianta et al. (2007).

Exploratory factor analysis suggested that scores on the individual codes of the CLASS instrument formed two primary dimensions: classroom organization (CLASS – Class Organization) and teacher emotional and instructional support (CLASS – Support). To generate teacher scores for each of these two dimensions, we estimate the following multilevel lesson-level equation:

$$CLASS_{jk} = \beta_0 + \mu_k + \epsilon_{jk}$$

where the outcome of interest, $CLASS_{jk}$, represents teacher k's Class Organization or Support score for lesson j. From the equation parameter μ_k, we estimate each teacher k's shrunken CLASS score, adjusting for differences in the reliability of estimates from teacher to teacher due to differences in total number of lessons scored.

MQI. The MQI observation instrument was developed to capture the quality of instruction on a set of mathematic-specific dimensions, including: the meaning orientation of the mathematics presented to students, the teacher's ability to work with students and mathematics, teacher errors and imprecisions, and whether students engagement in the classroom aligned with the Common Core. Each 7.5-minute segment of instruction was scored on every code of the MQI rubric by two raters. For a more complete description of the MQI, see Hill et al., 2008.

Exploratory factor analysis suggested that scores on the individual codes of the MQI instrument formed two primary dimensions: the errors and imprecision present in a teacher's instruction (MQI – Errors), and the extent to which instruction features mathematical richness and inquiry (MQI – Richness and Inquiry). To generate teacher scores for each of these two dimensions, we estimate the following multilevel lesson-level equation:

$$MQI_{jk} = \beta_0 + \mu_k + \epsilon_{jk}$$

where the outcome of interest, MQI_{jk}, represents teacher k's Errors or Richness/Inquiry score for lesson j. From the equation parameter μ_k, we estimate each teacher k's shrunken score, adjusting for differences in the reliability of estimates from teacher to teacher due to differences in total number of lessons scored.

Teacher knowledge. Teachers in our sample completed a fall survey and a spring survey distributed annually for three years. These surveys contained questions that were designed to test various conceptualizations of teacher knowledge, including teacher Mathematical Knowledge for

Teaching (MKT), teacher general mathematical knowledge (MTEL, derived from performance on items from the Massachusetts Test for Educator Licensure), and teacher knowledge of content and students (KCS).

MKT & MTEL. Each fall, teachers completed a survey that included a teacher mathematical knowledge section. That section contained a mix of MKT items and released items for the MTEL. We pooled teacher responses across years, giving us a total of 72 MKT items and 33 MTEL items.

Factor analysis of the polychoric correlations was ambiguous, even after removing 12 items with low item-rest correlations, and a 13th item that 98% of respondents answered correctly. As there was no clear factor structure dependent on item origin (MKT versus MTEL) nor item content (math knowledge specific to teaching versus common math knowledge, as judged by a blind panel of experts), we pooled all items and treated them as a single, unidimensional test of teacher mathematical knowledge. Missing responses were not penalized when a teacher skipped six or more contiguous items nearby; because MKT items contain testlets (a common stem producing several related items), we used a 1-parameter graded response model in IRTPRO.

Knowledge of content and students (KCS). Teachers in our sample were scored on a scale measuring their knowledge of the mathematics ability of their students (KCS). This measure was inspired by theories of teacher pedagogical content knowledge (PCK) and adapted from other instruments measuring knowledge of student misconceptions (KOSM, see Sadler et al., 2013). Teachers were presented with items from a low-stakes test given to their student by our project, then asked what percent of their students would answer the item correctly. To generate a score for a teacher, the teacher estimate and actual percent correct within the teacher's classroom were differenced. We then estimated the following multilevel student-level equation:

$$p_{jk} = \beta_0 + \mu_k + \epsilon_{jk}$$

where the outcome of interest, p_{jk}, is the absolute difference between teacher k's estimated percentage of his/her students answering item j on the low-stakes assessment correctly and the actual percentage of students answering item j correctly, populated for each student taught by teacher k answering item j. From the equation parameter μ_k, we estimate each teacher k's shrunken KCS score, adjusting for differences in the reliability of estimates from teacher to teacher due to differences in the total number of students answering each question.

