Leveraging Lotteries for School Value-Added: Testing and Estimation

Joshua Angrist	Peter Hull	Parag Pathak	Christopher Walters*
MIT and NBER	MIT	MIT and NBER	UC Berkeley and NBER

July 18, 2015

Abstract

We show how admissions lotteries can be used to assess and improve estimates of school value added. Widely-used ordinary least squares (OLS) estimates of value-added models (VAMs) are biased if the available controls are insufficient to capture differences in student ability across schools. We introduce a new test for VAM bias and apply this test to data from Boston. The results show evidence of bias in conventional VAMs. We improve on conventional estimates by developing a hybrid estimation strategy that combines biased but precise OLS estimates with unbiased but less precise lottery estimates. The hybrid approach uses a random coefficients model to describe the joint distribution of value-added, bias, and lottery compliance across schools and generates empirical Bayes posterior value-added predictions for individual schools. This procedure generates marked reductions in the mean squared error of value-added estimates for Boston schools. Simulations calibrated to the Boston data suggest that our hybrid strategy is likely to reduce errors in school accountability policies and that school expansion and closure decisions based on this strategy may boost achievement substantially more than decisions based on conventional estimates.

*We gratefully acknowledge funding from the National Science Foundation and the Laura and John Arnold Foundation. As always, we're indebted to SEII program manager Annice Correia for invaluable help and support. Thanks also go to Pat Kline, Rick Mansfield, Chris Nielson, Jesse Rothstein, Doug Staiger, and seminar participants at the 2104 All California Labor Economics Conference, the APPAM Fall 2014 research conference, the 2014 AEFP meeting, the 2015 ASSA annual meeting, and the 2015 SOLE/EALE annual meeting for suggestions and comments.

1 Introduction

An increasing number of public school districts use value-added models (VAMs) to assess teacher and school effectiveness. Conventional VAM estimates compare standardized test scores across classrooms or schools after regression-adjusting for students' demographic characteristics and earlier scores. Achievement differences remaining after adjustment are attributed to differences in teacher or school quality. Some districts use estimates of teacher value-added to inform personnel decisions, while others use VAMs to generate "report cards" that allow parents to compare schools.[1] Value-added estimation is a high-stakes statistical exercise: low VAM estimates can lead to punitive actions such as school closure and teacher dismissal, while a growing body of evidence suggests the achievement gains produced by effective teachers and schools can persist into adulthood (see, eg., Chetty et al., 2011 and Chetty et al., 2014b for teachers and Angrist et al., forthcoming and Dobbie and Fryer, forthcoming for schools).

Their growing importance notwithstanding, the use of VAMs for teacher and school assessment remains controversial. Critics note that VAM estimates are misleading if the available control variables are inadequate to ensure *ceteris paribus* comparisons. VAM estimates are also likely to reflect considerable sampling error. The accuracy of teacher value-added models is the focus of a large and expanding body of research, but this work has yet to generate a consensus on the predictive value of VAM estimates or guidelines for "best practice" VAM estimation (Kane and Staiger, 2008; Rothstein, 2010; Koedel and Betts, 2011; Kinsler, 2012; Kane et al., 2013; Chetty et al., 2014a; Baicher-Hicks et al., 2014; Rothstein, 2014). The VAM research agenda has also been tilted towards teachers. Although the social significance of school-level VAM estimates is similar to that of teacher VAMs, research on school effectiveness looks mostly at entire sectors, such as charter schools and Catholic schools, with the quality of individual schools receiving less attention.

The proliferation of partially-randomized school assignment systems in large urban school districts provides a new tool for measuring school effectiveness. Assignment mechanisms based on the theory of market design, including those used in Boston, Chicago, Denver, New Orleans, and New York, use information on parents' preferences over schools and schools' priorities over students to match students and schools. These assignment algorithms often break ties at oversubscribed schools (those with more prospective students than seats) with randomly assigned sequence numbers, thereby creating stratified student assignment lotteries. Similarly, independently-run charter schools often use admission lotteries when they're oversubscribed. Scholars increasingly use these lotteries to identify the causal effects of attendance in various sectors, including charter schools, pilot schools, small high schools, and magnet schools (Cullen et al., 2006; Abdulkadiroğlu et al., 2011; Angrist et al., 2013; Dobbie and Fryer, 2013; Bloom and Unterman, 2014; Deming et al., 2014). Lottery-based estimation of individual school effectiveness is less common.[2]

This paper develops econometric methods that leverage admissions lotteries to assess and improve school

[1] Alabama, Arizona, California, Florida, Indiana, Louisiana, Maine, Mississippi, New Mexico, North Carolina, Texas, Utah, and Virginia issue letter-grade report cards with grades determined at least in part by adjusted standardized test metrics. See `http://www.ecs.org/html/educationissues/accountability/stacc_intro.asp` for more information.

[2] Exceptions include Deutsch (2012) and Deming (2014), discussed below.

value-added estimates, applying these procedures to evaluate public schools in Boston. Our first contribution is the development of a new lottery-based test for VAM bias. This test builds on recent experimental and quasi-experimental VAM validation strategies, including the work of Kane and Staiger (2008), Deutsch (2012), Kane et al. (2013), Chetty et al. (2014a) and Deming (2014). In contrast with these earlier studies, however, the test developed here exploits the complete set of overidentifying restriction implict in an empirical VAM framework augmented with lotteries. The resulting test statistic, though just as easily computed as those reported previously, exhausts the available information and is therefore likely to be more powerful than the tests employed in earlier work.

Our test reveals evidence of bias in conventional value-added models. This finding motivates the development of an improved measure of school effectiveness that synthesizes the information provided by lottery and conventional estimates. OLS estimates may be biased by systematic student sorting, but they're likely to be statistically precise. Lottery-based instrumental variables (IV) estimates are consistent for causal effects but less precise. Lotteries are also unavailable for some schools; as a result, a value-added model based only on lotteries is underidentified. In the spirit of earlier work on combination estimators (e.g., Judge and Mittlehammer, 2004, 2005, 2007; Hahn and Hausman, 2002), we therefore propose to minimize the mean squared error of VAM estimates by combining the two estimators.

Our hybrid estimation framework starts with an empirical Bayes/random-coefficients model for the joint distribution of value-added, bias, and lottery compliance across schools. Bayesian hyperparameters are estimated using a simulated minimum distance (SMD) procedure that matches moments of the joint distribution of OLS, lottery reduced form, and first stage estimates to those predicted by the model. Finally, the SMD estimates are used to compute empirical Bayes posterior predictions of individual school value-added conditional on all available information. Importantly, this approach makes effective and efficient use of as much lottery information as is available, without requiring a lottery for every school. Hybrid estimates for schools that are undersubscribed are improved by the information on the *distribution* of bias contributed by schools with active lotteries.

We quantify the payoff to hybrid estimation in a Monte Carlo experiment that compares the consequences of policy decisions based on conventional and hybrid VAM estimates. This comparison reveals substantial gains in the accuracy of school assessment based on the hybrid procedure, with a consequent improvement in student outcomes. For example, the probability of misclassifying failing (lowest-quintile) district schools as satisfactory falls from 47% using conventional VAM estimates to 34% with the hybrid estimates. Use of hybrid VAM for purposes of school closure and expansion generates a gain of about 0.1 standard deviations in student achievement beyond the gains seen when these decisions are based on conventional estimates.

