



# EDUCATION

THE ARTS  
CHILD POLICY  
CIVIL JUSTICE  
EDUCATION  
ENERGY AND ENVIRONMENT  
HEALTH AND HEALTH CARE  
INTERNATIONAL AFFAIRS  
NATIONAL SECURITY  
POPULATION AND AGING  
PUBLIC SAFETY  
SCIENCE AND TECHNOLOGY  
SUBSTANCE ABUSE  
TERRORISM AND  
HOMELAND SECURITY  
TRANSPORTATION AND  
INFRASTRUCTURE  
WORKFORCE AND WORKPLACE

This PDF document was made available from [www.rand.org](http://www.rand.org) as a public service of the RAND Corporation.

[Jump down to document](#) ▼

The RAND Corporation is a nonprofit research organization providing objective analysis and effective solutions that address the challenges facing the public and private sectors around the world.

## Support RAND

[Browse Books & Publications](#)

[Make a charitable contribution](#)

## For More Information

Visit RAND at [www.rand.org](http://www.rand.org)

Explore [RAND Education](#)

View [document details](#)

This product is part of the RAND Corporation reprint series. RAND reprints present previously published journal articles, book chapters, and reports with the permission of the publisher. RAND reprints have been formally reviewed in accordance with the publisher's editorial policy, and are compliant with RAND's rigorous quality assurance standards for quality and objectivity.

The Sensitivity of Value -Added Teacher Effect Estimates to Different Mathematics  
Achievement Measures

J.R. Lockwood, Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher,  
Vi-Nhuan Le and Felipe Martinez

The RAND Corporation

July 6, 2006

This material is based on work supported by the National Science Foundation under Grant No. ESI-9986612 and the Department of Education Institute of Education Sciences under Grant No. R305U040005. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations. We thank the Editor and three reviewers for feedback that greatly improved the manuscript.

The Sensitivity of Value -Added Teacher Effect Estimates to Different Mathematics  
Achievement Measures

Abstract

Using longitudinal data from a cohort of middle school students from a large school district, we estimate separate “value-added” teacher effects for two subscales of a mathematics assessment under a variety of statistical models varying in form and degree of control for student background characteristics. We find that the variation in estimated effects resulting from the different mathematics achievement measures is large relative to variation resulting from choices about model specification, and that the variation within teachers across achievement measures is larger than the variation across teachers. These results suggest that conclusions about individual teachers’ performance based on value -added models can be sensitive to the ways in which student achievement is measured.

In response to the testing and accountability requirements of No Child Left Behind (NCLB), states and districts have been expanding their testing programs and improving their data systems. These actions have resulted in increasing reliance on student test score data for educational decision-making. One of the most rapidly advancing uses of test score data is value-added modeling (VAM), which capitalizes on longitudinal data on individual students to inform decisions about the effectiveness of teachers, schools, or programs. VAM is gaining favor because of the perception that longitudinal modeling of student test score data has the potential to distinguish the effects of teachers or schools from non-schooling inputs to student achievement. As such, proponents of VAM have advocated its use for school and teacher accountability measures (Hershberg, 2005). VAM is currently being used in a number of states including Ohio, Pennsylvania and Tennessee as well as in individual school districts, and is being incorporated (as “growth models”) into federal No Child Left Behind compliance strategies (U.S. Department of Education, 2005).

However, because VAM measures rely on tests of student achievement, researchers have raised concerns about whether the nature of the construct or constructs being measured might substantially affect the estimated effects (Martineau, 2006; Schmidt, Houang & McKnight, 2005; McCaffrey, Lockwood, Koretz & Hamilton, 2003). The relative weights given to each content area or skill, and the degree to which these weights are aligned with the emphases given to those topics in teachers’ instruction, are likely to affect the degree to which test scores accurately capture the effects of the instruction provided. Prior research suggests that even when a test is designed to measure a single, broad construct such as mathematics, and even when it displays empirical unidimensionality, conclusions about relationships between achievement and student, teacher, and school factors can be sensitive to different ways of weighting or combining items (Hamilton, 1998; Kupermintz et al., 1995). These issues become even more complex in value-

added settings with the possibility of construct weights varying over time or across grade levels, opening the possibility for inferences about educator impacts to be confounded by content shifts (Hamilton, McCaffrey and Koretz, 2006; Martin eau, 2006; McCaffrey et al., 2003).

Examinations of test content and curriculum in mathematics have shown that these content shifts are substantial (Schmidt, Houang & McKnight, 2005).

If VAM measures are highly sensitive to specific properties of the achievement measures, then educators and policy makers might conclude that VAM measures are too capricious to be used fairly for accountability. On the other hand, if the measures are robust to different measures of the same broad content area, then educators and policy makers might be more confident in their use. Thus, the literature has advocated empirical evaluations of VAM measures before they become formal components of accountability systems or are used to inform high stakes decisions about teachers or students (Braun, 2005; McCaffrey, Lockwood, Koretz, Louis and Hamilton, 2004b; AERA, APA and NCME, 1999). The empirical evaluations to date have considered the sensitivity of VAM measures of teacher effects to the form of the statistical model ( Lockwood, McCaffrey, Mariano and Setodji, forthcoming; McCaffrey, Lockwood, Mariano and Setodji, 2005; Rowan, Correnti and Miller, 2002) and to whether and how student background variables are controlled (Ballou, Sanders and Wright, 2004; McCaffrey, Lockwood, Koretz, Louis and Hamilton, 2004a), but have not directly compared VAM teacher effects obtained with different measures of the same broad content area.

In this paper we consider the sensitivity of estimated VAM teacher measures to two different subscales of a single mathematics achievement assessment. We conduct the comparisons under a suite of settings obtained by varying which statistical model is used to generate the measures, and whether and how student background characteristics are controlled. This provides the three-fold benefits of ensuring that the findings are not driven by a particular

choice of statistical model, adding to the literature on the robustness of VAM teacher measures to these other factors, and permitting a direct comparison of the relative influences of these factors and the achievement measure used to generate the VAM estimates.

### **Data**

The data used for this study consist of four years of longitudinally linked student -level data from one cohort of 3387 students from one of the nation's 100 largest school districts. The students were in grade 5 in spring 1999, to which we refer as "year 0" of the study. The students progressed through grade 8 in spring 2002, and we refer to grades 6, 7 and 8 as "year 1", "year 2" and "year 3", respectively. The cohort includes not only students who were in the district for the duration of the study, but also students who migrated into or out of the district and who were in the appropriate grade(s) during the appropriate year(s) for the cohort. These data were collected as part of a larger project examining the implementation of mathematics and science reforms in three districts (Le et al., forthcoming).