TRIPOD student surveys. Teachers in our sample were scored on a metric derived from the responses of their students to questions on the TRIPOD survey, designed to capture students' academic and social behaviors, goals, beliefs, and feelings. Student engagement is conceptually and empirically predicted by seven, multi-item measures on the survey, covering the following domains of teacher effectiveness:

- Caring about students (nurturing productive relationships);
- Controlling behavior (promoting cooperation and peer support);
- Clarifying ideas and lessons (making success seem feasible);
- Challenging students to work hard and think hard (pressing for effort and rigor);
- Captivating students (making learning interesting and relevant);

- Conferring (eliciting students' feedback and respecting their ideas);
- Consolidating (connecting and integrating ideas to support learning).

Exploratory factor analysis suggested that Tripod scores formed a single dimension capturing all of these domains. To generate teacher scores for this measure, we estimate the following multilevel lesson-level equation:

$$TRIPOD_{jsk} = \beta_0 + \mu_k + \rho_{sk} + \epsilon_{jsk}$$

where the outcome of interest, $TRIPOD_{jsk}$, represents student s's response on item j of the Tripod survey. From the equation parameter μ_k, we estimate each teacher k's shrunken TRIPOD score, adjusting for differences in the reliability of estimates from teacher to teacher due to differences in total number of students who responded to the survey.

Alternative Composite Creation

Because correlating four models with seven alternative indicators would result in an overwhelming number of tests and potentially obscure general patterns in our data, we elected to reduce the set of alternative indicators. We began by conducting exploratory and confirmatory factor analyses for the set of study teachers with data for each of the indicators ($N = 250$). Exploratory factor analyses suggested a two-factor structure for the data, with the following dimensionality:

- *Mathematics-related measures of effectiveness*: MQI Richness/Inquiry, MQI Errors, MKT/MTEL ability, KCS
- *Student-interaction related measures of effectiveness*: CLASS Support, CLASS Classroom Organization, TRIPOD

To assess the fit of the factor structure to our data, we explored the fit of the data for these 250 teachers in a two-factor confirmatory factor analysis; the theoretical model and the estimated loadings are presented in Figure 1.

Figure 1

The fit statistics suggested that the two-factor structure was a good fit for our observed data, with the comparative fit index (CFI) = .99, the Tucker-Lewis fit index (TLI) = .98, the standardized root mean squared residual (SRMR) = .04, and the RMSEA =.02, with an upper bound = .07. Using this model, we generated for each of the 250 teachers in our sample factor scores for both latent variables using regression scoring.

Standardization Procedure

Because TBAMs are calculated within district, and because policy-makers will most likely be making high stakes decisions only for the teachers in their respective districts, we standardized each teacher's alternative composite scores and TBAMs within district for our specific sample of 150 teachers, to compare teachers only to their most direct peers on these alternative indicators

Results

We present our results in two sections. In the first, we correlate teacher scores from different TBAM models and the alternative measures of teacher quality. In the second, we conduct a sensitivity analysis in which adjust the alternative composites for a classroom- and school- level student demographics.

Relationship of Alternative Indicators to Different Teacher Scores

Figures 2 and 3 display the correlations between different model's scores with the math and interaction composite alternative indicators (for the full table of correlation coefficients and *p*-values, see Table 1 in the appendix). In each figure, the correlation of the three-year model (A) is compared to the average correlation of its respective one year models (B). Several patterns are quickly apparent.