The next section describes the Boston data used here as motivation and testbed for new methods of VAM assessment and estimation. Section 3 describes the value-added framework. Section 4 derives our VAM validity test and discusses the test implementation and results. Section 5 outlines the random coefficients model and empirical Bayes framework used for hybrid estimation. Section 6 reports estimates of the model

and the resulting posterior predictions of value-added. Section 7 describes the policy simulations. Finally, Section 8 concludes and discusses how the theoretical framework outlined here might be used in other work.

2 Setting and Data

2.1 Boston Public Schools

Boston public school (BPS) students can choose from a large and diverse set of options, including traditional BPS district schools, charter schools, and pilot schools. Like charter schools in most districts, Boston charters are publicly funded but free to operate wthin the confines of their charters. For the most part, charter staff are not covered by collective bargaining agreements and other Boston Public School (BPS) regulations. The charter sector includes both "Commonwealth" charters, which are authorized by the state and run as independent school districts, and "in-district" charters, which are authorized and overseen by the Boston School Committee.

Charter applicants apply to each school separately, typically by early spring of the preceding school year. By Massachusetts law, oversubscribed charter schools select students in public admissions lotteries, with the exception of applicants with a sibling already enrolled in the charter (who are guaranteed a seat). Studies based on these lotteries show that Boston charter schools boost test scores and other outcomes (Abdulkadiroğlu et al., 2011; Angrist et al., forthcoming).

Boston's pilot school sector arose as a union-supported alternative to charter schools, developed jointly by the BPS district and the Boston Teachers Union. As with in-district charters, pilot schools are part of the district but typically operate with more flexibility over school budgets, scheduling, and curriculum than do traditional public schools. On the other hand, pilot school teachers work under collective bargaining provisions similar to those in force at traditional public schools. In a lottery-based evaluation of the pilot sector, Abdulkadiroğlu et al. (2011) find mostly small and statistically insignificant effects of pilot school attendance.

In contrast with the decentralized charter school application process, applicants to traditional public and pilot schools rank 3 to 10 schools as the first step in a centralized match (age-appropriate students who are happy to stay in the school in which they're already enrolled need not participate in the match). These applicants are then assigned to schools via a student-proposing deferred acceptance mechanism (Abdulkadiroğlu et al., 2006). This mechanism combines student preference rankings with a strict priority ranking over students for each school. Priorities are determined by whether an applicant is already enrolled at the school and therefore guaranteed admission, has a sibling enrolled at the school, or lives in the school's walk-zone. Ties within these coarse priority groups are broken by random sequence numbers, which we refer to as lottery numbers.

2.2 Data and Descriptive Statistics

The data analyzed here consist of a sample of roughly 28,000 6th grade students attending 51 Boston public, pilot, and charter schools in the 2006-2007 through 2013-2014 school years. Sixth is the most common transition grade between elementary and middle school, so most Boston 6th graders participate in the match. Outcome variables are 6th grade test scores from the Massachusetts Comprehensive Assessment System (MCAS). Other variables used in the empirical analysis are school enrollment, race, sex, subsidized lunch eligibility, special education status, English-language learner status, and suspensions and absences. Test scores are standardized to have mean zero and unit variance in the population of Boston charter, pilot, and traditional public schools separately by subject, grade, and year. The data appendix describes the administrative files and data processing conventions used to construct the working extract.

Our analysis combines data from the Boston match with charter lottery data for charter applicants (participants in the BPS match are free to apply to as many charters as they like). The charter lottery offer instrument indicates any charter offer made on the night of the admissions lottery at each charter school. These offers are randomly assigned within school and application year, conditional on the set of charters applied to (we call this a charter lottery "risk set"). The BPS lottery instrument codes offers at an applicant's first choice (that is, highest ranked school) in the match. In particular, the BPS lottery offer indicates applicants whose lottery number is at least as high as the worst number offered a seat at their first-choice school among those in the same priority group for that school. Conditional on application year, first-choice school, and an applicant's priority at that school, offers are randomly assigned.

The schools and students analyzed here are described in Table 1. Schools serving fewer than 25 6th graders in each year are omitted, leaving a total of 17 charter, 9 pilot, and 25 traditional public schools. Of these, 28 schools (5 charter, 7 pilot, and 16 traditional public schools) had at least 50 students subject to random assignment. OLS VAM estimates were computed in a sample of 27,864 Boston 6th graders with complete baseline demographic information; 8,718 of these are also in the lottery sample.

Lotteried students are slightly more likely to be African American and to have qualified for a subsidized lunch than are students in the general 6th grade population. They are also somewhat less likely to be white or to have been suspended or recorded as absent in 5th grade. These and other descriptive statistics appear in Table 2. Overall, the lotteried sample looks broadly similar to the larger BPS population. Table 2 also documents the comparability of students who were and were not offered seats in a lottery. These results, reported in columns (3) through (6), compare the baseline characteristics of lottery winners and losers, controlling for assignment strata. The estimated differences by offer status are small and not significantly different from zero, both overall and within sector.

3 Value-added Framework

As in earlier investigations of value-added, the analysis here builds on a constant-effects causal model. This reflects a fundamental premise in the VAM framework: internally valid treatment effects from earlier years and cohorts are presumed to have predictive value for future cohorts. Student i's potential test score at school j, Y_{ij}, is therefore written as the sum of two non-interacting components, specifically:

$$Y_{ij} = \mu_j + a_i, \tag{1}$$

where μ_j is the mean potential outcome at school j and a_i is student i's ability. This additively-separable form implies that causal effects are the same for all students. The constant effects framework focuses attention on the possibility of selection bias in VAM estimates rather than treatment effect heterogeneity, though we briefly explore the latter as well.[3]

A dummy variable, D_{ij}, is used to indicate whether student i attended school j. The observed outcome for student i can then be written

$$\begin{aligned} Y_i &= Y_{i0} + \sum_{j=1}^{J} (Y_{ij} - Y_{i0}) D_{ij} \\ &= \mu_0 + \sum_{j=1}^{J} \beta_j D_{ij} + a_i. \end{aligned} \tag{2}$$

The parameter $\beta_j = \mu_j - \mu_0$ measures the causal effect of school j relative to an omitted school (school 0). In other words, β_j is school j's value added.

Conventional value-added models use regression methods in an attempt to eliminate selection bias. Write

$$a_i = X_i'\gamma + \epsilon_i, \tag{3}$$

for the regression of a_i on a vector of controls, X_i, which includes lagged test scores. Note that $E[X_i\epsilon_i] = 0$ by definition of γ. This decomposition implies that observed outcomes can be written

$$Y_i = \mu_0 + \sum_{j=1}^{J} \beta_j D_{ij} + X_i'\gamma + \epsilon_i. \tag{4}$$

It bears emphasizing that equation (4) is a causal model: while ϵ_i is defined so as to be orthognal to X_i, this residual need not be uncorrelated with the school attendance indicators, D_{ij}.

We're interested in how agnostic regression estimates compare with the causal parameters in equation

[3]Appendix B briefly discusses the consequences of match-specific outcomes in the VAM framework. Section 4.3 explores the robustness of our tests for VAM bias to effect heterogeneity.

(4). We therefore define population regression coefficients in a model with the same conditioning variables:

$$Y_i = \alpha_0 + \sum_{j=1}^{J} \alpha_j D_{ij} + X_i'\Gamma + v_i. \tag{5}$$

This is the population projection, so here the residuals, v_i, are necessarily orthogonal to all right-hand-side variables.

