

Exploring Student-Teacher Interactions in Longitudinal Achievement Data

J.R. Lockwood

Daniel F. McCaffrey

RAND Corporation

April 7, 2008

PRELIMINARY DRAFT

NOT FOR DISTRIBUTION OR CITATION

This material is based on work supported by the Department of Education Institute of Education Sciences under Grant No. R305U040005, and has not undergone RAND peer review. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations.

Exploring Student-Teacher Interactions in Longitudinal Achievement Data

Abstract

This article develops a model for longitudinal student achievement data specifically designed to estimate student-teacher interactions. The model specifies interactions with respect to a student's general ability level, and estimates an average effect of teachers as well as a parameter that indicates whether individual teachers are more or less effective, relative to their average effect, with students of different ability levels. Using various longitudinal data sources, we find evidence of interactions of teacher effects with students' general ability levels that appear to be of relatively consistent but modest magnitude across different contexts, accounting for on the order of 10% of the total variation in teacher effects across all students. However, the amount that the interactions matter in practice depends on how different are the groups of students taught by different teachers. Using empirical estimates of the heterogeneity of students across teachers, we find that the interactions account for on the order of 2%-4% of total variation in teacher effects on different classes, suggesting that ignoring these interactions is not likely to lead to appreciable bias in estimated teacher effects for most teachers in most settings.

1 Introduction

Using longitudinal achievement data to estimate effects of educational interventions is increasingly commonplace (Bifulco & Ladd, 2004; Gill, Zimmer, Christman, & Blanc, 2007; Goldhaber & Anthony, 2004; Hanushek, Kain, & Rivkin, 2002; Le, Stecher, Lockwood, Hamilton, Robyn, Williams, Ryan, Kerr, Martinez, & Klein, 2006; Schacter & Thum, 2004; Zimmer, Buddin, Chau, Daley, Gill, Guarino, Hamilton, Krop, McCaffrey, Sandler, & Brewer, 2003). Though usage of the term varies, "value-added modeling" (VAM) generally refers to statistical analyses of longitudinal achievement data along with links of students to their teachers or schools to estimate the effects of individual teachers or schools on student learning (Ballou, Sanders, & Wright, 2004; Braun, 2005a; Harris & Sass, 2006; Jacob & Lefgren, 2006; Kane, Rockoff, & Staiger, 2006; Koedel & Betts, 2005; Lissitz, 2005; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders, Saxton, & Horn, 1997). The reputation of VAM is that with rich enough data and sophisticated modeling, teacher performance can be fairly compared across teachers even though students are not randomly assigned to teachers and classrooms are often not comparable in terms of student backgrounds and prior achievement. This reputation, along with increasingly available longitudinal data at the district and state levels due to NCLB-mandated testing and rapidly expanding data archiving capabilities, has led to calls for the use of VAM in teacher accountability. Estimates of teacher impacts on student achievement are now used in some places for pay for performance (e.g., Florida and Houston among others) and some researchers have even called for teacher hiring/firing decisions to be based on information obtained from VAM (Gordon, Kane, & Staiger, 2006).

Research to date on VAM has centered primarily on the internal validity of individual teacher effect estimates. That is, how confident can we be that the statistical or econometric methods applied to the data provide estimated effects of teachers that truly reflect the contributions of those teachers rather than the effects of other biasing factors, particularly the characteristics of the students or prior educational inputs? Some researchers have raised doubts about whether VAM can support estimates of true causal effects of teachers or whether the effects of teachers can be separated from classroom context (Ballou, 2004; Braun, 2005b; McCaffrey et al., 2003; Raudenbush, 2004; Rothstein, 2007; Rubin,

Stuart, & Zanuto, 2004). However, empirical research has consistently found evidence of differential teacher effects, even with methods offering controls for many potential biasing factors either through sophisticated panel data models (Aaronson, Barrow, & Sander, 2003; Harris & Sass, 2006; Koedel & Betts, 2005) or multivariate mixed model analyses (Ballou et al., 2004; Lockwood, McCaffrey, Mariano, & Setodji, 2007b; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Raudenbush & Bryk, 2002; Sanders et al., 1997). Moreover, the variation is not necessarily strongly correlated with student backgrounds and prior achievement and it is predictive of teachers' future students' outcomes. Analyses of experimental data with students randomized to classes corroborates these findings: these studies find variability among teachers that is similar to the value-added estimates (Nye, Konstantopoulos, & Hedges, 2004) and find that value-added estimates predict differences among teachers on randomly assigned classes (Kane & Staiger, 2008). In all, while the concerns about internal validity have not been solved definitively, the evidence suggests that teachers are differentially effective and that these differences can be measured with some fidelity with appropriate data and analyses.

Relatively less attention has been paid to the external validity of VAM estimates. Current models estimate a single effect for each teacher, perhaps separately by subject and by year when sufficient data are available. Even if VAM estimates are truly reflective of how a teacher performed at teaching a particular subject with a particular group of students at a particular point in time, to what extent does this provide a generalizable inference about that teacher's effectiveness? The use of VAM for high-stakes decisions about teachers increases the need for the estimates to reflect something about a teacher's performance that is stable across units (students), outcomes (subjects and tests) and settings (schools, school years, courses, and other contextual factors). If an individual teacher's effect is markedly heterogeneous across units, outcomes or settings, the credibility of VAM estimates is eroded because of their sensitivity to potentially idiosyncratic circumstantial factors.

The limited research on the stability of estimates across outcomes and settings has provided mixed results. In general, estimates appear moderately stable across time. For example, Lockwood, McCaffrey and Sass (2008) consider the stability of the effects of middle school mathematics teachers and find moderate correlations (on the order of 0.4) of

the estimated effects for the same teacher across different school years, but given the large sampling errors a correlation of 0.4 suggests fairly stable effects for teachers across time. McCaffrey, Han and Lockwood (2008) found similar results. Koedel and Betts (2005) find that quintile groupings of estimated math teacher effects for the same teachers across different years can be unstable but that teachers in the tails of the distribution demonstrate somewhat higher stability. In the only study of the stability of effects as teachers change context, Sanders, Wright, Springer and Langevin (2008) found that teacher effects were relatively stable when they moved across schools serving very different populations of students. Conversely, Lockwood et al. (2007a) found considerable sensitivity of effects to the test, with the correlation between teacher effects estimated from two subscales of a mathematics test to be on the order of 0.2 or less, for a small sample of middle school mathematics teachers.

Limited research has been done on the potential heterogeneity of teacher effects across different students. Both anecdotal evidence and experience, however, suggest that teachers may be differentially effective with students of different aptitudes or other characteristics (Dee, 2003; Hanushek, Kain, O'Brien, & Rivkin, 2005; Harris & Sass, 2006). For example, two teachers may be equally effective on average across all students, but one may be particularly effective with students who are generally high achieving and the other may be particularly effective with students who are generally low achieving. Sanders (personal communication) has claimed that through extensive analysis of Tennessee achievement data, it is possible to identify teachers with characteristic patterns ("sheds", "reverse sheds" and "teepees") of differential effectiveness across students with different average levels of achievement. On the other hand, Koedel and Betts (2005) tested for variation in individual teacher effects across groups of students with prior test scores above and below the median prior score and fail to reject the null hypothesis of no interactions, and Hanushek et al. (2005) find moderate correlations of between 0.3 and 0.6 for average gains made by groups of students sharing a teacher but stratified by their prior score. Using data from a randomized experiment, Dee (2003) finds a positive effect on achievement for students being paired with a same-race teacher.

If teachers are differentially effective with different students, understanding the nature and magnitude of these differences is important to VAM for several reasons. First, as

noted, heterogeneity of effects across students calls into question the generalizability of the inferences made from VAM estimates. Models that estimate a single teacher effect implicitly are estimating a teacher's effect on the particular group of students taught by that teacher, and thus teachers who would be equally effective on similar students may have different VAM estimates simply because their classrooms have different student compositions. This would strike at the foundation of VAM because its primary purpose is to provide fair comparisons of teachers despite the fact that teachers teach different types of students. Second, if heterogeneity in teacher effects across different types of students can be reliably measured, this could enhance the utility of VAM estimates for improving education. For example, if average teacher effects can be broken down into a more fine-grained assessment of teacher performance across different types of students, this could provide useful diagnostic information for targeted interventions. Such information could also lead to more efficient assignment of student/teacher pairings that leveraged each teacher's relative strengths.

