WORKING PAPER

# A Value-Added Modeling Approach for Examining the Relationship Between Reform Teaching and Mathematics Achievement

J.R. LOCKWOOD, VI-NHUAN LE, BRIAN STECHER, LAURA HAMILTON

WR-262-EDU

April 2005

RAND EDUCATION

DRAFT

Paper presented at the annual meeting of the American Education Research Association, Montreal, Canada, April 2005

This material is based on work supported by the National Science Foundation under Grant No. ESI-9986612. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

Recent large-scale efforts to improve mathematics and science education have focused on changing teachers' instructional practices to be more aligned with the teaching standards put forth by professional organizations such as the National Council of Teachers of Mathematics or the American Association for the Advancement of Science. These national organizations, as well as other local reform efforts, advocate an approach to mathematics and science instruction that places less emphasis on the acquisition of discrete skills and factual knowledge, and greater weight on conceptual understanding, inquiry, and application and communication of mathematical or scientific ideas (National Research Council, 1996; American Association for the Advancement of Science, 1993; National Council of Teachers of Mathematics, 1989; 2000). This approach, commonly referred to as reform-oriented instruction, is intended to engage students as active participants in their own learning and to promote the development of complex cognitive skills and processes. Although advocates of this approach do not dispute the importance of computational skills and factual knowledge, they argue that traditional curricula have often emphasized these outcomes to the exclusion of more complex problem-solving and reasoning skills, and as a result, students are often poorly prepared for careers that require the use of higher-level mathematics and science skills and knowledge.

Research shows that many teachers have begun to incorporate these approaches in their classrooms (Kim, Crasco, Blank, & Smithson, 2001), but the evidence supporting the efficacy of these practices in mathematics and science is relatively weak. Studies that have examined the relationship between student achievement and teachers' reports of the frequency with which they engaged in reform-based instruction suggest that these practices may improve student achievement, but in most cases, the effects appear to be quite small. Mayer (1998) observed small positive or null relationships between reform-based practices and student scores on a standardized multiple-choice test. Similar results are described by Wenglinsky in mathematics (2002) and Smerdon, Burkam, and Lee (1999) in science. A synthesis of data from eleven NSF-funded Systemic Reform Initiatives found a mixture of null and small positive results on both multiple-choice and open-response assessments (Hamilton et al., 2003).

Of particular interest is whether reform-oriented teaching can enhance performance of scientific or mathematical communication, problem solving, or other "higher-order thinking" skills. Advocates of reform teaching believe these are areas in which reform practices might be especially effective. There is some evidence that reform instruction is positively related to these types of skills, albeit weakly. A study by Cohen and Hill (2000) revealed an association between reform instruction and scores on the California Learning Assessment System (CLAS) mathematics test, a performance-based assessment designed to measure students' understanding of mathematics problems and procedures. Thompson and Senk (2001) found that reform practices, in conjunction with a reform-oriented curriculum, correlated positively with mathematics achievement, especially on multistep problems, and problems involving applications or graphical representations in mathematics. Similarly, Saxe, Gearhart, and Seltzer (1999) found that reform instruction was associated with student mathematics problem-solving scores.

The weak relationships between reform pedagogy and achievement may be partially attributable to several factors, including the use of achievement measures that were not well-aligned with reform practices or curricula. Because of cost and logistic constraints, many studies have relied on existing state tests, most of which consist of multiple-choice items, as indicators of achievement. However, state tests and the multiple-choice format have been criticized for focusing on low-level skills (Achieve, 2004; Resnick & Resnick, 1992), and advocates of reform teaching believe these measures do not adequately reflect the range of competencies that reform teaching is expected to develop. It is believed that tests requiring students to construct their own response and to engage in complex problem solving may be more sensitive to the effects of reform teaching. Indeed, Hamilton et al (2003) found that reform teaching had stronger relationships with open-ended measures than with multiple-choice tests.

Another factor underlying the weak relationships between reform pedagogy and student test scores is the limited timeframe in which reform teaching is examined. Most studies have examined relationships over a one-year period, but the use of particular instructional strategies in a single course during a single school year may not expected to lead to large effects, especially in comparison to other factors such as student background characteristics. Several years of exposure may be needed to achieve a reasonably large effect. This suggests the need for longitudinal studies that examines the relationships between achievement and reform for longer periods of time.

Finally, the weak relationships may have stemmed from inadequate measurement of reform-oriented instruction. Many of these studies used surveys that asked teachers about the frequency with which they engaged in particular practices. While researchers have successfully used these data to explore relationships between instructional practices and student achievement (Cohen & Hill, 2000; Gamoran, Porter, Smithson, & White, 1997; Wenglinsky, 2002), surveys nonetheless have problems that limit their utility (Rowan, Correnti, & Miller, 2002). Surveys, for example, are designed to be applicable to a range of settings and focus on long-term patterns of behavior. Thus, they often ignore important variations in teaching strategies related to grade level or content (Mullens & Gayler, 1999). Surveys also cannot capture subtleties in how teachers understand terminology or implement practices. In a validation study, Mayer (1999) concluded that survey data can distinguish between teachers who frequently use certain types of practices from teachers who do not, but it cannot provide information about how those practices are substantively implemented.

The purpose of this study is to explore the relationship between mathematics achievement and reform teaching over a three-year period. We use data gathered from various teacher sources (to be described later), and from district records of student test scores to explore this relationship. A unique feature of our analysis entails innovative vignette-based methods for measuring instructional practice in the hopes of capturing aspects of reform teaching that may not be well measured by traditional surveys. In addition, our analysis examines whether the relationship between reform instruction and achievement differs when achievement is measured using a multiple-choice versus an open-ended format.

## Methodology

### Sample Participants

Three school districts that were implementing NSF-sponsored Local Systemic Change (LSC) projects participated in this study. The LSC program is one of a series of initiatives designed to promote systemic reform of mathematics and science teaching. A large portion of the initiatives' funds is devoted to training to increase teachers' use of classroom practices that are consistent with the reform principles recognized by some national professional organizations (e.g., the National Council for Teachers of Mathematics, the American Association for the Advancement of Science). This paper focuses on the two districts with participating mathematics sites.

Students from these school districts have been grouped into several cohorts based on school district, grade level and subject content, and their student achievement data are being tracked for three years. The districts and subject-grade combinations were chosen to meet several criteria, including an established record of teacher participation in reform-related professional development, a willingness to administer student achievement tests as well as the instructional practice measures, and the availability of a data system that would allow us to track individual students over time and link students to their math or science teachers. The sample includes three cohorts, two cohorts for middle school mathematics, and one for elementary mathematics. This resulted in an initial year sample consisting of third-grade, sixth-grade, and seventh-grade mathematics. We followed these students for an additional two years, resulting in a longitudinal sample of third through fifth grade, sixth through eighth grade, and seventh through ninth grade.

Table 1 indicates the total number of responding teachers, and the number of students in the district-provided dataset who were linked to those teachers. Response rates varied from 73 to 100 percent for teachers, and sample sizes for students ranged from approximately 1650 to 3500.

Table 1. Description of Participating Cohorts in Year 1

| Student Cohorts | School District | Grades | Teacher Response Year 1 (n/N) | Teacher Response Year 2 (n/N) | Teacher Response Year 3 (n/N) | Student Sample Size Year 1 | Student Sample Size Year 2 | Student Sample Size Year 3 |
|---|---|---|---|---|---|---|---|---|
| Cohort 1 | District 1 | 3-5 | 68/81 | 80/85 | 71/72 | 1642 | 1928 | 1649 |
| Cohort 2 | District 1 | 7-9 | 43/59 | 64/96 | 57/59 | 3520 | 3511 | 3459 |
| Cohort 3 | District 2 | 6-8 | 64/79 | 43/43 | 38/38 | 1851 | 2451 | 3043 |

For each site, we received student demographic information, including racial/ethnic group, gender, and eligibility for free-or reduced-price (FRL) lunch. In many of the sites, we received additional information, such as gifted status, and limited English proficient (LEP) status. Student demographics for each of the sites are provided in Appendix A.

## Measures

For each subject-grade combination, different methods were used to measure classroom practice: a survey, a set of vignette-based questions, and classroom logs. The measures were designed to provide overlapping evidence about the extent of reform-oriented teaching that occurred in each class. For each measure, teachers were provided with a set of responses designed to capture a range of behaviors, from less reform-oriented to more reform-oriented. The measures are described in detail in the following sections.

*Teacher Survey*. The survey included questions about the teacher's educational background and experience, the mathematics curriculum taught in the class, and the use of a variety of teaching practices. Teachers indicated how much class time was spent on various mathematical topics (e.g., multiplication/division of whole numbers, patterns/functions/algebra). They indicated the frequency with which they engaged in particular instructional activities (e.g., lecture or introduce content through formal presentations, encourage students to explore alternative methods for solutions). They also indicated the frequency with which students took part in specific learning activities (e.g., practice computational skills, work on extended mathematical investigations). Questions of these types have been used extensively in research on mathematics instruction (see for example, Cohen & Hill, 2000; Hamilton et al., 2003; Wenglinsky, 2002).

*Teacher Logs*. Teachers were asked to fill out a daily log describing specific activities that occurred during their mathematics lesson. While the surveys focused on long-term patterns of behavior, the logs focused on activities during a specific two-day period. Teachers indicated how much time students spent on selected activities (e.g., use manipulatives to solve problems, complete worksheets or problem sets from text). Similarly, they indicated how much time they devoted to selected behaviors (e.g., monitor students as they work, ask questions of individuals to test for understanding). Finally, teachers indicated whether or not certain activities occurred at all during the lesson (e.g., students engaged in a debate or discussion of ways to solve a problem, teacher or student connected today's mathematics topic to another subject (e.g., social studies)).

*Vignette-Based Items*. As part of the survey, teachers responded to two vignettes that contained descriptions of realistic classroom settings and events and asked teachers to indicate how they would respond in each setting. Each vignette presented the teachers with four instructional problems that provided teachers with hypothetical situations at different stages within the unit. The first problem focused on the manner in which the teacher would introduce a new unit. The second problem dealt with how the teacher responds to mistakes from students. The third problem involved teachers' reactions to students who gave two approaches to solving a problem, both of which were correct, but differed in their efficiency. The final problem asked teachers about their emphasis on different learning objectives (see Appendix B for the vignettes from sixth-/seventh-grade mathematics).

For each problem, teachers were provided with a set of responses that were designed to capture a range of behaviors, from less reform-oriented to more reform-oriented. Teachers were asked to rate the likelihood of engaging in each option using a four-point scale from "very unlikely" to "very likely" or, in the case of questions of emphasis, from "no emphasis" to "great emphasis." Teachers' responses provided an indication of their intent to teach in a less-reformed or more-reformed manner.

The topics for the vignettes were selected to be appropriate to the grade and curriculum of the site. In creating the vignettes, we followed a template so that the hypothetical situations were roughly comparable within and across grades for a given subject. In other words, the vignettes were designed to be "parallel" in the sense that we presented teachers with similar instructional problems and response options, but situated in different topics. However, the different setting for each vignette meant that some response options had to be modified to fit the specific mathematical or scientific context. Thus, while the vignettes are highly similar, they cannot be considered strictly parallel.

*Achievement Measures*. In addition to information gathered from teachers, we obtained achievement scores for students. In this paper, we focus on the multiple-choice scores obtained on the mathematics section of the SAT-9. In addition, we received prior year scores on various subjects from the districts' state testing program, and open-ended SAT-9 scores for Cohort 3 (i.e., grades 6-8 cohort). The former was used in our models to control for prior student ability, and the latter allowed us to explore potential differential relationships between reform teaching and open-ended and multiple-choice formats.

## Analysis

### Scale Development

*Scales Derived from the Vignettes*. We scored the vignettes separately for each grade and subject. Our goal in scoring the vignette-based questions was to characterize teachers along a single dimension of reform from highly reform-oriented to not very reform-oriented. This necessitated deciding which of the options described behaviors that reflect reform teaching. We used a judgmental process to assign a value from 1 (low reform) to 4 (high reform) to each response option. Members of the research team independently rated each response option, then convened to reconcile differences. The panel of mathematics educators rated a comparable set of scenarios in a previous year, and we used the decision guidelines they established. For the purposes of analysis, we considered options that had been rated a 3 or 4 to be indicative of high-reform pedagogy, and options that had been rated a 1 or 2 to be indicative of low-reform teaching (See Appendix C for reform ratings for sixth/seventh-grade mathematics).

We used a multidimensional scaling process in which both the high- and low-reform items were included as part of the scoring procedures. From the *n*-teacher by *N*-item matrix, we created an *n* x *n* teacher similarity matrix, and used multidimensional scaling to plot the teachers in three-dimensional space. We added into this plot two additional points corresponding to a simulated "ideal" high-reform teacher (whose responses corresponded exactly to our judgments of reform orientation), and a simulated "ideal" low-reform teacher (whose responses were just the opposite). An examination of the teachers' responses showed that teachers were generally more similar to our idealized high-reform teacher than they were to our idealized low-reform teacher, although there was also variation in responses. We then used the *n*-teacher by *N* response option matrix to construct and plot a *N* by *N* response option similarity matrix in 3-dimensional space. The results suggested that teachers were consistent in how they were responding to the items. That is, high-reform teachers indicated that they were likely to engage in many of the high-reform options and few of the low-reform options, whereas low-reform teachers showed the opposite pattern.

On the basis of these similarity analyses, we created a scale of reform-oriented instruction by calculating the Euclidean distance between each teacher and the ideal high-reform teacher. We scaled this measure, which we refer to as Euclid, so that teachers who are closer to the ideal high-reform teacher receive higher scores. That is, larger values of Euclid are associated with teachers whose responses are more like our ideal high-reform teacher. (Readers interested in the relationships between teachers' responses to the vignettes and to the surveys and logs are referred to Le et al., 2003).

*Scales Derived from Survey and Daily Logs*. Several additional scales were derived from teachers' responses to the survey and the daily log. The scales were created using a combination of empirical analyses (e.g., factor analysis), prior results with similar items in other studies, and expert judgments based on the underlying response options. We created scales that we thought were key indicators of low- and high-reform instruction, scales that described teacher background, and scales that described classroom context. Each is described below.

Table 2 contains brief descriptions of the scales and examples of illustrative items. (The complete set of items that comprise each scale is presented in Appendix D.) Of the instructional practices scale, three were derived from the surveys. Mathematical Processes measured teachers' self-reported emphases on NCTM-endorsed processes, including proof and justifications, problem solving, mathematical communication, connections and mathematical representations. Reform Relative to Text assessed teachers' emphasis on certain reform practices in relation to their primary textbook. Reform Practices measured the frequency with which teachers engaged in nine specific reform-oriented instructional practices. The items on this latter scale are similar to those used in other research on mathematics and science reform, including some national longitudinal surveys (Cohen & Hill, 2000; Hamilton et al., 2003; Swanson & Stevenson, 2002; Wenglinsky, 2002).