*Outcome variables:* For grades 6, 7 and 8, the data contain student IRT scaled scores from the Stanford 9 mathematics assessment from levels Intermediate 3, Advanced 1 and Advanced 2 (Harcourt Brace Educational Measurement, 1997). In addition to the Total scaled scores, the data include scaled scores on two subscales, Problem Solving and Procedures, which are the basis of our investigation of the sensitivity of VAM teacher effects. Both subscales consist entirely of multiple-choice items with 30 Procedures items per grade and 48, 50 and 52 Problem Solving items for grades 6, 7 and 8, respectively. The subscales were designed to measure different aspects of mathematics achievement. Procedures items cover computation using symbolic notation, rounding, computation in context and thinking skills, whereas Problem Solving covers a broad range of more complex skills and knowledge in the areas of

measurement, estimation, problem solving strategies, number systems, patterns and functions, algebra, statistics, probability, and geometry. This subscale does not exclude calculations, but focuses on applying computational skills to problem-solving activities. The two sets of items are administered in separately timed sections.

Across forms and grades, the internal consistency reliability (KR -20) estimates from the publisher's nationally-representative norming sample are approximately 0.90 for both subscales (ranging from 0.88 to 0.91). These values are nearly as high as the estimates for the full test of approximately 0.94 across forms and grades (Harcourt Brace Educational Measurement, 1997). Also, the publisher's subscale reliabilities are consistent with those calculated from our item-level data, which are 0.93 for Problem Solving in each of years 1, 2 and 3 and 0.90, 0.89 and 0.91 for Procedures in years 1, 2 and 3, respectively.

In our data, the correlations of the Problem Solving and Procedures subscores within years within students are 0.76, 0.69 and 0.59 for years 1, 2 and 3, respectively. These correlations are somewhat lower, particular in year 3, than the values of 0.78, 0.78 and 0.79 reported for grades 6, 7 and 8 in the publisher's norming sample (Harcourt Brace Educational Measurement, 1997). The lower values in our sample could reflect the fact that the characteristics of the students in the district are markedly different than the norming sample. The students in our district are predominantly non-White, the majority participate in free and reduced-price lunch (FRL) programs, and the median Total score on the Stanford 9 mathematics assessment for the students in our sample is at about the 35<sup>th</sup> percentile of the national norming sample across years 1 to 3. Another possible explanation for the lower correlations may be the behavior of the Procedures subscores; the pairwise correlations across years within students are on the order of 0.7 for Problem Solving but only 0.6 for Procedures. That is, Procedures subscores are less highly correlated within student over time than Problem Solving subscores. In

addition, Procedures gain scores have about twice as much between-classroom variance in years 2 and 3 than the Problem Solving gain scores

*Control variables:* Our data include the following student background variables: FRL program participation, race/ethnicity (Asian, African-American, Hispanic, Native American and White), limited English proficiency status, special education status, gender, and age. Student age was used to construct an indicator of whether each student was behind his/her cohort, proxying for retention at some earlier grade. The data also include scores from grade 5 (year 0) on the mathematics and reading portions of the state-developed test designed to measure student progress toward state standards<sup>1</sup>. Both the student background variables and year 0 scores on the state tests are used as control variables for some of the value-added models.

*Teacher links:* The dataset links students to their grade 6 - 8 mathematics teachers, the key information allowing investigation of teacher-level value added measures (no teacher links are available in year 0). There are 58, 38, and 35 unique teacher links in grades 6, 7, and 8, respectively. Because teacher-student links exist only for teachers who participated in the larger study of reform implementation, the data include links for about 75% of the district's 6<sup>th</sup> grade mathematics teachers in year 1 and all but one or two of the district's 7<sup>th</sup> and 8<sup>th</sup> grade mathematics teachers in years 2 and 3, respectively. Our analyses focus on estimated teacher effects from years 2 and 3 only (estimates for year 1 teachers are not available under all models that we consider), and because the data were insufficient for estimating two teachers' effects with some models, the analyses include only the 37 year 2 and 34 year 3 teachers for whom estimates are available under all models.

*Missing data:* As is typical in longitudinal data, student achievement scores were unobserved for some students due to the exclusion of students from testing, absenteeism, and

---

<sup>1</sup> To maintain anonymity of the school district, we have withheld the identification of the state.

mobility into and out of the district. To facilitate the comparison of teacher measures made with the two alternative mathematics subtest scores, we constrained students to have either both the Problem Solving and Procedures subscores, or neither score, observed in each year. For students who had only one of the subscores reported in a given year (approximately 10% of students per year), we set that score to missing, making the student missing both subscores for that year. The result is that the longitudinal pattern of observed and missing scores for the Problem Solving and Procedures measures is identical for all students, ensuring that observed differences in teacher effects across achievement measures cannot be driven by a different sample of available student scores. The first row of Table 1 provides the tabulation of observation patterns after applying this procedure for the scores in years 1, 2 and 3 for the 3387 students. The 532 students with no observed scores in any year, predominantly transient students who were in the district for only one year of the study, were eliminated from all analyses. This leaves a total of 2855 students, most (nearly 71%) of whom do not have complete testing data.

#### TABLE 1 ABOUT HERE

About 27% of these 2855 students were missing test scores from year 0; this group is comprised primarily of students who entered the district in year 1 of the study or later. Plausible values for these test scores were imputed using a multi-stage multiple imputation procedure supporting the broader study for which these data were collected (Le et al, forthcoming). The results reported here are based on one realization of the imputed year 0 scores, so that for the purposes of this study, all students can be treated as having observed year 0 scores. We ensured that the findings reported here were not sensitive to the set of imputed year 0 scores used by re-running all analyses on a different set of imputations; the differences were negligible.

In addition to missing achievement data, some students were also missing links to teachers. Students who enter the district partway through the study are missing the teacher links

for the year(s) before they enter the district, and students who leave the district are missing teacher links for the year(s) after they leave. Also, as noted, teacher -student links are missing for students whose teachers did not participate in the study of reform implementation. The patterns of observed and missing teacher links are provided in the second row of Table 1. The methods for handling both missing achievement data from years 1 to 3 and missing links are discussed in the Appendix.

### Study Design

The primary comparison of the paper involves value -added measures obtained from the Procedures and Problem Solving subscores of the Stanford 9 mathematics assessment (the relationships of estimates based on the subscores to those based on the total scores are addressed in the Discussion section). As noted, we performed the comparison across settings varying with respect to the basic form of the value added model and the degree of control for student background characteristics. In this section we describe the four basic forms of value added model and the five different configurations of controls for student background characteristics that we considered.

*Form of value-added model* (“MODEL”; 4 levels): The general term “value -added” encompasses a variety of statistical models that can be used to estimate inputs to student progress, ranging from simple models of year -to-year gains, to more complex multivariate approaches that treat the entire longitudinal performance profile as the outcome. McCaffrey et al. (2004a) provide a typology of the most prominent models and demonstrate similarities and differences among them. Here we consider four models, listed roughly in order of increasing generality, that cover the most commonly -employed structures:

- *Gain score model*: considers achievement measures from two adjacent years (e.g.

6<sup>th</sup> and 7<sup>th</sup> grade or 7<sup>th</sup> and 8<sup>th</sup> grade), and uses as the outcome the gain in achievement from one year to the next;

- *Covariate adjustment model*: also considers two adjacent years, but regresses the achievement measure from the second year on that from the first;
- *Complete persistence model*: is a fully multivariate model specifying the three - year trajectory of achievement measures as a function of current and past teacher effects, and assumes that past teacher effects persist undiminished into future years;
- *Variable persistence model*: is equivalent to the complete persistence except that the data are used to inform the degree of persistence of past teacher effects into future years.