The goal of this article is to develop a model that allows teacher effects to vary across individual students, and to apply the model to a variety of longitudinal achievement datasets to examine the nature and magnitude of the student-teacher interaction effects. Of particular interest is understanding the consequences of ignoring these interactions given that VAM methods estimating a single effect for each teacher by subject by year are mostly likely to be used for high-stakes decisions about teachers. The interactions that we consider regard to what extent teachers are differentially effective with students of different general ability levels. We conceive of a student's "general ability level" as a latent characteristic that can be inferred from an extensive longitudinal achievement data series which provides measures of an individual student's achievement taken from different grades, subjects and contexts. While interactions with respect to this latent characteristic are less straightforward to examine than those with observable student characteristics, they are an intuitively plausible source of heterogeneity of teacher effects, and may lead to more actionable inferences than interactions with respect to observable student characteristics such as SES, race/ethnicity, or gender. They should also provide a more accurate and precise estimate of a student's ability than a single prior score, as used by previous studies that have examined teacher interactions with student ability

(Aaronson et al., 2003; Hanushek et al., 2005; Koedel & Betts, 2005).

The remainder of the article is organized as follows. Section 2 develops our basic model that allows teacher effects to depend on a student’s general level of achievement and discusses the specification of the model in a Bayesian framework. Section 3 discuss the data sources we use in our investigations. Section 4 presents model diagnostics and model selection criteria assessing the fit of the model and its performance relative to a sequence of simpler alternatives. Section 5 presents inferences about the parameters representing the student-teacher interactions and uses those estimates along with other features of the data to calibrate their magnitudes. Finally, Section 6 offers some concluding remarks and discussion points.

2 A Model for Student-Teacher Interactions

2.1 Basic Model

We begin with a simplified scenario in which I students are taught by J teachers, with each student being taught by a single teacher. We let $i = 1, \dots, I$ index the students and let Y_i , the outcome of interest, denote the measure of achievement for student i . We let $j = 1, \dots, J$ index teachers and use the notation $j(i)$ to indicate the teacher index j of the teacher who taught student i .

The basic model underlying our investigations is

$$Y_i = \mu + \delta_i + \theta_{0j(i)} + \theta_{1j(i)}\delta_i + \epsilon_i \quad (1)$$

The overall mean achievement is represented by μ . δ_i is the “general ability” of student i , scaled so that it has mean zero across all students in the data but otherwise is on the scale of the test. θ_{0j} the main effect of teacher j , defined as that teacher’s average effect¹ across all students, or equivalently, that teacher’s effect on students of average ability

¹We follow the convention of calling these parameters teacher effects even though they really represent only unexplained heterogeneity among students linked to the same teacher. Ideally this unexplained heterogeneity is primarily a result of differential teacher performance, but there might be many sources of this heterogeneity, including contextual effects and omitted student characteristics (McCaffrey et al., 2003; McCaffrey et al., 2004)

($\delta = 0$). These effects are scaled so that they have mean zero across the teachers in the data. θ_{1j} is the interaction term for teacher j that indicates whether this teacher is relatively more effective with students of higher general ability ($\theta_{1j} > 0$) or with students of lower general ability ($\theta_{1j} < 0$). Again these effects are scaled so that they have mean zero across all teachers in the data. Finally, the error terms ϵ_i are assumed to be mean zero and independent of the other terms in the model.

Thus, the basic model parameterizes teacher effects with two parameters: a main effect or intercept, and a slope indicating whether teachers are more or less effective with students of differing values of δ_i . In principle this linear restriction on the functional form of the teacher effect profile as a function of δ_i is not necessary but it is a first natural step and we do not consider more complex functional forms further in this article. If each δ_i were known, the teacher intercepts and slopes could be estimated by OLS by regressing $Y_i^* = Y_i - \delta_i$ on a grand mean, sum-to-zero constrained teacher main effects and sum-to-zero constrained teacher slopes on the δ_i . The intercept terms would be identified by the within-class means of Y_i^* and slope terms would be identified by a within-class regression of the Y_i^* on δ_i . Teachers who tended to have more positive values of Y_i^* (relative to their main effect) for students with $\delta_i > 0$ would have positive slope estimates.

As written, the model cannot be estimated because δ_i is not known. However, longitudinal data can be used to estimate it, because a student's scores on prior tests provide relatively strong predictive information about their general ability level as it pertains to likely performance on a given test (e.g., the Y_i above). Suppose that we had p prior scores on the students (e.g., coming from prior school years), denoted by Z_{ip} ², and that we modeled these scores by:

$$Z_{ip} = \mu_p + \beta_p \delta_i + \epsilon_{ip} \quad (2)$$

Here μ_p is a marginal mean for the p th prior score. δ_i is the same general ability term

²We use the notation Z_{ip} rather than Y_{ip} in Equation 2 to emphasize the fact that when fitting the model in practice, we standardize all prior scores using rank-based z-scores (Kirby, McCaffrey, Lockwood, McCombs, Naftel, & Barney, 2002) defined as $Z_{ip} = \Phi^{-1}(\hat{F}_p(Y_{ip}))$ where \hat{F}_p is the empirical CDF of the unstandardized scores Y_{ip} and Φ^{-1} is the inverse CDF of the standard normal distribution. This forces the Z_{ip} to be marginally normal, which improves the plausibility of the linear conditional relationships among scores assumed by Equation 2 given the variety of scales on which achievement is reported.

appearing in Equation 1 but now scaled by the parameter β_p to allow for the possibility that the test depends on δ_i differently than does the target outcome Y_i . The ϵ_{ip} are error terms treated as independent both within and across students and as independent of the error terms ϵ_i in Equation 1 (assumptions that we discuss in more detail later in the section) with a variance σ_p^2 that varies across the p different scores. The notion is that the prior scores depend on the same latent quantity δ_i as the target score Y_i , but differentially through the scale factors β_p . In this sense, δ_i can be thought of as a generalized average prior score, put onto the scale of the target score Y_i . The idea is similar to other student achievement modeling research in which latent effects for levels and/or growth of achievement are introduced and the relationships between other covariates and achievement are specified through these parameters (Raudenbush & Bryk, 2002; Thum, 2003; Seltzer, Choi, & Thum, 2003, 2002; Choi, 2001).

So, our general strategy is to use past test score data on individual students to estimate δ_i for each student through Equation 2, and then to estimate Equation 1 to produce estimates of the teacher intercepts and slopes. These two steps are carried out simultaneously in the context of a joint model for the prior scores and current scores, in which the teacher effects and student effects are estimated simultaneously. The model is estimated in a Bayesian framework, and additional details on this specification are provided in Section 2.3 and in the Appendix.

Equation 2 is potentially misspecified because of the assumption of the independence of the residuals ϵ_{ip} both among themselves and with the ϵ_i of Equation 1. The residuals may be related within students due to similar performance on particular types of tests (e.g. scores from reading tests may be more correlated with one another than they are with math tests). They may also be related within and across students because of omitted teacher effects or other contextual effects. Within a student our model is consistent with a student's teachers being independent across subjects and grades and their effects having no persistence after the current year. A model with less restrictive assumptions specifies current achievement as a function of an accumulation of past and current teacher effects, possibly with downweighting of prior teacher effects (Lockwood et al., 2007b; McCaffrey et al., 2004; Ballou et al., 2004; Sanders et al., 1997; Harris & Sass, 2006).

In the current application we use a more restrictive model to improve the compu-

tational efficiency of our models and to improve the stability of the estimates of our parameters of interest (θ_{0j} and θ_{1j} for the teachers in a target year). In this way our model for the prior scores is not intended to be structural but rather descriptive, and is chosen to allow us to extract information from prior scores for use in examining teacher effects and interactions in the target year. The model appears to do an adequate job of extracting this information, and we present results of a specification test in Section 4.3 that examines how much bias might be introduced into our results about teacher effects in the target year by using this simple single factor structure with independent residuals rather than a more complex specification that came closer to a structural model for the prior achievement outcomes.

2.2 Extensions for Nonlinearity and Heteroskedasticity

When applying the model to various longitudinal data sources, we found that allowing for non-linearity and heteroskedasticity in Equation 1 was beneficial. We allow for non-linearity in Equation 1 using a piecewise quadratic depending on whether or not $\delta_i < 0$, as:

$$Y_i = \mu + \delta_i + (\lambda_{m1}1_{\delta_i < 0} + \lambda_{m2}1_{\delta_i \geq 0})\delta_i^2 + \theta_{0j(i)} + \theta_{1j(i)}\delta_i + \epsilon_i \quad (3)$$

where 1_A is the indicator function of the event A . This form was suggested by exploratory analyses and is flexible enough to capture convex, concave or sigmoidal relationships between Y_i and δ_i .

