How Sensitive Are School Value-Added Estimates to Methodology?

Julian R. Betts* Sam M. Young*

June 1, 2017

Abstract

Education scholars frequently use "value-added" models to evaluate the effectiveness of teachers or schools, to help account for differences in inputs. While the idea behind value added is simple, in practice there are many variations of the model that researchers use, which may affect the resulting estimates. This paper uses panel data from San Diego Unified School District to estimate school value-added using a variety of models to examine sensitivity to both the control specification and overall class of value-added model.

Our results suggest a modest sensitivity to the choice of covariates, and a somewhat larger sensitivity to the overall class of value-added model. Results from a gain-score model and an Arellano-Bond model tend to show low correlation with one another, while an uninstrumented level-on-lag model shows moderate correlation with both. Precision of estimates appears sensitive to the inclusion of school-level controls, and whether researchers use panel instrument techniques to address dynamic panel bias. Finally, we observe that value-added estimates show a moderate positive correlation with schools' average levels of test scores, but less so under the gain-score model than the other two classes of models.

*Department of Economics, University of California, San Diego. We are grateful to Andrew Zau, Karen Bachofer, and administrators at San Diego Unified School District for their essential support on data and administrative issues. We thank Andrew Chamberlain, Julie Cullen, Gordon Dahl, Itzik Fadlon, Matt Gibson, Mark Jacobsen, Craig McIntosh, Marc Muendler, Karthik Muralidharan, Paul Niehaus, Sally Sadoff, Krislert Samphantharak, and other participants at UC San Diego for comments made during presentation.

1 Introduction

The most readily available information on school performance comes in levels of test scores. Yet levels of test scores do not necessarily identify how effective schools are, as schools differ on many external factors over which they have little control — such as students' socioeconomic status or family attitudes toward schooling.

Rather than looking only at raw test scores, the effectiveness of schools is better characterized by how well they perform *given* their students' background and academic histories. Toward this goal, there has been an increase in value-added modeling (VAM) for both teachers and schools during the past 10–20 years, which measures effectiveness based on within-student gains in achievement. And while many applications of school accountability still rely on levels of performance, rather than value-added, this could change in upcoming years as VAM continues to gain popularity.

But even with growing popularity, there is not a uniform standard for how to conduct value-added modeling. This may sometimes affect results. In a real world example, in 2010 the Los Angeles Times newspaper published a relative ranking of local schools by value-added, based on research by Buddin (2010). About seven months later, the Los Angeles Unified School District provided its own estimates of schools' value-added. Schools often did not rank similarly between the two analyses.[1]

If VAM becomes a key policy tool, it is essential to know whether its results are sensitive to the methodology used. While there have been some attempts to investigate this issue, most of these studies have looked at teacher value-added. It is unclear whether their findings necessarily carry over to school value-added, since the issues that present econometric challenges differ somewhat between the two cases. Thus, in this paper, we estimate school value-added using a variety of models to examine sensitivity to both the set of controls used and overall class of value-added model.

Our results indicate a modest sensitivity of value-added estimates to the control specification, and a somewhat larger sensitivity to the overall class of value-added model used. Correlations between individual model classes are heterogeneous, but in ways that can largely be explained by the assumed degree of persistence in each model. We also find the precision of estimates to be sensitive to the inclusion of school-level controls, and whether researchers use panel instrument techniques to address dynamic panel bias.

Finally, we compare how well school value-added estimates correlate with schools' average levels of test score. We observe an interesting pattern that although school-average levels show almost

[1] Imberman and Lovenheim (2016) provide additional details about this example from Los Angeles. Figure 1 in their appendix shows scatterplots of schools' value-added estimates in the various analyses.

no correlation with school-average gains (when using minimal covariates), in most specifications value-added shows a moderate positive correlation with average test score levels.

The rest of the paper proceeds as follows. Section 2 reviews the relevant literature, Section 3 summarizes the methods used, while Section 4 summarizes the data used. Section 5 previews results, focusing on the relative precision of various models, sensitivity of school rankings to covariates and model class, and comparing the value-added models to results from a naive analysis of test score levels, which is the typical approach used in states' accountability systems. Section 6 concludes.

2 Value-Added Overview and Related Literature

This section provides a basic overview of VAM approaches to help motivate the models that we assess. Equations are written for the context of school value-added, though the patterns that we describe are also largely true of teacher value-added. We then the review the segment of the literature that is most closely related to the current piece. A more complete literature review can be found in Koedel, Mihaly and Rockoff (2015).

2.1 General Overview

VAM typically models students' test score advancement over time while including a set of school indicators as regressors. An example specification is

$$\Delta S_{ist} = \alpha_s + \epsilon_{ist}, \tag{1}$$

where ΔS_{ist} denotes the one-year change in test score for student i at school s at time t, and α_s represents the value-added parameter for school s. This model is often referred to as the "gain-score" model.

A second common class of value-added model takes the form

$$S_{ist} = \rho S_{is,t-1} + \alpha_s + \epsilon_{ist}, \tag{2}$$

which we refer to as the "level-on-lag" model. Instead of using the change in score as the outcome, this model regresses the level of test score on the lagged level, along with school dummies and possibly covariates. This model is identical to the gain-score model when $\rho = 1$, but is otherwise more general, since it allows ρ to take possibly different values — as may be appropriate if past learning depreciates over time or otherwise does not carry over perfectly between years.[2]

[2]In recent years, a majority of studies seem to favor the "level-on-lag" specification of Equation 2 over the "gain" specification of Equation 1 (Koedel, Mihaly and Rockoff, 2015). However, there remain many earlier value-added studies that use the gain-score model.

Researchers sometimes differ on what covariates to include. While the preceding equations are written without covariates, for simplicity, in practice these models often feature an additional set of control variables, which varies in parsimony between studies. For instance, the Tennessee Value-Added Assessment System (Sanders and Horn, 1994, 1998) advocates a parsimonious specification with few covariates — relying instead on the gain structure of VAM to (implicitly) control for student background factors through the prior test score. The argument is that, to the extent students do not sort randomly to schools, inclusion of some variables like socioeconomic status may not just capture the intended control effects, but may also proxy for school effectiveness. When this is the case, part of a school's actual effectiveness may attach to the control variables instead, resulting in biased estimates of value-added. Nevertheless, many researchers prefer to use explicit controls for student background and other factors during VAM. The most common of these controls are basic student characteristics, such as race and socioeconomic status, followed by aggregate characteristics at the school- or classroom-level.

It is also common to see Equations 1 and 2 specified with or without student fixed effects. We note, however, that the inclusion of student fixed effects may be better-suited for teacher value-added settings (which to date have been the majority of value-added studies) than school value-added settings. The reason is that, once student fixed effects are included, identification of the value-added parameters comes only from students who change teachers or schools. But whereas most students change teachers from year to year, most students do not change schools between years, except at set transitional grades (like going from middle school to high school).[3] Student fixed effects thus remove a large degree of variation in school value-added settings. This issue is somewhat more innocuous in teacher value-added settings, but in both cases, the reduction in variation may result in noisier estimates of key parameters.

This highlights a few other important differences between school value-added and teacher value-added settings. While non-random sorting is a common issue to both, it likely poses a greater danger in teacher value-added models because students are sorted not only between schools, but also likely across classrooms due to ability grouping or tracking. Limited sample size is also an issue in the teacher value-added literature, where a given teacher may only be observed teaching a small number of students. Schools, on the other hand, typically generate a larger average sample of students, but may still be affected by size in a different manner, since schools themselves vary in size. Typically, changes in average test scores tend to be more volatile for smaller schools (Chay, McEwan and Urqiola, 2005; Kane and Staiger, 2002).

That the econometric challenges differ somewhat between the two cases is important, given that

[3]In our data, roughly 80% of students attend only one elementary school. Moreover, the minority of students who do change schools may be fundamentally dissimilar from the rest.

the vast majority of the value-added literature has focused on teachers, rather than schools. We revisit this issue toward the end of our literature review.

2.2 Similar Papers

Having described the basics of value-added, we next review the segment of the literature that is most closely related to our paper. We focus on papers that estimate value-added using a variety of model specifications and/or estimation techniques, with the specific intent comparing their performance.

One subset of papers studies the impact of student demographic controls on teacher value-added estimates. This has sometimes been a point of contention among researchers, due to the absence of covariates in the Tennessee Value-Added Assessment System (TVAAS). Papers studying this issue have reached somewhat mixed conclusions. Ballou, Sanders and Wright (2004) find that the inclusion of demographics controls affects teacher value-added estimates under a simple fixed effects estimator, but not under the TVAAS.[4] McCaffrey et al. (2004), meanwhile, use a simulation approach and find that the sensitivity to covariates is lower when students sort homogeneously students sort to teachers and schools, on these observed characteristics.

A second group of papers considers more general variations in the model specification, often changing both the set of covariates and overall class of value-added model. We highlight three papers in this group: Lockwood et al. (2007), Newton et al. (2010), and Tekwe et al. (2004).