The remaining instructional practices scales were derived from the daily logs. The Discussion scale was based on the amount of time teachers reported that the class was engaged in a dialogue about mathematical/scientific thinking and understanding. Groupwork entailed the amount of time the class spent in groups as a whole. Two other scales, Mixed-ability Groupwork and Problem-Solving Groupwork were created from the Groupwork scale. Mixed-ability Groupwork measured the proportion of groupwork time in which students worked in mixed-ability groups. Problem-solving Groupwork assessed the proportion of groupwork time in which students collaboratively solved new problems.

The Reform Activities scale measured the number of specific reform behaviors that occurred during the lesson, and the Seatwork scale described the amount of time students spent reading from the textbook, completing worksheets or problem sets, or other low-reform activities.

There were two scales relating to curriculum coverage. Operations measured the extent to which teaches covered operations with whole numbers, which is thought by proponents of reformed pedagogy to be a more traditional area of math, and Proof and Patterns assessed the extent to which they focused on proof/justification/verification and patterns/function/algebra.  According to the advocates of reformed teaching, this latter topic represents more reform-oriented areas of math.

There were also six scales, all derived from the surveys, about teacher background.  The Certification variable indicated whether the teacher held a standard certification, and the Confidence scale assessed whether teachers felt very confident in their mathematics or science knowledge that they were asked to teach.  Masters assessed whether teachers held a masters degree in any subject, and Math Degree indicated whether the teacher held a major or minor in mathematics.  Professional Development measured the amount of subject-specific in-service training received in the past 12 months.  The Experience scale indicated the total number of years teachers taught on a full-time basis, and Experience at Grade scale indicated the total number of years teachers taught at grade level.

The final set of scales was designed to provide context about classroom conditions.  The Hours of Weekly Instruction represented teachers' estimation of the number of hours of math instruction students received in a typical week, and the Time on Task scale measured teachers' self-report of how much of the class time was spent effectively on mathematics instruction.  Class size was a measure of the number of students in the class, and Classroom Heterogeneity was an indicator variable denoting whether or not the classroom consisted of a mix of student abilities.

Table 2.  Descriptions of Scales

| Scale | Description | Illustrative Items | Number of Scale Items | Source | Variable Name |
|---|---|---|---|---|---|
| Instructional Practices | | | | | |
| Reform Inclinations | Standardized sum of teachers' answers to the high-reform response options across the two scenarios | Have a classroom discussion about the differences between the two approaches | 27 | Vignette | Allhigh |
| Euclid | Euclidean distance of the teacher from the ideal high reform teacher | Tell them they are both right and move on to the next problem | 51 | Vignette | Euclid |
| Reform Practices | Frequency with which the teacher engaged in reformed instructional practices | How often does the teacher encourage students to explore alternative methods for solutions? | 9 | Survey | Reform |
| Mathematical Processes | Extent to which teachers emphasized NCTM-endorsed mathematical processes | How much emphasis does the teacher place on proof and justification/verification? | 5 | Survey | NCTM |
| Reform Relative to Text | Average emphasis on reform activities relative to textbook | How much does your emphasis on solving real-life problems compare to that of your primary textbook or published curriculum material? | 5 | Survey | Text |
| Discussion | Amount of time the class engaged in dialogue about mathematical thinking and understanding | How much time did students spend explaining their thinking about mathematical problems? | 4 | Log | Discuss |
| Groupwork | Amount of time the class spent in groupwork | How long did students work in groups during today's mathematics lesson? | 1 | Log | Groupwrk |
| Mixed-ability Groupwork | Amount of time the class spent working in mixed-ability groups | If groups were used, what share of the group time was used working in groups of mixed ability? | 1 | Log | Absmxd |
| Problem-solving Groupwork | Amount of time the class spent solving new problems together as a group | If groups were used, what share of the group time was used solving new problems together as a group? | 1 | Log | Absprob |
| Reform Activities | Presence of selected reform activities | In today's mathematics lesson, did a student restate another student's ideas in different words? | 6 | Log | Refact |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| Seatwork | Extent to which students engaged in seatwork and other low-reform activities | How much time did students spend reading from textbook or other materials? | 2 | Log | Traditional |
| Curriculum |  |  |  |  |  |
| Operations | Weeks spent on operations with whole numbers | Indicate the approximate amount of time you will spend on operations with signed whole numbers | 1-2 | Survey | Operations |
| Proofs and Patterns | Weeks spent on proof/justification/verification and patterns/functions/algebra | Indicate the approximate amount of time you will spend on patterns/functions? | 3 | Survey | Proof.patterns |
| Teacher Background |  |  |  |  |  |
| Certification | Whether the teacher holds standard certification | Which type of teaching certification do you hold? | 1 | Survey | -- |
| Confidence | Whether the teacher is very confident in his/her mathematics knowledge | With respect to the mathematics that you are asked to teach, how confident are you in your mathematical knowledge? | 1 | Survey | Confidence |
| Masters | Whether the teacher holds at least a masters degree (any subject) | What is the highest degree you hold? | 1 | Survey | Mastdeg |
| Math Degree | Whether the teacher holds a math-intensive undergraduate degree (major or minor) | Did you major in mathematics or a mathematics-intensive field for your Bachelor's degree? | 2 | Survey | Mathdeg |
| Professional Development | Amount of professional development received | In the past 12 months, how much time have you spent on professional development activities that focused on in-depth study of mathematics content? | 7 | Survey | Pdtot |
| Experience | Number of years of experience | Including this year, how many years have you taught on a full-time basis? | 1 | Survey | Yrstch |
| Experience at Grade | Number of years teaching at grade level | Including this year, how many years have you taught third grade? (third grade version) | 1 | Survey | Yrs.grade |
| Classroom Context |  |  |  |  |  |
| Heterogeneous Classroom | Whether the classroom is of mixed ability levels | How would you describe the variation in mathematics ability of your class? | 1 | Survey | Hetero |

| | | | | | |
|---|---|---|---|---|---|
| Class size | Number of students in the class | How many students were present? | 1 | Log | Class_size |
| Hours of Weekly Instruction | Number of hours of math instruction per week | In a typical week, how many hours of mathematics instruction do students in your class receive? | 1 | Survey | Hrs_instruct |
| Time on Task | Number of minutes effectively spent on mathematics (i.e., disregarding disruptions) | How long was today's mathematics lesson? | 2 | Log | Efftime |

Most of the scales were dichotomous or continuous, but Reform Practices, Discussion, Professional Development, and the scales relating to curriculum (i.e., Operations and Proof and Patterns) were on a 5-point metric, and Reform Relative to Text was on a 3-point scale. For these scales, the score was the average response across items, with higher scores denoting more frequent use of the practices measured by that scale. For example, a score of 5 on the Reform Practices scale indicates that teachers spent much time on reform-oriented activities. In contrast, a score of 1 on a scale such as Discussion indicates that little class time was devoted to talking about mathematical thinking.

**Modeling Approach**

*Overview of the Data*. Our analysis used data from three cohorts of students from two different school districts. The data from each cohort had largely parallel structures. In general we had longitudinally linked student test scores on the SAT-9 for 2001-2003 (which we call years 1, 2 and 3 respectively throughout the discussion). In addition to the SAT-9 mathematics scores from years 1-3, which are our primary outcomes, we also had test scores from "year 0," the year prior to data collection on the teachers. These test scores are used as fixed-effects regressors in our models, removing a large part of the variation among students in initial achievement. In addition to student test scores, we also had student demographic variables that were specific to each cohort but in general included information such as race, gender, FRL eligibility, LEP status, and age. Students were linked over time to their teachers in years 1-3, for whom we have measures of teaching practices and background characteristics derived from the surveys and logs.

*Model Structure.* The longitudinal design of our data collection offers opportunities for powerful analyses, but simultaneously presents conceptual and practical challenges. Students are changing teacher and classroom contexts over time, and the primary interest is in cumulative effects of exposure to different teaching practices on mathematics achievement. That is, the achievement at time *t* is in part a function of the history of exposure to teaching practices up to and including time *t*. More generally, both observed and unobserved factors specific to students affect achievement, and both observed and unobserved factors specific to educational experience affect achievement. Accounting for all of these factors to make reasonable inferences about the effects of exposure to teaching practices poses challenging statistical issues.

The modeling approach we took is an implementation of the general multivariate linear mixed model presented by McCaffrey et al (2004). We jointly model the outcomes from years 1-3, expressing student scores as a linear function of overall means, adjustments for student background variables and year 0 performance, a function of the teaching practice exposure history, a function of unobserved random teacher effects, and residual errors that are allowed to be correlated across time within student. The model is as follows:

$$Y_{i1} = (u_1 + X_i^T B_1 + C_{11} Y_{i0} + C_{12} Y_{i0}^2) + g_{11} P_{j1(i)} + T_{j1(i)} + e_{i1}$$

$$Y_{i2} = (u_2 + X_i^T B_2 + C_{21} Y_{i0} + C_{22} Y_{i0}^2) + (g_{12} P_{j1(i)} + g_{22} P_{j2(i)}) + (a_{12} T_{j1(i)} + T_{j2(i)}) + e_{i2}$$

$Y_{i3} = (u_3 + X_i^T B_3 + C_{31} Y_{i0} + C_{32} Y_{i0}^2) + (g_{13} P_{j1(i)} + g_{23} P_{j2(i)} + g_{33} P_{j3(i)}) +$
$(a_{13} T_{j1(i)} + a_{23} T_{j2(i)} + T_{j3(i)}) + e_{i3}$

*Notation description*: i indexes students. $Y_{it}$ is the score for student i in year t. $u_t$ are overall yearly means. $X_i$ are student covariates which as noted vary by cohort, with coefficients $B_t$ that are allowed to depend on t. Note that the student characteristics are treated as time invariant (that is, they are not indexed by t). While some of the variables actually may change over time (e.g., LEP status and FRL participation), the fraction of students for which time variation was observed was generally less than 1%. For such variables, we defined the time-invariant version as whether or not the student was ever classified as LEP or ever participated in FRL programs. This strategy seemed reasonable given the fact that so few students exhibited time variation, and that allowing for time variation in the covariates would have added additional complication to the missing data imputations because a model for time variation would have been required.

The student index "jt(i)" refers to the index j of the teacher linked to student i in year t (t=1,2,3). The key variables of interest are the exposure variables $P_{jt(i)}$. This refers to any one of the various teacher practice variables under consideration; we examined these variables one at a time in the context of the above model. The coefficients $g_{t1t2}$ for t1 <= t2 determine the impacts of these exposures on student achievement after controlling for the other variables in the model. $g_{t1t2}$ is the effect of the exposure experienced in t1 on the t2 score. By this construction, we are assuming that each year, the effect of exposure to a practice is a linear function of the current and past exposures, where the coefficients are allowed to be different for each future year in which a past exposure may contribute. This leads to 6 coefficients parameterizing the effect of exposure history on student achievement. As discussed below, this parameterization is flexible and subsumes other simpler models that might be considered given our data structure.

The remaining terms of the model capture two types of residual error. The terms $T_{jt}$ are random teacher effects (or more properly classroom effects) with variance components that vary by year. These are intended to capture systematic deviations of classroom-level mean residuals after controlling for the other variables in the models, which might reflect the effect of an unmeasured teacher or classroom variable. Moreover, we allow for the possibility that such effects persist into the future, which provides a mechanism for modeling residual covariance from current year outcomes of students who shared a teacher in the past. The coefficients $a_{t1t2}$ determine how much a teacher/classroom effect from year t1 persists in year t2. These are analogous to the terms $g_{t1t2}$ that measure the persistence of the effects of exposure to teaching practices with the restriction that $a_{t1t2} = 1$. Finally, the terms $e_{it}$ are residual terms that are assumed to have mean zero and unrestricted covariance matrix (different variances each year and different pairwise correlations) over time within student. This helps to capture unobserved student characteristics that affect achievement.

With no regressors in the model and under the restriction of $a_{t1t2} = 1$, the model is equivalent to the "layered model" (Sanders, Saxton, & Horn, 1997) used by the TVAAS and now SAS for value-added modeling of teacher effects. We have shown in other contexts (McCaffrey et al, 2004; Lockwood et al. 2004) that a model that allows $a_{t1t2}$ to be estimated by the data provides a better fit to the data.

The linear parameterization of the effects of the teaching practices subsumes several submodels that might be of interest. First, setting $g_{t1t2} = 0$ for $t1 < t2$ obtains a model where current year exposure affects current year level scores. Alternatively, the constraints $g_{11} = g_{12} = g_{13}$ and $g_{22} = g_{23}$ obtain a model where, after the first year, exposure affects current year gain scores. It is difficult to specify *a priori* which of these two specifications is more realistic. A key advantage of the full six-parameter specification is that both of these extreme cases are covered, as are more general models where both levels and gains are affected by past and current exposure. We are currently looking into allowing interaction terms in our models by relaxing the restriction of a linear function of past exposure history to a quadratic function.

*Implementation Issues.* Because the students change teachers over time, traditional HLM methods that assume nested data structures are not appropriate. Fitting the models above require specialized methods for cross-classified random effects. Under some restrictions, particularly on the accumulation of the teacher random effects, and provided that the data sets are small, it is possible to estimate the parameters of the model with the nlme package (Pinheiro & Bates, 2000) for the R statistical environment (Ihaka and Gentleman. 1996). The particular methods for doing so are described in Lockwood, Doran, & McCaffrey (2003).

However, fitting the model in its full generality and achieving the computational efficiency required to fit the model to realistically sized data sets necessitated building specialized algorithms and software. We developed a Bayesian specification of the model, and an associated Markov Chain Monte Carlo algorithm for estimating that model, that is effective for even very large data sets. This algorithm is described in Lockwood, McCaffrey, Mariano, & Setodji (2004).

In addition to facing computational challenges, we also faced challenges of missing data. Less than 20% of the students in each cohort had fully observed covariates, test scores, and links to teachers for whom we had survey and log data. The interest in cumulative effects makes handling and interpreting outcomes from students not observed and linked to teachers for all three years difficult. We developed a multistage multiple imputation process for dealing with missing student covariate information, missing student test scores, item-level non-response for teacher variables, unit-level non-response for teachers, and missing student-teacher links. This is described in detail in Appendix E. All inferences presented here are based on aggregating model results across multiply imputed data sets using standard procedures (Schafer, 1997).

## Results

*Distribution of the Scales*. Table 3 provides the percent of teachers answering "yes" to a dichotomous item. Tables 4-6 present the descriptive statistics for selected scales for years 1, 2, and 3, respectively.