We used a similar approach to modeling heteroskedasticity in the error terms ϵ_i in Equation 1. A common byproduct of the design and scaling of standardized achievement tests is that the measurement error of high or low scores is larger than intermediate scores. To approximate this relationship³ we estimated a variance function that allowed $\text{Var}(\epsilon_i)$ to depend on δ_i as

$$\text{Var}(\epsilon_i|\delta_i) = \sigma^2 \exp[\delta(\lambda_{v1}1_{\delta_i < 0} + \lambda_{v2}1_{\delta_i \geq 0})] \quad (4)$$

³Classical test theory would motivate a model where the “true score” u_i would be given by the entire right-hand side of Equation 3 except for the error term ϵ_i , and the variance of ϵ_i would be a function of u_i . We considered such models but they led to convergence problems in the model for some of our datasets, so we opted to keep the simpler specification where the variance of ϵ_i depends on δ_i .

This variance function is non-negative. With $\lambda_{v1} = \lambda_{v2} = 0$ it reduces to homoskedasticity. Values of $\lambda_{v1} < 0$ and $\lambda_{v2} > 0$ generate various convex error functions with minimum value σ^2 at the average general achievement value $\delta_i = 0$.

The term “complete model” in the remainder of the article refers to the model that allows for both nonlinearity as specified in Equation 3 and heteroskedasticity of the error term as specified in Equation 4, along with the model for the prior scores in Equation 2. In Section 4.1 we show that the complete model is more appropriate for the data than a sequence of simpler alternatives, and so all of the substantive results about interactions that we report are based on the complete model.

2.3 Bayesian Specification

We estimate the complete model using a Bayesian specification (Carlin & Louis, 2000; Gilks, Richardson, & Spiegelhalter, 1996; Gelman, Carlin, Stern, & Rubin, 1995) implemented in the Bayesian modeling software WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). The Bayesian framework specifies a conditional probability distribution of the data given all unknown parameters (the “likelihood”) and a probability distribution for the unknown parameters (the “prior distribution”). These lead to a conditional probability distribution for all of the unknown parameters given the observed data (the “posterior distribution”), from which all inferences about the unknown parameters are derived. In this section we discuss the Bayesian specification of the model. Additional details on the prior distributions are provided in the Appendix. The WinBUGS code we used to fit the model is available from the authors upon request.

The modeling assumptions for the student and teacher effects are that δ_i are iid $N(0, \nu)$ and $(\theta_{0j}, \theta_{1j})'$ are iid $N(\mathbf{0}, \mathbf{G})$ and independent of δ_i , where both ν and \mathbf{G} have prior distributions that allow their estimates to be driven by the data. \mathbf{G} is the (2×2) covariance matrix of the teacher main effects and teacher slopes with the variance of the main effects denoted by τ_0^2 , the variance of the slopes denoted τ_1^2 , and correlation r so that the covariance between the slopes and intercepts is $r\tau_0\tau_1$. Most of the important inferences about interactions that we report later are based on functions of ν and \mathbf{G} .

The assumption of independence between the student and teacher effects is question-

able, but the results of Lockwood and McCaffrey (2007) indicate that when many prior scores are available for estimating δ_i (as is the case in our applications) the independence assumption is not consequential and the student and teacher effects can be estimated with minimal bias even in the presence of selection of students to teachers (e.g., that students with higher values of δ_i are more likely to be assigned to teachers with higher values of θ_{0j}).

For the remaining terms in the model for the target year scores, the overall mean μ and the parameters λ_{m1} and λ_{m2} governing the non-linearity are modeled with minimally informative independent normal priors. The error terms ϵ_i are modeled as independent mean-zero normals with variance given by Equation 4, where σ^2 and the parameters λ_{v1} and λ_{v2} governing the heteroskedasticity are given minimally informative priors. For the parameters in the distribution of the prior scores in Equation 2, we modeled the μ_p with minimally informative normal priors, and the β_p as $\Gamma(1, 1)$ which has mean 1, a specification that made sense given that both the current and prior scores were normed to have the same marginal variance and so the coefficients on the prior scores should be on the order of 1. The error terms ϵ_{ip} are modeled as independent mean zero normals with variances σ_p^2 , with the σ_p^2 given minimally informative priors.

Given the large number of prior scores that are used in our applications, many students do not have all prior scores observed due to mobility and missed testing. Thus, some students have a lot of information from which to estimate δ_i and others relatively less. The missing prior scores are handled via data augmentation (van Dyk & Meng, 2001; Schafer, 1997; Tanner & Wong, 1987), which is automatically implemented in WinBUGS under an assumption that the missing scores are missing at random (Little & Rubin, 1987). This algorithm accommodates arbitrary missing data patterns for the prior scores, and allows students to contribute to the estimated teacher effects and interactions in proportion to how much information the data provide about their individual δ_i .

3 Data

We use three different longitudinal achievement datasets to investigate the interaction model presented in Section 2. The datasets come from three different large urban school

districts referred to as A, B and C throughout the remainder of the article. The datasets cover different grade ranges and achievement outcomes but otherwise have similar structures. District A data are from a single cohort of about 9200 students followed from grades 1 through 5, with students linked to their teachers each year and with students tested in math and reading at each grade. For the analyses we focus on teachers from grades 3 to 5. District B data are from a single cohort of about 3400 students followed from grades 5 to 8 with students linked to their math teachers and students tested in math in each grade, and with a variety of other test scores available including science and reading. For the analysis we focus on math teachers in grades 7 and 8. Finally District C data are from four cohorts of students who were in grades 5 to 8 during the 2006-2007 school year. The approximately 26000 students are linked to their mathematics teachers from 2006-2007, and those students' achievement measures in math, science, reading and social studies are available back to the 2002-2003 school year ⁴ and as early as grade 3. This leads to between 8 and 14 prior scores available for students who were continuously enrolled in the district and who did not miss any testing in prior grades.

As noted in Section 2 we consider the interaction model for students linked to a set of target teachers in a single target grade, and treat all scores prior to that grade as prior scores under Equation 2. To be included in an analysis, a student must be linked to a teacher in the target grade and must have the target grade achievement outcome observed. We also enforce that the student must have at least two observed test scores prior to the target grade so that at least some information about their δ_i is available. In most cases, the vast majority of students have many more than two prior scores available.

In total from districts A, B and C we consider twelve groups of target teachers and their associated effects. Some summaries of these twelve groups are presented in Table 1, including the district, the target grade, the target subject (math or reading), the number of teachers, the number of students, and the maximum and mean numbers of prior scores available for the students. The target teacher groups from District B bridge an interesting gap between those from Districts A and C because the math achievement outcome from district B is the same as that from District A (same test developer, test edition and scale) but the grade ranges available from District B overlap with district C. As discussed further

⁴Social studies and science tests are not available for the 2002-2003 school year.

in the results, this helps to resolve one of the differences in the results about interactions between Districts A and C.

[Table 1 about here.]

4 Model Assessments

4.1 Comparisons of Complete Model to Simpler Alternatives

We were first interested in establishing whether the model that includes the student-teacher interaction terms is a more appropriate model for the data than simpler alternatives. To do this we fit a sequence of five increasingly complex models to each of the twelve target teacher groups, culminating in the complete model that allows for nonlinearity, heteroskedasticity, teacher main effects, and teacher-student interactions. We then compared the models with the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin, & van der Linde, 2002) which is a model comparison criterion for complex Bayesian models that heuristically combines a measure of model fit and model complexity to indicate which, among a set of models being compared, is preferred (as indicated by the smallest DIC value).

The five models that we compared start with the simplest case (“Model 1”) in which there are no allowances for nonlinearity and heteroskedasticity ($\lambda_{m1} = \lambda_{m2} = \lambda_{v1} = \lambda_{v2} = 0$) and no teacher effects at all in the target year ($\tau_0^2 = \tau_1^2 = 0$). Model 2 retains these restrictions but allows for nonlinearity. Model 3 allows for both nonlinearity and heteroskedasticity, but again has no teacher effects. Model 4 allows for nonlinearity, heteroskedasticity, and teacher main effects only ($\tau_1^2 = 0$). Finally, Model 5 is the complete model that allows for nonlinearity, heteroskedasticity, teacher main effects, and teacher-student interactions.

The DIC values for each of the models are presented in Figure 1 and demonstrate that the complete model is preferred in all twelve target teacher groups. To facilitate comparisons across the twelve groups, DIC values for Model n are presented as $(DIC_n - DIC_1)/DIC_1$ so that all groups have a value of 0 for Model 1. In general, each successively more complex model is preferred over the simpler alternatives as indicated by decreasing

DIC values. The largest improvements generally occur between Models 3 and 4 with the introduction of teacher main effects. The student-teacher interaction terms provide an additional benefit which is generally smaller than the incremental improvements provided by the other model features, but is consistent across all twelve target teacher groups.