Figure 2

Figure 3

One pattern that emerges from the figures is that TBAMs from three-year models outperform the average of their single-year counterparts in terms of correlational strength. This is true for both alternative indicators. The difference in average correlation between single-year and three-year TBAMs to the mathematics composite is 0.05, whereas the difference in average correlation to the interaction composite is 0.03. Furthermore, Table 1 shows that the relationship of single-year TBAMs and alternative indicators vary from year to year and model to model in terms of magnitude (*r*, range: 0.05-0.27) and statistical significance. If more years of classroom data have the effect of making TBAMs more stable, as many suggest (Goldhaber & Hansen, 2012; Koedel & Betts, 2011), this may help explain the stronger correlation with three-year scores.

We also see that TBAMs from models with more stringent controls tend to yield lower correlations. None of the TBAMs derived from models controlling for school fixed effects were estimated as significantly correlated with either alternative indicator. For the math composite, scores from the SGP model and simple model were most strongly correlated. For the interaction composite, simple model scores again most strongly correlated. In fact, when excluding SGPs, the order of correlational strength with regards to model specification is identical: simple, peer, then school fixed effect.

This ordering may arise because more stringent controls assign some of each teacher's impact on student achievement to other influencing factors, such as the demographic make-up of a student's peers (peer model) or the characteristics of the teacher's school (school fixed effect). If this is the case, these controls may cause the estimates of teacher TBAMs to decrease. Yet given the unequal distribution of teachers into schools, it is difficult to disentangle teacher and student characteristics with regards to impact on student achievement. If less advantaged students learn more slowly independent of teacher quality, the more controlled models may be more accurate.

Teacher and teaching effectiveness as represented by our alternative composites, however, may also be confounded with the demographic make-up of the students in a teacher's classroom or school. If this is the case, our alternative indicators may be overestimating each teacher's actual effectiveness. As a result, unadjusted alternative composites will correlate more strongly with TBAMs that also fail to control for these factors (e.g. simple VAM or SGP) compared to those that do control for them (e.g. peer or school fixed effect), because such TBAMs assign less impact on student achievement to each teacher. If adjusting alternative indicators for similar controls results in stronger correlations to more strictly controlled TBAM models than less strictly controlled models, it may then be more appropriate to use such controls for measuring teacher effectiveness from non-test-based measures, in order to arrive at more accurate scores for teachers.

We explore this possibility in the next set of sensitivity analyses. For this analysis, we focus only on 3-year TBAMs due to their consistently higher performance compared to their single-year counterparts.

Adjusting the Composite Indicators for Classroom or School Composition

It is not hard to imagine why teachers' value-added scores may be influenced by student characteristics. However, some alternative indicators may be influenced by students as well. Teachers' CLASS and TRIPOD scores, for instance, may be partially governed by the assignment of students with behavioral problems to classrooms; scores on some dimensions of the MQI, such as student reasoning and explanation, may be partially governed by students' prior experiences in and aptitude for mathematics. To adjust for this potential bias, we take two potential paths. In the first, we adjust teachers' scores for their classroom composition:

$$ALT_k = \beta_0 + \beta_1 C_k + \epsilon_k$$

Where ALT_k represents teacher k's alternative indicator score. C_k represents a vector of district-standardized classroom aggregates, including class size, student prior achievement, and student demographics, averaged to teacher k across all his or her classrooms.

However, controlling at the classroom level may overlook the wider environment in which teachers work. Thus in the second approach, we adjust teachers' scores for the school composition:

$$ALT_k = \beta_0 + \beta_1 S_k + \epsilon_k$$

Where S_k represents a vector of district-standardized school- and grade-level aggregates, including size, student prior achievement, and student demographics, averaged to teacher k. We recover from these equations ϵ_k, which represents the deviation of teacher k on the math composite as compared to the baseline β_0 and controlling for the classroom and school factors. We replace the math and interaction composite score used in the original correlations to different TBAMs with these recovered estimates.

Figure 4 and 5 display the correlations between different 3-year model scores with the alternative indicators, adjusting for both class and school composition.