Alphas values ranged from .30 to .91 for the first year, -.15 to .90 for the second year, and .31 to .87 for the third year.  Although some scales had low internal consistency estimates, the majority had adequate reliability, typically above .60.  Most of the scales also showed moderate variation, with the exception of Certification (which was subsequently dropped from further analysis).  There were also some differences across sites in the score ranges and variability.  Although these differences could influence the likelihood of detecting relationships with achievement, the results we discuss in later sections show no clear patterns with respect to these differences.

Table 3. Distribution of “Yes” Responses for Selected Dichotomous Scales

| Scale | Cohort 1 (grades 3-5) | | | | | | Cohort 2 (grades 7-9) | | | | | | Cohort 3 (grades 6-8) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Year 1 | | Year 2 | | Year 3 | | Year 1 | | Year 2 | | Year 3 | | Year 1 | | Year 2 | | Year 3 | |
| | N | Pct | N | Pct | N | Pct | N | Pct | N | Pct | N | Pct | N | Pct | N | Pct | N | Pct |
| Teacher Background | | | | | | | | | | | | | | | | | | |
| Certification | 68 | 89.55 | 80 | 89.61 | 70 | 92.96 | 43 | 90.70 | 64 | 83.87 | 57 | 90.91 | 64 | 89.09 | 43 | 85.00 | 38 | 94.44 |
| Masters | 68 | 52.94 | 80 | 48.05 | 70 | 50.00 | 43 | 48.84 | 64 | 63.49 | 57 | 65.45 | 62 | 66.13 | 43 | 39.53 | 38 | 42.11 |
| Confidence | 66 | 50.00 | 80 | 59.74 | 70 | 60.00 | 43 | 83.72 | 63 | 85.48 | 57 | 87.27 | 62 | 59.68 | 43 | 83.72 | 38 | 89.47 |
| Classroom Context | | | | | | | | | | | | | | | | | | |
| Heterogeneous Classroom | 67 | 61.19 | 80 | 71.61 | 71 | 56.34 | 43 | 39.53 | 64 | 43.75 | 57 | 36.84 | 56 | 71.43 | 43 | 65.12 | 38 | 63.16 |

Table 4. Descriptive Statistics for Selected Mathematics Scales in Year 1

| Scale | Scale Metric | Cohort 1 (grades 3-5) | | | | Cohort 2 (grades 7-9) | | | | Cohort 3 (grades 6-8) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Mean | SD | Alpha | N | Mean | SD | Alpha | N | Mean | SD | Alpha |
| Instructional Practices | | | | | | | | | | | | | |
| Reform Inclinations | Continuous | 68 | 0.00 | 1.00 | .75 | 43 | 0.00 | 1.00 | .78 | 64 | 0.00 | 1.00 | .81 |
| Euclid | Continuous | 68 | 1.93 | .59 | .85 | 43 | 2.00 | .57 | .84 | 64 | 1.27 | .55 | .91 |
| Reform Practices | 5-point Likert | 68 | 3.81 | .47 | .72 | 43 | 3.54 | .39 | .57 | 64 | 4.14 | .54 | .82 |
| Reform Relative to Text | 3-point Likert | 68 | 2.38 | .41 | .63 | 41 | 2.26 | .36 | 48 | 59 | 2.18 | .34 | .51 |
| Mathematical Processes [b] | 4-point Likert | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| Discussion | 5-point Likert | 63 | 2.45 | .59 | .77 | 35 | 1.96 | .45 | .79 | 63 | 2.50 | 13.23 | .76 |
| Groupwork | Continuous | 63 | 11.27 | 9.26 | -- | 35 | 4.18 | 5.48 | -- | 63 | 22.59 | 13.23 | -- |
| Mixed-ability Groupwork | Continuous | 62 | 6.38 | 7.92 | -- | 35 | 1.96 | 5.07 | -- | 61 | 14.97 | 14.38 | -- |
| Problem-solving Groupwork | Continuous | 62 | 5.51 | 7.76 | -- | 35 | 1.41 | 4.11 | -- | 62 | 13.14 | 11.42 | -- |
| Reform Activities | 6-point scale | 63 | 3.39 | 1.33 | .54 | 34 | 2.65 | 1.47 | .38 | 63 | 4.05 | 1.44 | .63 |
| Seatwork | 5-point Likert | | 2.58 | .77 | .58 | 35 | 2.42 | .78 | .67 | 63 | 2.74 | .73 | .55 |
| Curriculum | | | | | | | | | | | | | |
| Operations | 5-point Likert | 67 | 3.78 | .82 | .64 | 41 | 3.27 | .87 | -- | 64 | 3.76 | .90 | -- |
| Proofs and Patterns | 5-point Likert | 67 | 2.61 | .95 | .72 | 41 | 2.79 | .64 | .30 | 64 | 2.91 | 1.01 | .35 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Teacher Background | | | | | | | | | | | | | |
| Professional Development | 5-point Likert | 68 | 2.53 | .87 | .88 | 43 | 2.83 | .91 | .85 | 64 | 2.85 | 1.07 | .89 |
| Experience | Continuous | 68 | 13.28 | 9.61 | -- | 43 | 14.70 | 10.18 | -- | 63 | 10.51 | 8.52 | -- |
| Experience at Grade | Continuous | 68 | 5.54 | 4.51 | -- | 43 | 9.65 | 8.25 | -- | 63 | 5.98 | 5.38 | -- |
| Math Degree [a] | 2-point scale | -- | -- | -- | -- | 43 | .63 | .79 | -- | 64 | .35 | .75 | -- |
| Classroom Context | | | | | | | | | | | | | |
| Hours of Weekly Instruction [b] | Continuous | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- | -- |
| Class size | Continuous | 63 | 22.95 | 3.66 | -- | 35 | 25.61 | 8.53 | -- | 63 | 18.85 | 6.17 | -- |
| Time on Task | Continuous | 63 | 43.03 | 11.91 | -- | 35 | 26.47 | 15.20 | -- | 63 | 52.17 | 17.05 | -- |

Note.

[a] Math degree takes on a value of 2 if the teacher holds a major, 1 if the teacher holds a minor, and 0 otherwise.

[b] Scale not present the first year.

Table 5. Descriptive Statistics for Selected Mathematics Scales in Year 2

| Scale | Scale Metric | Cohort 1 (grades 3-5) | | | | Cohort 2 (grades 7-9) | | | | Cohort 3 (grades 6-8) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Mean | SD | Alpha | N | Mean | SD | Alpha | N | Mean | SD | Alpha |
| Instructional Practices | | | | | | | | | | | | | |
| Reform Inclinations | Continuous | 80 | 0.00 | 1.00 | .80 | 64 | 0.00 | 1.00 | .78 | 43 | 0.00 | 1.00 | .81 |
| Euclid | Continuous | 77 | 1.47 | .54 | .86 | 64 | 1.59 | .59 | .82 | 41 | 1.94 | .62 | .88 |
| Reform Practices | 5-point Likert | 80 | 3.71 | .47 | .73 | 64 | 3.51 | .47 | .69 | 43 | 4.18 | .37 | .65 |
| Reform Relative to Text | 3-point Likert | 80 | 3.91 | .47 | .64 | 64 | 3.64 | 1.20 | .54 | 43 | 4.36 | .76 | .61 |
| Mathematical Processes [b] | 4-point Likert | 80 | 2.36 | .44 | .75 | 64 | 2.15 | .32 | .56 | 42 | 2.18 | .34 | .78 |
| Discussion | 5-point Likert | 77 | 2.59 | .67 | .80 | 62 | 2.16 | .49 | .51 | 42 | 2.83 | .73 | .84 |
| Groupwork | Continuous | 77 | 15.45 | 14.79 | -- | 62 | 9.56 | 9.57 | -- | 42 | 28.43 | 16.18 | -- |
| Mixed-ability Groupwork | Continuous | 76 | 12.02 | 13.27 | -- | 60 | 5.84 | 8.33 | -- | 42 | 22.02 | 16.71 | -- |
| Problem-solving Groupwork | Continuous | 75 | 10.55 | 12.19 | -- | 58 | 6.02 | 7.49 | -- | 42 | 28.43 | 16.18 | -- |
| Reform Activities | 6-point scale | 78 | 4.87 | 1.17 | .64 | 63 | 3.98 | 1.31 | .54 | 43 | 5.35 | .81 | -.15 |
| Seatwork | 5-point Likert | 77 | 2.46 | .80 | .60 | 62 | 2.59 | .73 | .28 | 42 | 2.76 | .66 | .55 |
| Curriculum | | | | | | | | | | | | | |
| Operations | 5-point Likert | 79 | 3.85 | .84 | .63 | 64 | 3.44 | 1.04 | -- | 43 | 3.07 | .99 | -- |
| Proofs and Patterns | 5-point Likert | 80 | 2.88 | .73 | .33 | 64 | 2.92 | .70 | .87 | 43 | 3.37 | .76 | .34 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Teacher Background | | | | | | | | | | | | | |
| Professional Development | 5-point Likert | 80 | 2.32 | .82 | .87 | 64 | 2.94 | .94 | .86 | 43 | 3.26 | 1.03 | .90 |
| Experience | Continuous | 80 | 13.33 | 9.67 | -- | 63 | 12.22 | 9.83 | -- | 43 | 10.47 | 9.56 | -- |
| Experience at Grade | Continuous | 80 | 5.54 | 5.14 | -- | 63 | 9.13 | 8.90 | -- | 43 | 4.91 | 4.70 | -- |
| Math Degree [a] | 2-point scale | -- | -- | -- | -- | 64 | .78 | .83 | -- | 42 | 1.02 | .95 | -- |
| Classroom Context | | | | | | | | | | | | | |
| Hours of Weekly Instruction [b] | Continuous | 80 | 6.70 | 2.32 | -- | 64 | 4.64 | 4.24 | -- | 43 | 5.70 | 7.08 | -- |
| Class size | Continuous | 78 | 25.03 | 3.51 | -- | 62 | 25.35 | 4.61 | -- | 42 | 22.46 | 4.49 | -- |
| Time on Task | Continuous | 78 | 46.56 | 12.40 | -- | 63 | 37.90 | 12.35 | -- | 43 | 52.81 | 18.27 | -- |

Table 6. Descriptive Statistics for Selected Mathematics Scales in Year 3

| Scale | Scale Metric | Cohort 1 (grades 3-5) | | | | Cohort 2 (grades 7-9) | | | | Cohort 3 (grades 6-8) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Mean | SD | Alpha | N | Mean | SD | Alpha | N | Mean | SD | Alpha |
| Instructional Practices | | | | | | | | | | | | | |
| Reform Inclinations | Continuous | 71 | 0.00 | 1.00 | .82 | 57 | 0.00 | 1.00 | .77 | 38 | 0.00 | 1.00 | .81 |
| Euclid | Continuous | 71 | 1.87 | .59 | .85 | 57 | 1.68 | .61 | .83 | 38 | 2.33 | .64 | .85 |
| Reform Practices | 5-point Likert | 71 | 3.84 | .58 | .83 | 57 | 3.70 | .53 | .79 | 38 | 4.21 | .44 | .73 |
| Reform Relative to Text | 3-point Likert | 70 | 3.76 | 1.20 | .70 | 57 | 3.98 | 1.13 | .65 | 36 | 4.24 | 1.12 | .71 |
| Mathematical Processes [b] | 4-point Likert | 70 | 2.32 | .43 | .58 | 57 | 2.10 | .45 | .64 | 37 | 2.16 | .36 | .67 |
| Discussion | 5-point Likert | 69 | 2.41 | .64 | .74 | 57 | 2.20 | .49 | .68 | 37 | 2.87 | .70 | .69 |
| Groupwork | Continuous | 69 | 13.29 | 12.96 | -- | 56 | 11.94 | 12.08 | -- | 37 | 27.35 | 13.78 | -- |
| Mixed-ability Groupwork | Continuous | 68 | 10.32 | 12.13 | -- | 55 | 9.72 | 12.33 | -- | 35 | 21.20 | 15.81 | -- |
| Problem-solving Groupwork | Continuous | 67 | 8.48 | 9.32 | -- | 54 | 7.06 | 9.04 | -- | 34 | 18.40 | 12.58 | -- |
| Reform Activities | 6-point scale | 69 | 4.43 | 1.39 | .77 | 57 | 4.39 | 1.13 | .57 | 37 | 5.03 | .99 | .70 |
| Seatwork | 5-point Likert | 69 | 2.52 | .82 | .31 | 57 | 2.20 | .80 | .53 | 37 | 2.58 | .90 | .64 |
| Curriculum | | | | | | | | | | | | | |
| Operations | 5-point Likert | 70 | 3.56 | 1.20 | -- | 57 | 2.68 | .98 | -- | 37 | 2.49 | .90 | -- |
| Proofs and Patterns | 5-point Likert | 70 | 2.85 | .76 | .50 | 57 | 3.27 | .44 | .65 | 37 | 3.11 | .77 | .82 |
| Teacher Background | | | | | | | | | | | | | |
| Professional Development | 5-point Likert | 71 | 2.54 | .88 | .87 | 57 | 2.99 | 1.02 | .87 | 38 | 2.91 | .86 | .85 |
| Experience | Continuous | 70 | 12.29 | 8.77 | -- | 57 | 14.30 | 10.33 | -- | 38 | 6.13 | 5.25 | -- |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Experience at Grade | Continuous | 71 | 6.24 | 5.29 | -- | 57 | 10.05 | 8.68 | -- | 38 | 5.74 | 5.10 | -- |
| Math Degree [a] | 2-point scale | 70 | -- | -- | -- | 57 | 1.04 | .87 | -- | 38 | .37 | .75 | -- |
| Classroom Context |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Hours of Weekly Instruction [b] | Continuous | 70 | 6.21 | 1.54 | -- | 57 | 4.58 | 1.75 | -- | 36 | 7.06 | 10.87 | -- |
| Class size | Continuous | 69 | 26.20 | 3.57 | -- | 57 | 25.61 | 8.53 | -- | 37 | 22.53 | 5.22 | -- |
| Time on Task | Continuous | 69 | 50.53 | 11.62 | -- | 55 | 40.16 | 11.93 | -- | 37 | 50.11 | 19.88 | -- |

*Relationships between Teacher-Reported Practices and Student Achievement*. As indicated earlier, we examined relationships between teacher-reported instructional practices and student achievement using a value-added modeling approach that controlled for prior achievement and student background characteristics. We estimated separate models for each of our sites, and report effects in terms of standardized NCE points. That is, the effect size represents the expected difference in test score standard deviations for a one standard deviation unit increase in scores on the instructional practices scales.