The formal model comparisons are based on incremental improvements in DIC on an absolute rather than relative scale. The smallest incremental improvement between Models 4 and 5 is about 13 DIC points, for District C grade 5 math, while the largest improvement is about 275 DIC points for District A grade 3 math. The median improvement across the groups is about 83 DIC points. A typical rule of thumb is that DIC improvements of between 5 and 10 points are substantial, so it appears that the complete model allowing student-teacher interactions is uniformly more appropriate for the data than a model that includes teacher main effects alone.

[Figure 1 about here.]

4.2 Model Fit

For each of the twelve target teacher groups we performed various posterior predictive checks (Gelman, Meng, & Stern, 1996; Gilks et al., 1996) that assessed how well the complete model captures important features in the data. Posterior predictive checks use the posterior distribution of the model parameters to generate new hypothetical data that in principle should look like the actual observed data if the model is adequate. The idea is that if the model is a close approximation to the data generating mechanism, then the observed data should look like a typical realization from this mechanism.

The simplest posterior predictive check that we examined was whether the model was able to capture the relationship between the average prior scores for the students (the average of whichever of Z_{ip} in Equation 2 are observed for student i) and Y_i . For each parameter vector in the MCMC sample, we generated values of Z_{ip} for each student using the current estimate of their δ_i and the other parameters governing the distributions of the Z_{ip} . We then took the average \bar{Z}_i of these over the p components that were actually observed for student i . Similarly, we generated a value of Y_i for each student under the model for the scores in the target year, including accounting for whatever nonlinearity,

heteroskedasticity, and the teacher main effects and interactions are implied by the values of the parameters in the particular MCMC iteration. For each iteration of the MCMC algorithm, this results in a realization of (\bar{Z}_i, Y_i) sampled from the posterior predictive distribution of the data for these students. From these samples, pooled across MCMC iterations, we calculated pointwise 0.025, 0.50 and 0.975 quantiles of the conditional distribution of Y_i given \bar{Z}_i , and compared these bounds to the observed data.

Figure 2 shows representative examples of the results for a subset of the twelve target teacher groups; results for groups not shown are similar. Each frame of the figure plots the posterior predictive bounds described above, along with the corresponding (\bar{Z}_i, Y_i) pairs for the observed data. The close correspondence between the bounds and the actual data indicates that the model is adequately capturing features of the relationship between the average prior scores and the scores in the target year, including nonlinearity and heteroskedasticity. The numbers in parentheses at the top of each figure give the percentage of the observed data pairs that fall outside of the predictive bounds. These percentages are very close to the target values of 5%.

[Figure 2 about here.]

We carried out similar checks at the level of individual teachers to ascertain that the model was capturing features of the achievement of the students in each teacher's class. In particular we were interested in seeing whether the estimated values of main effects and intercepts for teachers implied a profile of effectiveness as a function of δ that was consistent with the achievement patterns evident in the data. Figure 3 provides an example for selected teachers from District B's grade 8 math teachers. The gray lines indicate the model estimate of $E(Y|\delta)$ in the absence of any teacher effects, while the black lines indicate the model estimate of $E(Y|\delta, j)$ for a given teacher j accounting for the estimated main effect and slope for that teacher. The points in each frame are the actual Y_i values for students linked to that teacher, plotted as a function of posterior mean of δ_i for those students. These particular teachers were chosen to illustrate a teacher with a positive estimated slope (top frame), a zero estimated slope (middle frame) and a negative estimated slope (bottom frame), and these estimates seem to capture the achievement patterns of the students linked to each class. For the teacher with the negative slope

estimate, the students with higher values of δ_i are scoring generally less well in this class than would be predicted in the absence of teacher effects compared to students with lower values of the δ_i . The opposite is true for the teacher with the positive slope estimate. Analogous checks carried out for all teachers in the twelve target teacher groups showed similar correspondence between model estimates and data.

[Figure 3 about here.]

4.3 Assessing Potential for Misspecification Bias

Given the simple structure of our model for prior scores in Equation 2, it was important to assess whether the model appeared to be ignoring information contained in the prior scores that might be biasing the estimates of the teacher effects in the target year. For example, since the prior scores are from tests across a mixture of subjects, specific information about math achievement, for example, might be omitted from the estimate of δ_i and thus may lead to biased estimates of the teacher effects in the target year if it clusters at the teacher level. We wanted to ensure that this bias was sufficiently small to leave our substantive conclusions unaffected.

We investigated this for the nine target teacher groups where math was the target outcome as follows. For each student, we obtained the posterior mean of δ_i under the model, regressed the math score from the year immediately prior to the target year on these values⁵, and obtained the residuals from this regression. The idea is that if there is important information about math achievement in the target year that could have been predicted from past scores but is not already captured by δ_i , the math score from the immediate prior year is probably the best source of that information, and the residuals from the regression on δ_i isolate it. In order to indicate a source of bias, these residuals need both to be related to target year scores and to vary across teachers. To assess this, we regressed the target year math scores on these residuals and fixed effects for target year teachers to obtain a within-teacher estimate of the relationship between the residuals and target year math scores, and we also estimated the between-teacher variance component of the residuals. The squared regression coefficient times the between-teacher variance

⁵The R^2 from this regression varied between 0.72 and 0.81 across the twelve groups.

component indicates the variance in mean target year scores across classes that could be due to the omitted biasing factor. We compared this value to the estimate of the teacher main effect variance τ_0^2 to calibrate how much of the estimated between teacher variance might be due to the omitted factor.

The results indicate that the size of this bias is quite small. The worst cases are in Districts B and C, grade 8, where the bias could account for about 2% of the estimated between-teacher variance estimated from the model. The values in the other target teacher groups were all 1% or less. This suggests that our restricted model for the prior scores is not leading to appreciable bias in our estimates of teacher effects in the target year.

5 Results on Teacher Effects and Interactions

5.1 How Large are Teacher Effects and Interactions?

Table 2 summarizes the main results regarding the teacher effects and interactions, based on the complete model fit to each of the twelve target teacher groups. The first column contains the posterior means of the teacher main effect variances τ_0^2 along with 0.025 and 0.975 quantiles of its posterior distribution. This 95% credible interval is the approximate analog to a 95% confidence interval in a classical analysis but in the Bayesian framework has the interpretation of an interval in which the parameter has 0.95 posterior probability of being. When fitting the model to each target teacher group, the target year outcomes Y_i were standardized by subtracting their mean and dividing by their standard deviation so that the target year outcomes have marginal variance 1. This standardization of the outcomes is different than the rank-based z-scores applied to the prior scores because it does not force the scores to have marginal normality; rather it forces a marginal variance of 1 without distorting other properties of the achievement scale which we assume are meaningful and take at face value. Forcing marginal variance of 1 makes the values of τ_0^2 interpretable as the fraction of the marginal variance of the scores accounted for by the variance of the teacher main effects.

The results indicate that teacher main effects account for on the order of 10% of the marginal variance of the target year outcomes, with somewhat higher values in the early

elementary grades of District A and somewhat lower values in District C. These results are roughly consistent with the teacher variance percentages reported in other analyses (Rowan, Correnti, & Miller, 2002; Nye et al., 2004).

[Table 2 about here.]

Calibrating the magnitudes of the interaction terms is more difficult because their magnitudes depend both on τ_1^2 as well as ν , the marginal variance of the general ability parameters. The simplest way to calibrate the size of the interactions is to imagine a teacher with main effect θ_0 and slope θ_1 , a student with general ability δ , and to calculate the total effect that the teacher would have on this student, which under the model is $\theta_0 + \delta\theta_1$. The variance of this quantity under random sampling (i.e. independent sampling assuming no selection) of both the teacher and the student is $\tau_0^2 + \nu\tau_1^2$. The part of this total variance that is due to the interaction effects is $\nu\tau_1^2$, which motivates the quantity $\gamma = \frac{\nu\tau_1^2}{\tau_0^2 + \nu\tau_1^2}$ loosely interpreted as the fraction of the total teacher effect variance that is due to interactions.