Regression model (5) has a causal interpretation when the parameters in this equation coincide with those in the causal model, equation (4). This in turn requires that school choices be unrelated to the unobserved component of student ability, a requirement that can be expressed as:

$$E\left[\epsilon_i | D_{ij}\right] = 0; j = 1, ..., J. \tag{6}$$

Restriction (6), sometimes called "selection-on-observables," means that $\alpha_j = \beta_j \ \forall j$. In practice, of course, regression estimates need need have a causal interpretation; rather, they may be biased. This possibility is represented by writing

$$\alpha_j = \beta_j + b_j,$$

where the parameter b_j is defined as the difference between regression and causal parameters for school j.

4 Testing Conventional VAM

4.1 Test Procedure

We use the variation in school attendance generated by admission lotteries at oversubscribed schools to assess the causal interpretation of conventional VAM estimates. A vector of dummy variables, $Z_i = (Z_{i1}, .., Z_{iL})'$, indicates lottery offers to student i for seats at L oversubscribed schools. Offers at school ℓ are randomly assigned conditional on a set of lottery-specific stratification variables, $C_{i\ell}$. These include an indicator for applying to school ℓ and possibly other variables such as application cohort and walk zone status. The vector $C_i = (C_{i1}', ..., C_{iL}')'$ collects stratification variables for all lotteries.

We maintain the assumption that lottery offers are (conditionally) mean-independent of student ability. In other words,

$$E[\epsilon_i | C_i, Z_i] = \lambda_0 + C_i'\lambda_c, \tag{7}$$

for a vector of parameters λ_0 and λ_c. This means that admission offers are valid instruments for school attendance after controlling for assignment strata.

We have fewer lotteries than schools ($L < J$), so the restrictions in (7) are insufficient to identify the parameters of the causal model, equation (4). At the same time, these restrictions can be used to test implications of the conventional value-added framework. Selection-on-observables implies that $v_i = \epsilon_i$, so

under assumption (6) the moment conditions (7) are equivalent to

$$E[v_i|C_i, Z_i] = \lambda_0 + C_i'\lambda_c. \tag{8}$$

We test this implication in an overidentification test of the sort widely used with IV estimators. Specifically, we look for coefficients of zero in the regression:

$$v_i = \phi_0 + C_i'\phi_c + Z_i'\phi_z + \omega_i. \tag{9}$$

The restriction (8) implies that $\phi_z = 0$. Note that a conventional IV overidentification test statistic has degrees of freedom given by the degree of overidentification; the orthogonality restrictions motivating a just-identifed IV model can't be tested. Here, however, the relevant test procedure has L degrees of freedom. Because conventional VAM ignores lotteries, we bring extra testable information to the table even with a single lottery.

Two variations on the test procedure based on equation (9) help us furthur understand the nature of the restrictions provided by randomized admissions lotteries. First, note that the VAM regression residual, v_i, is necessarily the difference between observed achievement and the fitted values generated by conventional OLS VAM estimation, $\hat{Y}_i$. We can therefore conclude that the two regressions

$$Y_i = \rho_0 + C_i'\rho_c + Z_i'\rho_z + \eta_i, \tag{10}$$

and

$$\hat{Y}_i = \psi_0 + C_i'\psi_c + Z_i'\psi_z + u_i, \tag{11}$$

should produce the same result. In other words, testing the vector equality (with L restrictions)

$$\rho_z = \psi_z \tag{12}$$

is the same as testing $\phi_z = 0$ in (9). This version of the test captures the intuition that effects of lottery offers on test scores should equal their effects on predicted scores from the value-added model.

A further interesting implication of the selection-on-observables restriction builds on (12). Specifically, we can treat (10) and (11) as the reduced form and first stage equations associated with a two-stage least squares procedure that uses lottery offers to instrument a model with Y_i on the left-hand side and $\hat{Y}_i$, treated as endogenous, on the right. Using all lottery offers as instruments for this model, or using them one at a time, we can expect each of the resulting IV estimates to equal 1. This too amounts to a test of L restrictions.

This interpretation links our approach to the "forecast bias" tests implemented in previous efforts to validate VAMs (Kane and Staiger, 2008; Kane et al., 2013; Deming, 2014; Chetty et al., 2014a). These tests ask whether the coefficient on predicted value-added equals one in procedures similar to the IV model

described here. Such tests have one degree of freedom, though the underlying models imply Q restrictions, where Q is the number of available quasi-experiments. Our test leverages these additional overidentifying restrictions, thereby increasing power.

4.2 Test Results

We test for bias in three OLS value-added models, differentiated by the control variables included in X_i. Outcomes are 6th grade math scores, and baseline covariates are measured in 5th grade.[4] Model 1 controls only for outcome year indicators; estimates from this model are essentially school average test score levels. Model 2 adds a vector of student demographics, including indicators for sex, race, subsidized lunch eligibility, special education status, limited English proficiency, and counts of baseline absences and suspensions. Finally, Model 3 includes cubic polynomials in baseline math and ELA test scores. Our richest model includes the same student-level control variables used in recent value-added studies by Kane et al. (2008) and Chetty et al. (2014a; 2014b).[5]

Figure 1 summarizes math value-added estimates from the three models. Each column reports an estimated standard deviation of α_j across schools, expressed in test score standard deviation units and adjusted for estimation error.[6] Adding controls for demographics and previous scores reduces the standard deviation of α_j from 0.5σ in Model 1 to 0.23σ in Model 3. This implies that observed student characteristics explain a large portion of the variation in school test score levels. The last three columns in Figure 1 report estimates of within-sector value-added standard deviations, constructed using residuals from regressions of $\hat{\alpha}_j$ on charter and pilot school indicators. Including sector effects reduces variation in value-added, suggesting large differences in school effects across sectors.

Table 3 describes the results of tests for bias in conventional VAMs. The first row shows value-added coefficients from overidentified IV models instrumenting VAM fitted values with lottery offers as in equations (10) and (11). These models are estimated via a two-stage optimal IV procedure that is efficient with arbitrary heteroskedasticity (White, 1982). The second row reports first stage F-statistics for these models, which measure the strength of the relationship between lottery offers and predicted value-added. A strong first stage is an important requirement for the IV version of the test: with a weak first stage, IV estimates are biased towards OLS, which in this case is one by construction (Bound et al., 1995).[7] Weakness in the first stage will therefore bias the test against rejection of value-added. The F-statistics in columns (1) through

[4]We focus on math scores because variation in value-added is larger for math than for ELA, resulting in more precise estimates. Results of bias tests for ELA, presented in Appendix Table A1, yield qualitatively similar results.

[5]These studies measure teacher value-added using class mean residuals from regressions of test scores on controls, with class mean covariates included as additional controls. This is appropriate under the assumption that value-added is a random effect independent of covariates, an additional assumption that is not required for our fixed effects approach. In Appendix Table A2 we report results of tests based on mean residuals, which also result in rejection of the restrictions in (8).

[6]The estimated standard deviations are given by $\hat{\sigma}_\alpha = (\frac{1}{J}\sum_j[(\hat{\alpha}_j - \hat{\mu}_\alpha)^2 - SE(\hat{\alpha}_j)^2]^{1/2}$, where $\hat{\mu}_\alpha$ is mean value-added and $SE(\hat{\alpha}_j)$ is the standard error of $\hat{\alpha}_j$.

[7]The OLS version of this model is a regression of test scores on VAM fitted values. If estimated in the same sample with no additional controls, this regression produces a coefficient that is algebraically one. The OLS analogues of our IV models differ slightly from one because the IV models control for lottery strata and exclude some students from the OLS sample.

(3) are between 30 and 45, well above the standard rule-of-thumb value of 10 (Staiger and Stock, 1997). This implies that lottery offers are strong instruments for OLS value-added.