We estimated both current year and cumulative exposure effects. For the current year effect, we report the average effect of the current year's exposure on current year achievement. Using the notation of the model, this is estimated as (g_11 + g_22 + g_33)/3. For the cumulative exposure effect, we report the expected difference in year 3 scores between two groups of students, one of whom was one standard deviation below average with respect to exposure to reform teaching across the three years, and the other who was exactly average with respect to exposure to reform pedagogy. This is estimated by (g_31 + g_32 + g_33). In addition, we estimated the effects of both current and cumulative exposure in comparison to student background and residual error.

Figure 1 presents the percentage of total variance in test scores explained by student background, residual error, and the sequence of classrooms experienced by the students. As shown in Figure 1, student background explains the majority of the variance. Student background explains approximately 50 percent in the lower grades (i.e., Cohort 1) and on the open-ended scores, and between 50 and 75 percent in the upper grades (i.e., Cohort 2 and Cohort 3). Much of the variance can also be attributed to residual error, which accounts for 25 to 50 percent, depending on the grade and cohort. Variation that can be attributed to past and current classrooms accounts for at most about 10 percent of the total variance in outcomes, and substantially less for some grades and cohorts. This places an upper bound on the variance that could possibly be explained by any of our teacher-level measures, making it clear that no individual reform practice is likely to explain much of the variation in levels of test scores.

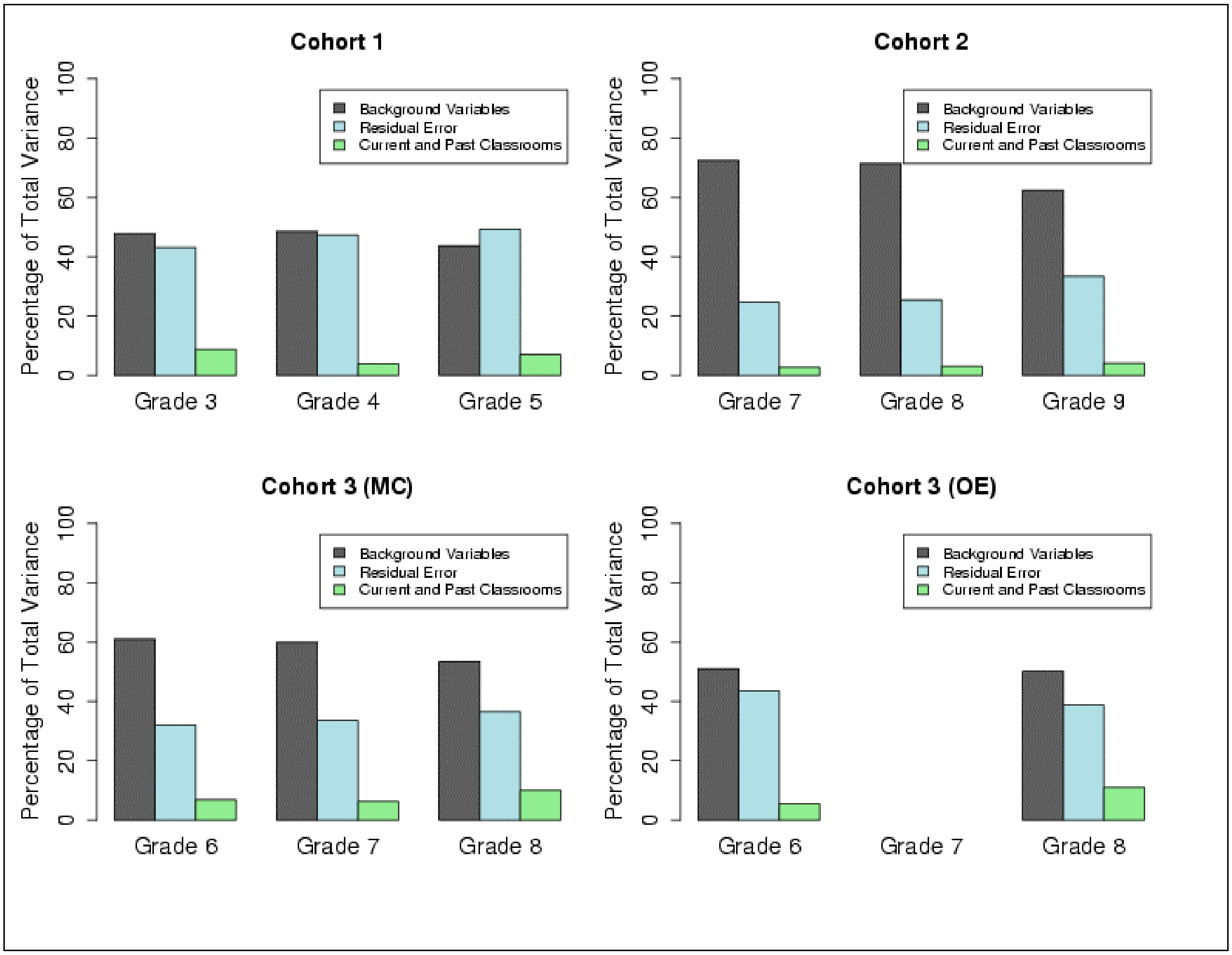

Figure 1. Percentage of Variance in Test Scores Explained by Student Background, Residual Error, and Exposure to Reform Teaching
Note: In Cohort 3, open-ended scores were not available for Grade 7.

This is confirmed by Figures 2 and 3, which present a graphical representation of the relationships between mathematics achievement and current and cumulative effects of reform teaching, respectively. In Figures 2 and 3, the red dot represents negative relationships with achievement, and the green dot represents positive relationships with achievement. The magnitude of the relationships is indicated by the area of the dot, with the larger dot representing stronger effects. Appendix F provides the standardized coefficients for each scale within each cohort.

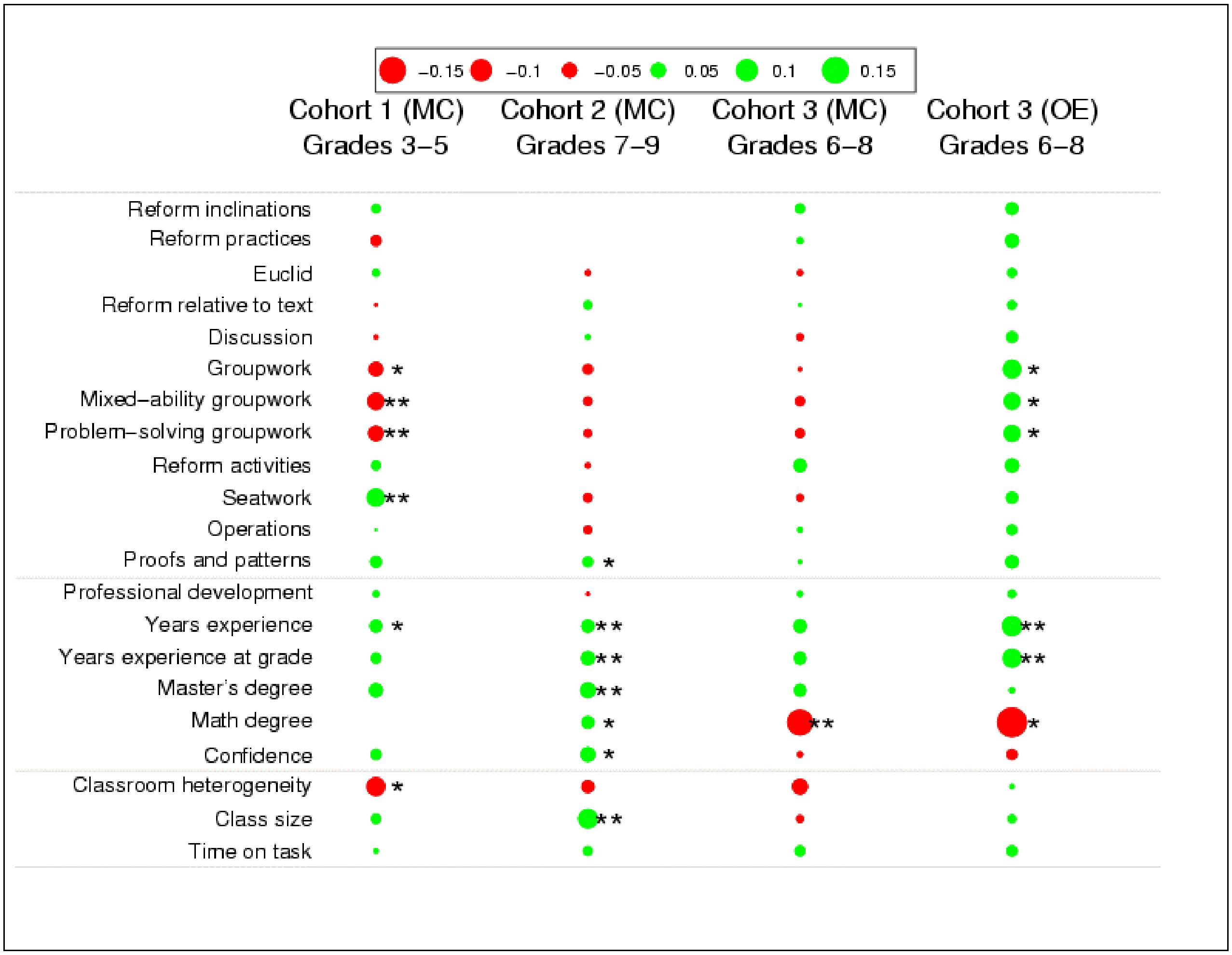

Figure 2. Relationship between Reform Teaching and Student Achievement (Current Year Effects)

Notes:

* significant at the .05 level.

** significant at the .01 level.

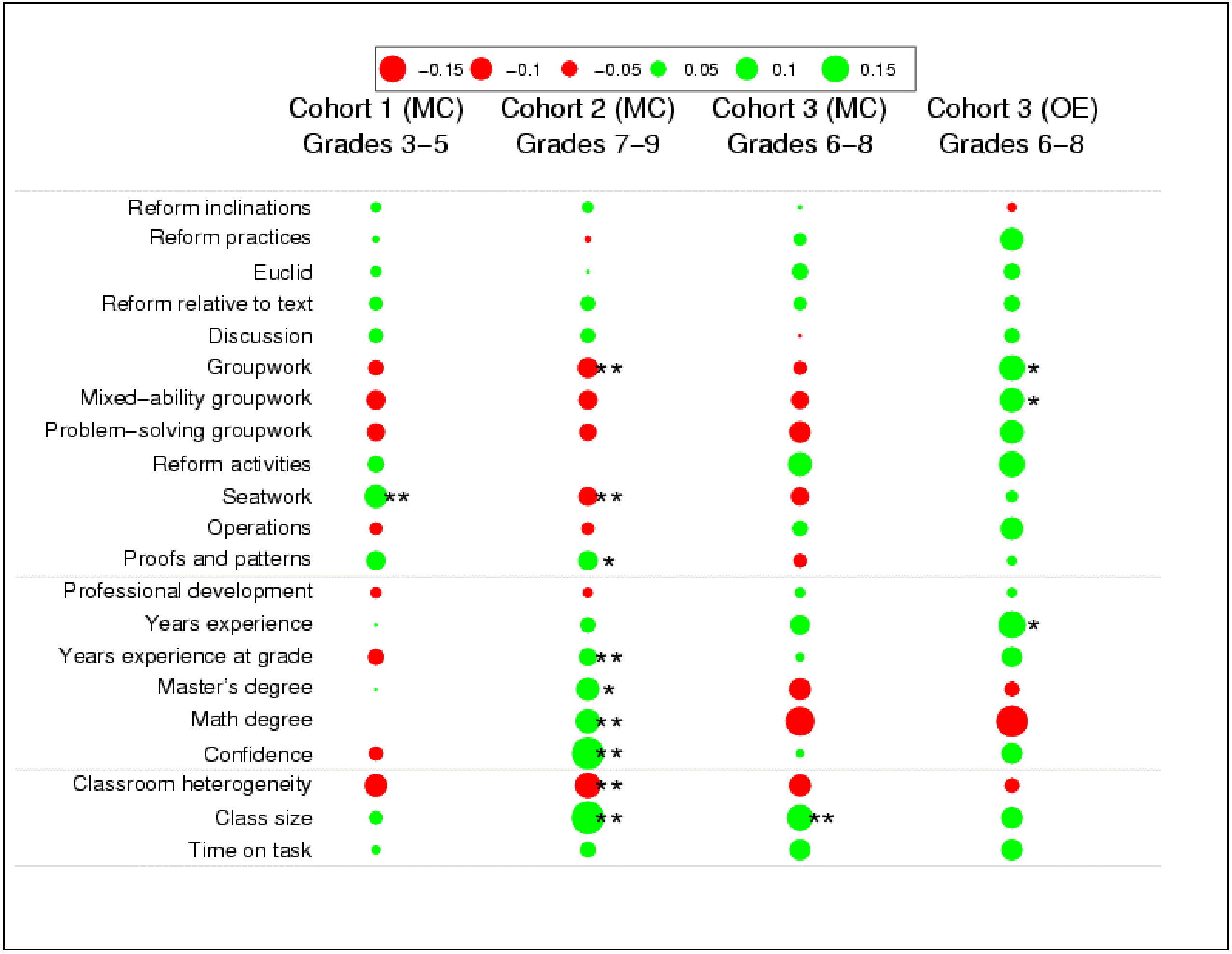

Figure 3. Relationships between Reform Teaching and Student Achievement (Cumulative Effects)

Notes:
* significant at the .05 level.
** significant at the .01 level.

As shown in Figures 2 and 3, few of the teacher-related variables reached statistical significance, including our more innovative vignette-based measures. For Cohort 1, more time spent in groups, in mixed-ability groups, and in groups solving new problems was negatively related to achievement. These were single year effects, such that greater time spent in groupwork-related activities did not appear to be associated with mathematics performance in subsequent years. In contrast, more time spent reading from the textbook and engaging in other seatwork-related activities (e.g., completing worksheets) was positively associated with test scores. Greater focus on seatwork not only had positive relationships with the current year scores, but the effects persisted to show positive associations with test scores in later years as well. Additionally, classroom composition was related to achievement, such that heterogeneous classrooms were associated with lower mathematics performance.

Most of the instructional practices variables that reached statistical significance in Cohort 2 had cumulative effects on performance. More time spent in groupwork-related activities was negatively related to mathematics scores in subsequent years. A similar pattern was observed for Seatwork such that greater emphasis on low-reform activities was negatively associated with mathematics scores in later years. On the other hand, more time spent covering proofs and patterns appeared to be positively related with mathematics achievement, both within the current year and in future years. Of the teacher characteristics, teaching experience, both in total and at grade level, were associated with test scores, as was holding a masters degree, having a mathematics-specific undergraduate degree, and being very confident in their mathematics knowledge. These characteristics had positive effects, both within a given year and beyond. Larger class size was also positively related to mathematics scores, both within the current year, as well as on subsequent years. Additionally, teaching heterogeneous classrooms was associated with lower mathematics achievement. This finding is similar to that observed for Cohort 1.