These papers perform similar exercises, though the first two focus on teacher value-added, while the third focuses, like we do, on school value-added. The papers estimate value-added using a variety of specifications and/or estimation techniques, before computing correlations of how well these estimates align across cases. Lockwood et al. (2007) use the most extensive menu of models — featuring 5 sets of covariates and 4 classes of models, for 20 specifications total — to study the sensitivity of teacher value-added estimates. They find generally high correlations across models when changing either covariates, model class, or both. Newton et al. (2010) use a somewhat smaller array of models, but find similar results. Both of these preceding papers suggest only a mild degree of sensitivity to the model specification.[5] However, this is not the case for Tekwe et al. (2004), who study school value-added, and find results sensitive to the inclusion of student-level covariates. In

[4]Ballou, Sanders and Wright (2004) argue that, in practice, the importance of covariates may depend on how fully a model makes use of the information contained in students' test scores. The more information about a student's academic traejectory that a model gleans from test scores themselves (perhaps by using covariances between scores in different subjects and years, like in the TVAAS), the less additional information is gained by covariates, which would otherwise proxy for similar information — and the less the results change.

[5]These authors do not, however, conclude that value-added estimates are universally robust to methodology. Model specification aside, Lockwood et al. (2007) highlight sensitivity to using different math achievement measures, while Newton et al. (2010) highlight temporal instability of teacher value-added estimates, and that teachers can have different value-added by course taught.

their paper, school value-added estimates from the model with student controls show the lowest correlations with all the other models. This is especially true for third grade math, where the correlation with estimates from other models is typically around 0.6.

(We note that in this last finding, Tekwe et al. (2004) introduce covariates only at the same time as a change in the overall structure of the value-added model, making it unclear what role each respective component plays. But the end finding of sensitivity still differs from the earlier papers. Our paper extends upon Tekwe et al. (2004) by separately exploring sensitivity to covariates and to the class of value-added model.)

Sass, Semykina and Harris (2014) is another similar paper that estimates teacher value-added using a variety of models. Their paper initially focuses on testing the underlying assumptions of value-added, which they typically reject, before also comparing the similarity of estimates across models.[6] They do so in two ways: (1) reporting how often the same teacher is estimated as being in the top 10% or bottom 10% across separate models, and (2) computing rank correlations of the teacher value-added coefficients across models. Doing so, they find a wide range in the similarity of estimates across models. While some models give similar results to others, other combinations give very different estimates of teacher value-added. The presence of student fixed effects, in particular, seems to make a big difference. While estimates from models without student fixed effects tend to show high correlations with those of other models without student fixed effects, this often is not the case when comparing two models that both have student fixed effects, or when comparing a model with student fixed effects to one without. The authors comment that this may be due to the reduction in variation that comes from using student fixed effects, which results in noisier estimates.

Guarino, Reckase and Wooldridge (2015) use a simulation approach to assess the performance of various teacher value-added models and estimation techniques while setting the data-generating process. Their simulations consider multiple variations for how students are assigned to teachers (not always randomly), the degree of persistence between years (ρ), and whether teacher effects themselves are large or small. They then assess the performance of each model/estimation approach under each of the different scenarios. Unlike the other papers mentioned, they do not compare the similarity of value-added estimates from each estimation approach against those of the other approaches, but against the underlying truth that they specify. Unsurprisingly, they find some methods of estimating value-added to be more reliable at capturing actual teacher effectiveness than others, across the various data-generating scenarios. What is more surprising is that ordinary least squares (while controlling for lagged test score level) appears to be the best

[6] See Todd and Wolpin (2003) for an excellent primer on value-added that motivates and explains the underlying assumptions.

overall performer — outperforming several alternatives that offer theoretical improvements based on structural considerations.

To recap, there have been somewhat mixed findings from the literature about the sensitivity of value-added estimates to the methodology — with some papers reporting higher degrees of sensitivity to covariates and/or model class than others. Further complicating matters is the fact that most of these papers have studied teacher value-added, and not school value-added. (Among the papers mentioned in our literature review, only Tekwe et al. (2004) studies school value-added.) Due to some differences in the econometric pitfalls between the two cases, it is unclear which results also apply to school value-added.

3 Methods

This paper features a menu of nine (9) value-added models, consisting of three model classes and three sets of covariate specifications each. Our main analysis estimates value-added using these nine specifications, before computing correlations across models to see how sensitive results are to methodology. We then explore a number of extensions, described in the Results.

Our three classes of value-added models are listed below:

$$\Delta S_{igst} = \alpha_s + \gamma_g + \gamma_t + x'_{igst}\beta + \epsilon_{igst} \tag{3}$$

$$S_{igst} = \rho S_{igs,t-1} + \alpha_s + \gamma_g + \gamma_t + x'_{igst}\beta + \epsilon_{igst} \tag{4}$$

$$\begin{aligned} \Delta S_{igst} = \rho \Delta S_{igs,t-1} + \alpha_s \Delta 1\{\text{school } s\} + \gamma_g \Delta 1\{\text{grade } g\} \\ + \gamma_t \Delta 1\{\text{year } t\} + \Delta x'_{igst}\beta + \Delta \epsilon_{igst} \end{aligned} \tag{5}$$

where S_{igst} denotes the standardized test score of student i in grade g at school s at time t; ΔS_{igst} denotes the one-year change in test score; $(\alpha_s, \gamma_g, \gamma_t)$ denote school, grade, and year fixed effects respectively; x'_{igst} denotes a vector of other control variables; and ϵ_{igst} denotes an error term.

Equations 3 and 4 correspond to the gain and level-on-lag models introduced previously. Equation 5 is a transformed version of the level-on-lag model, in first difference, that allows for and sweeps out a student fixed effect. We estimate this equation using an Arellano-Bond (1991) estimator that instruments the change in lagged test score, as well as the changes in the school dummies, which we treat as endogenous.

The Arellano-Bond estimator helps address the issue of dynamic panel bias (Nickell, 1981), which arises in equations featuring both student fixed effects and lagged dependent variables as regressors. An undesired result is that, if left unchecked, this may prevent our uninstrumented

level-on-lag model from returning consistent estimates.[7] Interestingly, this issue is only sometimes addressed by the existing value-added literature.[8]

Equations 3 and 4 are estimated using ordinary least squares (OLS), with robust standard errors clustered by school. We otherwise use the xtabond2 command in Stata (Roodman, 2009) to estimate our Arellano-Bond regressions, which allows us to preserve school-level clustering, and also allows fine control of the instrument matrix. Additional details about how we perform this estimation are provided in the appendix.

Within each model class, we also feature three different specifications of controls:

Control Specification A: school, grade, and year fixed effects

Control Specification B: specification A + student characteristics

Control Specification C: specification B + school demographics

where the exact variables that comprise 'student characteristics' and 'school demographics' are listed in Table 1. These specifications are chosen to mirror the value-added literature, where student characteristics are the most commonly used controls, and school demographics are the next most common.

4 Data

To estimate these value-added models, we use data from San Diego Unified School District spanning a five-year period from the 2006-07 to 2010-11 school years. The dataset records detailed information at the student-year level, and tracks these students over time to form a panel.

Key variables include a student's grade and primary school attended each year, race, parental education, English learner status, special education status, and school-level demographics. Student achievement is measured using scores on the California Standards Test (CST) in Math and English Language Arts (henceforth 'Reading'), which we convert to Z-scores relative to the state mean and standard deviation, for the respective year and grade level. CST scores are available from grade 2 onward, when standardized testing begins. These tests are administered each year in the spring, toward the end of the school year.

Although the district records these data for students of all grade levels, for tractability reasons, we restrict our analysis to students in elementary schools.[9] Keeping only cases with available

[7] The key issue is that in short (small T) panels, any unmodeled shocks affect the *estimated* value of the fixed effect, creating a correlation between the lagged dependent variable and the error term.

[8] Some example papers that use panel instrument methods include Andrabi et al. (2011), Guarino, Reckase and Wooldridge (2015), and Sass, Semykina and Harris (2014). Many other papers seem to ignore or overlook issue.

[9] A key issue for secondary grades is that not all students take the same CST Math test, even within a given grade.

outcomes, we have a total of 162,921 student-year observations, comprised of 77,647 unique students across 120 elementary schools.[10] Table 1 provides summary statistics for this group of students.

5 Results

5.1 Distribution of Value-Added Coefficients and Their Precision

Before showing correlations of estimates across specifications, we first examine basic patterns in the distribution of school coefficients and their standard errors, as this information may help inform some of the later results. Table 2 provides summary information (mean and standard deviation) about the estimated value-added coefficients and their standard errors, under each specification.

Looking first at the standard error columns, two patterns can be clearly seen. The first pattern is that, within any class of model, standard errors are much larger under control specification C, which introduces school-level demographics, than either control specifications A or B. The most likely explanation seems to be collinearity between the school demographic variables and the school fixed effects. Values of school demographics, such as the percentage of African-American students, typically do not drift much during the 5-year period that we study, such that there is not much within-school variation for these variables. Standard errors increase dramatically when the model nevertheless attempt to estimate coefficients for these variables in addition to school fixed effects.

The second pattern is that the Arellano-Bond model gives much larger standard errors than either the gain-score model or uninstrumented level-on-lag model. This pattern is not entirely surprising, since the Arellano-Bond (1991) estimator first-differences the data, meaning that school coefficients are identified by the subsample of students who switch schools — only about one-fifth of students. This is a major limitation of this approach. Precision is also be affected by the use of an instrumental variables procedure.

Both of these patterns on standard errors also carry over to the coefficient estimates. Looking at the coefficient columns in Table 2, coefficients themselves are also most widely dispersed under control specification C (for any class of model), and more widely dispersed under the Arellano-Bond model than either the gain-score or level-on-lag model. This has systematic effects on any resulting correlations. Other things the same, we typically expect lower correlations for less precise estimates.