*Controls for student background variables* (“CONTROLS”; 5 levels): The goal of VAM is to distinguish educational inputs from non -schooling inputs to student achievement. However, there is considerable debate about whether or not statistical modeling with test score data alone is sufficient to achieve this goal, or whether models that explicitly account for student background variables are required to remove the effects of non -schooling inputs from estimated teacher effects. In applications, models have ranged from those with no controls for student background variables (Sanders, Saxton and Horn, 1997) to models that include extensive controls for such variables (Webster and Mendro, 1997). In this study we consider five different configurations of controls:

- *None*: includes no controls for student background variables;
- *Demographics*: includes all individual-level demographic information (e.g. FRL participation, race/ethnicity, etc. listed previously);
- *Scores*: includes individual-level year 0 test scores;

- *Both*: includes both individual-level demographics and year 0 test scores;
- *Aggregates*: includes three teacher-level aggregates of student characteristics (percentage of students participating in the FRL program, the total percentage of African-American and Hispanic students, and the average year 0 math score)

The consideration of the aggregate variables addresses a specific concern about the impact of contextual factors on estimated teacher effects (McCaffrey et al., 2004a; Ballou, Sanders and Wright, 2004; Ballou, 2005). Additional details on the model and covariate specifications are provided in the Appendix.

For each of the 20 cells defined by the full crossing of these two factors (MODEL and CONTROLS), each teacher receives one estimated VAM measure based on the Procedures achievement outcomes and one based on the Problem Solving achievement outcomes, for a total of 40 estimated effects per teacher. Because the gain score and covariate adjustment models provide estimated teacher effects for only year 2 and year 3 teachers, we consider the estimated effects for only these teachers in our comparisons.

A final clarification is that the student records available for the gain score and covariate adjustment models are a subset of those available for the multivariate models because the former require observed scores in adjacent pairs of years and observed teacher links in the second year of each pair, while the latter can handle arbitrary patterns of observed and missing scores as well as missing teacher links. All 2855 students with at least one observed score were used for the multivariate models, while 1155 and 1104 students were used for the gain score and covariate adjustment models for years 2 and 3, respectively. We examined the results using only the subset of students who had scores available in all three years and teacher links available in years 2 and 3, which ensures that all models use precisely the same students. The findings from this restricted analysis were nearly identical to those presented here.

## Results

Consistent with the descriptive information provided in the Data section, the data provide evidence of score variation at the teacher level, and this share of the variance varies notably across the two outcomes. For the Problem Solving scores, the estimated teacher value-added variance components (see the Appendix) account for about 5% of the total year 2 variance and about 7% of the total year 3 variance, averaging across all levels of MODEL and CONTROLS. The analogous percentages for the Procedures scores are 13% for year 2 and 27% for year 3, indicating that Procedures scores exhibit stronger variation among teachers than do the Problem Solving scores. These values for the teacher's share of the total variance in scores are consistent with, and for the Procedures scores go somewhat beyond, those reported in other settings (Rowan, Correnti, and Miller, 2002; McCaffrey et al., 2004a; Nye, Konstantopoulos, and Hedges, 2004).

In addition to having different variation, the teacher effects from the two outcomes are only weakly correlated. Table 2 presents the correlations between the estimates from the two different outcomes, holding the levels of MODEL and CONTROLS constant. The rows indicate the model and the columns indicate the covariate configuration used with both outcomes to estimate the effects. For example, in the rows labeled "Gain Score," the column labeled "None" contains the correlation between estimated teacher effects based on the Problem Solving score from the gain score model without controls and the estimated effects based on the Procedures score under the same conditions. These correlations are uniformly low, with a maximum value of 0.46 in year 2 and 0.27 in year 3. The correlations are particularly low when the models include aggregate covariates. In year 3 the estimates from these models fit to the two outcomes are essentially uncorrelated ranging from .01 to .11 depending on the model. The Spearman rank correlations (not shown) are also low, averaging only about 0.06 larger than the Pearson

correlations in the table. Thus the two achievement outcomes lead to distinctly different estimates of teacher effects.

TABLE 2 ABOUT HERE

However, the story is quite different when we compare the value -added estimates for the same achievement outcome, but based on different models or degrees of control for student covariates. In these cases correlations of the teacher effects are generally high. For each year and outcome we calculated the (20 x 20) correlation matrix of the estimated teacher effects across the levels of MODEL and CONTROLS, containing 190 unique pairwise correlations for each year and outcome. These 190 correlations can be broken into three categories: 40 are for a given MODEL with different levels of CONTROLS, 30 are for different MODELS with a given level of CONTROLS, and the remaining 120 are from design points varying on both MODEL and CONTROLS. For each year and outcome, Table 3 summarizes these correlations by category. The full correlation matrices are available from the authors upon request.

As indicated by the final column of Table 3, the average correlation when MODEL is held fixed and the level of CONTROLS is varied ranges from 0.92 to 0.98 across years and outcomes. Based on the full suite of correlations (not shown), the correlations were generally highest among the levels of CONTROLS that include only student -level variables. Each of the minimum correlations in Table 3 (first column) when CONTROLS are varied is obtained for a model with controls for teacher-level aggregates compared to the same model with one of the student-level control settings. This indicates a greater sensitivity of the estimates to the inclusion of aggregate-level covariates compared to individual-level covariates, but the high average correlations indicate a general robustness to both types of controls.

The estimates are slightly more sensitive to different levels of MODEL than to different levels of CONTROLS, but are still quite robust. The average correlation when MODEL is

varied and the level of CONTROLS is held fixed ranges from 0.87 to 0.92 across years and outcomes. Certain pairs of models tend to show more consistent differences; for example, each of the minimum correlations in Table 3 when MODEL is varied for fixed CONTROLS occurs for the variable persistence model compared to the gain score model. As is to be expected, the correlations when both MODEL and CONTROLS differ are generally lower than those obtained when one factor is held constant, but even then the average correlations substantially exceed 0.8.

Overall, the sensitivity of the estimates to MODEL and CONTROLS is only slight compared to their sensitivity to the achievement outcome. The smallest of any of the 760 (=190 x 2 outcomes x 2 years) correlations related to changing MODEL or CONTROLS is 0.49 (first column of Table 3), which is larger than the largest correlation between teacher effects from the Procedures and Problem Solving outcomes (0.46 from Table 2) under any of the combinations of MODEL and CONTROLS.

#### TABLE 3 ABOUT HERE

Table 4 further quantifies the strong influence of the achievement outcome on estimated teacher effects relative to MODEL and CONTROLS. The table provides analysis of variance (ANOVA) decompositions of the variability of the 1480 teacher effect estimates from year 2 (37 teachers times 40 estimated effects per teacher), and for the 1360 teacher effect estimates from year 3 (34 teachers times 40 estimated effects per teacher). Terms included in the decomposition are variability due to teachers and to the interactions between teachers and each of the factors. There are no main effects for the factors because estimated effects were pre-centered to have mean zero by design cell.<sup>2</sup>

---

<sup>2</sup> For the gain score and covariate adjustment models, the estimated effects for a given year have mean zero. For the multivariate models, the estimated teacher effects for the teachers have nonzero means that depend on design cell. This results from a complex interplay of the methods used to deal with missing teacher links and the fact that students missing teacher links are generally lower scoring. This variation in mean effect across cells is nuisance for the desired comparisons of this study, and thus for each design cell using the multivariate model, the teacher effects were centered to have mean zero.