For Cohort 3, none of the instructional practices variables were related to SAT-9 multiple-choice scores, but there were relationships between class size and mathematics-specific undergraduate degree and mathematics achievement. Unlike Cohort 2, students of teachers who had a mathematics-specific undergraduate degree showed poorer performance than their peers whose teachers had less mathematics training. However, similar to Cohort 2, teaching a larger class was positively related to mathematics scores during subsequent years.

Some instructional practice measures reached statistical significance when scores on an open-ended test served as the outcome measure. More time spent in groups, in groups comprised of mixed-ability students, and in groups solving new problems was positively associated with SAT-9 open-ended scores. All three had effects on current year scores, and the former two practices had a cumulative effect such that greater frequency of groupwork and of mixed-ability groupwork was associated with better mathematics performance in later years.

Total teaching experience and grade-level teaching had significant positive relationships with performance on the open-ended test, with the former also having cumulative effect. Additionally, having a mathematics-specific degree was associated with poorer performance on open-ended items. This is a counterintuitive finding and merits more attention.

## Discussion

The purpose of this study was to examine the relationships between teachers' reported use of instructional practices in mathematics and student achievement. Previous research that has explored this issue has found that reform teaching is positively associated with student achievement, but the effects are weak. Two possible explanations for the weak effects from past studies are the limitations arising from the use of surveys to measure instruction, and the brief exposure to reform teaching. In our study, we examined the relationship between instruction and achievement using a variety of measures, including vignettes, which we hoped captured aspects of teaching that could not be captured via surveys. Additionally, we examined relationships over a three-year period so as to better capture the effects of reform teaching.

Regardless of whether instruction was assessed via surveys or vignettes, few of the teacher-level measures showed significant relationships with student achievement. Across all three cohorts, total teaching experience was positively associated with test scores. Grade-level teaching experience, masters degree, and confidence in mathematics knowledge also tended to show positive associations, but mathematics-specific degree had inconsistent relationships, showing positive relationships in some cohorts, and negative relationships in others. The negative associations may represent a purposive decision by districts or schools to place their most qualified teachers in lower-performing classrooms in an attempt to raise achievement.

Time spent working in groups, in groups with students of mixed abilities, and in groups solving new problems also showed mixed relationships, as they were negatively associated with multiple-choice scores (although not always reaching statistical significance), but positively associated with open-ended scores. This trend is particularly compelling given that the effects of groupwork, mixed-ability groupwork, and groupwork involving problem solving are reversed for multiple-choice and open-ended tests in Cohort 3, where the same group of students take both formats. This finding attests to the potential differential sensitivity of the different formats, and supports the premise that reform teaching may have larger effects on open-ended measures than on multiple-choice tests.

One of the key features of this study is the exploration of cumulative effects of reform teaching. It is widely believed that students must be exposed to reform practices for more than a single year before the effects of these practices on achievement can become clearly evident. Our study found some evidence of cumulative effects. In the middle school years, greater emphasis on Seatwork had negative relationships with mathematics performance in later years. This result, however, was reversed from that observed for elementary grades, where greater time spent on Seatwork was positively related with subsequent mathematics performance. This suggests that reform teaching may have differential effects at different stages within students' learning experiences. At the middle grades, working in groups had negative relationships with subsequent multiple-choice scores, but had positive relationships with subsequent open-ended scores. Additionally, working in mixed-ability groups and emphasis on proof and patterns (a topic thought by advocates of reformed pedagogy to be more reform oriented) had positive relationships with test scores in later years.

Taken together, the findings provide some support that certain aspects of reform pedagogy may be associated with improved student achievement, but the effects are quite weak, may take several years before the relationships are manifested, and may be evident only on certain formats. Several reasons may help explain the weak effects. As noted by Rowan, Correnti, and Miller (2002), if variation in test scores is largely attributable to factors other than teaching practices, it is unlikely that any individual teaching practice measure will show large relationships. Indeed, in our study, student background and residual error accounts for the vast majority of the variation in test scores, leaving little variance that can be attributable to reform pedagogy.

Weak effects may also stem from lack of measurement quality from our indicators of reform instruction. Despite our attempts to use vignettes to measure instruction in innovative ways, we could not address how reform instruction was actually implemented. Other methods that provide more refined measures of instruction, such as observations, may find stronger relationships between reform practices and student outcomes. We are currently undertaking an analysis in which we explore how observers' ratings of the "reformedness" of the classrooms relate to student achievement.

Finally, the weak relationships may be attributed to the lack of alignment between the achievement measures and reform pedagogy and curricula. Our findings suggest that open-ended measures, which are thought to better capture the effects of reform teaching, may indeed be more sensitive to reform teaching than multiple-choice measures. However, it is important to recognize that our open-ended measure is still somewhat limited and probably fails to capture all the aspects of higher-order thinking that reform pedagogy is intended to promote. We are currently pursuing analysis conducted on subscores, including performances that reflect problem solving and computation. This will allow us to explore whether reform teaching is particularly effective at enhancing "higher order" thinking skills.

# References

Achieve (2004). Do Graduation Tests Measure Up? Washington, DC: Author.

American Association for the Advancement of Science (1993). Benchmarks for science literacy. New York: Oxford University Press.

Cohen, D.K., & Hill, H.C. (2000). Instructional policy and classroom performance: The mathematics reform in California. Teachers College Record, 102(2), 294-343.

Gamoran, A., Porter, A.C., Smithson, J., & White, P. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. Educational Evaluation and Policy Analysis, 19 (4), 325-338.

Hamilton, L.S., McCaffrey, D.F., Stecher, B.M., Klein, S.P., Robyn, A., & Bugliari, D. (2003). Studying large-scale reforms of instructional practice: An example from mathematics and science. Educational Evaluation and Policy Analysis, 25(1), 1-29.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics. 5: p 299-314.

Kim, J., Crasco, L., Blank, R., Smithson, J. (2001) Survey results of urban school classroom practices in mathematics and science: 2000 Report. Norwood: Systemic Research, Inc.

Le, V., Stecher, B.S., Hamilton, L.S., Ryan, G., Williams, V., Robyn, A., Alonzo, A. (2003). Vignette-based surveys and the Mosaic II project. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Lockwood J.R., Doran H. and McCaffrey D.F. (2003). Using R for estimating longitudinal student achievement models. The R Newsletter, 3:3 17-23.

Lockwood J.R., McCaffrey D.F., Mariano L.T. and Setodji, C. (2004). Bayesian methods for scalable multi-subject value-added assessment. Currently under review by Journal of Educational and Behavioral Statistics.

Mayer, D.P. (1998). Do new teaching standards undermine performance on old tests? Educational Evaluation and Policy Analysis, 20, 53-73.

McCaffrey D.F., Lockwood J.R., Koretz D., Louis T.A. and Hamilton L. (2004). Models for value-added modeling of teacher effects. Journal of Educational and Behavioral Statistics, 29:1 67-101.

McCaffrey D.F., Lockwood J.R., Mariano L.T. and Setodji, C. (2005). Challenges for value-added assessment of teacher effects, proceedings of the

University of Maryland Conference on Value Added Assessment.

Mullens, J., & Gayler, K. (1999). Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction. (NCES 1999-08). Washington, DC: National Center for Education Statistics.

National Council of Teachers of Mathematics. (1989). Curriculum and Evaluation Standards for School Mathematics. Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics (2000). Principles and Standards for School Mathematics. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1996). National Science Standards. Washington, DC. National Academy Press.

Pinheiro, J. and Bates, D.M. (2000). Mixed-Effects Models in S and S-PLUS. New York: Springer.

Resnick, L., & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M.O' Conner (Eds.). Changing assessments: Alternative views of aptitude, achievement, and instruction. Boston, Kluwer, 37-75.

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. The Teachers College Record, 104 (8), 1525-1567.

Sanders, W.L. and Saxton, A.M. and Horn, B.P. (1997). The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment. In Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure? J. Millman (ed). Corwin Press, Inc., Thousand Oaks, CA, p 137-162.

Saxe, G.B., Gearhart, M., & Seltzer, M (1999). Relations between classroom practices and student learning in the domain of fractions. Cognition and Instruction, 17(1), 1-24.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. New York: Chapman & Hall.

Stein, M.K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. Educational Research and Evaluation, 2, 50-80.

Swanson, C.B., & Stevenson, D.L. (2002). Standards-based reform in practice: Evidence on state policy and classroom instruction from the NAEP state assessments. Educational Evaluation and Policy Analysis, 24(1), 1-27.

Thompson, D, & Senk, S.  (2001). The Effects of Curriculum on Achievement in Second Year Algebra: The Example of the University of Chicago Mathematics Project. Journal for Research in Mathematics Education, 32 (1), 58-84.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. Education Policy Analysis Archives, 10(12), http://epaa.asu.edu/epaa/v10n12/.

## Appendix A: Student Demographic Characteristics for Each Site

Year 1:

| Demographic Characteristics | Cohort 1 | Cohort 2 | Cohort 3 |
|---|---|---|---|
| Racial/Ethnic Group | | | |
| African-American | 3.78 | 3.27 | 45.60 |
| Hispanic | 38.73 | 23.47 | 8.00 |
| Other | 5.05 | 6.36 | 22.00 |
| White | 52.44 | 66.90 | 24.40 |
| Female | 49.51 | 50.00 | 51.03 |
| Limited English Proficient | 21.86 | 9.46 | 20.63 |
| Two-Parent Family | 73.93 | 76.53 | -- |
| Eligible for Free or Reduced Price Lunches | 56.88 | 36.22 | 71.94 |
| Special education | -- | -- | 11.09 |
| Gifted | -- | -- | 18.34 |

Year 2:

| Demographic Characteristics | Cohort 1 | Cohort 2 | Cohort 3 |
|---|---|---|---|
| Racial/Ethnic Group | | | |
| African-American | 4.20 | 3.16 | 44.09 |
| Hispanic | 42.43 | 22.10 | 9.68 |
| Other | 6.90 | 5.61 | 20.22 |
| White | 46.47 | 69.13 | 26.01 |
| Female | 50.73 | 50.50 | 51.23 |
| Limited English Proficient | 17.06 | 4.53 | 20.18 |
| Two-Parent Family | 76.13 | 78.12 | -- |
| Eligible for Free or Reduced Price Lunches | 62.34 | 33.21 | 68.69 |
| Special education | -- | -- | 10.79 |

Year 3:

| Demographic Characteristics | Cohort 1 | Cohort 2 | Cohort 3 |
|---|---|---|---|
| Racial/Ethnic Group | | | |
| African-American | 3.88 | 2.25 | 46.47 |
| Hispanic | 42.69 | 20.12 | 15.26 |
| Other | 6.8 | 5.09 | 15.82 |
| White | 46.63 | 72.54 | 22.45 |
| Female | 50.09 | 50.16 | 53.07 |
| Limited English Proficient | 14.61 | 4.48 | 78.42 |
| Two-Parent Family | 76.47 | 79.65 | -- |
| Eligible for Free or Reduced Price Lunches | 59.25 | 29.03 | 65.73 |
| Special education | -- | -- | 10.88 |

## Appendix B: Example of Sixth- and Seventh-Grade Mathematics Vignette

### Teaching Scenarios

**Instructions.** The following questions contain brief "scenarios" or stories that describe teaching situations and ask how you would respond in each case. We know there are many ways to teach mathematics, and you may not organize your lessons in the manner that is presented. Please answer as if you were in the situation that is described.

The scenarios are brief and do not describe every detail. Assume that other features are similar to your current school and your current students.

Please do the following:

a. Read the scenario.
b. Read the first possible option.
c. Circle the response that shows how likely you would be to do this option.
d. Read the next option and circle your response.
e. Repeat the process until you have responded to all the options.
f. Please evaluate each of the options independently of the others. In other words, you may select as many 1's (or 2's or 3's or 4's) as you like.

### SCENARIO I: U.S. STANDARD MEASUREMENT UNITS (4 QUESTIONS)

Imagine you are teaching a sixth-grade class. You are about to begin a week-long unit on converting units of length within the U.S. standard measurement system. Your students have had experience using rulers to measure objects in feet and inches, and are also familiar with yards and miles as units of measurement.

1. You are ready to start the unit on conversion. How likely are you to do each of the following activities **to introduce** the unit?

***(Circle One Response in Each Row)***

| | **Very unlikely** | **Somewhat unlikely** | **Somewhat likely** | **Very likely** |
|---|---|---|---|---|
| a. Ask students what they know about inches and feet | 1 | 2 | 3 | 4 |
| b. Have students use rulers / yardsticks to measure lengths of objects in the classroom(e.g., desks or chairs) | 1 | 2 | 3 | 4 |
| c. Demonstrate how to solve problems such as converting 22 inches into feet and inches | 1 | 2 | 3 | 4 |

|  |  |  |  |  |
|---|---|---|---|---|
| d. Display an equivalence table on the board that provides conversions among inches, feet, yards, and miles | 1 | 2 | 3 | 4 |
| e. Have students solve a problem such as estimating the width of the classroom in inches | 1 | 2 | 3 | 4 |
| f. Explain the procedures for converting units (e.g., multiply by 12 when converting feet into inches) | 1 | 2 | 3 | 4 |
| g. Lead a classroom discussion about the problems of measuring if you only had one unit of measurement (e.g., foot) | 1 | 2 | 3 | 4 |
| h. Have students work in pairs or groups to measure the size of each other's feet | 1 | 2 | 3 | 4 |

2. You are at the midpoint of your unit on conversion, and most students appear to understand the procedures. Next, you pose more complex problems. You ask your students how many inches are in 9 yards, 2 feet.

   When most students appear to have completed the task, you ask Joey if he will share his solution. He replies that 9 yards, 2 feet is close to 10 yards, which is 360 inches, so he subtracted, and found the answer to be 358 inches.

   You know, however, that the correct answer is 348 inches.

   After praising Joey for knowing that 9 yards, 2 feet is close to 10 yards, what do you do next? How likely are you to do each of the following?