Figure 4

Figure 5

From these two figures, common trends emerge. When adjusting a teacher's score on the alternative composites for either classroom- or school- composition of students, correlations to the corresponding 3-year TBAM increase. For example, adjusting the mathematics composite for classroom composition results in more similar correlations between the peer VA score, the simple VA score, and the SGP score, though this was not the case before. Similarly, adjusting either the mathematics or interaction composite for school composition results in increased correlational strength between the teacher scores from the school fixed effects model and these alternative indicators.

Despite this finding, however, we continue to see that TBAM models that control for fewer student and school characteristics display the highest correlations with the alternative indicators. With regards to the mathematics composite, the SGP and simple scores perform equally well if not better compared to the peer and school fixed effect scores, regardless of the adjustment made. Similarly, with regard to the interaction composite, the simple model score correlates most strongly, regardless of adjustment, just like our original findings.

If correlations between our composites and TBAMs that control for peer or school fixed effects did not change when measuring performance on the alternative indicator adjusting for similar controls, we would conclude that classroom- and school-level student demographics do matter and that the unadjusted non-test-based measures did not overstate teacher and teaching effectiveness. Instead, the results suggest that this may be the case, because correlations between TBAMs that control for peer or school fixed effects and adjusted composites do increase. Yet because correlations between less controlled models still outperform those of more stringently controlled models, it appears that the impact of student composition on teacher performance on non-test-based indicators is smaller than that of the impact on TBAMs.

Conclusion

Our investigation uncovered fluctuations in correlations between alternative indicators and TBAMs both across years (in single-year models) and across model specifications. This suggests model choice is important to consider for researchers trying to validate such indicators, and for district policy-makers hoping to present a consistent message to teachers from multiple components of teacher evaluations.

One lesson for researchers is to use multiple years of value-added scores when conducting validity research into classroom and TBAM indicators. One-year estimates are notably noisy for both indicators, ranging from 0.05 (math indicator, school fixed effects in 2012) to 0.27 (math indicator, 2011 SGP) with obvious implications for statistical significance. One-year TBAM estimates also tend to have lower correlations, on average, with the alternative indicators. Studies that rely on single-TBAMs may modestly mislead research consumers. It also suggests that districts using these models are more likely to provide their teachers with disparate messages regarding quality, and to also have, on average, performance evaluation metrics consisting of elements with conflicting conclusions.

Recommending a 3-year VAM model for such inquiries makes sense because there are strong reasons to believe this returns a more stable estimate of teachers' underlying propensity to improve students' scores (Goldhaber & Hansen, 2012; Koedel & Betts, 2011, Schochet & Chiang, 2010). However, the choice of TBAM model – and whether to control for classroom- or school-level student characteristics when estimating alternative indicators – is less straightforward. For a district intent on maximizing the confluence between its different indicators, choosing a simple model or, in the case of mathematics, the SGP would accomplish this aim. However, as many authors have noted (Goldhaber & Theobald, 2012; McCaffrey et al., 2003), the TBAM model that most validly represents teachers' skills is unknown.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*,*25*(1), 95-135.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational researcher*, *37*(2), 65-75.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37-65.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... & Tsai, Y. M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, *47*(1), 133-180.

Betebenner, D. W. (2008). A primer on student growth percentiles. *Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved February*, *18*, 2011.

Betebenner, D. W. (2009). Growth, Standards and Accountability. Dover, NH: National Center for the Improvement of Educational Assessment.

Betebenner, D. W. (2010). SGP: Student growth percentile and percentile growth projection/trajectory functions [R package version 0.0–6].

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, *38*(2), 190-215.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics*, *126*(4), 1593-1660.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*,*41*(4), 778-820.

Daley, G., & Kim, L. (2010). A teacher evaluation system that works. *National Institute for Excellence in Teaching. Santa Monica CA.*

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8-15.