The remaining rows of Table 3 report p-values for three tests. First, we test the null hypothesis that the IV value-added coefficient equals one. Second, we test the IV model's overidentifying restrictions, which require that value-added coefficients are the same in all lotteries but not necessarily equal to one. Finally, we test the full set of restrictions by regressing VAM residuals on lottery offers as in equation (9). Since asymptotic critical values for overidentification tests can be inaccurate, we also report p-values for the third test based on the bootstrap procedure described by Hall and Horowitz (1996), which has been shown to yield higher-order refinements.

The results of these tests establish that OLS value-added models are biased. The IV value-added coefficients for models 1 and 2 are 0.40 and 0.65, and all three tests reject at conventional levels for these models ($p \leq 0.01$). Model 3 adds cubics in previous test scores, powerful control variables in models measuring test score value-added. The addition of lagged scores boosts the IV coefficient to 0.86, and this estimate only differs from one at marginal significance levels ($p = 0.071$). Nonetheless, the overidentifying restrictions for Model 3 are rejected ($p = 0.003$), and the null hypothesis of joint predictive accuracy across all lotteries is also rejected (bootstrap-refined $p = 0.03$). These findings show that OLS value-added estimates fail to predict lottery effects even after controlling for a rich set of baseline covariates.[8]

Figure 2 presents a "visual instrumental variables" representation of the bias tests. This figure plots lottery-specific reduced form effects on test scores against first stage effects on OLS fitted values, separately by subject. Each panel also displays a line through the origin with slope equal to the relevant IV coefficient from Table 3 (solid lines) along with the 45-degree line (dashed lines). If VAM estimates perfectly predict lottery effects, the 45-degree line should fit all points up to sampling error. Shaded points indicate estimates that differ from the 45-degree line at the 10-percent confidence level.

Consistent with the results in Table 3, the clouds of points for models 1 and 2 do not seem to follow the prescribed trend, with most quasi-experiments rejecting VAM validity at the 10-percent level. The joint rejection is less visually obvious in Model 3, though a plot (Figure 3) of each lottery's t-statistic shows that 4 of 28 tests reject Model 3 at the 5-percent level, 8 reject at the 10-percent level, and more than half produce a t-statistic of at least one. Figure 2 also reveals that much of the value-added models' predictive power is generated by charter school lotteries, which contribute large first stage and reduced form effects. The relationship between OLS value-added and lottery effects is less clear in the pilot and traditional public school sectors. This impression is confirmed in column (4) of Table 3, which reports results for Model 3 excluding charter lotteries. The value-added coefficient for this model is further from one (0.55), though the absence of charter lotteries reduces its precision.[9]

[8]To check whether this finding generalizes beyond Boston we also applied our test procedure to the Charlotte-Mecklenberg lottery data analyzed by Deming (2014). For a VAM including controls for past achievement, our test generated a bootstrap-refined joint p-value of 0.00, implying bias in value-added estimates for Charlotte. Further details are available upon request.

[9]The first stage F-statistic for this model is 11.2, on the border of the range conventionally thought to indicate weak instruments (Staiger and Stock, 1997). To explore this possibility we also estimated limited information maximum likelihood

The results in Figure 2 highlight the advantage of our approach over previous efforts to validate VAMs. It is visually evident that much of our power to detect bias comes from lotteries that generate points near the vertical axis, implying small value-added first stages and large impacts on test scores. Such lotteries contribute weak instruments to the IV model and therefore have little influence on the value-added coefficient. These points clearly indicate that the predictions of the value-added model are violated, however, a finding that is captured by the overidentification test results. Previous efforts to validate VAMs focus only on the value-added coefficient in similar models, ignoring overidentifying restrictions. In our Boston data, testing these extra restrictions results in a more decisive rejection of value-added than a test based on the IV coefficient alone.

4.3 Heterogeneity vs. Bias

The tests in Table 3 reject because OLS VAM estimates fail to predict the effects of randomly generated offers of admission. Such failures signal bias in OLS estimates, denoted by the terms labeled b_j in our constant-effects model. In a world of heterogeneous causal effects, however, these test results might instead signal divergence between the local average treatment effects (LATEs) identified by lottery instruments and possibly more representative treatment effects captured by OLS (Imbens and Angrist, 1994; Angrist et al., 1996).

It bears emphasizing that rejections driven by effect heterogeneity also pose a problem for the value-added framework. In the presence of heterogeneous effects OLS identifies a variance-of-treatment weighted average causal effect that need not have predictive value for specific individuals (Angrist, 1998). The value-added enterprise is built on a foundation of limited variation in causal effects: the goal here is prediction, and value-added estimates are of little policy relevance if they fail to reliably predict the effects of changing school assignments.[10] Nevertheless, it is useful to know whether our rejection of value-added is due to bias or effect heterogeneity. We therefore present a brief exploration of heterogeneity in value-added estimates.

Our analysis of effect heterogeneity asks whether differences between IV and OLS estimates are explained by differences in observed characteristics between lottery compliers and other students, or by differences in the samples used to construct both sets of estimates. We conduct two analyses to explore these possibilities. First, we repeat our bias tests using OLS VAM specifications that allow school effects to differ across covariate-defined "types" of students (e.g. special education students or those with low levels of baseline achievement). Intuitively, this approach accounts for variation in school effects across covariate cells that may be weighted differently by IV and OLS. Appendix B provides a formal justification for this technique. Second, we conduct tests based on OLS VAMs estimated in the quasi-experimental sample. This approach checks whether differences between IV and OLS are caused by differences between students subject to lottery

(LIML) models which are robust to weak instruments (Stock and Yogo, 2005). The LIML estimate corresponding to column (4) equals 0.54 with a standard error of 0.19, very similar to the IV estimate.

[10] See Condie et al. (2014) for a discussion of the hazards of teacher value-added estimation in the presence of heterogeneous effects.

assignment and the general student population.

Panel A of Table 4 reports test results for a version of Model 3 with school effects that vary across student types. Column (2) allows value-added to differ by school year to address the concern that school effects may "drift" over time (Chetty et al., 2014a). Although the IV coefficient rises relative to the unadjusted specification repeated in column (1), the joint test for VAM bias continues to reject with a bootstrap-refined p-value of 0.038. Columns (3) through (5) define types based on subsidized lunch, special education, and baseline test score tercile. Results from these specifications are very similar to the pooled results in column (1). Column (6) allows value-added to differ across cells constructed by fully interacting six covariates: race, sex, subsidized lunch eligibility, special education, English language learner status, and baseline score tercile. Although stratifying by this large set of variables reduces the precision of conventional VAM estimates, the test for VAM validity rejects at the 5-percent level ($p = 0.036$).

Panel B of Table 4 reports results based on OLS VAMs estimated within the quasi-experimental sample. These tests again yield similar p-values, despite noisier VAM estimates from a smaller sample. This suggests that differences between lottery applicants and other students are not the source of differences between OLS and lottery estimates. Overall, the results in Table 4 reinforce the interpretation that our rejection of value-added is due to bias in OLS estimates rather than effect heterogeneity.

4.4 Quantitative Importance of Bias

While our results indicate the presence of bias in conventional value-added models, our baseline value-added coefficient of 0.86 also shows that VAM estimates have substantial predictive power. It's worth asking whether the magnitude of bias is large enough to meaningfully affect accountability policies that rely on value-added models.