As shown in the table, including teachers and the interaction of teachers with each of the factors in the design accounts for most of the observed variance in the estimated teacher effects ( $R^2 = 0.97$  for year 2 and  $0.96$  for year 3). However, teachers and their interaction with outcome account for the majority of this explained variability ( $R^2 = 0.89$  for year 2 and  $0.89$  for year 3), corroborating the correlation findings that MODEL and CONTROLS have relatively little impact on estimated teacher effects. While teachers have the highest mean square for both years, part of this observed variation among teacher means is due to the contributions of the other factors. The variance component estimates (final column of Table 4) separate these alternative sources of variance. For both years, and particularly for year 3, the largest variance component is for the teacher by outcome interaction, which is substantially larger than even the main effect for teachers. This indicates that in these data, the variation across achievement outcomes within teachers is larger than the overall variation among teachers.

TABLE 4 ABOUT HERE

### **Discussion**

In response to the pressing need to empirically study the validity of VAM measures of teacher effects for educational decision-making and accountability, this study examined the sensitivity of estimated teacher effects to different subscales of a mathematics assessment. Across a range of model specifications, estimated VAM teacher effects were extremely sensitive to the achievement outcome used to create them. The variation resulting from the achievement outcome was substantially larger than that due to either model form or degree of control for student covariates, factors that have been raised in the literature as potentially influential. And the variation within teachers across outcomes was substantially larger than the variation among teachers.

Our results provide a clear example that caution is needed when interpreting estimated teacher effects because there is the potential for teacher performance to depend on the skills that are measured by the achievement tests. Although our findings are consistent with the warnings about the potential sensitivity of value-added estimates to properties of the achievement measures (Martineau, 2006; Schmidt, Houang & McKnight, 2005), we must be careful not to over-interpret results from a single dataset examining about 70 teachers on a single set of tests. The subscales behave somewhat differently in our data than in the national norming sample, and the lower student-level correlations between the subscale scores, particularly at grade 8, could be strongly related to our findings about the sensitivity of estimated teacher effects. The low student-level correlations and the lack of correspondence of the teacher effects from the subscores both could result from two distinctly different scenarios: 1) one or both of the subscales is behaving poorly in our data, so that subscores at any level of aggregation show low correlation; or 2) real phenomena at the classroom level are differentially affecting the two subscales. While we cannot definitively establish which scenario is closer to the truth, the fact that our estimated subscale reliabilities are consistent with the reasonably high values reported by the publisher suggests that differential classroom or teacher effects on the subscales in our dataset are more likely to be a source of the low marginal correlations rather than a symptom. However, regardless of the true nature of the relationship, the differences we find in our sample relative to the norming sample could indicate that our results might not generalize to other contexts.

On the other hand, our district is similar to many large urban districts seeking innovative ways to improve student outcomes. It seems plausible that local conditions (in terms of student populations, curriculum characteristics, instructional practices, assessment properties, or other policies), like those that may have led to the low correlation between subscales and the resulting

teacher effects in this district, could exist in any given district. If this school district were to use Procedures scores to evaluate its middle school mathematics teachers, it would come to conclusions that were substantially different than evaluations based on Problem Solving scores. Although these two outcomes are intended to measure different constructs within the broader domain of mathematics, they are from the same testing program and use the same multiple-choice format. The use of other measures of middle school mathematics achievement might reveal an even greater sensitivity of teacher effects to choice of outcome, particularly if the format is varied to include open-ended measures.

In practice, it is unlikely that separate teacher effects would be estimated from the Procedures and Problem Solving outcomes, or more generally from subscores intended to capture performance on different constructs. This would require groups of items forming subscales to be explicitly identified each year, subscale scores to be computed and reported, and separate value-added measures to be computed and reported for the subscales. While such detailed information could be a valuable part of growing efforts to use student test score data to improve educational decisionmaking, it is more plausible (and more consistent with existing practice such as the Tennessee Value Added Assessment System (Sanders, Saxton, and Horn, 1997) and Florida's E-Comp bonus plan (<http://www.floridaecomp.com>)) that value-added measures for a particular subject would be based on a single assessment that measures a number of constructs within the relevant domain. For example, the Stanford 9 Total mathematics score is based on a combination of the performance on the Procedures and Problem Solving subscales, and most mathematics achievement tests that would be used in a value-added context address both procedures and problem solving even if groups of items forming the subscales are not explicitly identified and separately scored.

The results of this study indicate that value-added teacher effect estimates calculated from

total scores may be sensitive to the relative contributions of each construct to the total scores. To explore this issue further, we used the Procedures and Problem Solving scores to estimate teacher effects based on hypothetical aggregate outcomes that weight the two subscales differently. In particular, we used the Procedures and Problem Solving score data to create aggregate outcomes of the form  $\alpha$ Procedures + (1- $\alpha$ )Problem Solving for values of  $\alpha$  ranging from 0 to 1 in increments of 0.2.  $\alpha=0$  corresponds to the Problem Solving outcome and  $\alpha=1$  to the Procedures outcome, while intermediate values correspond to unequally weighted combinations of the two subscales. We then estimated teacher effects using each of the resulting six hypothetical outcomes, using the complete persistence model and including controls for student demographics and year 0 scores.

The analysis shows that inferences about teacher effects can be sensitive to the content mix of the test. Figure 1 plots the VAM measures estimated for the 6 hypothetical outcomes for each teacher connected by a light gray line, with year 2 teachers in the top frame and year 3 teachers in the bottom frame. Black dots indicate effects that are detectably different from the average effect and gray dots indicate effects that are not. There is a large amount of crossing of the lines for teachers, indicating that differentially weighting the subscales changes the ordering of the teacher effects and their statistical significance. The spread widens as  $\alpha$  approaches 1, reflecting the larger variation in teacher effects for Procedures subscores. Importantly, the composite scores with  $\alpha=0.4$  correlate greater than 0.99 with the Stanford 9 Total scaled scores each year, so that this analysis effectively includes a comparison of the subscale-specific estimates to those based on the Total score as a special case. As shown in Table 5, inferences remain constant for about 62% of year 2 teachers and 38% of year 3 teachers; for the remaining teachers the classification of the teacher effect is sensitive to the weighting of the subscores. Moreover, the substantial majority of the consistent effects are those that are not detectably

different from zero. Restricting attention to only those effects that are ever classified as detectably different, only 26% for year 2 and 16% for year 3 have classifications insensitive to the weighting.