Ferguson, R. F. (2012). Can student surveys measure teaching quality?. *Phi Delta Kappan*, *94*(3), 24-28.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). Evaluating Teachers: The Important Role of Value-Added. *Brookings Institution*.

Goe, L., & Holdheide, L. (2011). Measuring Teachers' Contributions to Student Learning Growth for Nontested Grades and Subjects. Research & Policy Brief.*National Comprehensive Center for Teacher Quality*.

Goldhaber, D., & Hansen, M. (2012). Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. *Economica*.

Goldhaber, D., & Theobald, R. (2012). Do Different Value-Added Models Tell Us the Same Things? Retrieved from http://www.carnegieknowledgenetwork.org/wp -content/uploads/2012/10/CKN_2012-10_Goldhaber.pdf.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Washington, DC: Brookings Institution.

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *The American Economic Review*,*61*(2), 280-288.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement?. *Journal of Applied Econometrics*,*18*(5), 527-544.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers’ mathematical knowledge for teaching on student achievement. *American educational research journal*, *42*(2), 371-406.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, 372-400.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430-511.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831.

Hull, J. (2013). Trends in Teacher Evaluation: How States are Measuring Teacher Performance. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance/Trends-in-Teacher-Evaluation-Full-Report-PDF.pdf.

Koedel, C., & Betts, J. R. (2011). Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy*, *6*(1), 18-42.

Koenker, R. (2005). *Quantile regression* (No. 38). Cambridge university press.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational evaluation and policy analysis*, *25*(3), 287-298.

Loeb, S., & Reininger, M. (2004). Public Policy and Teacher Labor Markets. What We Know and Why It Matters. *Education Policy Center*.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003).*Evaluating Value-Added Models for Teacher Accountability*. *Monograph*. RAND Corporation. PO Box 2138, Santa Monica, CA 90407-2138.

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

Milanowski, A. T. (2011, April). Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching. Paper presented at the annual meeting of the American Education Research Association, New Orleans, L.A. Abstract retrieved from http://eric.ed.gov/?id=ED520519.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. *Educational evaluation and policy analysis*, *26*(3), 237-257.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. Corwin-volume discounts.

Polikoff, M. S. (2013). The Stability of Observational and Student Survey Measures of Teaching Effectiveness Association for Education Finance and Policy annual conference March, 2013 New Orleans, LA.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2007). Classroom Assessment Scoring System (CLASS) Manual. Baltimore, MD: Brookes Publishing.

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

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N., & Miller, J. L. (2013). The Influence of Teachers' Knowledge on Student Learning in Middle School Physical Science Classrooms. *American Educational Research Journal*.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256.

Schochet, P. Z., & Chiang, H. S. (2010). Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains. NCEE 2010-4004.*National Center for Education Evaluation and Regional Assistance*.

Tekwe, C. D., Carter, R. L., Ma, C. X., Algina, J., Lucas, M. E., Roth, J., ... & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11-36.

Uribe, C., Murnane, R. J., Willett, J. B., & Somers, M. A. (2005). *Expanding school enrollment by subsidizing private schools: Lessons from Bogota* (No. w11670). National Bureau of Economic Research.

Appendix

Table 1

Correlation Coefficients between Teacher TBAMs and Unadjusted Alternative Indicators

Model	Math	Interaction
3-Year SGP	.25**	.14~
3-Year Simple	.24**	.17*
3-Year Peer	.21**	.15~
3-Year School Fixed Effect	.12	.11
2010 SGP	.14~	.10
2010 Simple	.15~	.11
2010 Peer	.11	.08
2010 School Fixed Effect	.10	.08
2011 SGP	.27**	.13
2011 Simple	.23**	.16*
2011 Peer	.21**	.15~
2011 School Fixed Effect	.15~	.13
2012 SGP	.19*	.10
2012 Simple	.20*	.18*
2012 Peer	.16*	.13
2012 School Fixed Effect	.05	.07