This question cannot be answered on the basis of the estimates in Table 3 alone. Chetty et al. (2014a) note that models instrumenting OLS value-added predictions with quasi-experimental assignments produce estimates of the forecast bias coefficient, the regression of causal value-added β_j on OLS parameters α_j. This coefficient is

$$\frac{Cov\left(\beta_j, \alpha_j\right)}{Var\left(\alpha_j\right)} = \frac{\sigma_\beta^2 + \rho\sigma_\beta\sigma_b}{\sigma_\beta^2 + \sigma_b^2 + 2\sigma_\beta\sigma_b},$$

where σ_β and σ_b are standard deviations of β_j and b_j, and ρ is their correlation. The degree of forecast bias is defined as one minus this coefficient.

If value-added and bias are assumed to be uncorrelated ($\rho = 0$), the forecast bias coefficient equals one if and only if there is no bias ($\sigma_b = 0$), and a coefficient close to one implies little bias. Without this assumption, however, the forecast bias coefficient is not a reliable guide to the extent of bias: this coefficient also equals one when $\rho = -\sigma_b/\sigma_\beta$. There is no reason to expect value-added and bias to be uncorrelated, so it is impossible to assess the quantitative importance of bias on the basis of forecast bias tests. In the next section we develop methods to directly estimate the joint distribution of value-added and bias.

5 The Distribution of School Effectiveness

5.1 Model

We use a random coefficients model to describe the joint distribution of value-added, bias, and lottery compliance across schools. The random coefficients analysis serves two purposes. First, it allows us to estimate the relative magnitudes of variation in value-added and bias, as well as the correlation between the two. This yields a clearer picture of the quantitative importance of bias in OLS estimates. Second, the model allows us to construct empirical Bayes (EB; Morris, 1983) posterior predictions of individual school value-added that condition on both OLS and lottery estimates. These posteriors use lottery information to improve the accuracy of value-added predictions relative to conventional posteriors that use OLS estimates alone.

The basis for our model is a set of observed OLS, lottery reduced form, and first stage coefficient estimates. Let ρ_z^ℓ denote the element of ρ_z, the reduced form coefficient vector from equation (10), corresponding to $Z_{i\ell}$. Equations (4) and (7) imply that

$$\rho_z^\ell = \sum_{j=1}^{J} \pi_j^\ell \beta_j,$$

where π_j^ℓ is the first-stage coefficient on $Z_{i\ell}$ from a regression of D_{ij} on Z_i and C_i.[11] This expression shows that the lottery at school ℓ identifies a linear combination of value-added parameters, with coefficients equal to shares of students shifted into (or out of) each school by the lottery offer. Estimates of the π_j^ℓ are obtained by substituting D_{ij} for Y_i on the left-hand side of (10).

The OLS, reduced form and first stage estimates are modeled as noisy measures of school-specific parameters, which are in turn modeled as draws from a random coefficient distribution in the population of schools. The observed estimates can be written:

$$\begin{aligned} \hat{\alpha}_j &= \beta_j + b_j + e_j^\alpha, \\ \hat{\rho}_z^\ell &= \sum_j \pi_j^\ell \beta_j + e_\ell^\rho, \\ \hat{\pi}_j^\ell &= \pi_j^\ell + e_{\ell j}^\pi. \end{aligned} \tag{13}$$

Here the e_j reflect mean-zero estimation errors that vanish as within-school and within-lottery samples tend to infinity. Subject to the usual asymptotic approximations these errors can be modeled as normally distributed with a known covariance structure. Table 1 shows that the OLS and lottery estimation samples used here typically include hundreds of students per school, so these asymptotic approximations seem reasonable.

[11] Conditional random assignment of admission offers implies that Z_i is conditionally independent of baseline covariates X_i in addition to ϵ_i. Z_i is therefore conditionally independent of all terms in equation (4) except $\sum_j \beta_j D_{ij}$, so a regression of this quantity on Z_i controlling for C_i produces ρ_z.

The second level of the model treats school-specific parameters as draws from a joint distribution of causal effects, bias, and lottery compliance patterns. We parametrize the first-stage effect of admission at school ℓ on the probability of attending school this school as

$$\pi_\ell^\ell = \frac{\exp(\delta_\ell)}{1+\exp(\delta_\ell)}. \tag{14}$$

The parameter δ_ℓ can be viewed as the mean utility in a binary logit model predicting compliance with a random offer at school ℓ. Likewise, the first-stage effect of an offer to attend school $\ell \neq j$ on attendance at school j is parametrized as

$$\pi_j^\ell = -\pi_\ell^\ell \cdot \frac{\exp\left(\xi_j + \nu_j^\ell\right)}{1+\sum_{k\neq\ell}\exp\left(\xi_k + \nu_k^\ell\right)}. \tag{15}$$

The quantity $\xi_j + \nu_j^\ell$ is the mean utility for school j in a multinomial logit model predicting fallback school choices among students that comply with offers in lottery ℓ. The parameter ξ_j is constant across lotteries, which allows for the possibility that some schools are consistently likely to serve as fallback options for lottery losers. ν_j^ℓ is a random utility shock specific to school j in the lottery at school ℓ. Together, the parametrization in (14) and (15) guarantees that lottery offers increase the probability of enrollment at the target school and reduce enrollment probabilities at all other schools, and that effects on all probabilities are between zero and one in absolute value.[12]

Each school is characterized by a value-added coefficient β_j, a bias term b_j, an offer compliance utility δ_j, and a mean fallback utility ξ_j. We model these parameters as draws from a multivariate normal distribution:

$$(\beta_j, b_j, \delta_j, \xi_j) \sim N\left((\beta_0, b_0, \delta_0, \xi_0), \Sigma\right). \tag{16}$$

In some specifications we also allow the mean parameters $(\beta_0, b_0, \delta_0, \xi_0)$ to vary across school sectors (traditional, charter or pilot). Finally, the lottery and school-specific utility shocks are also modeled as normal:

$$\nu_j^\ell \sim N\left(0, \sigma_\nu^2\right). \tag{17}$$

Equations (16) and (17) can be viewed as a prior distribution in a hierarchical Bayesian framework, with $\theta \equiv (\beta_0, b_0, \delta_0, \xi_0, \Sigma, \sigma_\nu^2)$ the vector of hyperparameters governing the distribution of parameters across schools. We take an EB approach by estimating the hyperparameters then using the estimated prior to compute posterior value-added predictions for individual schools.

[12] This parameterization implies that $0 < \pi_\ell^\ell < 1$, $-1 < \pi_j^\ell < 0$ for $j \neq \ell$, and $\pi_\ell^\ell > -\sum_{j\neq\ell}\pi_j^\ell$. The residual probability $\pi_\ell^\ell + \sum_{j\neq\ell}\pi_j^\ell$ is minus the effect of an offer at the omitted school zero.

5.2 Simulated Minimum Distance Estimation

We estimate hyperparameters by simulated minimum distance (SMD), a variant of the method of simulated moments (McFadden, 1989). SMD focuses on moments that are informative about the parameters of interest, minimizing the distance between sample moments and corresponding model predictions. Our implementation uses means, variances, and covariances of functions of the OLS value-added estimates $\hat{\alpha}_j$, lottery reduced forms $\hat{\rho}_z^\ell$, and first stage coefficients $\hat{\pi}_j^\ell$. Appendix C provides a complete list of the moments used for SMD estimation.

Simulation allows us to compute model predictions in a straightforward manner, in spite of the fact that the moments of interest are complicated functions of the model's hyperparameters.[13] We simulate moments by drawing vectors of school-level parameters using equations (16) and (17), and vectors of the estimation errors in (13) from the joint asymptotic distribution of the OLS, reduced form and first stage estimates. Let $\hat{m}$ denote the vector of observed moments, and let $\tilde{m}^s(\theta)$ denote a corresponding vector of moments constructed from a simulated sample drawn in simulation s. This vector depends on the hyperparameter vector θ through the simulation draws. We average over simulated samples to construct the final vector of predicted moments:

$$\tilde{m}(\theta) = \frac{1}{S}\sum_{s=1}^{S} \tilde{m}^s(\theta).$$

Our baseline estimates set $S = 500$.