One way to mitigate this issue is to apply shrinkage, or adjust schools' value-added estimates

This makes it difficult to define a school's overall value-added for Math. Students in elementary grades, meanwhile, all take the same Math CST test, such that this issue does not apply.

[10] We ultimately omit six elementary schools with only a few students during this time. These omitted schools typically have fewer than 10 gain-scores available, where ΔS_{igst} is well-defined, in either subject. Of the remaining 114 schools, the average number of gain-scores is about 720 in each subject.

toward the baseline average, depending on their precision. This is a common, though not universal, practice in the value-added literature.[11] Following Morris (1983), we use a shrinkage estimator of the form

$$\hat{\alpha}_s^{shrunk} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + V_s}\right)\hat{\alpha}_s + \left(1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + V_s}\right)\bar{\alpha},$$

where $\hat{\sigma}^2$ denotes the sample variance of the estimated school coefficients, V_s denotes the variance of school s's coefficient estimate, and $\bar{\alpha}$ denotes the overall average of the school value-added coefficients, for the given specification. The more noisily that $\hat{\alpha}_s$ is estimated, the more weight is given to the baseline average, rather than the unadjusted estimate, and vice versa.

The effects of shrinkage can be seen in Figures 1 and 2, which show histograms of the value-added coefficients for Math and Reading respectively, for both shrunk and unshrunk cases. For many specifications, the shrunk estimates are nearly identical to their unshrunk counterparts. But shrinkage has a pronounced impact on specifications with larger baseline standard errors — specifically anything that uses control set C or the Arellano-Bond (1991) estimator. In these cases, the shrunk coefficients typically still have a similar mean as the unshrunk case, but are much more tightly packed around this mean.

In a few cases, the shrunk estimates also exhibit more skewness than the unshrunk estimates. (See, for example, the gain-score model and level-on-lag model for Reading under control specification C.) However, we believe these to be idiosyncratic to our results in those particular specifications, and not a general property of the shrinkage estimator.[12]

5.2 Sensitivity to Covariates and Model Class

We use three ways of showing the sensitivity of estimates across value-added models: (1) showing correlations between the various specifications; (2) showing scatterplots of school coefficients; and (3) showing transition matrices for how schools' quintile rankings compare under different specifications. For each of these three cases, tables and figures in the main text use unshrunk estimates throughout, though we discuss the effects of shrinkage and provide analogous tables/figures using shrunk estimates in the appendix.

[11] Among non-economist researchers, value-added models are often estimated using hierarchical linear modeling (HLM) or empirical Bayes estimators, which intrinsically "shrink" a school's value-added parameter toward the baseline average, depending on its precision. Economists, on the other hand, typically do not use HLM as the initial estimation technique, but sometimes achieve a similar shrinkage effect after the fact. See, for example, Chetty, Friedman and Rockoff (2014) and Lefgren and Sims (2012).

[12] Imprecise estimates (or those most affected by shrinkage) are somewhat more likely to be in the left or right tails of the initial unshrunk distribution, since the imprecision of these estimates helps them take more extreme values. Skewness may result after shrinkage if, for some reason, a disproportionate share of these imprecise values come from the right tail rather than the left, or vice versa.

Table 3 begins by showing Spearman rank correlations of value-added estimates across control specifications for each of the three classes of value-added models that we use.[13] We henceforth refer to these as "cross-control" correlations, while correlations across model classes, for a given set of controls, are called "cross-model class" correlations.

We generally find only a mild degree of sensitivity to the covariates used. One exception is that the gain-score model appears somewhat sensitive to the use of school-level demographics — where estimated coefficients from models that include these variables show rank correlations of 0.52–0.64 compared to other versions of the gain-score model. But cross-control correlations are otherwise generally above 0.7 in all three models, for both Math and Reading. The overall average correlation when changing covariates alone is about 0.781 in Table 3.[14]

Table 4, meanwhile, shows rank correlations of value-added estimates across model classes for each of control specifications A, B, and C. As a whole, these correlations are considerably lower than those seen in Table 3, suggesting that value-added estimates are much more sensitive to the overall class of model than to the control specification. Strikingly, correlations between the gain-score model and the Arellano-Bond model are as low as about 0.1 for both Math and Reading, under control specification A. As an overall average, the average across-model class correlation is about 0.501 in Table 4, but with substantial heterogeneity between model classes.

The value of the autoregressive parameter, ρ, may explain much of the patterns observed between model classes. Whereas the gain-score model imposes $\rho = 1$, both versions of the level-on-lag model estimate ρ to be significantly lower. In our data, the uninstrumented level-on-lag model typically estimates $\hat{\rho} \approx 0.7$, while the Arellano-Bond model estimates $\hat{\rho} \approx 0.25$, for both Math and Reading. This ordering for the perceived value of ρ is consistent with the correlations between model classes. Specifically, we see that the gain-score model and Arellano-Bond model show the lowest correlation with one another under any control specification (and about 0.222 on average within Table 4), while the uninstrumented level-on-lag model shows at least moderate correlation with either of the others (0.741 with the gain-score model, and 0.540 with the Arellano-Bond model when averaging across Table 4).

It is also plausible that imprecision may affect these correlations to a degree. Other things the same, less precise estimates typically give lower correlation values, and we have seen some specifications have systematically larger standard errors than others. We thus check how shrinkage affects the observed correlations.

[13] We use Spearman rather than Pearson correlations because in practice the main policy use of the school coefficients is to rank schools. When using Pearson correlations instead (results available upon request), cross-control correlations are slightly higher than those displayed, while cross-model class correlations are slightly lower, on average.

[14] This calculation reports a simple average of the non-trivial correlations in Table 3, averaged across both subjects (Reading and Math) and all three panels. Subsequent calculations are performed similarly.

Tables A2 and A3 of the appendix show cross-control and cross-model correlations using the shrunk coefficients instead. By and large, these correlations are similar to their values in the unshrunk case. Rank correlations across control specifications remain right around 0.78, on average, using the shrunk coefficients. Across-model class correlations increase slightly to about 0.514 on average, though we see both increases and decreases in the degree to which the Arellano-Bond model (which generates the least precise estimates) correlates with the others. So although we have seen that shrinkage (sometimes) has a pronounced effect on the distribution of coefficient estimates, rank correlations do not change much as a result.

Figures 3 and 4 next show scatterplots for Math and Reading respectively. Each scatterplot shows school coefficient estimates from one class of model on the X-axis, plotted against school coefficients from another class of model on the Y-axis, for the same set of covariates. Plots are arranged so that each row compares the same two classes of model throughout, while the columns show results under control specifications A, B, and C respectively.

These scatterplots make it easy to see any nonlinearities or outliers, which may affect how to view the earlier correlations. For our data, we do not observe any major nonlinearities worth noting. However, do we observe a few outliers, especially with the Arellano-Bond model. These outliers are somewhat less pronounced after shrinkage (see Figures A1 and A2 in the appendix), but as noted, rank correlations do not change drastically.

Another pattern that stands out in these scatterplots is that the results of the gain-score model and level-on-lag model become increasingly similar as the number of controls increases. In the top row of either Figure 3 or 4, the scatterplots become much closer to a straight line as one proceeds rightward across columns. This, however, does not appear to be a general pattern for the other model classes.

Our final way of presenting these results is to show how often a school's relative ranking changes depending on the specification used. This is more than just a statistical exercise, as school districts sometimes target additional resources to bottom-performing schools based on their percentile ranking. (For example, San Diego Unified School District at one point targeted additional resources to elementary schools whose average test scores ranked within the bottom two deciles of the state as a whole.)

To do this, we rank schools in quintiles based on their value-added coefficients, under various specifications, and generate "transition matrices" for how these quintiles compare. Tables 5–6 display results in this format, comparing each of the three model classes that we use, for control specification B. In both tables, the top panel compares rankings under the gain-score model against those of the level-on-lag model; the middle panel compares rankings under the gain-score model against those of the Arellano-Bond model; and the bottom panel compares rankings under the

level-on-lag model against those of the Arellano-Bond model. In all cases, the diagonal elements of each matrix show how often schools remain in the same quintile, across the two models, while off-diagonal elements show how often schools change quintiles.

Our interest in these tables is seeing what percentage of the time schools take (at least roughly) the same relative ranking across models. To facilitate this task, all panels report the sum of diagonal entries in the upper left hand corner (just outside the transition matrix itself). Looking at these values, it can be seen that these rates are far from 100 percent, as would be the case if rankings remained unchanged between models. In even the "closest" pairing, which for both subjects is the gain-score model and the level-on-lag model, schools change quintiles slightly over half the time. Meanwhile, schools change quintiles as much as 75–80 percent of the time when comparing the least similar models (the gain-score model and the Arellano-Bond model, for both subjects).

We next consider how shrinkage affects these results. Tables A4 and A5 show analogous transition matrices using the shrunk value-added estimates. Overall, we find a similar picture as occurs in the unshrunk case. Because the gain-score and level-on-lag school coefficients are estimated fairly precisely, the school quintile assignments are virtually the same in these two models. Where shrinkage matters, for control specification B, is the Arellano-Bond model. But even after shrinkage, schools' quintile rankings continue to differ between this model and the others at generally similar rates as before.[15] Altogether, shrinkage appears to do little to resolve the variability of schools' rankings between models.