#### TABLE 5 ABOUT HERE

Interpreting the differences arising from the alternative subscores is difficult. It could mean that teachers are differentially effective at helping students develop skills and knowledge across the two subdomains. If so, valid use of information from VAM systems might require the estimation of multiple teacher effects rather than a single summative measure. Teacher effect estimates broken down by subdomain would provide the additional benefit of more fine-grained diagnostic feedback that could be used for targeted professional development or instructional interventions. The differences across subscores might also indicate the impacts of other factors at the classroom level that are not accounted for in the models. For example, if students are taking different courses and the curriculum varies across courses this could result in students with different levels of achievement growth on the two subscales. Whether this difference is truly a curriculum difference or teacher by curriculum difference would be hard to determine and it would make the interpretation of effects challenging.

Although this study did not examine subjects other than mathematics, prior research has shown that inferences about factors that affect achievement in both math and science are dependent on features of the outcome measure (Hamilton 2004), and there is no reason to believe that other subjects would be immune to this problem. The findings of this study are particularly important given the high-stakes nature of many state and district testing systems and the likelihood that value-added methods will be incorporated into those systems in the future. To the extent that these findings are indicative of what might occur in other settings, they raise concerns about the generalizability of inferences drawn from estimated teacher effects.

In contrast to the sensitivity of estimated teacher effects to the mathematics subscores, the relative insensitivity to model form and degree of control for student covariates is encouraging. The finding of robustness to differing controls for student -level covariates corroborates the findings of Ballou, Sanders and Wright (2004) and McCaffrey et al. (2004a), and is particularly encouraging because in our data there is substantial variation across teachers in the characteristics of the students that they teach. For example, in each of years 1 to 3, the percentage of African American students linked to each teacher ranges from less than 10% to more than 80%, and the percentage of students participating in free lunch programs covers nearly the full range from 0% to 100%. Thus, the robustness found here suggests that value -added methods are living up to their promise of removing the effects student background variables that are beyond the control of the teachers whose effects we are interested in estimating. On the other hand, while the covariates are notably related to levels of student achievement, they explain only a tiny fraction of the variation of year -to-year gain scores ( $R^2$  on the order of 0.05). It is thus not entirely surprising that these covariates do not strongly influence the value -added estimates in this study, and this robustness may not necessarily hold in other contexts or with other less coarse student-level information. On balance, however, this study adds to a growing body of empirical evidence that suggests that value -added methods are robust to omitted student variables in some contexts.

The situation with the aggregate student variables, as well as model form, requires somewhat more cautious optimism. Generally, estimates are highly correlated across model forms, which cover a representative set of model specifications, and across models with and without aggregate student variables. But the correlations are not uniformly high (recall that the minimum correlation from Table 3 is 0.49). There are particular combinations of model forms, most notably the gain score model versus the variable persistence model, in which real

differences are evident. This is similar to the findings of Lockwood et al. (forthcoming), who report that in an empirical study using elementary school data, the complete persistence model and the variable persistence model agreed on whether or not teacher effects were above, below, or not detectably different from the average for only about 2/3 of the teachers. Thus, although the influences of model form and the inclusion of aggregate variables are small in this study, issues surrounding model specification still require careful consideration.

A number of unanswered questions remain, and this analysis suggests several steps that should be taken in future investigations of the effects of student achievement measure on value-added estimates. First, as noted this study was limited to a single test, a single subject, and a limited grade span, and used only multiple-choice items. The effects of varying the outcome measure in other contexts (e.g., elementary reading), and of expanding the range of measures to include additional test publishers or a broader range of item formats, should be explored. Second, this study relies on predefined subscales, but other research shows that subscales based on test specifications might not capture all of the important distinctions among items within tests. For example, one important factor appears to be the extent to which the test content was included in school curricula: a science subscale created from items closely resembling problems presented in textbooks showed greater sensitivity to school and classroom factors than a subscale that emphasized material that was less well-aligned with school curricula (Hamilton 1998). The use of existing test specifications on the test used in this study would not have revealed this difference. Although this finding is unsurprising, it has important implications for high-stakes uses of information from tests. Given our finding of sensitivity of value-added estimates to the subscales, empirical analysis to examine alternative ways of creating subscales could be especially informative in the context of value-added modeling of teacher effects.

In short, the results of this study suggest that conclusions about whether a teacher has

improved student achievement in mathematics can be relatively insensitive to assumptions about model form or the inclusion of student controls, compared to differences in how mathematics achievement is measured. This study shows that even subscales of the same test, by the same test developer, can yield different results, as can different weighting among subscales in a composite score. Although the specific findings from this district might not be replicated in other contexts, they provide evidence that inferences based on VAM can, at least in some cases, be affected by the characteristics of the outcome measure. Additional research is needed to understand how our findings would be affected by changes in the student or teacher population or in the outcome measure used, but these findings do suggest reason for caution. Users of VAM must resist the temptation to interpret estimates as pure, stable measures of teacher effectiveness. Application of VAM, particularly for high-stakes purposes, should be accompanied by an examination of both the test and its alignment with the desired curriculum and instructional approach. And to the extent possible, analyses should explore the sensitivity of the estimates to different ways of combining information from test items.

## References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing*. APA, Washington DC.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value -added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 21, 37-66.

Ballou, D. (2005, April) *Distinguishing Teachers from Classroom Context*. Paper presented at the annual conference of the American Education Research Association, Montreal, Canada.

Braun, H. (2005). Value -added modeling: What does due diligence require. In R. Lissitz (Ed.), *Value added models in education: Theory and practice* (pp. 19-38). Maple Grove, MN: JAM Press.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Goldstein, H. (2003). *Multilevel Statistical Models* (3rd ed.). London: Edward Arnold.

Hamilton, L.S. (2004). Improving inferences about student achievement: A multi -dimensional perspective. In S. J. Paik (Ed.), *Advancing educational productivity: Policy implications from national databases* (pp. 175-201). Greenwich, CT: Information Age Publishing.

Hamilton, L.S. (1998). Gender differences on high school science achievement tests: Do format and content matter? *Educational Evaluation and Policy Analysis*, 20, 179-195.

Hamilton, L.S., McCaffrey, D.F., & Koretz, D.M. (2006). Validating achievement gains in cohort-to-cohort and individual growth-based modeling contexts. In R. Lissitz (Ed.), *Longitudinal and value added modeling of student performance*. Maple Grove, MN: JAM Press.

Harcourt Brace Educational Measurement (1997). *Stanford Achievement Test Series (9<sup>th</sup> Edition): Technical Data Report*. San Antonio: Harcourt Brace and Company.

Hershberg, T. (2005, February). Value -added assessment and systemic reform: A response to America's human capital development challenge. Paper prepared for the Aspen Institute's Congressional Institute, Cancun, Mexico. Available at <http://www.cgp.upenn.edu/pdf/aspen.pdf> (retrieved 1/12/06).

Kupermintz, H., Ennis, M.M., Hamilton, L.S., Talbert, J.E., & Snow, R.E. (1995). Enhancing the validity and usefulness of large -scale educational assessments: I. NELS:88 mathematics achievement. *American Educational Research Journal*, 32, 525-554.

Le, V., Stecher, B., Lockwood, J.,R., Hamilton, L.S., Robyn, A., Williams, V., Ryan, G., Kerr, K., Martinez, F., & Klein, S. (forthcoming). *Improving Mathematics and Science Education: A Longitudinal Investigation of the Relationship between Reform-Oriented Instruction and Student Achievement*. MG-480-EDU. Santa Monica, CA: RAND.