***(Circle One Response in Each Row)***

| | **Very unlikely** | **Somewhat unlikely** | **Somewhat likely** | **Very likely** |
|---|---|---|---|---|
| a. Ask Joey, "How did you get from 10 yards to 358 inches?" | 1 | 2 | 3 | 4 |
| b. Pose another similar problem for the class | 1 | 2 | 3 | 4 |
| c. Suggest that Joey use a ruler to solve the problem | 1 | 2 | 3 | 4 |
| d. Tell Joey that he was close, but the answer is 348 | 1 | 2 | 3 | 4 |
| e. Call on another student who you expect will give you the right answer | 1 | 2 | 3 | 4 |
| f. Tell Joey that his answer is close, and ask if anyone can help him with his | 1 | 2 | 3 | 4 |
| g. Ask the class, "Did anyone else use a similar method but get a different | 1 | 2 | 3 | 4 |
| h. Explain that one foot (12 inches) should have been subtracted | 1 | 2 | 3 | 4 |
| i. Ask the class, "Are there any other answers?" | 1 | 2 | 3 | 4 |
| j. Give Joey another problem similar to this one, and ask him to solve it | 1 | 2 | 3 | 4 |

3. You are almost at the end of the unit on conversion. You ask students to work in pairs or groups to solve the following problem.

$$\begin{array}{r} 5\text{ ft }\ 3\text{ in} \\ -\ \underline{3\text{ ft }\ 6\text{ in}} \end{array}$$

After working on the problem for a while, you ask each group if they will share their work.

The first group responds that the answer is 1 feet 9 inches. They explain that they converted 5 feet 3 inches to 4 feet 15 inches, then subtracted.

The second group gives the same answer, and explains that they drew the distances on the floor using a yardstick and measured the non-overlapping portion.

How likely are you to do each of the following in response to these two explanations?

***(Circle One Response in Each Row)***

| | **Very unlikely** | **Somewhat unlikely** | **Somewhat likely** | **Very likely** |
|---|---|---|---|---|
| a. Ask the class if they can think of other ways to solve the problem | 1 | 2 | 3 | 4 |
| b. Think of a new problem in which the two methods are not equally effective and ask the groups to solve it | 1 | 2 | 3 | 4 |
| c. Tell them that they are both right and move on to the next problem | 1 | 2 | 3 | 4 |
| d. Tell them that it is better to use the first group's method because it can be applied to any similar distance problems | 1 | 2 | 3 | 4 |
| e. Have a classroom discussion about the differences between the two approaches | 1 | 2 | 3 | 4 |

4. If you were to teach a unit on conversion of lengths to the target class, how much emphasis would you place on each of the following learning objectives?

***(Circle One Response in Each Row)***

| | **No emphasis** | **Slight emphasis** | **Moderate emphasis** | **Great emphasis** |
|---|---|---|---|---|
| a. Students will understand that similar principles of conversion apply in other situations (e.g., when measuring area, volume) | 1 | 2 | 3 | 4 |
| b. Students will be able to use rulers and yardsticks to solve conversion problems (e.g., show why there are 48 inches in 1 yard, 1 foot) | 1 | 2 | 3 | 4 |
| c. Students will be able to solve mixed-unit problems (e.g., converting 1 yard 2 feet to inches) | 1 | 2 | 3 | 4 |
| d. Students will be able to estimate the lengths of objects in their neighborhoods (e.g., cars) | 1 | 2 | 3 | 4 |
| e. Students will know how to convert among inches, feet, and yards | 1 | 2 | 3 | 4 |
| f. Students will know which units of measurement are appropriate for measuring objects or distances of differing length | 1 | 2 | 3 | 4 |

## Appendix C: Example of Judgmental Reform Ratings for the Response Options from Sixth-and Seventh-Grade Mathematics

| Item No. | Reform Rating |
|---|---|
| Vignette 1<br>(Measurement System) | |
| 1a | 3 |
| 1b | 3 |
| 1c | 1 |
| 1d | 1 |
| 1e | 3 |
| 1f | 1 |
| 1g | 4 |
| 1h | 3 |
| 2a | 4 |
| 2b | 2 |
| 2c | 2 |
| 2d | 1 |
| 2e | 1 |
| 2f | 2 |
| 2g | 3 |
| 2h | 1 |
| 2i | -- |
| 2j | 2 |
| 3a | 3 |
| 3b | 4 |
| 3c | 1 |
| 3d | 2 |
| 3e | 4 |
| 4a | 4 |
| 4b | 2 |
| 4c | -- |
| 4d | 3 |
| 4e | -- |
| 4f | 4 |

# Appendix D: Survey and Log Scale Items

Survey

**Instructional Practices**

Reform Practices

*On average throughout the year, approximately how often do you employ the following teaching strategies during your mathematics lessons?*

Use open-ended questions
Require students to explain their reasoning when giving an answer
Encourage students to communicate mathematically
Encourage students to explore alternative methods for solutions
Help students see connections between mathematics and other disciplines

*On average throughout the year, approximately how often do your students take part in the following activities as part of their mathematics lessons?*

Share ideas or solve problems with each other in small groups
Engage in hands-on mathematics activities
Work on extended mathematics investigations (a week or more in duration)
Record, represent, or analyze data

Text Reform

*How does your emphasis on each of the following student mathematics activities compare to that of our primary textbook or published curriculum?*

Solving real-life problems
Engaging in hands-on activities (e.g., working with manipulatives)
Explaining answers, justifying solutions
Discussing multiple solutions
Solving open-ended problems

Mathematical Processes

*Below is a selected list of processes that you may emphasize in teaching mathematics to your fourth-grade students. How much emphasis do you place on each of the following?*

Proof and justification/verification (e.g., using logical argument to demonstrate correctness of mathematical relationship)
Problem solving (e.g., finding solutions that require more than merely applying rules in a familiar situations)
Communication (e.g., expressing mathematical ideas orally and in writing)
Connections (e.g., linking one mathematical concept with another; applying math ideas in contexts outside of math)
Representations (e.g., using tables, graphs and other ways of illustrating mathematical relationships)

**Teacher Background**

Certification

*What type of teaching certification do you hold?* (Circle one)

Not certified
Temporary, provisional, or emergency certification (requires additional coursework before regular certification can be obtained)
Probationary certification (the initial certification issued after satisfying all requirements except the completion of a probationary period)
Regular or standard certification

Confidence

*With respect to the mathematics/science that you are asked to teach, how confident are you in your mathematical/scientific knowledge?* (Circle one)

Not confident at all
Somewhat confident

Moderately confident
Very confident

Masters Degree
*What is the highest degree you hold?*
BA or BS
MA or MS
Multiple MA or MS
PhD or EdD
Other

Math Degree
*Did you major in mathematics/science or a mathematics/science-intensive field for your Bachelor's degree?*
*Did you minor in mathematics or a mathematics/science-intensive field for your Bachelor's degree?*

Professional Development
*In the past 12 months, how much time have you spent on professional development activities that focused on the following aspects of teaching mathematics?*
In-depth study of mathematics content
Methods of teaching mathematics
Use of particular mathematics curricula or curriculum materials
Students' mathematical thinking
Mathematics standards or framework
Mathematics assessment/testing
Use of educational technology for mathematics/science instruction

Experience
*Including this year, how many years have you taught on a full-time basis?*

Experience at Grade
*Including this year, how many years have you taught third-graders?* (*third-grade version)3*

**Curriculum**

Operations
*Indicate the approximate amount of time you will spend on each content area this school year*?
*(Third-grade version)*
Addition/subtraction of whole numbers
Multiplication/division of whole numbers

*(Sixth- and seventh-grade version)*
Operations with signed whole numbers

Proof and Patterns
*Indicate the approximate amount of time you will spend on each content area this school year*?
*(Third- and sixth-grade version)*
Proof and justification/verification
Patterns/functions/algebra
*(Seventh-grade version)*
Proof and justification/verification
Patterns/functions
Algebra

Log

**Instructional Practices**

Discussion

*How much time did students spend on each of these activities during today's mathematics lesson?*
Explain their thinking about mathematical problems
Lead discussion of a mathematics topic
*How much time did you spend on each of these activities during today's mathematics lesson?*
Ask open-ended questions and discuss solutions to them
Ask questions of individuals to test for understanding

Groupwork
*How long did students work in groups during today's mathematics/science lesson?*

Mixed-ability Groupwork
*If groups were used, what share of the group time was used in the following ways?*
Working in groups of mixed ability
Problem-solving Groupwork
*If groups were used, what share of the group time was used in the following ways?*
Solving new problems together as a group

Reform Activities
*How often did the following occur during today's mathematics lesson?* (Mathematics version)
Students engaged in debate/discussion about ways to solve a problem
Student restated another student's ideas in different words
Students demonstrated different ways to solve a problem
Students explored a problem different from any they had solved previously
Students worked on an activity or problem that will take more than one period
Teacher or student connected today's math topic to another subject (e.g., social studies)

Seatwork
*How much time did students spend on each of these activities during today's mathematics lesson?*
Read from textbook or other materials
Complete worksheets or problem sets from text

**Classroom Context**
Hours of Weekly Instruction
*In a typical week, how many hours of mathematics instruction do students in your class receive?*

Classroom Heterogeneity
*How would you describe the variation in mathematics ability of students in your class? (Circle one)*
Fairly homogenous and low in ability
Fairly homogenous and average in ability
Fairly homogenous and high in ability
Heterogeneous with a mixture of two or more ability levels

Time on Task
*How long was today's mathematics/science lesson?*
*How much mathematics/science time was lost to discipline issues?*

## Appendix E: Strategy for Missing Data in Longitudinal Analyses

Our longitudinal analyses required making decisions about various kinds of missing data. This document presents a framework for the missing data and the multistage process by which missing data were imputed. All of the discussion below pertains to a single cohort; the procedure was carried out analogously for all cohorts.

*Notation*

$\theta$ represents the entire collection of unknown parameters of interest under some statistical model, such as coefficients of practice variables and variance components.

$\boldsymbol{y}_o$ and $\boldsymbol{y}_m$ are, respectively, the observed and missing test scores. Throughout this discussion we use $\boldsymbol{y}_o$ and $\boldsymbol{y}_m$ to refer to the test scores on the left hand side of the model equations; that is, those test scores that are treated as response variables. The scores from year 0 that are used as controls are treated as student covariates.

$\boldsymbol{x}_0$ and $\boldsymbol{x}_m$ are the observed and missing student-level covariates (including year 0 scores). For simplicity we assume that all student-level covariates are time-invariant. This is not exactly true for a few of the covariates such as free lunch status and LEP status, but the percentage of students with observed time variation on these covariates was extremely small and so developing an explicit model for time variation seemed both difficult and unnecessary. However, the general methods discussed below can be adapted to handle time-varying covariates for students.

$\boldsymbol{v}_0$ and $\boldsymbol{v}_m$ are the observed and missing teacher-level variables (such as practices and teacher background variables).

$\boldsymbol{z}_o$ and $\boldsymbol{z}_m$ are the observed and missing student-teacher links. That is, an observed link is the information that a particular observed student was linked to a particular observed, responding teacher in a given year. Missing links occur when the student was not observed in a given year (e.g. migrated into or out the cohort). Missing links also occur when students are linked to non-responding teachers, because we do not observe classroom groupings unless teachers respond to surveys.

We structure the missing data issues and solutions in the Bayesian framework because it is clearer (indeed multiple imputation approaches to missing data used in classical inference rely on Bayesian arguments). In the Bayesian framework we are interested in the posterior distribution of the parameter given the observed data. For notational convenience let $\boldsymbol{d}_o = (\boldsymbol{y}_o, \boldsymbol{x}_o, \boldsymbol{v}_o, \boldsymbol{z}_o)$ be all of the observed data. Then we are interested in $p(\boldsymbol{\theta}|\boldsymbol{d}_0)$. This "observed data posterior distribution" can be thought of as the "complete data posterior distribution" (the posterior distribution we would have obtained had all the data been observed) integrated over the predictive distribution of the missing data:

$$p(\boldsymbol{\theta}|\boldsymbol{d}_0) = \iiiint p(\boldsymbol{\theta}, \boldsymbol{y}_m, \boldsymbol{x}_m, \boldsymbol{v}_m, \boldsymbol{z}_m|\boldsymbol{d}_0)d\boldsymbol{y}_m d\boldsymbol{x}_m d\boldsymbol{v}_m d\boldsymbol{z}_m \qquad (1)$$

$$= \iiiint p(\boldsymbol{\theta}, \boldsymbol{y}_m, \boldsymbol{x}_m, \boldsymbol{v}_m, \boldsymbol{z}_m, \boldsymbol{d}_0)p(\boldsymbol{y}_m, \boldsymbol{x}_m, \boldsymbol{v}_m, \boldsymbol{z}_m|\boldsymbol{d}_0)d\boldsymbol{y}_m d\boldsymbol{x}_{\boldsymbol{m}} d\boldsymbol{v}_{\boldsymbol{m}} d\boldsymbol{z}_{\boldsymbol{m}} \qquad (2)$$

Equation 2 is intuitively the same idea as multiple imputation from the classical perspective -- in that case, the complete-data likelihood inferences are averaged over some number of realizations from the predictive distribution of the missing data given the observed data. The first term in the integrand in Equation 2 is the complete data posterior, and that needs to be averaged over the distribution of the missing data given the observed data (the second term in the integrand).

A fully Bayesian solution to the missing data problem would specify a full joint distribution for the parameters, the missing data and the observed data. Given the complexities of the models we would be fitting to the complete data, this fully Bayesian approach was not feasible. The logic of multiple imputation is to simplify this fully Bayesian approach somewhat to produce "complete" data sets outside of the model fitting procedure, to which standard modeling tools can be applied. Then the resulting inferences are combined over the imputations to result in inferences that account for the missing data. We pursue that same strategy here.

As noted, we used specialized software for fitting Bayesian value added models for our analyses (Lockwood, McCaffrey, Mariano and Setodji, 2004). That software has built-in capability of carrying out fully Bayesian data augmentation for missing test scores. Thus our imputation procedures need to deal only with ($\boldsymbol{x}_m, \boldsymbol{v}_m, \boldsymbol{z}_m$); the missing test scores are handled naturally as part of the algorithm used to sample the posterior distribution of $\theta$. That is, it is more convenient for us to write the observed data posterior distribution as:

$$p(\theta | d_0) = \int\int\int \left[ \int p(\theta \,| x_m, v_m, z_m, d_0) d\, y_m \right] p(x_m, v_m, z_m | d_0) dx_m dv_m dz_m \qquad (3)$$

The inner-most integral inside the square brackets is handled by the software, with the missing scores being imputed by data augmentation. Thus the multiple imputation scheme needs only to produce samples from $p(\boldsymbol{x}_m, \boldsymbol{v}_m, \boldsymbol{z}_m | \boldsymbol{d}_0)$, each of which provides a complete realization of the data aside from the missing test scores. The Bayesian analysis can be applied to these realizations, and the resulting inferences combined. We now discuss a strategy for specifying and sampling from $p(\boldsymbol{x}_m, \boldsymbol{v}_m, \boldsymbol{z}_m | \boldsymbol{d}_0)$. We can achieve this by sampling sequentially from $p(\boldsymbol{z}_m | \boldsymbol{d}_0)$, $p(\boldsymbol{x}_m | \boldsymbol{d}_0, \boldsymbol{z}_m)$ and $p(\boldsymbol{v}_m | \boldsymbol{d}_0, \boldsymbol{z}_m, \boldsymbol{x}_m)$. For each cohort we obtained 25 such samples, and all of the inferences that we report are aggregated across these samples. Using 25 imputed data sets is larger than the 5 or 10 that is typically used. However, given the complexity of the missing data structure, it seemed reasonable to use a larger number of imputed datasets to better represent the missing data space.