~p<.10
**p<.05*
***p<.01*

Table 2

Correlation Coefficients between Teacher TBAMs and Adjusted Alternative Indicators

	Classroom-Adjusted		School-Adjusted	
Model	Math	Interaction	Math	Interaction
3-Year SGP	.21*	.25**	.13~	.14~
3-Year Simple	.21**	.25**	.16*	.18*
3-Year Peer	.21**	.22**	.16~	.16*
3-Year School Fixed Effect	.09	.17*	.10	.17*

~p<.10
**p<.05*
***p<.01*

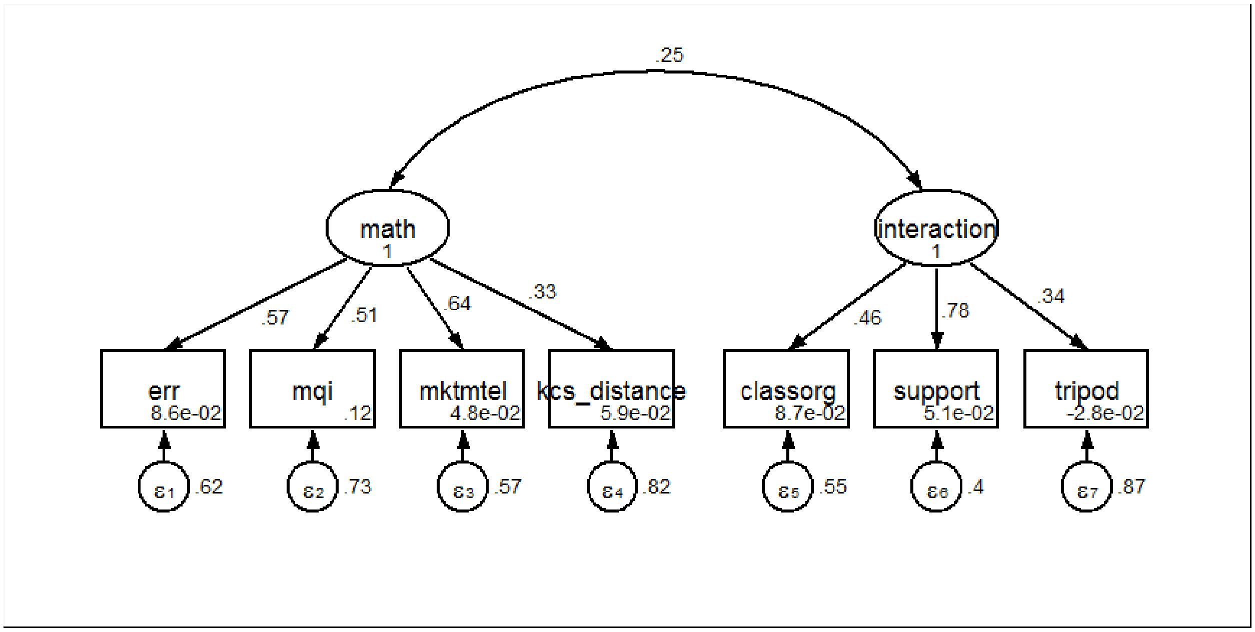

Figure 1. Factor loadings from two dimension confirmatory factor analysis of alternative indicators.