Little, R. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2<sup>nd</sup> ed.). New York: John Wiley & Sons.

Lockwood, J.R., McCaffrey, D.F., Mariano, L.T. & Setodji, C. (forthcoming). Bayesian methods for scalable multivariate value -added assessment. *Journal of Educational and Behavioral Statistics*.

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

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability*, MG-158-EDU. Santa Monica, CA: RAND.

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., Louis, T.A., & Hamilton, L.S. (2004a). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29, 67-101.

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., Louis, T.A., & Hamilton, L.S. (2004b). Let's see more empirical studies of value -added models of teacher effects: A reply to Raudenbush, Rubin, Stuart and Zanuto. *Journal of Educational and Behavioral Statistics*, 29, 139-144.

McCaffrey, D.F., Lockwood, J.R., Mariano, L.T. & Setodji, C. (2005). Challenges for value -

- added assessment of teacher effects. In R. Lissitz (Ed.), *Value added models in education: Theory and practice* (pp. 272-297). Maple Grove, MN: JAM Press.
- Nye, B., Konstantopoulos, S. & Hedges, L. V. (2004). How large are teacher effects?, *Educational Evaluation and Policy Analysis*, 26, 237-257.
- Pinheiro, J.C. & Bates, D.M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.
- Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2<sup>nd</sup> Edition). Thousand Oaks: Sage Publications.
- R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Robinson, G.K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, 6, 15-51.
- Rowan, B., Correnti, R., & Miller, R.J. (2002). What large -scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, 104, 1525-1567.
- Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee Value -Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment. In J. Millman (Ed.),

*Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137- 162). Thousand Oaks, CA: Corwin Press, Inc.

Schmidt, W.H., Houang, R.T., & McKnight, C.C. (2005). Value -added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value added models in education: Theory and practice* (pp. 272-297). Maple Grove, MN: JAM Press.

U.S. Department of Education (2005). *Secretary Spellings announces growth model pilot, addresses Chief State School Officers' Annual Policy Forum in Richmond* (press release, November 18, 2005). Retrieved November 25, 2005 from <http://www.ed.gov/news/pressreleases/2005/11/11182005.html>.

Webster, W. & Mendro, R. (1997). The Dallas Value -Added Accountability System. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?*, (pp. 81-99). Thousand Oaks, CA: Corwin Press, Inc.

### Appendix: Model Specification and Estimation

In this appendix we describe the statistical models we used to create VAM teacher measures. As described in the text, we considered four different models: two that examine gains across adjacent pairs of years (gain score model and covariate adjustment model) and two multivariate models for the three-year trajectories of scores (complete persistence model and variable persistence model). We keep the descriptions of the models brief because all of the models we consider have been discussed in detail elsewhere in the literature (see, e.g., McCaffrey et al., 2003, 2004a). We provide more specific references where appropriate.

We use the following notation in describing the models. We let  $Y_{it}$  be an achievement score for student  $i$  in year  $t$ ; this is used generally to refer to either the Problem Solving or Procedures score depending on which achievement measure is being modeled. We let  $\mathbf{DEMO}_i$  denote the vector of student-level demographic variables for student  $i$ , which as noted in the text include free or reduced priced lunch program participation, race/ethnicity, limited English proficiency status, special education status, gender, and an indicator of being behind cohort. We let  $\mathbf{SCORE0}_i$  denote a vector consisting of the mathematics and reading test scores from year 0 for student  $i$ , as well as the squares of those scores, to account for a slightly non-linear relationship between their scales and those of the Stanford 9. Teachers are indexed by  $j$  ranging from 1 to the total number of teachers in all years. We let  $\mathbf{TAGG}_j$  denote the vector of aggregates of some of the student-level variables for the students linked to teacher  $j$ . The aggregates include the percentage of students participating in the FRL program, the total percentage of African-American and Hispanic students, and the average year 0 math score. In order to accommodate student-teacher linkages in both the year-to-year models and the more complex multivariate models with consistent notation, the subscript  $j(i,t)$  is used to denote the

index  $j$  of the teacher to whom student  $i$  is linked in year  $t$ .

*Year-to-Year Modeling of Gains:* This approach involves fitting one model to estimate teacher effects in year 2 using the gains students make in seventh grade, and then fitting a separate model to estimate teacher effects in year 3 using the gains students make in eighth grade. For each year  $t = 2, 3$  and for each of the Procedures and Problem Solving achievement measures, we fit the following hierarchical linear model:

$$Y_{it} = \mu_t + \lambda_t Y_{it-1} + \beta_t' \text{DEMO}_i + \gamma_t' \text{SCORE0}_i + \delta_t' \text{TAGG}_{j(i,t)} + \theta_{j(i,t)} + \varepsilon_{it} \quad (1)$$

The coefficients of all predictor variables are subscripted with  $t$  to make explicit the fact the coefficients are estimated separately for each successive pair of annual scores. The teacher effects  $\theta_j$  are modeled as independent, mean zero normal random variables with a common variance estimated from the data.

We consider two variants of this model. The first forces  $\lambda_t = 1$ . For this variant simple algebra shows that Equation 1 can be rewritten as:

$$(Y_{it} - Y_{it-1}) = \mu_t + \beta_t' \text{DEMO}_i + \gamma_t' \text{SCORE0}_i + \delta_t' \text{TAGG}_{j(i,t)} + \theta_{j(i,t)} + \varepsilon_{it} \quad (2)$$

This variant of the model is a hierarchical linear model for gain scores ( $Y_{it} - Y_{it-1}$ ) and we thus refer to it as the *gain score* model. The second variant estimates  $\lambda_t$  along with all the other model parameters. Because this model treats the prior year score as a covariate we call this the *covariate adjustment* model. Both models are discussed in detail in McCaffrey et al (2004a).

As discussed in the text, we replicated fitting the models in Equations (1) and (2) with subsets of  $(\beta_t, \gamma_t, \delta_t)$  set to zero corresponding to the different levels of CONTROLS. Separate

estimates were made for both the Procedures and Problem Solving scores and for years 2 and 3, using only the students who had scores in both the current year and prior year and who had an observed link to a teacher in the current year. In all cases model parameters and teacher effects were estimated using the hierarchical linear model routines available in the nlme package (Pinheiro and Bates 2000) for the R Statistics Environment (R Development Core Team 2005). As is standard for hierarchical linear models, teacher effects were estimated using the Best Linear Unbiased Predictor (BLUP) of the random effects  $\theta_j$  (Robinson 1991).

*Multivariate Modeling:* The second approach to modeling teacher effects simultaneously considers the year 1, 2, and 3 scores for a given achievement measure using a multivariate model that includes terms for the correlations over time among scores for the same student. Because students switch classes each year, the simple nesting structure of the traditional hierarchical models does not hold. Rather the data are cross-classified with multi-membership as students are crossed with multiple teachers (Goldstein, 2003; Raudenbush and Bryk, 2002).