5.3 Correlations between Value-Added and Level of Test Score

Much of the motivation behind VAM stems from the idea that levels of test score alone do not necessarily convey school effectiveness, due to differences in students' socioeconomic status (e.g.) or other factors that have little to do with the schools themselves. Yet raw test scores are very readily available to the public, and moreover, many aspects of school accountability continue to rely on levels of test score rather than value-added. For example, the federal No Child Left Behind (NCLB) law, which was in place from 2002 until its replacement in 2015, required states that received federal aid for K-12 education to test students in specified grades and subject areas. Schools that had lower than required percentages of students proficient for more than two years in a row (overall or for student subgroups) were subject to an escalating set of accountability sanctions. This approach did not take into account students' past test scores, and thus had the potential to produce misleading

[15] Consider the bottom panels of Tables A4–A5, which compare the level-on-lag and Arellano-Bond results. The sum of schools on the diagonals, representing schools that stay in the same quintiles, is 36.8 percent in the unshrunk estimates, and 37.8 percent in the shrunk models. For reading, the percentagse of schools on the diagonals is 35.1 and 38.6 in the unshrunk and shrunk models respectively.

pictures of school effectiveness.

With this being the case, the empirical degree to which value-added aligns with test score levels is a crucial policy question. To explore this question, we estimate an additional model of the form

$$S_{igst} = \alpha_s^L + \gamma_g^L + \gamma_t^L + \epsilon_{igst}^L. \tag{6}$$

This equation is not a value-added model, since it does not control for past achievement. Instead, it is a deliberately "naive" levels model, such that the $\hat{\alpha}_s^L$ coefficients essentially just rate schools by their average level of test scores.

Table 7 reports correlations between the results of this naive levels model and value-added coefficients from each of the earlier specifications. One interesting pattern is that, prior to the use of school-level controls, in covariate specifications A and B the the gain-score model shows very little correlation with school coefficients from the naive levels model, for either Math or Reading. In most other specifications, value-added shows a moderate correlation with the naive levels model: between 0.20–0.76 for Math, and 0.45–0.86 for Reading.

Earlier, we described that ρ, the value of the autoregressive parameter, was able to explain many of the correlation patterns between model classes. It is worth revisiting this issue in the context of the naive levels model.

Notice that the levels model is equivalent to the level-on-lag model when $\rho = 0$. In contrast, the gain-score model assumes $\rho = 1$, the empirical (uninstrumented) level-on-lag model estimates ρ to be about 0.7, and the Arellano-Bond model estimates ρ to be about 0.25, for both subjects. In this light, there is a natural ordering between the naive model and each of the value-added models. If ρ is the only key difference between these models, then we should generally expect the naive levels model to show the highest correlation with the Arellano-Bond model, followed by the level-on-lag model, and finally the lowest correlation with the gain model.

The pattern does not hold in Table 7, but comes fairly close. The naive model actually shows the highest correlation with the level-on-lag model, rather than the Arellano-Bond model (although the Arellano-Bond model still shows a higher correlation than does the gain-score model). This discrepancy likely reflects the fact the Arellano-Bond estimator differs in two other important ways from the naive model: it instruments the changes in lagged test score and the school dummies, and it also identifies school coefficients from the subsample of students who switch elementary schools.

6 Discussion

This paper uses data from San Diego Unified School District to examine the sensitivity of school value-added estimates to both the control specification and overall class of value-added model.

We consider three classes of value-added models — a gain-score model, a level-on-lag model, and an Arellano-Bond model that allows for a student fixed or random effect — and three control specifications each, for a total of nine specifications.

Overall, we find a relatively mild sensitivity of results to the control specification, and a somewhat greater sensitivity to the class of value-added model. Precision of estimates appears sensitive to the inclusion of school-level controls, and whether researchers use panel instrument techniques to address dynamic panel bias. Value-added estimates also typically show a moderate positive correlation with schools' average level of test scores. We recap these results in greater detail below, while offering additional commentary. While we note that a study based on any one set of schools and testing system cannot yield definitive best practices, some tentative conclusions about preferred approaches do emerge.

An important subsidiary goal of the paper has been to compare how school rankings based on value-added compare with rankings based on average levels of test scores.[16] In most cases, we find a moderate positive correlation between the two. There are both optimistic and pessimistic ways to view this result. On one hand, it may be a reassuring sign that, at least on average, schools with higher average achievement also appear to have higher value-added. On the other hand, our analysis using transition matrices suggests that even models that show a correlation of about 0.7 with one another can rank schools in different quintiles over half the time. We thus take a somewhat more pessimistic view on this issue, that although the common practice among policymakers of ranking schools by test score levels is expedient, it is unlikely to be accurate of what these rankings are often intended to capture. Thus, the naive method of ranking schools by test scores, as the U.S. federal government required states to do for a dozen years, is unlikely to produce unbiased estimates of school effectiveness, and should probably be avoided.

This still leaves us with nine value-added models. Our results indicate a modest sensitivity of value-added estimates to the control specification. With an exception that the gain-score model appears somewhat sensitive to the use of school-level controls, we generally find correlations of 0.7 or higher across covariate specifications. Sensitivity to the overall class of value-added model appears somewhat larger. For the models that we consider, the average correlation across model classes (while holding covariates fixed) is about 0.5 — albeit with substantial heterogeneity. Results of the gain-score model and Arellano-Bond model tend to show low correlation with one another, perhaps due to differences in the supposed degree of persistence (ρ), while the level-on-lag model shows moderate correlation with either of the others.

[16]The latter case is a restricted version of value-added models in which $\rho = 0$. Such a restriction, however, is strongly rejected by the data. In all of our regressions that use a level-on-lag format, whether instrumented or not, the estimated value of ρ can be statistically distinguished from both 0 and 1 at all conventional significance levels. P-values below 0.001 in all cases.

With the less-than-perfect agreement across the nine specifications, analysis of the results leads to tentative suggestions about how to narrow down the sorts of models one should estimate. The gain-score model assumes that ρ is exactly one. This assumption is strongly rejected by the data, as has been the case for some other researchers as well. Thus, the level-on-lag model (estimated by OLS) or the Arellano-Bond model (which allows for student fixed effects) are likely the most promising approaches for estimating school value-added. Due to concerns about lowered precision when one controls for school demographics, however, researchers attempting to estimate school effectiveness should exercise caution in using school-level controls. Thus, a tentative suggestion is to narrow our nine candidate specifications to four: the level-on-lag and Arellano-Bond estimators, with control specification A, which controls for school, grade and year fixed effects, or control specification B, which additionally controls for student's own demographics.

The Arellano-Bond model has the attractive property of potentially delivering more consistent estimates, due to its ability to address both individual heterogeneity and the related dynamic panel bias, but at the cost of dramatically lower precision. Thus the choice between the two models is not entirely clear, and provides yet another illustration of the trade-off between estimators on bias versus precision. Either of these methods, though, is likely to yield a more accurate picture of underlying rates of learning than naive levels models frequently used by governments or the gain-score approach that has been widely used by researchers in the recent past.

But even after having whittled nine candidate value-added models down to four, we note that school rankings are highly sensitive to the method. For instance, using control specification B, we showed that if we rank schools into quintiles based on their estimated effectiveness using the Arellano-Bond approach, of the 20 percent of schools in the lowest quintile (for Math), only 7.9 percent — well under half — are also in the bottom quintile of schools according to the level-on-lag model. Perhaps the main lesson from the analysis is that the choice among even the most reasonable models can have dramatic effects on school rankings. Identifying bottom-performing schools for interventions remains a very difficult task, subject to considerable uncertainty.

7 References

References

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja Tristan Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics*, 3(3): 29–54.

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specification for Panel Data:

Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies*, 58(2): 277–297.

Ballou, Dale, William Sanders, and Paul Wright. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics*, 29(1): 37–65.

Buddin, Richard. 2010. "How effective are Los Angeles elementary teachers and schools?", unpublished manuscript.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urqiola. 2005. "The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools." *American Economic Review*, 95(4): 1237–1258.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." 104(9): 2593–2632, American Economic Review.

Guarino, Cassandra M., Mark D. Reckase, and Jeffrey M. Wooldridge. 2015. "Can Value-Added Measures of Teacher Performance Be Trusted?" *Education Finance and Policy*, 10(1): 117–156.

Imberman, Scott A., and Michael F. Lovenheim. 2016. "Does the market value value-added? Evidence from housing prices after a public release of school and teacher value-added." *Journal of Urban Economics*, 91 104–121.

Kane, Thomas J., and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives*, 16(4): 91–114.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. "Value-added modeling: A review." *Economics of Education Review*, 47 180–195.

Lefgren, Lars, and David Sims. 2012. "Using Subject Test Scores Efficiently to Predict Teacher Value-Added." *Educational Evaluation and Policy Analysis*, 34 109–121.

Lockwood, J.R., Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le, and Jose-Felipe Martinez. 2007. "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures." *Journal of Education Measurement*, 44(1): 47–67.

McCaffrey, Daniel F., J.R. Lockwood, Daniel M. Koretz, Thomas A. Louis, and Laura Hamilton. 2004. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics*, 29(1): 67–101.

Morris, Carl N. 1983. "Parametric empirical Bayes inference: theory and applications." *Journal of American Statistical Association*, 78(381): 47–55.

Newton, Xiaoxia A., Linda Darling-Hammond, Edward Haertel, and Ewart Thomas. 2010. "Value-Added Modeling of Teacher Effectiveness: An Exploration of Stability across Models and Contexts." *Education Policy Analysis Archives*, 18(23): .

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica*, 49(6): 1417–1426.

Roodman, David. 2009. "How to do xtabond2: An introduction to difference and system GMM in Stata." *Stata Journal*, 9(1): 86–136.