The posterior means and 95% credible intervals for γ are reported in the second column of Table 2. These values were obtained by calculating the value of γ for each iteration of the MCMC sample and then calculating summaries of the distribution of this quantity. For the most part, the values are quite consistent across the target teacher groups, on the order of 0.10 or slightly less. A value of 0.10 can be interpreted as 10% of the total variance in teacher effects across all students is due to the interaction terms, with 90% due to main effect variance across teachers. This is not large, but the consistency across different data contexts does suggest the presence of interaction effects. Target groups C7 and C8 have markedly higher values of 0.25 and 0.15, respectively. Sensitivity analyses (not shown) based on dropping students with the lowest possible score, and based on considering teachers with no fewer than 10 linked students, did not lead to appreciable changes in these values. The fact that B7 and B8 are more closely aligned with the values in earlier grades does not suggest that the differences are due to something systematic about middle school math, but it is possible that some combination of the characteristics of the achievement tests and the curricular foci in grades 7 and 8 and district C is leading to larger interaction effects.

The amount that interaction effects might matter in practice, however, depends on how heterogeneously students are grouped across classes. Under the model the average effect that a teacher has on a class is $\theta_0 + \theta_1 \bar{\delta}$ where $\bar{\delta}$ is the average value of δ_i across the students linked to that teacher. A model that ignored the interactions and estimated only teacher main effects, but was otherwise correctly specified, is likely to produce an estimate of the main effect that is close to $\theta_0 + \theta_1 \bar{\delta}$. If all classes had the same value of $\bar{\delta}$, then the resulting comparisons of the teacher effects would not be misleading in the sense that teachers with equal values of θ_0 and θ_1 would not get systematically different estimates due to having different types of students. On the other hand, if $\bar{\delta}$ varies substantially across classes, there is potential for misleading comparisons. Thus, the degree to which the interaction effects matter in practice depends not only on their plausible magnitude (addressed previously), but also on how different $\bar{\delta}$ is across classes.

To investigate this, we began by using the model results to gauge how heterogeneous are groupings of students linked to different teachers. If δ_i for each student were known, this would be a simple matter of calculating the between-teacher variance component of the δ_i . In practice δ_i is not known, but the model estimates can be used in their place. So for each iteration of the MCMC, we estimated the fraction of the total variance of the δ_i that was between classes using a one-way random effects model, which gives rise to the posterior distribution of this quantity. The posterior means and 95% credible intervals of the percentage of the variance of δ_i that is between classes are given in column 3 of Table 2. The results indicate that about 20-30% of the variance in general student ability lies between teachers for Districts A and B, with notably larger values between about 50 and 60% for District C.

Thus students appear to be heterogeneously grouped, and so it made sense to examine the magnitude of the interactions with respect to the empirical results about this heterogeneity. Our strategy for doing this was the following. For each iteration of the MCMC algorithm, we calculated $\bar{\delta}_k$ for each of the K classes in the data that had 10 or more students. For each teacher, we then calculated $\theta_{0j} + \theta_{1j} \bar{\delta}_k$ - that is, the effect that each teacher would have on each of these K classes of 10 or more students, given the current values of all model parameters. If students were not heterogeneously grouped, $\bar{\delta}_k$ would be roughly constant (and equal to zero, since the mean of δ_i is zero) across classes and

so these plausible effects on different classes would all be nearly equal to θ_{0j} . On the other hand, if students are strongly heterogeneously grouped, then $\bar{\delta}_k$ would vary substantially across classes and thus the plausible effects $\theta_{0j} + \theta_{1j}\bar{\delta}_k$ would be more dissimilar within teachers. For each MCMC iteration, we thus calculated JK plausible effects of the J teachers across the K classes in the data with 10 or more students, and calculated the fraction of the variance of these JK plausible effects that was within teachers using a method of moments variance component calculation appropriate for balanced data (Searle, 1971). This quantity is analogous to γ discussed earlier, but rather than using ν , the marginal variance of δ_i , uses something more akin to the between-class variance component of δ_i but which is more closely tied to the actual class groupings of students.

We label this quantity γ^* and its posterior mean and 95% credible interval are given in column 4 of Table 2. The values reflect the patterns evident with γ but are smaller reflecting the fact that class means of the δ_i are not as variable as the δ_i themselves. The values for all target teacher groups are from 0.02 to 0.04 except C7 and C8 which again are substantially higher at 0.15 and 0.09, respectively. With the exception of C7 and C8, these values are small and suggest that estimates made from VAMs that do not account for interactions are probably not grossly misleading for most teachers because of ignoring student-teacher interactions. The variation in main effects is large enough relative to the magnitudes of the interactions and the amount that those interactions would manifest due to heterogeneous classroom groupings that the variation due to ignoring these interactions would not lead to substantive differences in inferences about most teachers. However, teachers with relatively large interaction effects, and who are assigned to classes with $\bar{\delta}_k$ in the relative extremes of the $\bar{\delta}_k$ distribution, could receive estimates that are substantially different from those that might have been obtained had the teacher taught a much different group of students.

5.2 Sensitivities to Scaling

The final column of Table 2 gives the posterior mean and 95% credible interval for r , the correlation between the teacher main effects and teacher slopes. A positive value of r is interpreted as teachers who are effective on average are particularly effective with

students of above-average general abilities, while a negative value indicates that teachers who are effective on average are particularly effective with students of below-average general abilities. From a substantive standpoint, this correlation is potentially interesting because it provides insights into the nature of the effects that teachers have on students.

Unfortunately it appears to be difficult to learn about this correlation in a way that is not strongly tied to the way the tests are scaled. Note that the correlations for Districts A and B are all positive, while the correlations for District C are either indistinguishable from zero or negative. Recall from Section 3 that the test in Districts A and B are from the same test developer and on the same scale. In fact District B was included in the analysis specifically to resolve the discrepancy between Districts A and C because it was impossible to disentangle whether the different correlations in those districts were due to the different grade ranges (note that grade 5 in District C is actually a middle school grade, while grade 5 in District A is an elementary school grade) or different tests. District B has grade ranges overlapping with District C but the same test as District A, and the fact that the correlations more closely resemble those in District A suggests that the difference is due to the nature and scale of the test.

The posterior predictive plots in Figure 2 demonstrate some of the differences in the tests; the test in Districts A and B has a more pronounced non-linear relationship with δ and less pronounced heteroskedasticity than the test in District C. Further evidence that the differences are due to the test was provided by refitting the complete model in District A using a rank-based z-score transformation of Y_i which forces the scale to be marginally normal. The estimated variances due to the interactions in this case were similar to those obtained from the original scale, but the correlations changed from being positive to either zero or slightly negative. And the estimates of the slopes for individual teachers were markedly different depending how the data were scaled, while the estimates for the main effects were virtually identical (correlations of approximately 1). This indicates that inferences about the relationship of the interaction terms to the main effects, and, accordingly, the inferences about the slopes for individual teachers, can be highly sensitive to the properties of the test scale.

6 Summary and Discussion

As discussed in the Introduction, a primary goal of VAM is to provide fair comparisons of teacher performance using statistical adjustments to account for differences in the abilities and other characteristics of students taught by different teachers. Using VAM estimates for some types of high-stakes decisions about teachers is likely to require that the estimates provide inferences about teacher performance that are generalizable across settings, outcomes and units (students). If how a teacher performs with one group of students is not indicative of their likely performance with another group of students with different characteristics, then comparisons of teacher performance based on value-added information are potentially misleading and undermine the goal of trying to provide fair comparisons of teachers teaching different types of students.

This article develops a value-added model specifically designed to estimate student-teacher interactions. The model specifies interactions with respect to a student's general ability level, and estimates an average effect of teachers as well as a parameter that indicates whether individual teachers are more or less effective, relative to their average effect, with students of different ability levels. Interactions with respect to general ability level are an intuitively sensible source of heterogeneity in teacher effects, and if present would provide both challenges and opportunities for the use of VAM estimates.

Using various longitudinal data sources, we find evidence of interactions of teacher effects with students' general ability levels that appear to be of relatively consistent magnitude across different contexts. The magnitude is modest, accounting for on the order of 10% of the total variation in teacher effects across all students. The amount that ignoring these interactions could be biasing VAM estimates of teacher effects depends on how different are the groups of students taught by different teachers. Using empirical estimates of the heterogeneity of students across teachers, we find that the interactions account for on the order of 2%-4% of total variation in teacher effects on different classes, suggesting that ignoring these interactions is not likely to lead to appreciable bias in estimated teacher effects for most teachers in most settings. However, the interactions are estimated to be markedly larger in two of our target teacher groups, and even with small interactions, estimates for teachers with particularly strong interaction effects who are teaching classes

in the extremes of the distribution of average class ability could receive estimates that are not indicative of how these teachers might perform on different classes. Thus, further research on interaction effects in other contexts is warranted, and the results underscore the notion that using any type of statistical adjustment to compare teachers teaching very different types of students is potentially error-prone.