To accommodate this complex data structure we adopt the general multivariate value-added model of McCaffrey et al (2004a). The model begins by writing  $Y_{it} = U_{it} + Z_{it}$  where  $U_{it}$  does not depend on the student's teachers whereas  $Z_{it}$  depends on the student's current teacher and possibly also former teachers. As with the year-to-year models (Equations 1 and 2), the student component  $U_{it}$  depends on a grand mean, the student's background variables and year 0 scores, the classroom aggregates of the student variables, and residual errors:

$$U_{it} = \mu_t + \beta_t' \text{DEMO}_i + \gamma_t' \text{SCORE0}_i + \delta_t' \text{TAGG}_{j(i,t)} + \varepsilon_{it} \quad (3)$$

The model assumes that  $\varepsilon_{i1}$ ,  $\varepsilon_{i2}$ , and  $\varepsilon_{i3}$  are multivariate normal variables with mean zero and an unstructured covariance (i.e., variance parameters vary by year and correlations are

unconstrained).

The model for teacher effects assumes that teachers impact their students' scores both when they have them in class and in subsequent years although at a possibly diminished level, specified as:

$$\begin{aligned} Z_{i1} &= \theta_{j(i,1)} \\ Z_{i2} &= \alpha_{21}\theta_{j(i,1)} + \theta_{j(i,2)} \\ Z_{i3} &= \alpha_{31}\theta_{j(i,1)} + \alpha_{32}\theta_{j(i,2)} + \theta_{j(i,3)} \end{aligned} \quad (4)$$

The contribution in year 1 includes only student  $i$ 's year 1 teacher effect,  $\theta_{j(i,1)}$ . Later years contain teacher effect contributions from not only the current year, but also prior years. The strengths of the contributions of prior teacher effects to current scores are controlled by the persistence parameters ( $\alpha_{21}$ ,  $\alpha_{31}$  and  $\alpha_{32}$ ). If the persistence parameters are less than 1, then the effects of prior teachers on current outcomes are diminished relative to the teachers' initial contributions. Alternatively, if the persistence parameters equal one, then past teacher effects persist and accumulate with undiminished strength.

As with the year-to-year models, we employed two variants of the multivariate modeling approach. In the first variant the persistence parameters ( $\alpha_{21}$ ,  $\alpha_{31}$  and  $\alpha_{32}$ ) are treated as unknown and are estimated from the data. We call this the *variable persistence* model to signify that the persistence parameters are estimated. The second variant of the multivariate modeling approach forces all the persistence parameters to be equal to 1. We refer to this model as the *complete persistence* model. As shown by McCaffrey et al. (2004a), when all the components of  $(\beta_b, \gamma_b, \delta_i)$  in Equation 3 are equal to zero, the complete persistence model is equivalent to the layered model of the Tennessee Value Added Assessment System (Sanders, Saxton, and Horn, 1997).

Model parameters (including the regression coefficients, the persistence parameters, the

teacher effect variance components, and the variance-covariance matrix of the residual error terms) and teacher effects were estimated using the Bayesian approach of Lockwood et al. (forthcoming).<sup>3</sup> Separate models were fit for the Procedures and Problem Solving scores. Teacher effects were estimated by their posterior means, which are the analogs to the BLUPS used in the year-to-year gains analyses.

An advantage of the multivariate approach is that data from all students, even those missing some test scores, contribute to the estimation of model parameters and teacher effects. The missing data are effectively imputed as part of the computational algorithm under the assumption that the test scores are missing at random (Little and Rubin, 2002); i.e., missing scores for each year do not differ systematically from the observed scores conditional on the students' observed scores from other years. Inclusion of students with incomplete data does, however, require specialized methods for handling missing student-teacher links. We used the "pseudo-teacher" approach of Lockwood et al. (forthcoming) where each missing link is assigned to its own teacher to serve as a placeholder in the layering of the effects in Equation 4. The pseudo-teacher effects are estimated with the real teacher effects but are later discarded. The study by Lockwood et al. (forthcoming) indicates that estimated effects for actual teachers were essentially invariant to diverse methods for handling missing teacher links, including the pseudo-teacher method used here.

---

<sup>3</sup> We used standard non-informative prior distributions in the current study, allowing the estimated teacher effects to be driven by the data. Markov Chain Monte Carlo methods are used to sample the posterior distribution of all parameters (Gelman, Carlin, Stern and Rubin, 1995). For each teacher effect for each instance of the multivariate model, the posterior means and standard deviations were estimated from 5000 Markov Chain Monte Carlo samples after 1000 burn-in iterations. This sample size was sufficient to make Monte Carlo error in the estimates negligible.

**Table 1**

**Cross tabulation of number of students by observation pattern of test scores (row 1) or teacher links (row 2) in years 1, 2 and 3**

Observation Pattern	000	001	010	011	100	101	110	111
Count of Students with Given Pattern of Scores	532	280	259	314	439	288	436	839
Count of Students with Given Pattern of Teacher Links	10	382	231	542	289	146	260	995

4 0 indicates unobserved scores or links and 1 indicates observed data; e.g., pattern “101” indicate students who had observed scores or links in years 1 and 3 but not year 2. The counts of teacher linkage patterns in row 2 exclude the 532 students from row 1 with no observed scores.

**Table 2****Correlations of estimated teacher effects across Procedures and Problem Solving outcomes conditional on year, MODEL and CONTROLS<sup>5</sup>**

<b>Year 2</b>	<b>CONTROLS</b>				
<b>MODEL</b>	None	Demographics	Scores	Both	Aggregates
Gain Score	0.24	0.30	0.29	0.32	0.29
Covariate Adjustment	0.43	0.39	0.37	0.38	0.26
Complete Persistence	0.38	0.35	0.30	0.30	0.28
Variable Persistence	0.46	0.43	0.40	0.40	0.29
<b>Year 3</b>					
Gain Score	0.24	0.24	0.23	0.24	0.11
Covariate Adjustment	0.25	0.27	0.16	0.21	0.09
Complete Persistence	0.20	0.20	0.13	0.14	0.06
Variable Persistence	0.21	0.19	0.06	0.09	0.01

<sup>5</sup> Each cell contains the correlation between teacher effects estimated from Procedures and Problem Solving outcomes for the given combination of MODEL (rows) and CONTROLS (columns).

**Table 3**

**Summaries of correlations of teacher effects obtained from conditions varying on either CONTROLS, MODEL, or both, conditional on year and outcome<sup>6</sup>**

		Min	Q1	Q2	Q3	Max	Mean
<b>Year 2, Procedures</b>	Varying CONTROLS	0.92	0.96	0.98	1.00	1.00	0.98
	Varying MODEL	0.81	0.88	0.93	0.95	0.98	0.91
	Varying both	0.79	0.87	0.91	0.94	0.98	0.90
<b>Year 2, Problem Solving</b>	Varying CONTROLS	0.72	0.89	0.94	0.97	0.99	0.92
	Varying MODEL	0.60	0.82	0.88	0.94	0.98	0.87
	Varying both	0.49	0.76	0.84	0.90	0.96	0.82
<b>Year 3, Procedures</b>	Varying CONTROLS	0.78	0.90	0.99	1.00	1.00	0.94
	Varying MODEL	0.73	0.88	0.92	0.96	0.98	0.91
	Varying both	0.57	0.81	0.89	0.94	0.98	0.86
<b>Year 3, Problem Solving</b>	Varying CONTROLS	0.92	0.95	0.98	0.99	1.00	0.97
	Varying MODEL	0.76	0.91	0.94	0.95	0.98	0.92
	Varying both	0.79	0.89	0.92	0.94	0.96	0.91

<sup>6</sup> For each year and outcome, the three rows provide summaries of the 190 unique correlations obtained by varying MODEL and COVARIATE, partitioned as described in the text.