The simulated minimum distance estimator is defined by

$$\hat{\theta}_{SMD} = \arg\min_{\theta} \; J\left(\hat{m} - \tilde{m}(\theta)\right)' W \left(\hat{m} - \tilde{m}(\theta)\right),$$

where W is a weighting matrix. An efficient estimator sets this matrix equal to a consistent estimate of the inverse of the asymptotic covariance matrix of the moment conditions (Sargan, 1958; Hansen, 1982). We produce a first-step estimate with an identity weighting matrix, then use this estimate to compute a model-based optimal covariance matrix by simulation (see Appendix C for details). We then recompute $\hat{\theta}_{SMD}$ using the optimal weighting matrix.[14]

5.3 Empirical Bayes Posteriors

After estimating the random coefficients model we construct EB posterior predictions of the individual school value-added parameters β_j. Studies of teacher and school value-added typically shrink noisy teacher-

[13] For example, two of the moments are the mean and variance of the reduced form across lotteries. The probability limit of the reduced form is $\rho_\ell^z = \sum_j \pi_{\ell j}\beta_j$. Equations (14) through (17) imply that each term in ρ_ℓ^z is a normal random variate multiplied by a ratio involving sums of correlated log-normal random variates. There is no closed form expression for the mean or variance of this random variable.

[14] Altonji and Segal (1996) show that estimation error in the weighting matrix can generate finite-sample bias in two-step optimal minimum distance estimates. This bias is caused by correlation between the observations used to compute the moment conditions and those used to construct the weighting matrix. We compute the model-based weighting matrix using a second set of simulation draws independent of the draws used to compute the moments, which breaks the correlation between moment conditions and weights.

and school-specific estimates towards the grand mean in inverse proportion to the signal-to-noise ratio for these estimates, reducing mean squared error (see, e.g., Kane et al., 2008 and Jacob and Lefgren, 2008). Specifically, the conventional posterior mean for school j is

$$E\left[\alpha_j|\hat{\alpha}_j\right] = \lambda_j\hat{\alpha}_j + (1-\lambda_j)\mu_\alpha,$$

where $\lambda_j = \sigma_\alpha^2(\sigma_\alpha^2 + Var(e_j^\alpha))^{-1}$, and μ_α and σ_α^2 are the mean and variance of α_j. An EB version of the posterior mean plugs in estimates of the hyperparameters σ_α^2 and μ_α. We extend this approach to a setting where $\hat{\alpha}_j$ is biased and lottery estimates are available.

A just-identified example

To gain intuition for our lottery-based EB procedure, first consider a scenario where every school is oversubscribed, so $J = L$. In this case it is possible to estimate a value-added model that uses only variation from lotteries. Specifically, we can estimate a just-identified IV model treating the J school indicators D_{ij} as endogenous regressors and using the J lottery offer indicators Z_{ij} as instruments. This procedure yields a lottery-based value-added estimate for each school, $\hat{\beta}_j$.

Unlike OLS estimates, lottery-based IV estimates are not biased by student sorting. A researcher with access to both sets of estimates might nevertheless choose to put some weight on the OLS, however. The price for eliminating bias is a loss of precision: IV uses only the variation generated by random assignment, and the resulting estimates are likely to be considerably less precise than corresponding OLS estimates. The empirical Bayes approach optimally trades off bias in OLS estimates with variance in IV estimates to generate minimum mean squared error (MMSE) estimates of value-added (Judge and Mittlehammer, 2004; 2005; 2007).

In the just-identified case, the posterior mean for β_j conditional on the OLS estimate $\hat{\alpha}_j$ and the lottery estimate $\hat{\beta}_j$ is given by

$$E\left[\beta_j|\hat{\beta}_j, \hat{\alpha}_j\right] = w_{1j}\hat{\beta}_j + w_{2j}(\hat{\alpha}_j - b_0) + (1 - w_{1j} - w_{2j})\beta_0, \tag{18}$$

where

$$w_{1j} = \frac{\iota_{11}^j(\iota_{22}^j + \upsilon_{22}) + \iota_{12}^j(\upsilon_{22} - \upsilon_{12} - \iota_{12}^j)}{(\iota_{11}^j + 2\iota_{12}^j + \iota_{22}^j + \upsilon_{11})(\iota_{22}^j + \upsilon_{22}) - (\iota_{12}^j + \iota_{22}^j + \upsilon_{12})^2},$$

and

$$w_{2j} = \frac{\upsilon_{22}(\iota_{22}^j + \iota_{12}^j) - \upsilon_{12}\iota_{22}^j}{(\iota_{11}^j + 2\iota_{12}^j + \iota_{22}^j + \upsilon_{11})(\iota_{22}^j + \upsilon_{22}) - (\iota_{12}^j + \iota_{22}^j + \upsilon_{12})^2}.$$

Here the ι_{mn}^j are elements of the inverse covariance matrix of the sampling errors $(e_j^\beta, e_j^\alpha)'$, and the υ_{mn} are elements of Σ^{-1}.[15] As with conventional EB posteriors, an empirical Bayes version of the posterior mean plugs first-step estimates of b_0, β_0, w_{1j} and w_{2j} into equation (18).

[15]The formula in (18) applies under the assumption that estimation errors in $\hat{\beta}_j$ and $\hat{\alpha}_j$ are correlated within but not across schools. With correlation across schools the formula becomes a matrix-weighted average of the vectors of OLS and IV estimates.

The posterior mean in (18) is a weighted average of three objects: the unbiased IV estimate, the biased OLS estimate net of the mean bias, and the mean value-added.[16] This estimate jointly shrinks the IV and OLS estimates toward the mean to account for sampling error in both estimates and the bias of OLS. The shrinkage factors w_{1j} and w_{2j} are chosen to minimize the mean squared error of the resulting prediction.

Empirical Bayes posteriors in the underidentified case

In practice some schools are undersubscribed, so it is not feasible to obtain direct estimates of the β_j by IV. When $L < J$, there are more endogenous variables than instruments, and the order condition for identification of equation (4) is not satisfied (Hausman, 1983). The intuition developed in the just-identified case still applies, however: the posterior mean for β_j optimally combines information from OLS estimates and lotteries to minimize mean squared error.

Though lotteries are unavailable for some schools, OLS estimates are always available. The posterior mean for β_j conditional on the OLS estimate $\hat{\alpha}_j$ is given by

$$E\left[\beta_j|\hat{\alpha}_j\right] = \left(\frac{\sigma_\beta^2 + \sigma_{\beta b}}{\sigma_\beta^2 + \sigma_b^2 + 2\sigma_{\beta b} + Var\left(e_j^\alpha\right)}\right) \cdot (\hat{\alpha}_j - b_0) + \left(\frac{\sigma_b^2 + \sigma_{\beta b} + Var\left(e_j^\alpha\right)}{\sigma_\beta^2 + \sigma_b^2 + 2\sigma_{\beta b} + Var\left(e_j^\alpha\right)}\right) \cdot \beta_0. \tag{19}$$

It is straightforward to verify that equation (19) is the limit of (18) as the variance of the lottery estimate approaches infinity, which implies $\iota_{11}^j = \iota_{12}^j = 0$. Even with no lottery information available for school j, this prediction is an improvement over a conventional posterior mean that erroneously imposes $\sigma_{\beta b} = \sigma_b = 0$.