Sanders, William L., and Sandra P. Horn. 1994. "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Attainment." *Journal of Personnel Evaluation in Education*, 8 299–311.

Sanders, William L., and Sandra P. Horn. 1998. "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research." *Journal of Personnel Evaluation in Education*, 12(3): 247–256.

Sass, Tim R., Anastasia Semykina, and Douglas N. Harris. 2014. "Value-added models and the measurement of teacher productivity." *Economics of Education Review*, 38 9–23.

Tekwe, Carmen D., Randy L. Carter, Chang-Xing Ma, James Algina, Maurice E. Luas, Jeffrey Roth, Mario Ariet, Thomas Fisher, and Michael B. Resnick. 2004. "An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance." *Journal of Educational and Behavioral Statistics*, 29(1): 11–36.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal*, 113(485): F3–F33.

8 Tables and Figures

Table 1: Descriptive Statistics

Variable	**N**	**Mean**	**S.D.**	**Min**	**Max**
Outcomes:					
Math CST Score (Z)	162,569	0.03	1.01	-4.79	3.72
English Language Arts CST Score (Z)	162,754	0.06	1.05	-6.40	4.85
Student Characteristics:					
Female	162,921	0.49	0.50	0	1
African-American	162,921	0.12	0.33	0	1
Hispanic	162,921	0.45	0.50	0	1
Asian/Pacific Islander	162,921	0.17	0.37	0	1
Other/Unknown Race	162,921	0.01	0.10	0	1
Parent Ed. $<$ High School Diploma	162,921	0.12	0.32	0	1
Parent Ed. = High School Diploma	162,921	0.19	0.39	0	1
Parent Ed. = Some College	162,921	0.19	0.39	0	1
Parent Ed. = College Degree	162,921	0.16	0.36	0	1
Parent Ed. = Post-Graduate	162,921	0.10	0.30	0	1
Parent Ed. = Missing/Unknown	162,921	0.24	0.43	0	1
English Learner†	162,921	0.42	0.49	0	1
Fluent English Proficient†	162,921	0.00	0.04	0	1
Special Education†	162,921	0.10	0.30	0	1
School Demographics:					
% Students on Free or Reduced Priced Lunch	162,921	63.66	32.00	0	100
% African-American	162,921	12.52	11.26	0	90.21
% Hispanic	162,921	44.54	26.28	0	100
% Asian/Pacific Islander	162,921	16.38	13.50	0	54.86
% Other/Unknown Race	162,921	1.17	1.42	0	33.33
% English Learner†	162,921	41.41	24.72	0	100
% Fluent English Proficient†	162,921	0.15	0.31	0	2.97

Notes: Each observation corresponds to a student-year. Statistics are calculated for elementary school students with either Math or English Language Arts (ELA) score observed on the California Standards Test (CST) in a given year. Math and ELA scores are in standard deviation units relative to the state mean in the given year, for the student's respective grade level. Variables marked with dagger (†) are based on students' status in their first appearance in the district. We use first appearance status since subsequent status may be endogenous to school effectiveness. The dummy for white student ($\approx 25\%$), dummy for native English speaker ($\approx 57\%$), school percentage white, and school percentage of native English speakers are omitted to avoid collinearity.

Table 2: Summary Information about Coefficients and Standard Errors, before Shrinkage

	Coefficients		Standard Errors	
	Math	Reading	Math	Reading
Gain, Control A	-0.001	0.030	0.020	0.014
	(0.118)	(0.076)	(0.0017)	(0.0013)
Gain, Control B	-0.048	0.017	0.021	0.016
	(0.117)	(0.076)	(0.0017)	(0.0012)
Gain, Control C	-0.268	-0.245	0.367	0.305
	(0.172)	(0.131)	(0.1240)	(0.0956)
Level-on-Lag, Control A	-0.000	0.042	0.017	0.013
	(0.147)	(0.131)	(0.0015)	(0.0011)
Level-on-Lag, Control B	0.113	0.164	0.019	0.015
	(0.127)	(0.095)	(0.0021)	(0.0015)
Level-on-Lag, Control C	-0.044	0.015	0.322	0.299
	(0.193)	(0.138)	(0.1079)	(0.0935)
Arellano-Bond, Control A	-0.027	0.035	0.257	0.246
	(0.435)	(0.418)	(0.1714)	(0.1492)
Arellano-Bond, Control B	-0.067	0.017	0.250	0.231
	(0.615)	(0.543)	(0.1747)	(0.1278)
Arellano-Bond, Control C	-0.050	0.030	0.472	0.418
	(0.563)	(0.478)	(0.1448)	(0.1115)

Notes: Table reports the mean and standard deviation of school value-added coefficients and their standard errors, prior to shrinkage. Standard deviations are in parentheses.

Table 3: Spearman Rank Correlations of Value-Added Coefficients across Control Specifications

Gain Model

	Math				Reading		
	Control A	Control B	Control C		Control A	Control B	Control C
Control A	1			Control A	1		
Control B	0.993	1		Control B	0.994	1	
Control C	0.523	0.569	1	Control C	0.613	0.635	1

Level-on-Lag Model

	Math				Reading		
	Control A	Control B	Control C		Control A	Control B	Control C
Control A	1			Control A	1		
Control B	0.938	1		Control B	0.946	1	
Control C	0.881	0.765	1	Control C	0.913	0.856	1

Arellano-Bond Model

	Math				Reading		
	Control A	Control B	Control C		Control A	Control B	Control C
Control A	1			Control A	1		
Control B	0.815	1		Control B	0.829	1	
Control C	0.514	0.786	1	Control C	0.704	0.786	1

Table 4: Spearman Rank Correlations of Value-Added Coefficients across Model Classes

Covariate Specification A

	Math				Reading		
	Gain	LevLag	ABond		Gain	LevLag	ABond
Gain	1			Gain	1		
LevLag	0.599	1		LevLag	0.529	1	
ABond	0.112	0.570	1	ABond	0.075	0.613	1

Covariate Specification B

	Math				Reading		
	Gain	LevLag	ABond		Gain	LevLag	ABond
Gain	1			Gain	1		
LevLag	0.761	1		LevLag	0.684	1	
ABond	0.302	0.571	1	ABond	0.169	0.546	1

Covariate Specification C

	Math				Reading		
	Gain	LevLag	ABond		Gain	LevLag	ABond
Gain	1			Gain	1		
LevLag	0.932	1		LevLag	0.940	1	
ABond	0.231	0.351	1	ABond	0.441	0.586	1

Table 5: Relative Ranking of Schools, Compared across Models, for Math

Gain vs. Level-on-Lag

Sum Diagonal ≈ 46.5%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	**13.2**	5.3	0.9	0.9	0
Level-on-	2nd Quintile	4.4	**7.9**	5.3	1.8	0.9
Lag	3rd Quintile	1.8	4.4	**5.3**	7.9	0.9
Model	4th Quintile	0	1.8	**8.8**	6.1	3.5
	5th Quintile	0.9	0.9	0	3.5	**14.0**

Gain vs. Arellano-Bond

Sum Diagonal ≈ 23.6%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	4.4	5.3	**3.5**	4.4	2.6
Arellano-	2nd Quintile	6.1	**6.1**	4.4	2.6	0.9
Bond	3rd Quintile	6.1	2.6	**3.5**	4.4	3.5
Model	4th Quintile	1.8	4.4	**6.1**	2.6	5.3
	5th Quintile	1.8	1.8	2.6	**6.1**	7.0

Level-on-Lag vs. Arellano-Bond

Sum Diagonal ≈ 36.8%		Level-on-Lag Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	7.9	**3.5**	5.3	1.8	1.8
Arellano-	2nd Quintile	6.1	**7.0**	5.3	1.8	0
Bond	3rd Quintile	5.3	3.5	**5.3**	4.4	1.8
Model	4th Quintile	0.9	4.4	3.5	**6.1**	5.3
	5th Quintile	0	1.8	0.9	6.1	**10.5**

Notes: Percent frequencies are displayed. Within each row, the column entry in which the median observation occurs is in bold. The 1st quintile refers to the bottom 20 percent, while the 5th quintile refers to the top 20 percent. Results in this table use control specification B for both models.

Table 6: Relative Ranking of Schools, Compared across Models, for Reading

Gain vs. Level-on-Lag

Sum Diagonal ≈ 41.2%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	**12.3**	4.4	2.6	0.9	0
Level-on-	2nd Quintile	3.5	6.1	**7.0**	3.5	0
Lag	3rd Quintile	2.6	4.4	**5.3**	4.4	3.5
Model	4th Quintile	1.8	3.5	2.6	**7.0**	5.3
	5th Quintile	0	1.8	2.6	4.4	**10.5**

Gain vs. Arellano-Bond

Sum Diagonal ≈ 22.8%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	3.5	2.6	**6.1**	3.5	4.4
Arellano-	2nd Quintile	7.0	**5.3**	2.6	2.6	2.6
Bond	3rd Quintile	4.4	4.4	**3.5**	6.1	1.8
Model	4th Quintile	2.6	4.4	**4.4**	4.4	4.4
	5th Quintile	2.6	3.5	**3.5**	3.5	6.1

Level-on-Lag vs. Arellano-Bond

Sum Diagonal ≈ 35.1%		Level-on-Lag Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	7.0	**6.1**	2.6	2.6	1.8
Arellano-	2nd Quintile	7.0	**6.1**	4.4	2.6	0
Bond	3rd Quintile	1.8	5.3	**5.3**	7	0.9
Model	4th Quintile	3.5	1.8	4.4	**5.3**	5.3
	5th Quintile	0.9	0.9	3.5	2.6	**11.4**

Notes: Percent frequencies are displayed. Within each row, the column entry in which the median observation occurs is in bold. The 1st quintile refers to the bottom 20 percent, while the 5th quintile refers to the top 20 percent. Results in this table use control specification B for all models.