A further complication with estimating the interactions is the evident sensitivity of features of the interactions on idiosyncracies of the scale on which achievement is measured. Our investigations indicate that while the overall magnitude of the interaction effects is not overly sensitive to the scale of the test, the relationship of the interaction effects to teacher main effects is sensitive to the scale, as are the inferences that would be made about the interaction effects of individual teachers. This is in contrast to the findings about teacher main effects which are essentially invariant to rescaling of the tests. This suggests that obtaining generalizable information about the nature of student-teacher interaction effects, particularly of the kind investigated here, might lie beyond what the complications of achievement scales permit.

Future research might consider interaction effects with other student characteristics such as discipline information or other measures of student personality or disposition. Large-scale data on these factors are not generally available, but with the increasing scale and scope of student data collection, the potential to examine interactions with respect to other types of student characteristics will grow. Future research might also try to adapt the estimation of student-teacher interactions to more complex value-added models that account for the complete history of a student's teacher. The results obtained here using a stripped-down model for the prior scores is probably not leading to misleading inferences, but better models are possible.

7 References

- Aaronson, D., Barrow, L., & Sander, W. (2003). *Teachers and student achievement in the Chicago public high schools* (Technical report). Federal Reserve Bank of Chicago.
- Ballou, D. (2004). Rejoinder. *Journal of Educational and Behavioral Statistics*, 29(1), 131–134.

- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37–66.
- Best, N., Cowles, K., & Vines, S. (1995). CODA: convergence diagnosis and output analysis software for Gibbs sampling output, version 0.3. MRC Biostatistics Unit, Cambridge.
- Bifulco, R., & Ladd, H. (2004). *The impact of charter schools on student achievement: Evidence from north carolina* (Technical report). University of Connecticut and Duke University.
- Braun, H. (2005a). *Using student progress to evaluate teachers: A primer on value-added models* (Technical report). Educational Testing Service, Policy Information Center.
- Braun, H. (2005b). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.), *Value added models in education: Theory and practice* (pp. 19–38). JAM Press: Maple Grove, MN.
- Carlin, B., & Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press, second edition.
- Choi, K. (2001). Latent variable modeling in the hierarchical modeling framework in longitudinal studies: A fully Bayesian approach. *Asia Pacific Education Review*, 2(1), 44–55.
- Dee, T. (2003). *Teachers, race and student achievement in a randomized experiment* (Technical report). Swarthmore College.
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733–807.
- Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

- Gill, B., Zimmer, R., Christman, J., & Blanc, S. (2007). *State Takeover, School Restructuring, Private Management, and Student Achievement in Philadelphia (MG-533-ANF/WPF/SDP)*. Santa Monica, CA: RAND. Available from <http://www.rand.org/pubs/monographs/MG533>.
- Goldhaber, A., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Unpublished manuscript.
- Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performance on the job* (Technical report). The Brookings Institution. White Paper 2006-01.
- Hanushek, E., Kain, J., O'Brien, D., & Rivkin, S. (2005). *The market for teacher equality* (Technical report). National Bureau of Economic Research.
- Hanushek, E., Kain, J., & Rivkin, S. (2002). *The impact of charter schools on academic achievement* (Technical report). Hoover Institute.
- Harris, D., & Sass, T. (2006). *Value-added models and the measurement of teacher quality*. Unpublished manuscript.
- Jacob, B., & Lefgren, L. (2006). *When principals rate teachers* (Technical Report 2). Education Next.
- Kane, T., Rockoff, J., & Staiger, D. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. Unpublished manuscript.
- Kane, T. J., & Staiger, D. O. (2008). *Are teacher-level value-added estimates biased? An experimental validation of non-experimental estimates*. Unpublished manuscript.
- Kirby, S. N., McCaffrey, D. F., Lockwood, J., McCombs, J., Naftel, S., & Barney, H. (2002). Using state school accountability data to evaluate federal programs: A long uphill road. *The Peabody Journal of Education*, 99(4), 122–145.
- Koedel, C., & Betts, J. (2005). *Re-examining the role of teacher quality in the educational production function* (Technical report). Department of Economics, UCSD.
- Le, V., Stecher, B., Lockwood, J., Hamilton, L., Robyn, A., Williams, V., Ryan, G., Kerr, K., Martinez, J., & Klein, S. (2006). *Improving mathematics and science education: A longitudinal investigation of the relationship between reform-oriented instruction and student achievement (MG-480-EDU)*. Santa Monica, CA: RAND.

- Lissitz, R. (Ed.). (2005). *Value added models in education: Theory and applications*. JAM Press: Maple Grove, MN.
- Little, R., & Rubin, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, second edition.
- Lockwood, J., McCaffrey, D., Hamilton, L., Stecher, B., Le, V., & Martinez, J. (2007a). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.
- Lockwood, J., McCaffrey, D., Mariano, L., & Setodji, C. (2007b). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, *32*(2), 125–150.
- Lockwood, J., McCaffrey, D., & Sass, T. (2008). *The intertemporal stability of teacher effect estimates*. Paper presented at the Wisconsin Center for Educational Research National Conference on Value-Added Modeling, Madison, WI, April 2008.
- Lockwood, J., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electron. J. Statist.*, *1*, 223–252.
- Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *(10)*, 325–337.
- McCaffrey, D., Han, B., & Lockwood, J. (2008). *From data to bonuses: A case study of the issues related to awarding teachers pay on the basis of their students' progress*. Presented at National Center for Performance Incentives conference *Performance Incentives: Their Growing Impact on American K-12 Education*, Vanderbilt University, Nashville, TN, February 2008.
- McCaffrey, D., Lockwood, J., Koretz, D., Louis, T., & 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., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating*

- value-added models for teacher accountability, (MG-158-EDU)*. Santa Monica, CA: RAND.
- Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage Publications, second edition.
- Raudenbush, S. W. (2004). What are value added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.
- Rothstein, J. (2007). *Do value-added models add value? Teaching, fixed effects, and causal inference* (Technical report). Princeton University and National Bureau of Economic Research.
- Rowan, B., Correnti, R., & Miller, R. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104*, 1525–1567.
- Rubin, D. B., Stuart, E., & Zanuto, E. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*, 103–116.
- Sanders, W., Saxton, A., & Horn, B. (1997). The Tennessee Value-Added Assessment System: A Quantitative Outcomes-Based Approach to Educational Assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluational Measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press, Inc.
- Sanders, W., Wright, S., Springer, M., & Langevin, W. (2008). *Do teacher effects persist when teachers move to schools with different socioeconomic environments?* Presented at National Center for Performance Incentives conference *Performance Incentives: Their Growing Impact on American K-12 Education*, Vanderbilt University, Nashville, TN, February 2008.
- Schacter, J., & Thum, Y. M. (2004). Paying for high and low-quality teaching. *Economics of Education Review, 23*, 411–430.

- Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.
- Searle, S. (1971). *Linear Models*. New York: John Wiley & Sons.
- Seltzer, M., Choi, K., & Thum, Y. (2002). *Examining relationships between where students start and how rapidly they progress: Implications for constructing indicators that help illuminate the distribution of achievement within schools* (Technical report). CRESST/University of California, Los Angeles.
- Seltzer, M., Choi, K., & Thum, Y. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis*, 25(3), 263–286.
- Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 64, 583–639.
- Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82(398), 528–550.
- Thum, Y. (2003). Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods and Research*, 32(2), 153–207.
- van Dyk, D., & Meng, X. (2001). The art of data augmentation (with discussion). *Journal of Computational and Graphical Statistics*, 10(1), 1–111.
- Zimmer, R., Buddin, R., Chau, D., Daley, G., Gill, B., Guarino, C., Hamilton, L., Krop, C., McCaffrey, D., Sandler, M., & Brewer, D. (2003). *Charter school operations and performance: Evidence from California*. RAND: Santa Monica, CA.

8 Appendix - Additional Details on Prior Distributions and MCMC Results

All of the variance components in the model (ν , the variance of δ_i ; τ_0^2 , the variance of the teacher main effects; τ_1^2 , the variance of the teacher slopes; σ^2 , the variance of the residual error terms for the target outcome; and σ_p^2 , the variances of the residual error terms for the prior scores) were given prior distributions that were uniform on their square roots (standard deviations). The target year test scores were standardized to have mean zero and variance one and all prior distributions were chosen to be consistent with scores on this scale. $\sqrt{\nu}$ was modeled with a $U(0.5, 0.9)$ distribution, consistent with the student effects accounting for between about 25% and 80% of the marginal variance of the target year outcomes. τ_0 and τ_1 were modeled as independent $U(0.0, 0.7)$ consistent with teacher effects accounting for no more than 50% of the marginal variance of the target year scores. σ was modeled as $U(0, 0.7)$ consistent with factors other than student and teacher effects accounting for no more than 50% of the marginal variance of the target year scores. However, the σ_p were modeled with independent $U(0, 1)$ priors to allow the possibility that some prior scores (which were normed to have marginal variance one) were only weakly related to δ_i and thus essentially all of the variance is unexplained.