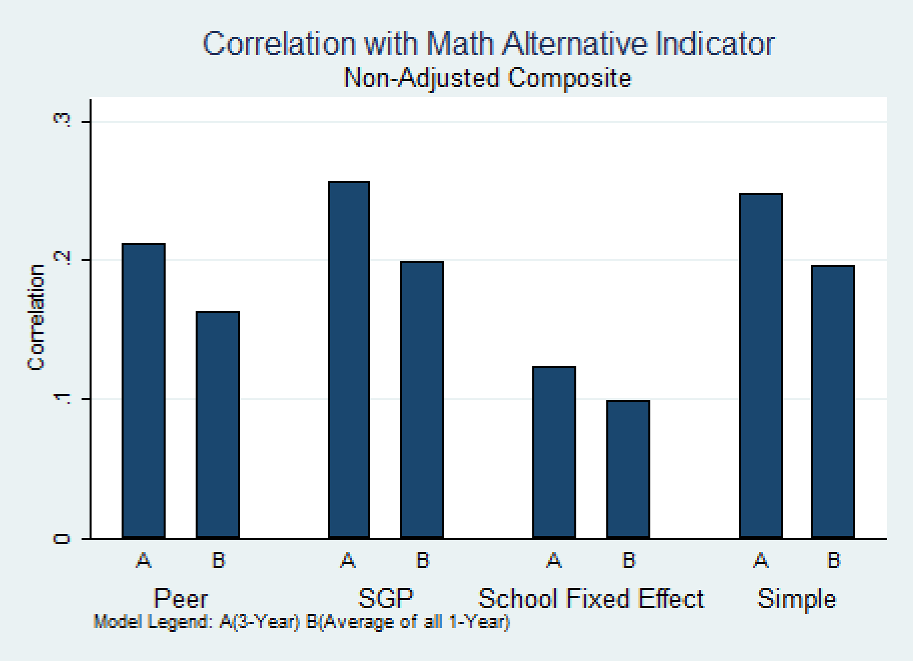

Figure 2. Correlation between different TBAMs to unadjusted mathematics alternative indicator.

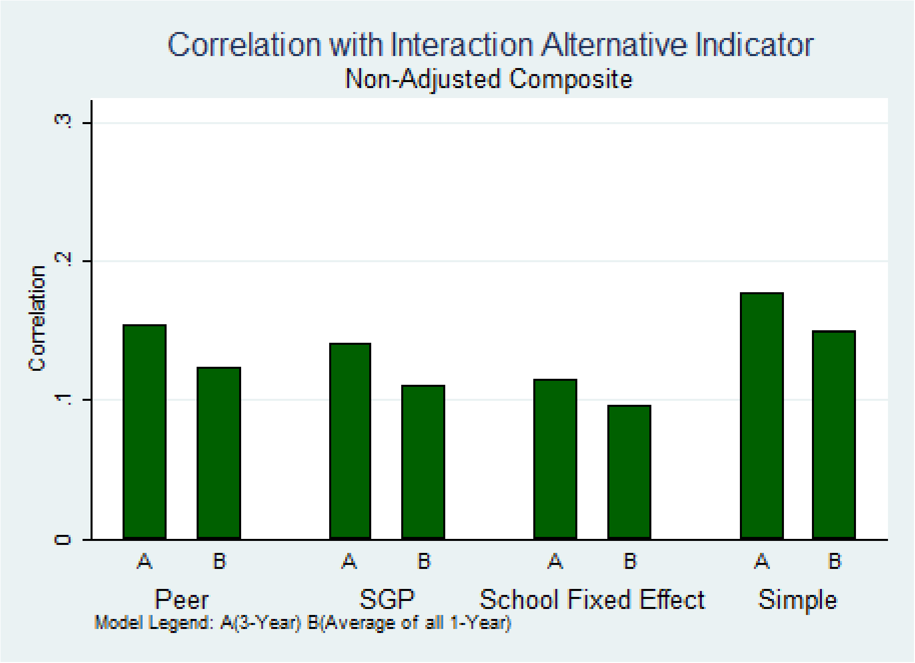

Figure 3. Correlation between different TBAMs to unadjusted interaction alternative indicator.