Table 7: Spearman Rank Correlations between Value-Added and Level of Test Score

	Math				Reading		
	Gain	LevLag	ABond		Gain	LevLag	ABond
Control A	-0.043	0.732	0.628	Control A	0.075	0.856	0.680
Control B	-0.013	0.530	0.427	Control B	0.107	0.706	0.454
Control C	0.632	0.764	0.203	Control C	0.645	0.785	0.605

Notes: Covariates change for the value-added models, but not for the levels model. The levels model includes only school, year, and grade fixed effects, with no additional covariates throughout.

Figure 1: Histograms of Value-Added Coefficients, Math

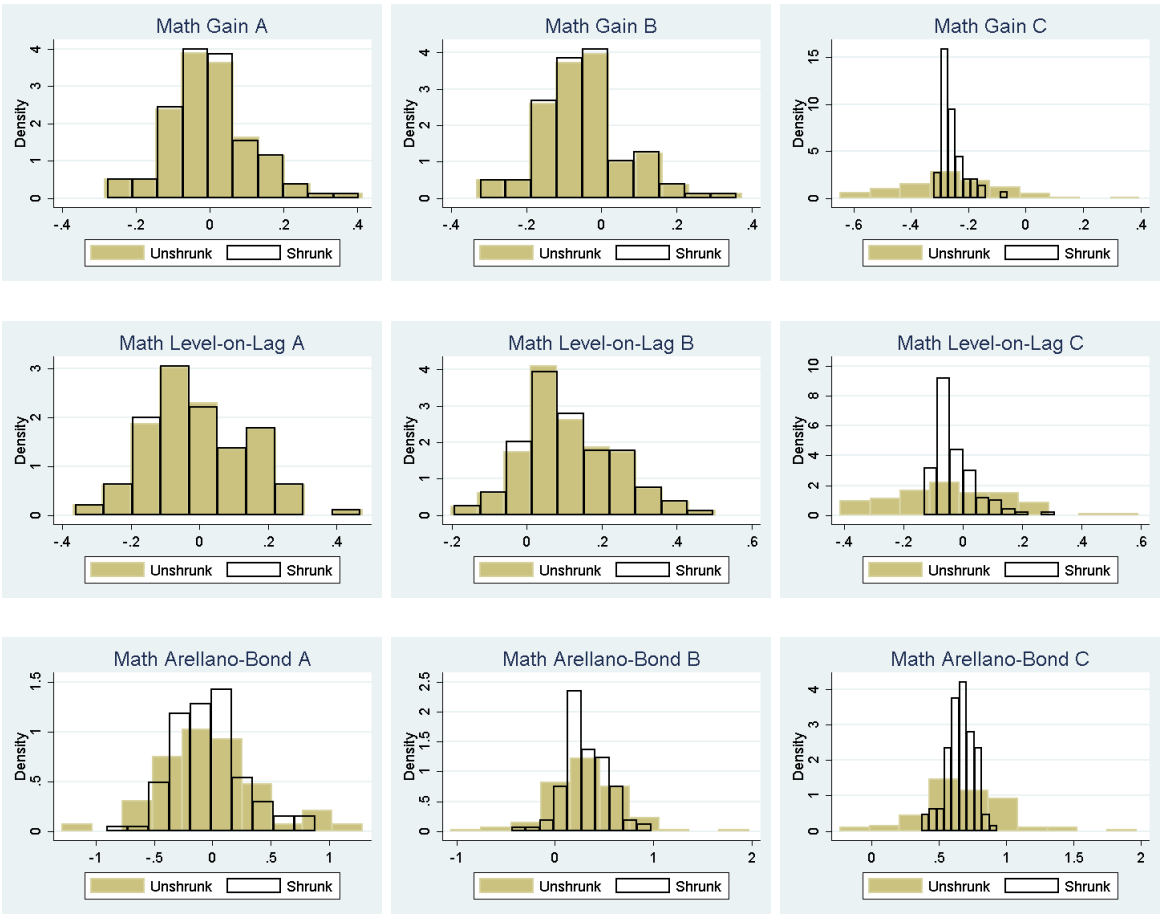


Figure 2: Histograms of Value-Added Coefficients, Reading

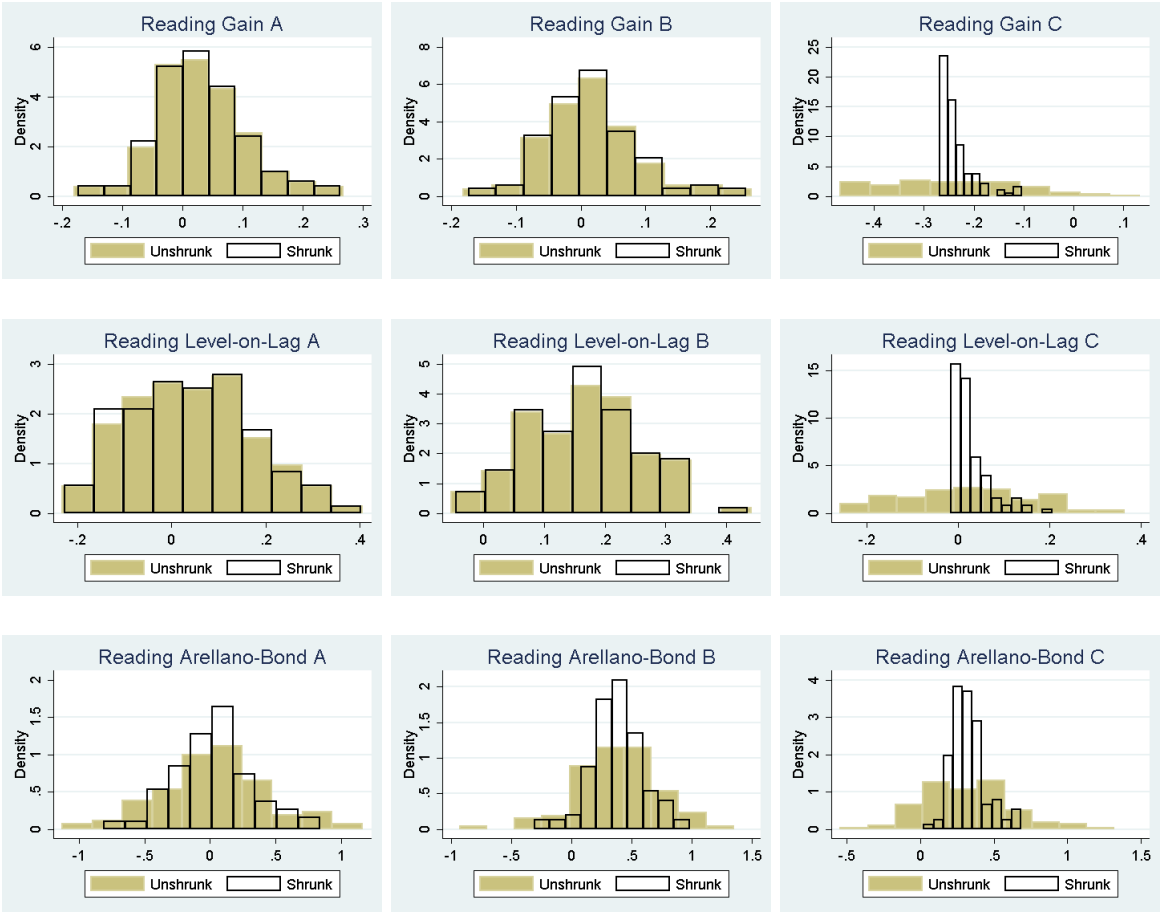


Figure 3: Scatterplots of Value-Added Coefficients, Math

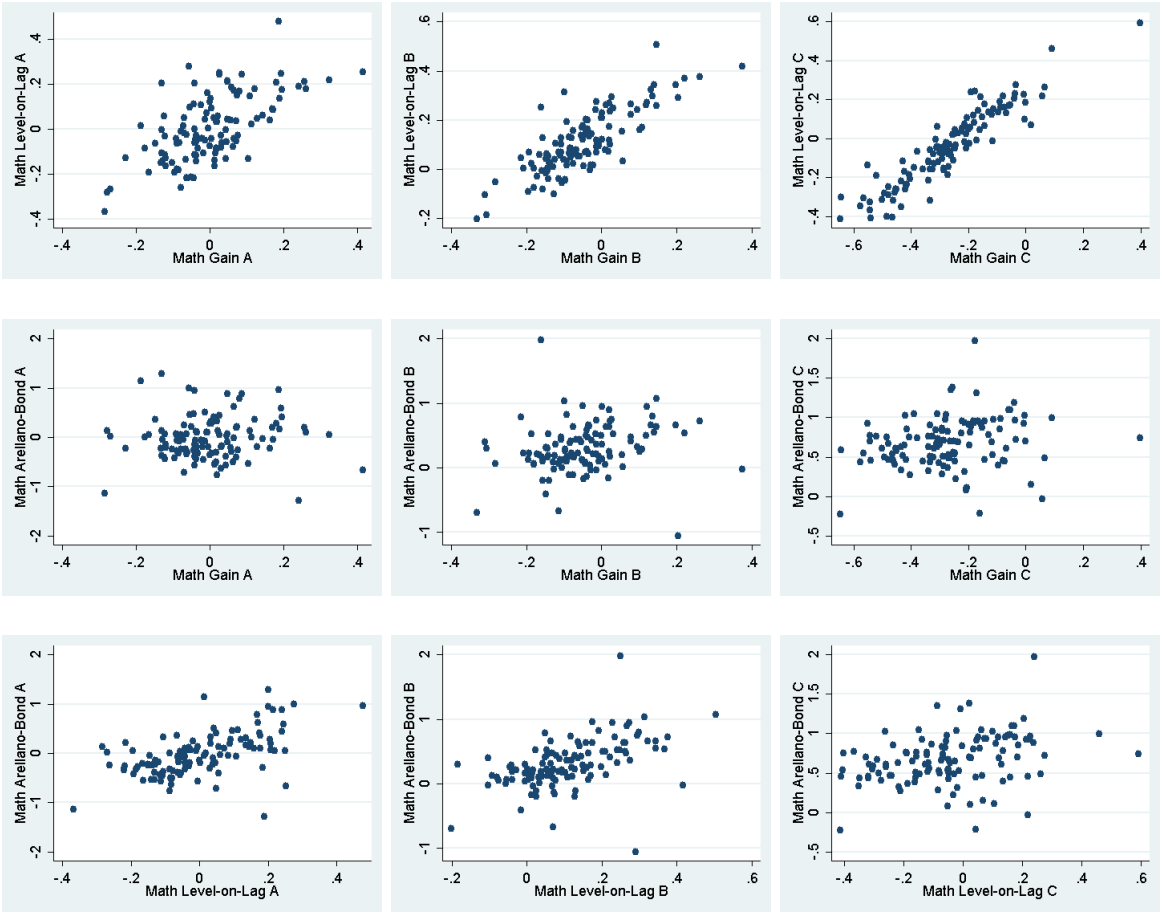
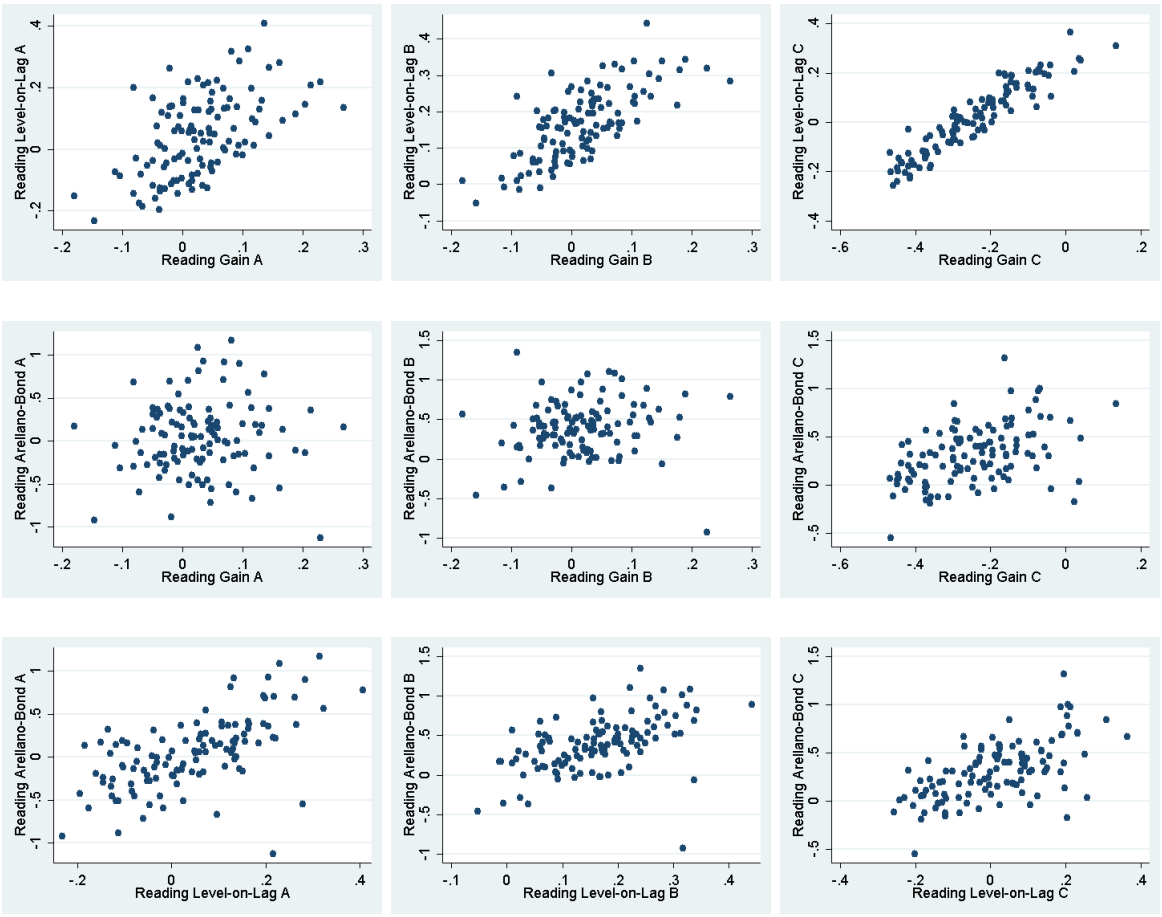


Figure 4: Scatterplots of Value-Added Coefficients, Reading



A Appendix

A.1 Arellano-Bond Estimation

Roodman (2009) provides a good overview of dynamic panel bias and techniques that have developed to address this issue. We explain the basics below, and then provide additional details about how we estimate regressions using an Arellano and Bond (1991) estimator.

Dynamic panel bias arises in equations that feature both fixed effects and lagged dependent variables as regressors (Nickell, 1981). A simple such equation may take the form

$$y_{it} = \rho^* y_{i,t-1} + x'_{it}\beta^* + \mu_i + u_{it}, \tag{A1}$$

where μ_i is an unobserved heterogeneity factor for person i, and $\epsilon^*_{it} = \mu_i + u_{it}$ is a composite error term.

The key issue is that the lagged dependent variable, $y_{i,t-1}$ is mechanically an endogenous variable. If Equation A1 is estimated using a conventional fixed effects estimator, any unmodeled shocks affect the *estimated* value of the fixed effect μ_i, creating a negative correlation between $y_{i,t-1}$ and ϵ^*_{it}, and returning downward-biased estimates of ρ^*. This problem is worse the smaller is T: the number of periods in the panel.

(It is worth remembering that in our context the key interest is not on the autoregressive parameter, ρ, but on the school value-added coefficients. Nevertheless, dynamic panel bias remains a possible concern, since biased estimates of ρ also affect the corresponding school value-added coefficients.)

Arellano and Bond (1991) propose a popular method to address dynamic panel bias by using using a first difference instrumental variables (FDIV) procedure. Equation A1 is transformed into a first-differenced version:

$$\Delta y_{it} = \rho^* \Delta y_{i,t-1} + \Delta x'_{it}\beta^* + \Delta u_{it} \tag{A2}$$

which no longer contains the person fixed effect, μ_i. Under appropriate conditions — most notably an assumption about no serial correlation in u_{it} — this equation can be consistently estimated using instrumental variables, with lagged values of regressors serving as the instruments.

Which lags are valid to use depend on what variables are exogenous, endogenous, or predetermined. Exogenous variables are assumed to be uncorrelated with past, present, and future disturbances. First differences of these variables are used as instruments in Equation A2. Predetermined variables are assumed to be uncorrelated with present and future disturbances, but may be correlated with past disturbances. Lagged values of one period or more of these variables are used as instruments. Endogenous variables are possibly correlated with both present and past disturbances. Lagged values of two periods or more of these variables are used as instruments.

Table A1 shows how we classify variable in the full specification of the Arellano-Bond model in our paper.

Table A1: Variable Classification in Arellano-Bond Regressions

Variable	Classification	Lags Used
Lagged Test Score $S_{igs,t-1}$	Predetermined	1–2
School Dummies	Endogenous	2–3
Grade Dummies	Predetermined	1–2
Year Dummies	Exogenous	N/A
Female Dummy	Exogenous	N/A
Race Dummies	Exogenous	N/A
Parent Ed Dummies	Predetermined	1–2
English Learner Status†	Predetermined	1–2
Fluent English Proficient Status†	Predetermined	1–2
Special Education Status†	Predetermined	1–2
% Free or Reduced Price Lunch	Predetermined	1–2
% African-American	Predetermined	1–2
% Hispanic	Predetermined	1–2
% Asian/Pacific Islander	Predetermined	1–2
% Native American	Predetermined	1–2
% English Learner†	Predetermined	1–2
% Fluent English Proficient†	Predetermined	1–2

Notes: Table reports the classification of variables in the full specification of the Arellano-Bond model. Exogenous variables are assumed to be uncorrelated with past, present, and future errors. Predetermined variables are assumed to be uncorrelated with present and future errors, but possibly correlated with past errors. Endogenous variables may be correlated with either present or past errors. Variables marked with dagger (†) are based on students' status in their first appearance in the district.