The parameters λ_{m0} and λ_{m1} governing the nonlinearity were given independent normal priors with mean zero and standard deviation 0.25, which made values larger than 0.5 in absolute value relatively unlikely under the prior distribution. Such a restriction was consistent with the likely strength of nonlinearities given the scaling of the data and δ_i . The parameters λ_{v0} and λ_{v1} governing the heteroskedasticity were given independent normal priors with mean zero and standard deviation 0.5, again essentially restricting the parameters from taking on values that would imply extreme heteroskedasticity inconsistent with what is expected with achievement scores. Finally, the correlation r between the teacher intercepts and slopes was modeled as $U(-1, 1)$ to allow the correlation to take on any possible value.

All models were fitted in WinBUGS (Lunn et al., 2000) using three parallel chains. Each chain was burned in for 7,500 iterations, a value chosen based on preliminary investigations and convergence diagnostics. Then, each chain was run for 10,000 iterations

and 1,000 evenly-spaced iterations were saved from each chain, totalling 3,000 posterior samples for each model fit to each target teacher group. All inferences reported in the article are based on summaries of these sets of 3,000 samples other than the DIC values, which were based on the full 30,000 post-burn in iterations from each model fit. Convergence was assessed using the diagnostic statistic for multiple parallel chains of Gelman and Rubin (1992) as implemented in the CODA package for the R Environment (Best, Cowles, & Vines, 1995).

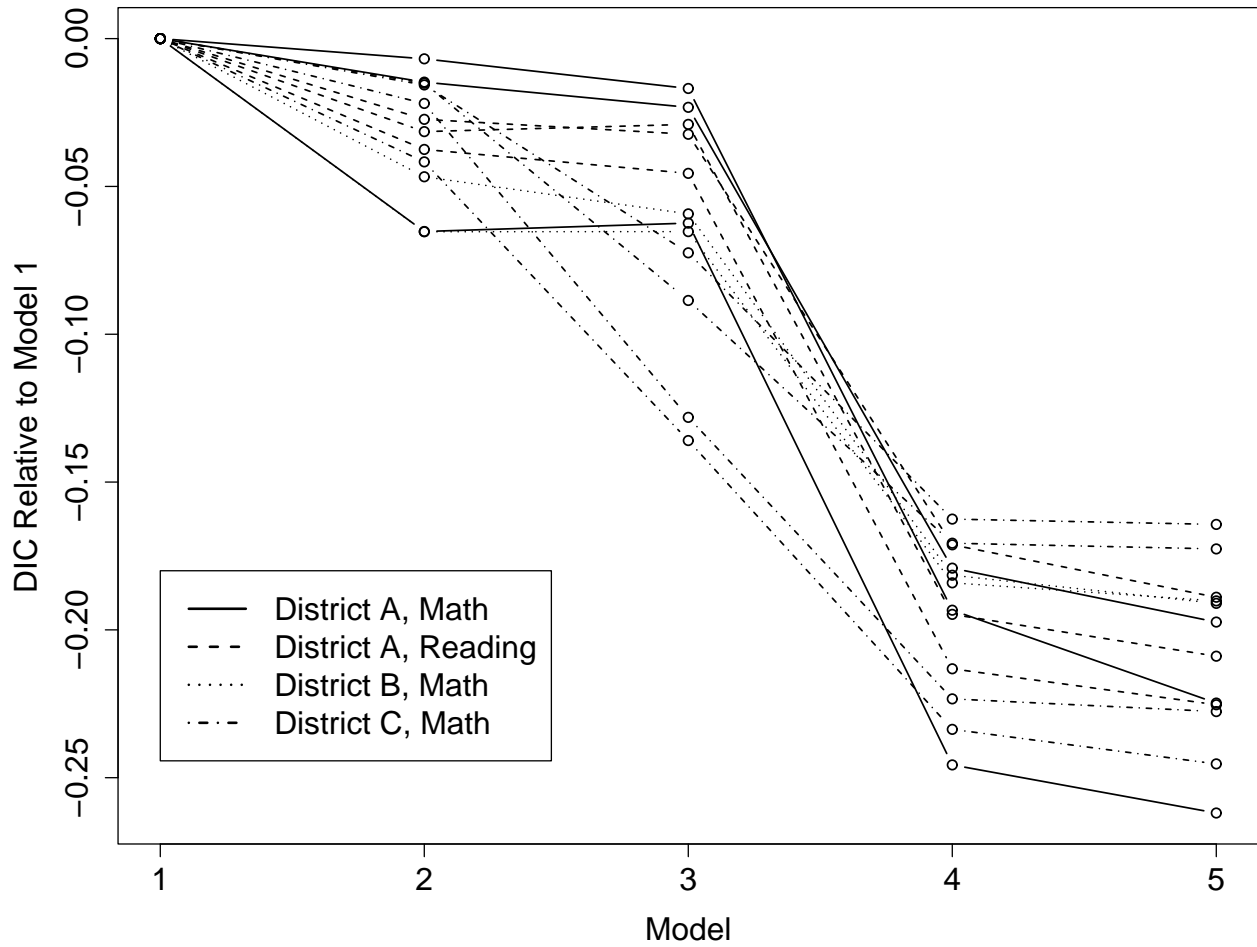


Figure 1: DIC for sequence of five increasingly complex models, scaled for Model n as $(DIC_n - DIC_1)/DIC_1$. Smaller (i.e. more negative) values indicate preferred models. Model 5 is the complete model that includes nonlinearity, heteroskedasticity, teacher main effects and teacher-student interaction terms.