**Table 4****Variance decomposition for estimated teacher effects by year<sup>7</sup>**

<b>Year 2</b>			
Term	Sums of Squares	Mean Squares	Estimated Variance Component
Teacher	63	1.8	0.02
Teacher * OUTCOME	37	1.0	0.05
Teacher * MODEL	6	0.1	0.01
Teacher * CONTROLS	2	0.0	0.00
Residual	3	0.0	0.00
<b>Year 3</b>			
Term	Sums of Squares	Mean Squares	Estimated Variance Component
Teacher	100	3.0	0.01
Teacher * OUTCOME	81	2.4	0.12
Teacher * MODEL	7	0.1	0.01
Teacher * CONTROLS	6	0.0	0.00
Residual	8	0.0	0.01

<sup>7</sup> Rows correspond to model terms of teacher main effect and interactions of teachers with each factor. Main effects for factors are zero due to centering of estimated effects in each design cell. Columns are sums of squares and mean squares from fixed effects regression model, and the estimated variance components from a random effects model.

**Table 5**

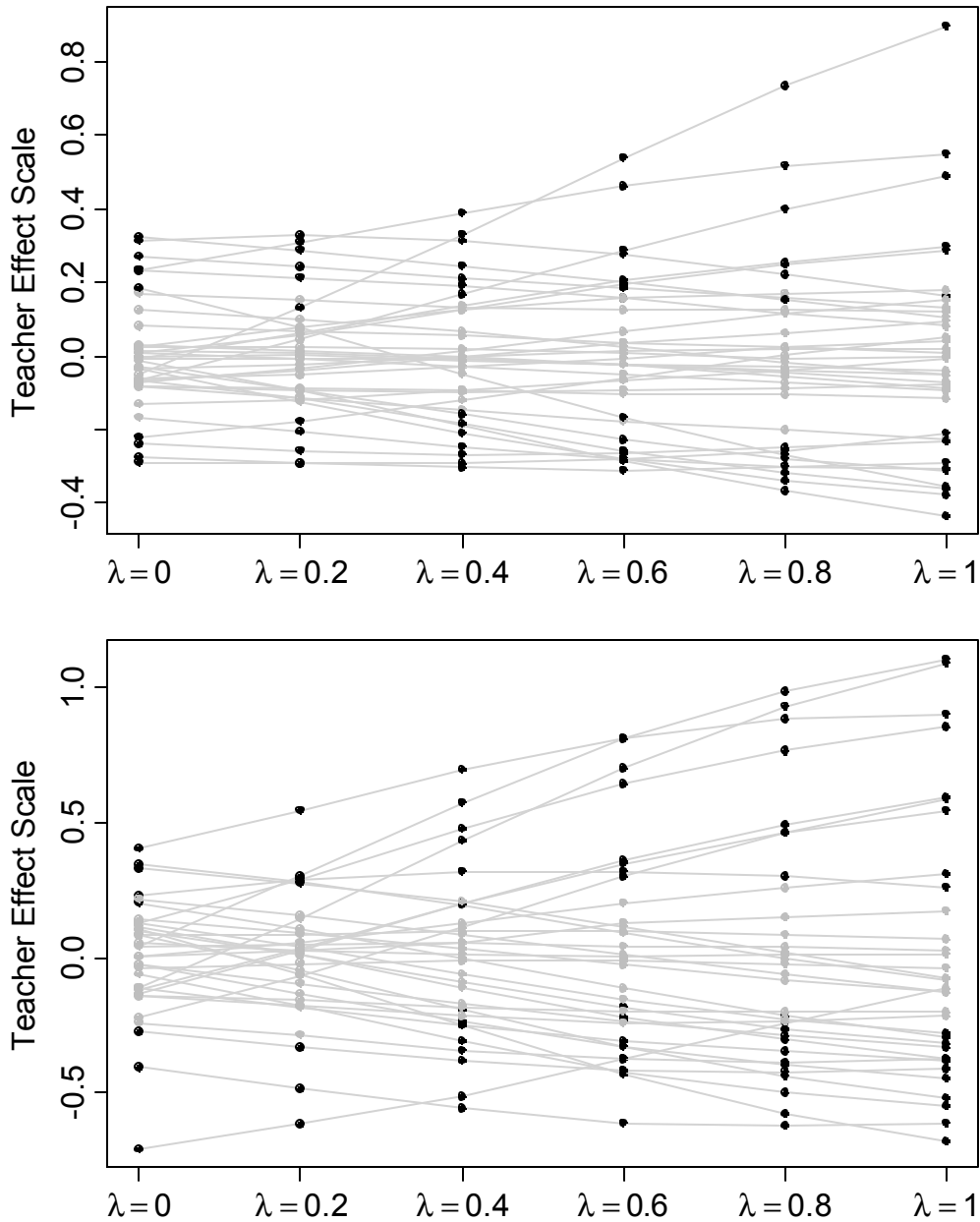
**Cross-tabulation of teacher effects by year according to sensitivity of the estimated effects to the composite outcome described in the text<sup>8</sup>**

	-	0	+	- 0	0 +	- 0 +
Year 2 (n=37)	3 (8%)	18 (49%)	2 (5%)	6 (16%)	7 (19%)	1 (3%)
Year 3 (n=34)	2 (6%)	9 (26%)	2 (6%)	11 (32%)	9 (26%)	1 (3%)

<sup>8</sup> Columns 1, 2 and 3 are teacher effects that were negative, not detectably different, or positive, respectively, across all composite outcomes. Column 4 represents teacher effects that were negative for some composite outcomes and not detectably different for others. Column 5 is analogous but for teacher effects that were positive in some cases. Finally Column 6 represents teacher effects that were estimated to be either negative, not detectably different, or positive depending on which composite outcome was considered.

Figure 1

Year 2 (top frame) and year 3 (bottom frame) teacher effects calculated from composite scores as described in text<sup>9</sup>



<sup>9</sup> Gray lines connect estimated effects for the same teacher. Black dots indicate effects that are detectably different from average, and gray dots indicate effects not detectably different from average. All estimates are based on the complete persistence model including controls for both student-level demographic variables and year 0 scores. For year 2, the correlations between the estimates based on the Procedures scores and those based on the composite scores as lambda goes from zero to one are (.30, .60, .83, .95, .99, 1.00) and the corresponding correlations for the Problem Solving scores are (1.00, .94, .77, .57, .42, .30). The analogous values for year 3 are (.14, .59, .86, .96, .99, 1.00) and (1.00, .88, .62, .39, .24, .14).