Table A2: Spearman Rank Correlations of Shrunk Value-Added Coefficients across Control Specifications

Gain Model

	Math				Reading		
	Control A	Control B	Control C		Control A	Control B	Control C
Control A	1			Control A	1		
Control B	0.993	1		Control B	0.995	1	
Control C	0.490	0.540	1	Control C	0.568	0.589	1

Level-on-Lag Model

	Math				Reading		
	Control A	Control B	Control C		Control A	Control B	Control C
Control A	1			Control A	1		
Control B	0.937	1		Control B	0.947	1	
Control C	0.889	0.770	1	Control C	0.916	0.857	1

Arellano-Bond Model

	Math				Reading		
	Control A	Control B	Control C		Control A	Control B	Control C
Control A	1			Control A	1		
Control B	0.811	1		Control B	0.832	1	
Control C	0.524	0.790	1	Control C	0.748	0.825	1

Table A3: Spearmank Rank Correlations of Shrunk Value-Added Coefficients across Model Classes

Control Specification A

Math				Reading			
	Gain	LevLag	ABond		Gain	LevLag	ABond
Gain	1			Gain	1		
LevLag	0.600	1		LevLag	0.530	1	
ABond	0.137	0.634	1	ABond	0.099	0.654	1

Control Specification B

Math				Reading			
	Gain	LevLag	ABond		Gain	LevLag	ABond
Gain	1			Gain	1		
LevLag	0.761	1		LevLag	0.683	1	
ABond	0.348	0.620	1	ABond	0.194	0.594	1

Control Specification C

Math				Reading			
	Gain	LevLag	ABond		Gain	LevLag	ABond
Gain	1			Gain	1		
LevLag	0.922	1		LevLag	0.929	1	
ABond	0.202	0.326	1	ABond	0.420	0.598	1

Table A4: Relative Ranking of Schools Using Shrunk Estimates, Compared across Models, for Math

Gain vs. Level-on-Lag

Sum Diagonal ≈ 46.5%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	**13.2**	5.3	0.9	0.9	0
Level-on-	2nd Quintile	4.4	**7.9**	5.3	1.8	0.9
Lag	3rd Quintile	1.8	4.4	**5.3**	7.9	0.9
Model	4th Quintile	0	1.8	8.8	**6.1**	3.5
	5th Quintile	0.9	0.9	0	3.5	**14.0**

Gain vs. Arellano-Bond

Sum Diagonal ≈ 25.4%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	4.4	5.3	**4.4**	4.4	1.8
Arellano-	2nd Quintile	7.0	**6.1**	3.5	3.5	0
Bond	3rd Quintile	5.3	2.6	**3.5**	3.5	5.3
Model	4th Quintile	1.8	4.4	**6.1**	3.5	4.4
	5th Quintile	1.8	1.8	2.6	**5.3**	7.9

Level-on-Lag vs. Arellano-Bond

Sum Diagonal ≈ 37.8%		Level-on-Lag Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	7.9	**5.3**	5.3	0.9	0.9
Arellano-	2nd Quintile	7.0	**5.3**	5.3	2.6	0
Bond	3rd Quintile	4.4	3.5	**5.3**	4.4	2.6
Model	4th Quintile	0.9	4.4	4.4	**7.0**	3.5
	5th Quintile	0	1.8	0	5.3	**12.3**

Notes: Percent frequencies are displayed. Within each row, the column entry in which the median observation occurs is in bold. The 1st quintile refers to the bottom 20 percent, while the 5th quintile refers to the top 20 percent. Results in this table use control specification B for all models.

Table A5: Relative Ranking of Schools Using Shrunk Estimates, Compared across Models, for Reading

Gain vs. Level-on-Lag

Sum Diagonal ≈ 41.2%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	**12.3**	4.4	2.6	0.9	0
Level-on-	2nd Quintile	3.5	6.1	**7.0**	3.5	0
Lag	3rd Quintile	2.6	4.4	**5.3**	4.4	3.5
Model	4th Quintile	1.8	3.5	2.6	**7.0**	5.3
	5th Quintile	0	1.8	2.6	4.4	**10.5**

Gain vs. Arellano-Bond

Sum Diagonal ≈ 21.1%		Gain Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	4.4	2.6	**6.1**	3.5	3.5
Arellano-	2nd Quintile	7.0	**5.3**	3.5	0.9	3.5
Bond	3rd Quintile	2.6	4.4	**3.5**	7.9	1.8
Model	4th Quintile	4.4	4.4	**3.5**	2.6	5.3
	5th Quintile	1.8	3.5	3.5	**5.3**	5.3

Level-on-Lag vs. Arellano-Bond

Sum Diagonal ≈ 38.6%		Level-on-Lag Model				
		1st Quintile	2nd Quintile	3rd Quintile	4th Quintile	5th Quintile
	1st Quintile	8.8	**6.1**	2.6	1.8	0.9
Arellano-	2nd Quintile	4.4	**7.0**	4.4	3.5	0.9
Bond	3rd Quintile	4.4	4.4	**4.4**	5.3	1.8
Model	4th Quintile	2.6	2.6	**5.3**	6.1	3.5
	5th Quintile	0	0	3.5	3.5	**12.3**

Notes: Percent frequencies are displayed. Within each row, the column entry in which the median observation occurs is in bold. The 1st quintile refers to the bottom 20 percent, while the 5th quintile refers to the top 20 percent. Results in this table use control specification B for all models.

Table A6: Spearman Rank Correlations between Shrunk Value-Added and Level of Test Score

	Math				Reading		
	Gain	LevLag	ABond		Gain	LevLag	ABond
Control A	-0.043	0.731	0.680	Control A	0.074	0.855	0.720
Control B	-0.012	0.530	0.464	Control B	0.106	0.706	0.505
Control C	0.655	0.793	0.224	Control C	0.657	0.789	0.628

Notes: Covariates change for the value-added models, but not for the levels model. The levels model includes only school, year, and grade fixed effects, with no additional covariates throughout. Shrinkage has been applied to the value-added coefficients, but not the coefficients of the levels model.

Figure A1: Scatterplots of Shrunk Value-Added Coefficients, Math

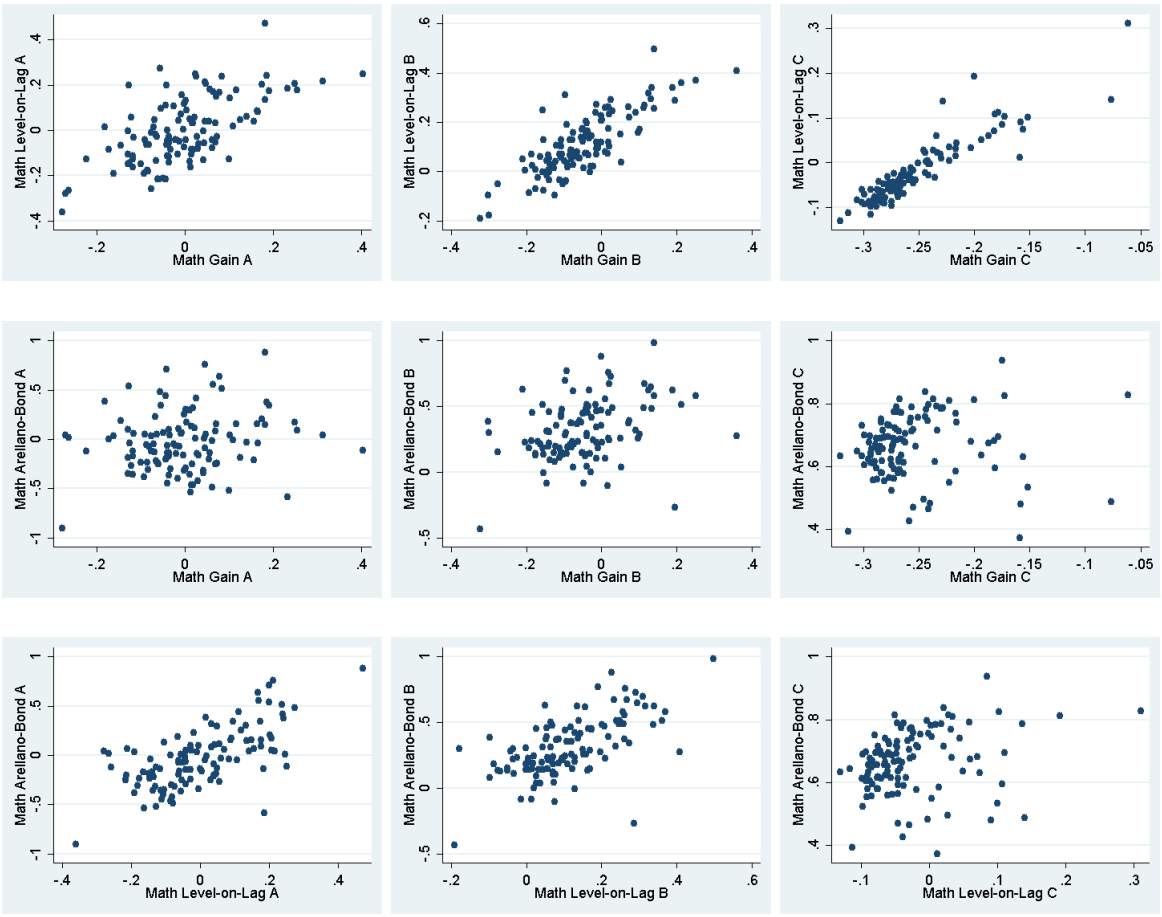


Figure A2: Scatterplots of Shrunk Value-Added Coefficients, Reading