*Missing links* $p(\boldsymbol{z}_m | \boldsymbol{d}_0)$:

All missing links are to teachers who are unknown or unobserved. That is, in no case is it appropriate to assign a missing link for a student to an observed teacher. In the case where we have observed scores for a student in a particular year but no teacher link, that student was linked to a non-responding teacher. We thus know the school and grade of the teacher but nothing else about them other than they did not respond to the survey. In the case where the student was not in the dataset at all for a particular year, it is likely that the student was not in the district at all during that year, and we know nothing about teachers outside of the district. Thus the observed data $d_0$ provide no information about assigning missing links and $p(\boldsymbol{z}_m|\boldsymbol{d}_0) = p(\boldsymbol{z}_m)$.

To deal with missing links, we need to consider the students' patterns of observed scores. Consider a single student. Because of the longitudinal and cumulative nature of the models we will be fitting, any missing links that occur prior to the last observed score for a student (i.e. "pre-dropout") must be assigned. This is because in order to specify the mean structure for the current year score, we need complete information from prior years. However, missing links that occur after the last observed score for a student are irrelevant to the desired inferences. For example, if a student has observed links and observed scores in years 1 and 2, but neither in year 3, the link in year 3 is irrelevant because the year 3 score is unobserved and thus does not need to be reconciled with any exposure to particular practices, etc. If the year 3 score were observed but the link missing, we would need to assign a link.

This suggests the following strategy, discussed in detail in McCaffrey et al, 2005. We assign all missing links that occur after the last observed score for a student to a ``dummy teacher" that has a zero teacher effect and values of zero for all teacher-level variables. On the other hand, missing links that occur prior to the last observed score are assigned to what we call a ``pseudo-teacher." Pseudo-teachers are unique to a particular missing link, occurring at the level of year within student. With respect to the model, they have all of the same qualities as observed teachers, with teacher effects and values of all teacher-level covariates (which will need to be imputed later; see below). We have found that in practice, models that assume that the teacher effects of the pseudo-teachers share a variance component with the actual observed teachers for that grade (and subject) perform better than models that estimate a separate variance component for these effects. Thus, to summarize, missing links will be assigned either to a null teacher with zero teacher effect, or to pseudo-teachers that function like actual teachers.

*Missing student covariates* $p(\boldsymbol{x}_m|\boldsymbol{d}_0, \boldsymbol{z}_m)$:

Recall that we are assuming that the student-level covariates are time-invariant, so that imputation occurs at the level of the student rather that at the student by year. We used the imputation package "norm" for the R statistical environment to do this imputation, which fits a large multivariate normal model to all of the observed data and uses that to predict values of the missing covariates. Dichotomous variables are treated continuously and are then rounded back to dichotomous values, and this has been shown to result in reasonable imputations even though the normal model does not hold. We fit a joint distribution to all student covariates, their average standardized observed score in year 1-3, as well as the averages of the teacher variables to which they are exposed. We used this estimated joint distribution to sample missing student covariates from their joint predictive distribution given the observed data.

*Missing teacher variables* $p(\boldsymbol{v}_m|\boldsymbol{x}_m, \boldsymbol{d}_0, \boldsymbol{z}_m)$:

This is the most difficult set of imputations because it requires two stages to be done most effectively. We needed to separate unit non-response (pseudo-teachers) from item non-response (imputing missing teacher variables for the observed, partially responding teachers). Fortunately the item non-response is pretty minimal, with only about 4 percent of responding teachers not responding completely. For the most part, teachers who responded at all responded relatively completely. For the small amount of item-nonresponse that we had, we again used the "norm" package to fit a joint distribution of the teacher variables at the teacher level, along with classroom aggregates of student scores and demographics. We used samples from this joint distribution to fill in plausible realizations of missing item-level data.

The more difficult part was specifying the entire suite of missing teacher-level variables for the pseudo-teachers. Below we discuss several options, ultimately concluding that a kind of hot-deck imputation method seemed most appropriate.

*Option 1: Do the same as with the item-level missing data for the real teachers:* This seemed particularly tenuous. We could not rely on classroom aggregates of student scores and demographics because each missing link has a unique pseudo-teacher. Directly plugging in an individual's student-level covariates as if they were classroom averages would likely tend to produce imputations of the teacher-level variables that are too extreme (because the imputations work just like regressions and we will be plugging in extreme covariates). Also we could not rely on the values of other observed teacher-level variables because there are none. In filling in the values of missing teacher variables for the observed teachers, we at least have some (and in most cases most) of the actual teacher level variables for that teacher upon which to make an educated guess. The fact that our model-based estimate of the joint distribution of all teacher-level variables is probably very noisy (since there is a large number of variables relative to the number of teachers) is probably not of too much consequence when filling in a few variables, but might behave erratically when trying to reproduce the entire suite of variables for a teacher. The good side of this approach is that it conditions as well as possible on the observed data and at least to some extent preserves the joint distribution of the teacher variables.

*Option 2: Impute missing teacher variables one at a time*. This is sort of like the above except it might be possible to make this approach more stable and result in individual teacher variable imputations that are better. On the other hand it destroys the correlation among the individual teacher variables -- the resulting vector of teacher variables for a pseudo-teacher may look nothing like an actual teacher. Hybrid approaches that include a small number of teacher variables in the predictive model for each of the other teacher variables are also possible.

*Option 3: Hot deck imputation*. This is used to refer to any one of a variety of ways of assigning to each pseudo-teacher (the ``recipient'') the entire vector of covariates from an observed teacher (the ``donor''). A pro of this method is that it preserves the joint distribution of the teacher variables by sampling from their empirical joint distribution based on the actual observed teachers. The simplest and most naive method of doing this would be pure random assignment without respect to anything. More structure could be imposed by restricting assignments to the same grade, school, or both. This is potentially misleading because it might underestimate the true variability of the unobserved teachers represented by the pseudo-teachers (indeed, even if we knew the actual grade and school for a missing teacher link, as might be possible in the case of non-responding teachers, it may not be prudent to assume that the characteristics of the non-responding teachers are represented by those who responded. This is a problem in general but potentially even more of a problem when we overly restrict the set of teachers who may serve as a donor of covariates for a pseudo-teacher).

A con of this method is that without additional structure relating assignments to test scores, we might attenuate correlations between teacher variables and outcomes. One way to at least partially sidestep this is to choose the donor based on the student scores and other student characteristics. That is, teachers who generally teach students more like the student in question are preferentially chosen as donors. The following is the procedure that we implemented to perform ``informed" hot deck imputation. For each real teacher, we calculate the average student characteristics (percentages of various ethnicities, of students participating in free- or reduced-price lunch program, of students classified as limited English proficient, etc) as well as the average standardized test scores *for all years* of students linked to that teacher. For example, for a third grade teacher, we calculate not only the mean test score for students in his/her third grade class, but also the mean test score for those same students in other years. This is important because most pseudo-teachers requiring a covariate donor do not have an associated observed score for that year, but always have at least one for another year (else the student would not be in the analysis at all). In addition to calculating the ``mean student profile" for each real teacher, we also calculate using the individual student-level data separate within-class covariance matrices by grade level that roughly quantify the level of within-classroom variance around the classroom-level aggregate values. At the end of this stage, for each real teacher $j$ in grade $g$, we have a mean $\mathbf{u}_{\text{gj}}$ and a covariance matrix $\mathbf{\Sigma}_g$. We can use this to decide which teachers are most reasonable covariate donors for the pseudo-teacher in question given the values of the available predictors for the associated student. These predictors always include the complete set of demographic variables (as these have been imputed in earlier stage if necessary) as well as at least one test score. For each pseudo-teacher we drop from the joint distributions the test scores that are not observed for the associated student. This leaves us with a restricted mean vector $\mathbf{u}^*_{\text{gj}}$ and a covariance matrix $\boldsymbol{\Sigma}^*_g$ for each teacher of the appropriate grade. We then use Bayes Rule and these values to place a probability distribution over the real teachers in grade $g$, where teachers with larger probabilities indicating that based on what we know about the student in question, we believe these teachers are more likely to have the characteristics of the unobserved teacher to who the student was linked in that year. In particular, Bayes Rule implies that

$$p(j|\boldsymbol{Z}) = \frac{p(\boldsymbol{Z}|j)p(j)}{p(\boldsymbol{Z})} \qquad (4)$$

where $j$ is the index of the real teacher in grade $g$, and $\boldsymbol{X}$ is the vector of student-level predictor variables. We assume that all grade $g$ teachers are equally likely to be donors prior to seeing the student's vector of covariates, so that $p(j)$ is a constant that does not depend on teachers. Also $p(\boldsymbol{Z})$ does not depend on teachers. Thus $p(j|\boldsymbol{Z}) \propto p(\boldsymbol{Z}|j)$. We evaluate the right hand side using a multivariate normal density with mean $\mathbf{u}^{*}_{\mathrm{gj}}$ and a covariance matrix $\boldsymbol{\Sigma}_{g}^{*}$, evaluated at the vector of observed predictors $\boldsymbol{Z}$ for the student. We then renormalize these density values to sum to one, which produces a multinomial distribution over the grade $g$ teachers, and a sample from this distribution provides the covariate donor. We replicate this sampling process for each pseudo-teacher, resulting in a complete assignment of donors to recipients for a single imputation.

# Appendix F: Standardized Coefficients for Each Scale By Cohort

Table F1. Point Estimates, Standard Errors, and 95% Confidence Intervals for the Relationships between Reform Teaching and Student Achievement (Current Year Effects)

| Cohort and Outcome | Variable Name | Lower | Point Estimate | Upper | Standard Error | P-Value |
|---|---|---|---|---|---|---|
| Cohort  mc | tallhigh | -0.016 | 0.017 | 0.051 | 0.017 | 0.307 |
| Cohort  mc | tclass.size | -0.018 | 0.021 | 0.059 | 0.019 | 0.285 |
| Cohort  mc | teuclid | -0.025 | 0.012 | 0.049 | 0.018 | 0.511 |
| Cohort  mc | tefftime | -0.037 | 0.005 | 0.048 | 0.021 | 0.795 |
| Cohort  mc | tdiscuss | -0.037 | -0.004 | 0.028 | 0.016 | 0.785 |
| Cohort  mc | ttraditional | 0.038 | 0.07 | 0.102 | 0.016 | 0 |
| Cohort  mc | tgroupwrk | -0.089 | -0.047 | -0.005 | 0.021 | 0.029 |
| Cohort  mc | trefact | -0.015 | 0.019 | 0.053 | 0.017 | 0.278 |
| Cohort  mc | tabsmxd | -0.099 | -0.062 | -0.026 | 0.018 | 0.001 |
| Cohort  mc | tabsprob | -0.089 | -0.053 | -0.016 | 0.018 | 0.005 |
| Cohort  mc | tpdtot | -0.022 | 0.009 | 0.039 | 0.015 | 0.57 |
| Cohort  mc | tconfidence | -0.05 | 0.025 | 0.101 | 0.037 | 0.5 |
| Cohort  mc | tmastdeg | -0.028 | 0.043 | 0.114 | 0.035 | 0.224 |
| Cohort  mc | treform | -0.059 | -0.025 | 0.009 | 0.017 | 0.143 |
| Cohort  mc | tproof.patterns | -0.001 | 0.029 | 0.059 | 0.015 | 0.054 |
| Cohort  mc | toperations | -0.037 | 0.001 | 0.039 | 0.019 | 0.964 |
| Cohort  mc | tyrstch | 0.002 | 0.035 | 0.067 | 0.016 | 0.035 |
| Cohort  mc | tyrs.grade | -0.01 | 0.023 | 0.057 | 0.017 | 0.172 |
| Cohort  mc | ttext | -0.036 | -0.002 | 0.033 | 0.017 | 0.929 |
| Cohort  mc | thetero | -0.138 | -0.075 | -0.011 | 0.032 | 0.022 |
| Cohort  mc | tallhigh | -0.016 | 0 | 0.015 | 0.008 | 0.958 |
| Cohort  mc | tclass.size | 0.063 | 0.081 | 0.099 | 0.009 | 0 |
| Cohort  mc | teuclid | -0.026 | -0.008 | 0.01 | 0.009 | 0.37 |
| Cohort  mc | tefftime | -0.009 | 0.019 | 0.047 | 0.013 | 0.169 |
| Cohort  mc | tdiscuss | -0.019 | 0.006 | 0.032 | 0.012 | 0.621 |
| Cohort  mc | ttraditional | -0.036 | -0.018 | 0 | 0.009 | 0.054 |
| Cohort  mc | tgroupwrk | -0.045 | -0.022 | 0.002 | 0.011 | 0.067 |
| Cohort  mc | trefact | -0.029 | -0.007 | 0.016 | 0.011 | 0.562 |
| Cohort  mc | tabsmxd | -0.045 | -0.017 | 0.01 | 0.013 | 0.203 |
| Cohort  mc | tabsprob | -0.043 | -0.014 | 0.016 | 0.014 | 0.353 |
| Cohort  mc | tpdtot | -0.018 | -0.003 | 0.012 | 0.008 | 0.678 |
| Cohort  mc | tconfidence | 0.001 | 0.049 | 0.098 | 0.024 | 0.045 |
| Cohort  mc | tmastdeg | 0.017 | 0.053 | 0.089 | 0.018 | 0.005 |
| Cohort  mc | treform | -0.02 | 0 | 0.021 | 0.01 | 0.99 |
| Cohort  mc | tproof.patterns | 0.001 | 0.023 | 0.044 | 0.011 | 0.039 |
| Cohort  mc | toperations | -0.041 | -0.015 | 0.01 | 0.013 | 0.227 |
| Cohort  mc | tyrstch | 0.021 | 0.037 | 0.054 | 0.008 | 0 |
| Cohort  mc | tyrs.grade | 0.026 | 0.043 | 0.059 | 0.008 | 0 |
| Cohort  mc | ttext | -0.001 | 0.016 | 0.034 | 0.009 | 0.062 |
| Cohort  mc | tanymathdeg | 0.002 | 0.034 | 0.066 | 0.016 | 0.036 |
| Cohort  mc | thetero | -0.068 | -0.034 | 0.001 | 0.017 | 0.054 |
| Cohort 3 mc | tabsmxd | -0.06 | -0.02 | 0.019 | 0.02 | 0.307 |
| Cohort 3 mc | tabsprob | -0.06 | -0.019 | 0.021 | 0.02 | 0.349 |
| Cohort 3 mc | tallhigh | -0.026 | 0.02 | 0.066 | 0.023 | 0.388 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Cohort 3 mc | tclass.size | -0.049 | -0.012 | 0.024 | 0.019 | 0.511 |
| Cohort 3 mc | tconfidence | -0.129 | -0.007 | 0.116 | 0.062 | 0.916 |
| Cohort 3 mc | tdiscuss | -0.051 | -0.011 | 0.029 | 0.02 | 0.582 |
| Cohort 3 mc | tefftime | -0.015 | 0.024 | 0.063 | 0.02 | 0.221 |
| Cohort 3 mc | teuclid | -0.045 | -0.008 | 0.029 | 0.019 | 0.672 |
| Cohort 3 mc | tgroupwrk | -0.044 | -0.003 | 0.039 | 0.021 | 0.9 |
| Cohort 3 mc | thetero | -0.142 | -0.055 | 0.032 | 0.044 | 0.208 |
| Cohort 3 mc | tmastdeg | -0.053 | 0.035 | 0.123 | 0.044 | 0.426 |
| Cohort 3 mc | tanymathdeg | -0.225 | -0.143 | -0.062 | 0.041 | 0.001 |
| Cohort 3 mc | toperations | -0.043 | 0.006 | 0.055 | 0.025 | 0.81 |
| Cohort 3 mc | tpdtot | -0.042 | 0.007 | 0.055 | 0.024 | 0.784 |
| Cohort 3 mc | tproof.patterns | -0.038 | 0.003 | 0.045 | 0.021 | 0.869 |
| Cohort 3 mc | trefact | -0.018 | 0.039 | 0.096 | 0.029 | 0.18 |
| Cohort 3 mc | treform | -0.043 | 0.009 | 0.062 | 0.026 | 0.72 |
| Cohort 3 mc | ttext | -0.036 | 0.002 | 0.041 | 0.02 | 0.903 |
| Cohort 3 mc | ttraditional | -0.051 | -0.011 | 0.03 | 0.02 | 0.599 |
| Cohort 3 mc | tyrs.grade | -0.005 | 0.035 | 0.074 | 0.02 | 0.085 |
| Cohort 3 mc | tyrstch | -0.001 | 0.04 | 0.081 | 0.021 | 0.056 |
| Cohort 3 oe | tabsmxd | 0.011 | 0.059 | 0.107 | 0.024 | 0.016 |
| Cohort 3 oe | tabsprob | 0.001 | 0.06 | 0.119 | 0.03 | 0.048 |
| Cohort 3 oe | tallhigh | -0.031 | 0.035 | 0.102 | 0.033 | 0.286 |
| Cohort 3 oe | tclass.size | -0.032 | 0.015 | 0.063 | 0.024 | 0.532 |
| Cohort 3 oe | tconfidence | -0.176 | -0.024 | 0.127 | 0.077 | 0.755 |
| Cohort 3 oe | tdiscuss | -0.023 | 0.028 | 0.078 | 0.025 | 0.279 |
| Cohort 3 oe | tefftime | -0.025 | 0.025 | 0.074 | 0.025 | 0.332 |
| Cohort 3 oe | teuclid | -0.035 | 0.018 | 0.072 | 0.027 | 0.5 |
| Cohort 3 oe | tgroupwrk | 0.01 | 0.072 | 0.133 | 0.031 | 0.023 |
| Cohort 3 oe | thetero | -0.106 | 0.005 | 0.116 | 0.056 | 0.925 |
| Cohort 3 oe | tmastdeg | -0.127 | 0.008 | 0.142 | 0.066 | 0.91 |
| Cohort 3 oe | tanymathdeg | -0.337 | -0.192 | -0.047 | 0.072 | 0.011 |
| Cohort 3 oe | toperations | -0.042 | 0.023 | 0.088 | 0.033 | 0.491 |
| Cohort 3 oe | tpdtot | -0.061 | 0.014 | 0.089 | 0.037 | 0.709 |
| Cohort 3 oe | tproof.patterns | -0.023 | 0.039 | 0.101 | 0.031 | 0.216 |
| Cohort 3 oe | trefact | -0.022 | 0.043 | 0.108 | 0.033 | 0.189 |
| Cohort 3 oe | treform | -0.019 | 0.042 | 0.103 | 0.031 | 0.171 |
| Cohort 3 oe | ttext | -0.044 | 0.017 | 0.078 | 0.031 | 0.575 |
| Cohort 3 oe | ttraditional | -0.023 | 0.03 | 0.082 | 0.026 | 0.262 |
| Cohort 3 oe | tyrs.grade | 0.028 | 0.076 | 0.124 | 0.024 | 0.002 |
| Cohort 3 oe | tyrstch | 0.023 | 0.086 | 0.149 | 0.032 | 0.007 |