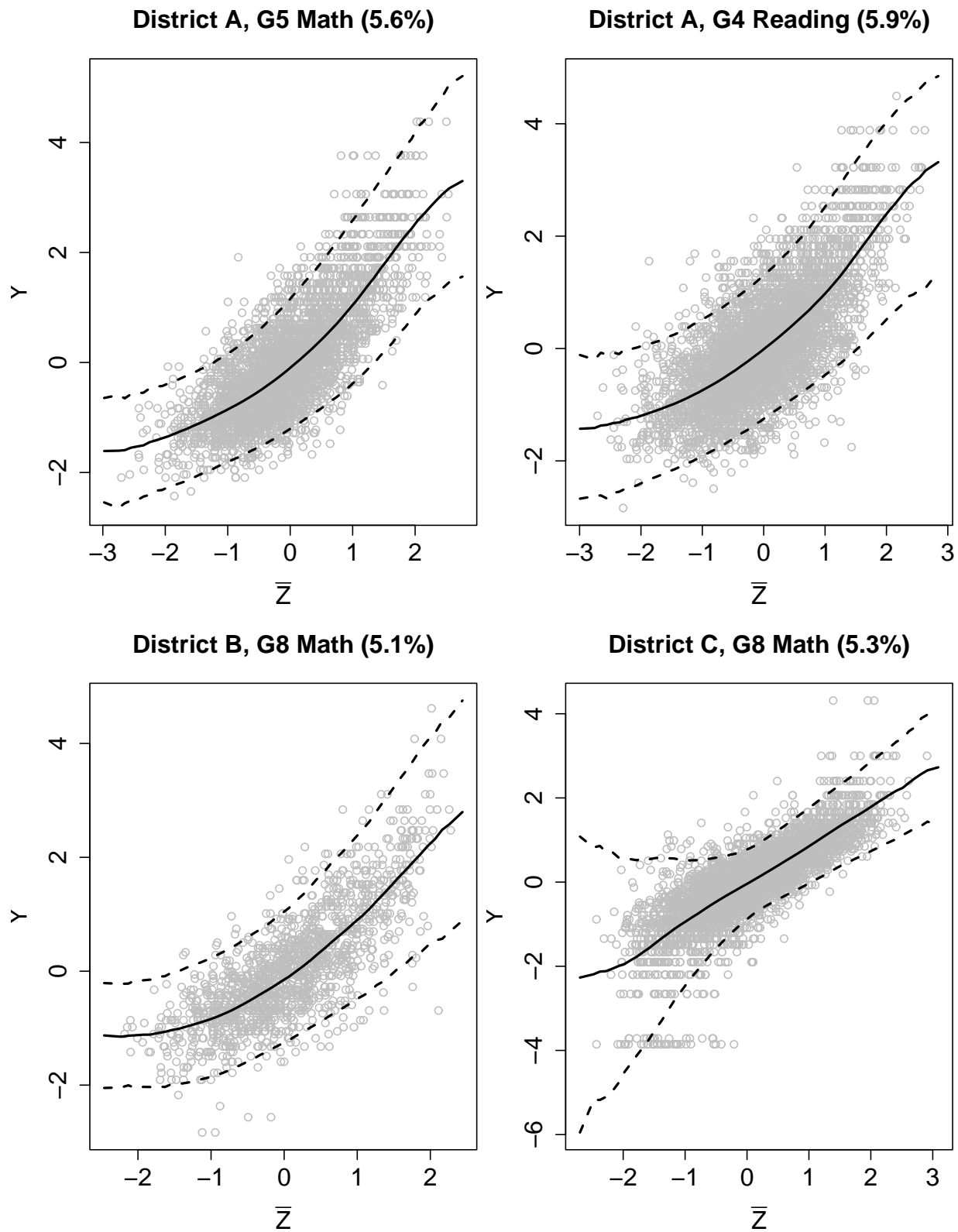


Figure 2: Posterior predictive checks on the relationship between the average prior score and Y_i for selected target groups. Solid line gives conditional median and dotted lines give conditional 0.025 and 0.975 quantiles. Numbers in parentheses at the top of each figure give the percentage of the observed data pairs that fall outside of the predictive bounds.

District B, G8 Math (Selected Teachers)

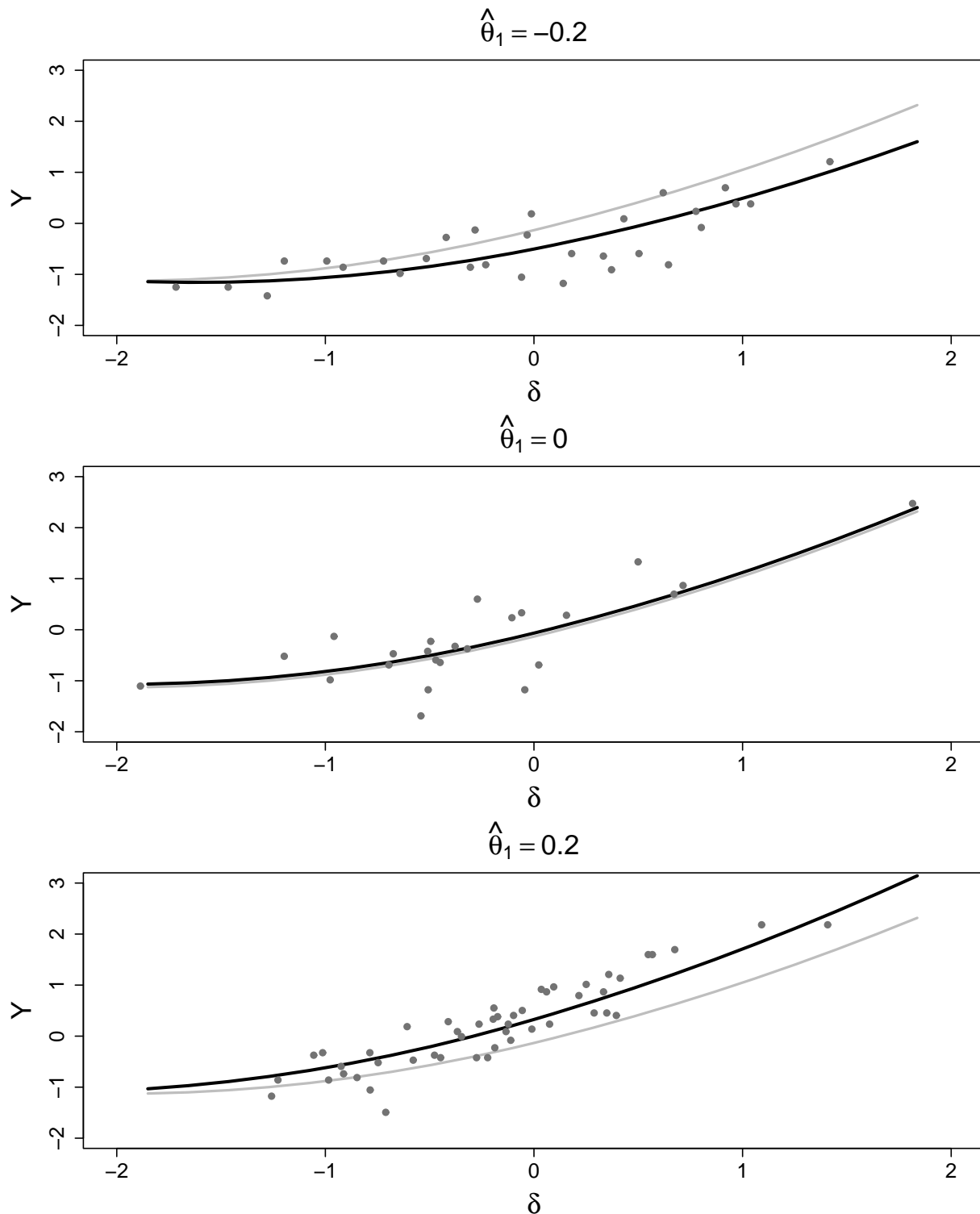


Figure 3: Examples of checks made on individual teachers. Gray lines indicate estimated expected value of Y_i as a function of δ in the absence of any teacher effects. Black lines give the estimated expected value of Y as a function of δ accounting for the effects of a particular teacher. The points in each frame are the actual Y_i values for students linked to that teacher, plotted as a function of posterior mean of δ_i for those students. Posterior means of the slope parameters for those teachers are given at the top of each plot.

Group	District	Grade	Subject	# Teachers	# Students	Max # Prior Scores	Mean # Prior Scores
A3M	A	3	math	302	4028	4	3.3
A4M	A	4	math	320	4539	6	4.6
A5M	A	5	math	254	3633	8	6.4
A3R	A	3	reading	301	3991	4	3.3
A4R	A	4	reading	319	4476	6	4.6
A5R	A	5	reading	254	3599	8	6.4
B7	B	7	math	39	1443	8	7.2
B8	B	8	math	35	1424	11	8.6
C5	C	5	math	193	3617	8	7.4
C6	C	6	math	170	3972	12	10.8
C7	C	7	math	142	4170	14	12.3
C8	C	8	math	133	4030	12	10.8

Table 1: Summary information about the twelve target teacher groups examined in the analyses. The descriptive labels in the first column used in the presentation of the results.

Group	τ_0^2	γ	% Var(δ_i) Between teachers	γ^*	r
A3M	.20 (.16,.24)	.10 (.07,.15)	21.8 (20.6,23.1)	.03 (.02,.04)	.23 (.03,.44)
A4M	.13 (.11,.16)	.11 (.06,.16)	30.3 (29.1,31.7)	.03 (.02,.05)	.50 (.31,.68)
A5M	.11 (.10,.14)	.10 (.05,.16)	28.2 (27.2,29.3)	.03 (.02,.05)	.65 (.42,.88)
A3R	.14 (.12,.17)	.11 (.06,.16)	22.4 (21.2,23.5)	.03 (.02,.05)	.31 (.08,.54)
A4R	.14 (.11,.17)	.10 (.05,.16)	31.0 (29.7,32.3)	.03 (.02,.05)	.41 (.18,.63)
A5R	.11 (.09,.14)	.07 (.03,.12)	28.8 (27.7,29.8)	.02 (.01,.04)	.60 (.21,.89)
B7	.09 (.05,.14)	.09 (.02,.21)	28.0 (26.6,29.3)	.03 (.01,.07)	.63 (.16,.96)
B8	.10 (.06,.17)	.11 (.03,.24)	30.9 (29.8,32.1)	.04 (.01,.09)	.65 (.23,.94)
C5	.08 (.06,.10)	.09 (.00,.20)	48.0 (47.0,49.0)	.04 (.00,.09)	-.21 (-.64,.31)
C6	.08 (.06,.11)	.07 (.00,.18)	51.0 (50.1,51.8)	.03 (.00,.08)	-.32 (-.77,.25)
C7	.05 (.04,.08)	.25 (.12,.39)	60.0 (59.2,60.8)	.15 (.08,.25)	-.41 (-.67,-.10)
C8	.07 (.05,.10)	.15 (.06,.28)	58.5 (57.7,59.3)	.09 (.03,.16)	-.63 (-.87,-.29)

Table 2: Posterior means and 95% credible intervals for key parameters for each of the twelve target teacher groups.