The lottery reduced form coefficients, $\hat{\rho}_z^\ell$, provide estimates of additional linear combinations of value-added parameters. If $\pi_{\ell j} \neq 0$, the reduced form for school ℓ puts weight on β_j and therefore contains a signal that can be used to improve the posterior prediction. The first stage coefficients, $\hat{\pi}_j^\ell$, reveal the weights in these linear combinations and provide further information due to the correlations between δ_j, ξ_j and β_j in (16). Note that lottery estimates improve posterior predictions for both oversubscribed and undersubscribed schools since schools without lotteries may serve as fallback enrollment options for lottery applicants.

In the underidentified case with both lottery and OLS estimates available and correlated estimation errors for all parameters, there is no closed form expression for the posterior mean of β_j and it is not possible to sample directly from the posterior distribution. The posterior mean can be approximated via Markov Chain Monte Carlo (MCMC) but the random coefficient vector is high-dimensional so this process is slow. We instead construct EB posterior modes (also known as maximum a posteriori estimates; see Gelman et al., 2013). The posterior modes are calculated by maximizing the log likelihood of the observed estimates conditional on the random coefficients plus the log likelihood of the random coefficients conditional on the estimated hyperparameters. Appendix D provides further details.

[16] We use the term "unbiased" here somewhat loosely. Just-identified IV estimates have no moments, so the mean of e_j^β is undefined. Still, as with OLS, IV estimation error disappears as the sample size grows. Note also that because the IV procedure sketched here is just-identified, the resulting estimates are approximately median-unbiased in the sense that the estimation error has a median of zero (Angrist and Pischke, 2009).

6 Parameter Estimates

6.1 Hyperparameters

We fit the random coefficients model to math value-added estimates from models 1 through 3, along with lottery reduced forms and first stages. Table 5 reports estimated distributions of value-added and OLS bias from each model.

The estimates reveal substantial heterogeneity in both causal value-added and bias across schools. The standard deviation of value-added is similar across models, between 0.19 test score standard deviations in Model 2 and 0.24σ standard deviations in Model 1. This is sensible since the set of controls included in OLS should not affect the distribution of school effectiveness. In contrast, the standard deviation of bias shrinks from 0.44σ in Model 1 to 0.16σ in Model 3. Evidently, controlling for observed student characteristics dramatically reduces the degree of bias in conventional value-added estimates. On the other hand, estimated standard deviations of bias are statistically significant in all models, and the estimates from Model 3 imply that variation in bias is comparable to variation in causal value-added even after controlling for demographics and baseline achievement. This implies that variation in conventional OLS value-added reflects school quality and student sorting in roughly equal measure.

The third row of Table 5 reports estimates of the correlation between value-added and bias across schools. Estimated correlations are close to zero for models 1 and 2. The estimated correlation betwen β_j and b_j is negative in model 3. This point estimate suggests that conventional models may overstate the effectiveness of low-quality schools and understate the effectiveness of high-quality schools, though it is too imprecise to be conclusive. The estimated variances of value-added and bias, together with their correlation, imply a value for the total variance of OLS value-added parameters, given by $\sigma_\alpha^2 = \sigma_\beta^2 + \sigma_b^2 + 2\rho\sigma_\beta\sigma_b$. The fourth row of Table 5 shows that the model's predictions for σ_α closely match the observed standard deviations of school effects displayed in Figure 1.

In the fifth row of Table 5 we report estimates of the forecast bias coefficient $(\sigma_{\alpha\beta}/\sigma_\alpha^2)$. The SMD estimates of the forecast bias coefficient range from 0.23 in Model 1 to 0.68 in Model 3. These estimates are smaller and less precise than the IV estimates in Table 3. This is likely due to differences in weighting: the SMD estimate is the coefficient from an unweighted regression in the population of schools, while the IV estimates implicitly weight lotteries by the precision of lottery-specific IV estimates (Angrist and Pischke, 2009). Consistent with this argument, the SMD forecast bias coefficient is between the pooled estimate in Table 3 and the estimate limited to the set of traditional public and pilot lotteries; this excludes charter schools, which contribute a small minority of lotteries but generate large value-added first stages.

The last row of Table 5 reports the J-statistic, which equals the minimized value of the SMD criterion function. Under the null hypothesis that the model is correctly specified the J-statistic follows a χ^2 distribution with degrees of freedom equal to the number of overidentifying restrictions (see Appendix C). We use 19 moments to estimate 15 parameters, so there are 4 overidentifying restrictions. The J-statistics suggest

that the random coefficients model provides a reasonable fit to the data: the overidentifying restrictions are not rejected at conventional levels in any model.

In previous work we've demonstrated differences in effectiveness between Boston's charter, pilot and traditional public sectors (Abdulkadiroğlu et al., 2011; Angrist et al., forthcoming). This suggests that accounting for sector effects may improve the predictive accuracy of school value-added models. Table 6 reports estimates from a version of Model 3 that allows the means of the four random coefficients to depend on school sector. Appendix Table A3 shows the full joint distribution of random coefficients from this model.

Consistent with earlier findings, the results in Table 6 imply a large charter advantage: charter school value-added exceeds traditional public school value-added by 0.38 standard deviations on average. The difference between pilot and traditional public schools is smaller and statistically insignificant. The estimates in column (2) indicate no systematic difference in bias across sectors. This also corroborates past work indicating that observational models with controls accurately reproduce lottery-based comparisons of the charter, pilot and traditional public sectors (Abdulkadiroğlu et al., 2011; Angrist et al., 2013). The bottom panel of Table 6 shows that sector effects do not eliminate variation in value-added and bias. Within-sector standard deviations of value-added and bias are 0.16σ and 0.13σ, slightly smaller than the corresponding estimates from Table 5.

6.2 Empirical Bayes Posteriors

Figure 4 summarizes empirical Bayes hybrid posterior mode predictions of school value-added for models 1 through 3. Each panel plots the EB posterior mode against conventional EB posterior means. The hybrid modes are highly correlated with the conventional means, implying that the posterior modes place significant weight on the OLS estimates. On the other hand, there are some rank reversals: rank correlations in versions of Model 3 with and without sector effects are 0.84 and 0.64. This suggests that incorporating information from lotteries may affect accountability decisions based on estimated value-added. In the next section we compare the effects of policies based on conventional and hybrid measures of school quality.

7 Consequences of Improved Value-added Predictions

Accurate value-added estimates are useful for parents wishing to choose high-quality schools, and for policymakers desiring to reward successful schools and censure unsuccessful ones. This section assesses the practical relevance of the improvements generated by our hybrid approach. We begin by measuring improvements in statistical error for value-added predictions. Next, we compute misclassification rates for accountability policies aimed at identifying high- and low-performing schools. Finally, we project achievement effects of school closure and expansion decisions based on value-added predictions.[17]

[17]In principle our model could be used to make optimal decisions for any specific loss function. Such decisions will generally use functions of the posterior distribution beyond the mean or mode. For example, a policy that seeks to minimize classification errors when sorting schools into categories should use posterior probabilities that a school falls into each category. Experimen-

These assessments are based on Monte Carlo trials with samples drawn from the random coefficient distributions described by the estimates in tables 5 and 6. In each trial we draw vectors of school effects and estimation error, and construct OLS, reduced form and first stage estimates. Next, we reestimate the random coefficients model on the simulated data and construct EB posterior predictions of value-added using the conventional and hybrid approaches. We then compare posterior predictions for each method to the corresponding true value-added parameters.