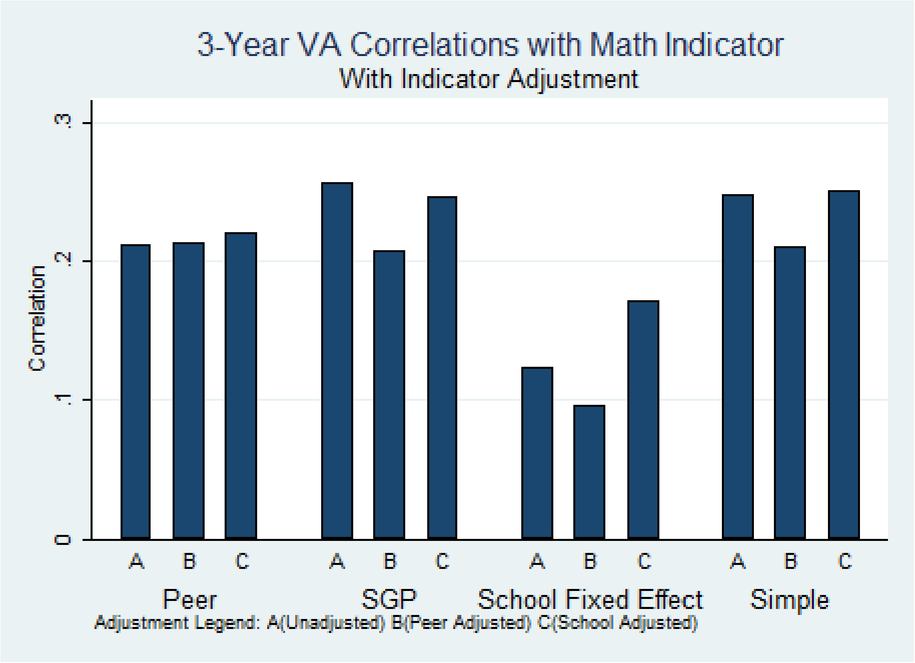

Figure 4. Correlation between different 3-Year TBAMs to adjusted mathematics alternative indicator.

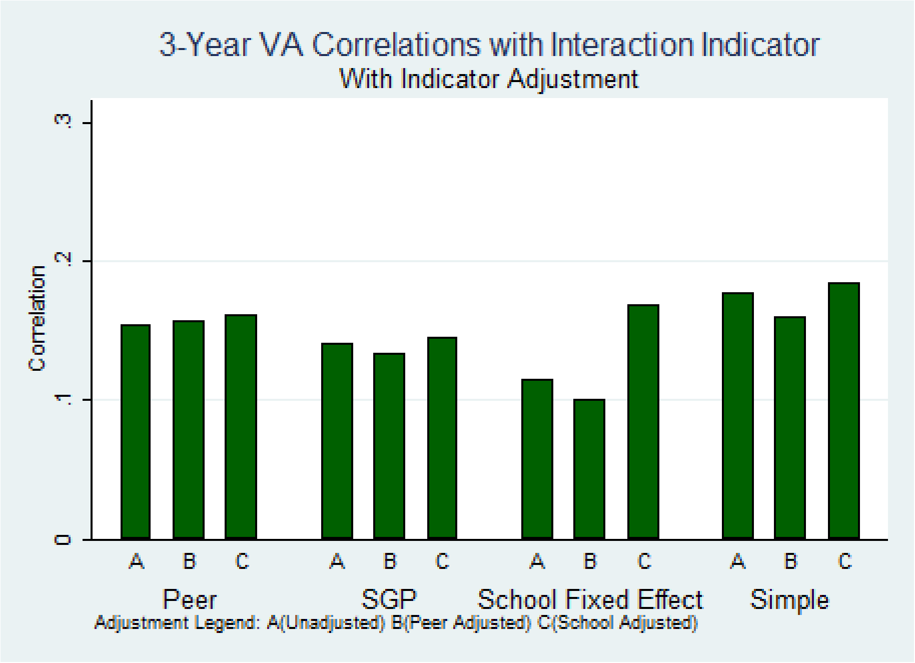

Figure 5. Correlation between different 3-Year TBAMs to adjusted interaction alternative indicator.