Note: The current year effect is estimated by (g_11 + g_22 + g_33)/3.

Table F2. Point Estimates, Standard Errors, and 95% Confidence Intervals for the Relationships between Reform Teaching and Student Achievement (Cumulative Effects)

| Cohort and Outcome | Variable Name | Lower | Point Estimate | Upper | Standard Error | P-Value |
|---|---|---|---|---|---|---|
| Cohort  mc | tallhigh | -0.057 | 0.019 | 0.095 | 0.038 | 0.623 |
| Cohort  mc | tclass.size | -0.06 | 0.036 | 0.132 | 0.048 | 0.453 |
| Cohort  mc | teuclid | -0.065 | 0.021 | 0.106 | 0.043 | 0.631 |
| Cohort  mc | tefftime | -0.069 | 0.013 | 0.094 | 0.041 | 0.756 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Cohort  mc | tdiscuss | -0.035 | 0.042 | 0.119 | 0.039 | 0.281 |
| Cohort  mc | ttraditional | 0.028 | 0.108 | 0.188 | 0.04 | 0.009 |
| Cohort  mc | tgroupwrk | -0.126 | -0.049 | 0.029 | 0.039 | 0.216 |
| Cohort  mc | trefact | -0.039 | 0.057 | 0.152 | 0.047 | 0.239 |
| Cohort  mc | tabsmxd | -0.15 | -0.074 | 0.001 | 0.039 | 0.054 |
| Cohort  mc | tabsprob | -0.141 | -0.065 | 0.012 | 0.039 | 0.097 |
| Cohort  mc | tpdtot | -0.099 | -0.022 | 0.055 | 0.039 | 0.57 |
| Cohort  mc | tconfidence | -0.186 | -0.041 | 0.103 | 0.073 | 0.572 |
| Cohort  mc | tmastdeg | -0.149 | 0.001 | 0.151 | 0.075 | 0.993 |
| Cohort  mc | treform | -0.07 | 0.007 | 0.085 | 0.039 | 0.849 |
| Cohort  mc | tproof.patterns | -0.001 | 0.075 | 0.151 | 0.039 | 0.053 |
| Cohort  mc | toperations | -0.111 | -0.031 | 0.049 | 0.04 | 0.447 |
| Cohort  mc | tyrstch | -0.073 | 0.001 | 0.075 | 0.037 | 0.987 |
| Cohort  mc | tyrs.grade | -0.139 | -0.054 | 0.031 | 0.042 | 0.21 |
| Cohort  mc | ttext | -0.043 | 0.037 | 0.118 | 0.04 | 0.354 |
| Cohort  mc | thetero | -0.25 | -0.106 | 0.038 | 0.073 | 0.149 |
| Cohort  mc | tallhigh | -0.021 | 0.024 | 0.07 | 0.023 | 0.294 |
| Cohort  mc | tclass.size | 0.174 | 0.228 | 0.282 | 0.027 | 0 |
| Cohort  mc | teuclid | -0.056 | 0.002 | 0.061 | 0.029 | 0.932 |
| Cohort  mc | tefftime | -0.021 | 0.051 | 0.124 | 0.035 | 0.159 |
| Cohort  mc | tdiscuss | -0.008 | 0.044 | 0.096 | 0.026 | 0.095 |
| Cohort  mc | ttraditional | -0.121 | -0.074 | -0.027 | 0.024 | 0.002 |
| Cohort  mc | tgroupwrk | -0.144 | -0.085 | -0.026 | 0.029 | 0.006 |
| Cohort  mc | trefact | -0.063 | 0 | 0.064 | 0.032 | 0.987 |
| Cohort  mc | tabsmxd | -0.149 | -0.074 | 0 | 0.036 | 0.051 |
| Cohort  mc | tabsprob | -0.135 | -0.059 | 0.017 | 0.037 | 0.122 |
| Cohort  mc | tpdtot | -0.063 | -0.019 | 0.025 | 0.022 | 0.39 |
| Cohort  mc | tconfidence | 0.095 | 0.218 | 0.341 | 0.062 | 0.001 |
| Cohort  mc | tmastdeg | 0.009 | 0.107 | 0.206 | 0.05 | 0.033 |
| Cohort  mc | treform | -0.062 | -0.007 | 0.048 | 0.027 | 0.807 |
| Cohort  mc | tproof.patterns | 0.018 | 0.077 | 0.136 | 0.03 | 0.012 |
| Cohort  mc | toperations | -0.097 | -0.032 | 0.032 | 0.032 | 0.316 |
| Cohort  mc | tyrstch | -0.001 | 0.047 | 0.096 | 0.024 | 0.054 |
| Cohort  mc | tyrs.grade | 0.017 | 0.064 | 0.11 | 0.023 | 0.008 |
| Cohort  mc | ttext | -0.005 | 0.045 | 0.095 | 0.025 | 0.078 |
| Cohort  mc | tanymathdeg | 0.035 | 0.119 | 0.203 | 0.043 | 0.006 |
| Cohort  mc | thetero | -0.226 | -0.135 | -0.045 | 0.046 | 0.004 |
| Cohort 3 mc | tabsmxd | -0.154 | -0.064 | 0.026 | 0.046 | 0.164 |
| Cohort 3 mc | tabsprob | -0.189 | -0.094 | 0.001 | 0.048 | 0.053 |
| Cohort 3 mc | tallhigh | -0.091 | 0.002 | 0.096 | 0.047 | 0.96 |
| Cohort 3 mc | tclass.size | 0.044 | 0.142 | 0.239 | 0.05 | 0.005 |
| Cohort 3 mc | tconfidence | -0.306 | 0.011 | 0.328 | 0.16 | 0.944 |
| Cohort 3 mc | tdiscuss | -0.099 | -0.001 | 0.097 | 0.049 | 0.983 |
| Cohort 3 mc | tefftime | -0.017 | 0.089 | 0.194 | 0.053 | 0.098 |
| Cohort 3 mc | teuclid | -0.04 | 0.055 | 0.15 | 0.048 | 0.251 |
| Cohort 3 mc | tgroupwrk | -0.13 | -0.038 | 0.054 | 0.047 | 0.422 |
| Cohort 3 mc | thetero | -0.306 | -0.099 | 0.109 | 0.105 | 0.349 |
| Cohort 3 mc | tmastdeg | -0.293 | -0.098 | 0.096 | 0.098 | 0.319 |
| Cohort 3 mc | tanymathdeg | -0.392 | -0.177 | 0.038 | 0.109 | 0.107 |
| Cohort 3 mc | toperations | -0.072 | 0.05 | 0.171 | 0.061 | 0.421 |
| Cohort 3 mc | tpdtot | -0.092 | 0.02 | 0.132 | 0.057 | 0.723 |
| Cohort 3 mc | tproof.patterns | -0.139 | -0.038 | 0.062 | 0.051 | 0.451 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Cohort 3 mc | trefact | -0.005 | 0.118 | 0.241 | 0.063 | 0.059 |
| Cohort 3 mc | treform | -0.085 | 0.031 | 0.148 | 0.059 | 0.595 |
| Cohort 3 mc | ttext | -0.076 | 0.035 | 0.146 | 0.056 | 0.533 |
| Cohort 3 mc | ttraditional | -0.163 | -0.068 | 0.027 | 0.048 | 0.16 |
| Cohort 3 mc | tyrs.grade | -0.068 | 0.013 | 0.094 | 0.041 | 0.749 |
| Cohort 3 mc | tyrstch | -0.027 | 0.078 | 0.183 | 0.053 | 0.145 |
| Cohort 3 oe | tabsmxd | 0.018 | 0.121 | 0.224 | 0.052 | 0.021 |
| Cohort 3 oe | tabsprob | -0.017 | 0.115 | 0.247 | 0.065 | 0.085 |
| Cohort 3 oe | tallhigh | -0.122 | -0.016 | 0.09 | 0.054 | 0.77 |
| Cohort 3 oe | tclass.size | -0.019 | 0.095 | 0.209 | 0.057 | 0.1 |
| Cohort 3 oe | tconfidence | -0.24 | 0.089 | 0.419 | 0.167 | 0.593 |
| Cohort 3 oe | tdiscuss | -0.067 | 0.047 | 0.161 | 0.057 | 0.415 |
| Cohort 3 oe | tefftime | -0.033 | 0.091 | 0.215 | 0.062 | 0.148 |
| Cohort 3 oe | teuclid | -0.052 | 0.055 | 0.162 | 0.054 | 0.314 |
| Cohort 3 oe | tgroupwrk | 0.018 | 0.136 | 0.253 | 0.059 | 0.024 |
| Cohort 3 oe | thetero | -0.286 | -0.044 | 0.198 | 0.122 | 0.72 |
| Cohort 3 oe | tmastdeg | -0.271 | -0.045 | 0.181 | 0.113 | 0.691 |
| Cohort 3 oe | tanymathdeg | -0.486 | -0.208 | 0.069 | 0.138 | 0.138 |
| Cohort 3 oe | toperations | -0.017 | 0.105 | 0.228 | 0.062 | 0.091 |
| Cohort 3 oe | tpdtot | -0.108 | 0.018 | 0.145 | 0.064 | 0.774 |
| Cohort 3 oe | tproof.patterns | -0.101 | 0.017 | 0.134 | 0.059 | 0.78 |
| Cohort 3 oe | trefact | -0.009 | 0.137 | 0.283 | 0.074 | 0.066 |
| Cohort 3 oe | treform | -0.015 | 0.109 | 0.233 | 0.063 | 0.085 |
| Cohort 3 oe | ttext | -0.07 | 0.053 | 0.175 | 0.061 | 0.395 |
| Cohort 3 oe | ttraditional | -0.096 | 0.028 | 0.152 | 0.062 | 0.65 |
| Cohort 3 oe | tyrs.grade | -0.007 | 0.085 | 0.178 | 0.047 | 0.071 |
| Cohort 3 oe | tyrstch | 0.032 | 0.151 | 0.27 | 0.06 | 0.013 |

Note: The cumulative effect after three years is estimated by (g_31 + g_32 + g_33).