7.1 Reduced Statistical Error

Figure 5 reports root mean squared error (RMSE) for conventional and hybrid posterior predictions in models 1 through 3. Variation in conventional posteriors is mostly driven by bias in models 1 and 2, leading to high RMSE (0.49σ and 0.37σ, respectively). The hybrid posterior mode reduces RMSE to 0.18σ in both models, thereby generating improvements of 63 and 52 percent. This illustrates the importance of accounting for bias when choosing the shrinkage factor to apply to conventional value-added estimates. The random coefficients model detects that much of the variation in $\hat{\alpha}_j$ is due to bias in models 1 and 2, and hybrid posterior predictions account for this fact.

Conventional models that control for past achievement generate much more accurate predictions of value-added. Nonetheless, the hybrid approach continues to yield improvements. RMSE is 0.20σ in conventional value-added Model 3 and falls to 0.15σ when sector effects are included. Hybrid posteriors further reduce RMSE to 0.16σ and 0.11σ, implying reductions of 20 and 27 percent relative to conventional posteriors.

Table 7 presents a more complete picture of the accuracy of value-added predictions for Model 3, with sector effects included. The first row reports value-added for the top-ranked school according to each measure (for this exercise we normalize mean value-added to zero). On average, the school ranked first by a conventional value-added model increases test scores by 0.46σ. Average value-added is 0.51σ for the school ranked first according to the hybrid posterior mode. Hybrid posteriors generate a smaller increase in the average value-added of schools ranked in the top quintile (0.02σ). Similarly, hybrid posteriors reduce the value-added of the lowest-ranked school by 0.05σ, and average value-added in the bottom quintile falls by 0.04σ.

These relatively modest improvements are driven partly by the sectoral patterns in value-added and bias exhibited in Table 6. Specifically, charter schools produce much higher average value-added than district (non-charter) schools, so the upper tail of the true value-added distribution is largely composed of charters. Since there is no systematic difference in bias between charter and district schools, conventional VAMs accurately rank charters relative to district schools and correctly classify most charters in the upper tail.

As can be seen in columns (4) through (7) of Table 7, conventional posteriors are less accurate within the set of district schools, which leaves more scope for improvement from the hybrid approach. Compared

tation with this approach yielded only small differences from the results presented here, so we focus on posterior means and modes for simplicity.

to conventional posteriors, hybrid posteriors raise average value-added of the highest-ranked district school by 0.09σ and boost average value-added in the highest quintile by 0.06σ. These changes represent 54 and 60 percent of the maximum improvements that could be obtained with perfect knowledge of value-added. Results at the bottom of the district school distribution are similar to corresponding results in the full sample. This is another consequence of the large difference between charter and district schools, which results in few charters near the bottom of either value-added measure.

The bottom row of Table 7 reports coefficients from regressions of school value-added on percentile rank according to each metric. These coefficients provide a summary measure of the benefits to moving up the value-added ranks. Hybrid posteriors close roughly half of the gap between regression coefficients for true value-added and conventional estimates. In the true value-added distribution a ten percentile point increase in rank translates into a 0.088 standard deviation increase in school quality. The corresponding coefficients for conventional and hybrid posteriors are 0.081 and 0.073. A parent who moves her child from the 25th to the 75th percentile of the conventional distribution might therefore expect an achievement benefit of $5 \times 0.073 = 0.37\sigma$. The corresponding benefit for the hybrid measure is 0.41σ. As before, hybrid posteriors generate larger gains within the district: a move from the 25th to the 75th percentile of the district school distribution generates an improvement of 0.18σ for the conventional measure and 0.24σ for the hybrid prediction. Together, the findings in Tables 6 and 7 suggest that parents deciding between charter and district schools would not be misled by conventional value-added estimates, but hybrid posteriors offer opportunities for more informed choices within the set of district schools.

7.2 Improved Accountability Targeting

Mistakes in school accountability schemes can lead to arbitrary punishments or rewards and dampen incentives for improved school effectiveness. In Table 8 we assess misclassification rates for simple accountability rules based on conventional and hybrid posterior predictions. Columns (1) and (2) evaluate a policy that seeks to identify failing schools, defined as those in the bottom quintile of true value-added. This policy mimics Massachusetts' Framework for School Accountability and Assistance, which places schools into five tiers based on four-year histories of test score levels and changes. Schools in the bottom quintile of this measure are automatically placed in Level 3 or above, and a subset are then classified in Levels 4 and 5, putting them at risk of restructuring or closure.[18] Our error metrics for this policy are the fraction of failing schools erroneously classified as not failing (type II errors), and the fraction of non-failing schools erroneously categorized as failing (type I errors). Columns (3) and (4) report corresponding results for a policy aimed at identifying successful (top-quintile) schools.

The simulation results show that hybrid posterior predictions reduce misclassification rates. In Model 1, conventional empirical Bayes posteriors misclassify 69 percent of bottom-quintile schools. As a result, 17

[18]The Massachusetts accountability system also uses information on graduation, dropout rates and school site visits to classify schools; see http://www.doe.mass.edu/apa/sss/turnaround/level5/schools/FAQ.html for more information. The methods developed here could be easily extended to measure school effects for these additional outcomes.

percent of schools outside the bottom quintile are mistakenly labeled as failing. Likewise, Model 1 classifies 74 percent of successful schools as unsuccessful and 19 percent of unsuccessful schools as successful. A purely random school ranking would generate type II and type I error rates of 80 and 20 percent, respectively; evidently, school rankings based on test score levels are only slightly better than random. The hybrid posterior mode misclassifies 54 percent of failing schools and 49 percent of successful schools in model 1. These error rates are high, but they represent reductions of 22 percent and 34 percent compared to the conventional posterior mean.

Adding controls for observed characteristics reduces misclassification rates for both value-added predictions but does not eliminate the utility of hybrid estimates. Conventional misclassification probabilities for Model 3 are 0.48 for failing schools and 0.46 for successful schools. Hybrid posteriors reduce these rates to 0.42 and 0.39, implying 13 and 15 percent declines in errors. The gap in error rates for failing schools is larger in simulations with sector effects. In this model, the hybrid method reduces the probability of type II error for failing schools from 0.40 to 0.28, a decline of 30 percent. The improvement for successful schools is more modest (23 percent); as in Table 7, this likely reflects the fact that both methods correctly conclude that the upper tail is largely composed of charter schools.

Finally, the bottom row of Table 8 shows misclassification rates for accountability policies limited to the set of district schools. Both conventional and hybrid posteriors produce less accurate rankings when charter schools are excluded. The hybrid approach reduces the Type II (I) error rate for failing schools from 0.46 (0.12) to 0.34 (0.08). Similarly, hybrid posteriors reduce Type II (I) error rates for successful schools from 0.52 (0.13) to 0.32 (0.08). These changes represent 26 and 39 percent declines in mistakes in the lower and upper tails, respectively.

7.3 Effects on Students

Taken together, Tables 7 and 8 show that supplementing OLS value-added models with lottery estimates can reduce statistical error and misclassification rates. To better interpret the magnitudes of these improvements, we next consider the effects on student achievement that would result from using hybrid rather than conventional posteriors for school closure and expansion decisions. Specifically, we compute the expected impact of closing a low-performing district school selected by a value-added prediction and sending the displaced students to a new school of either average quality, or with quality equivalent to a high-performing district school selected by the same prediction method. We also consider the effects of replacing low-performing district schools with charter schools, a policy that already exists in the form of Boston's in-district charter conversions (Abdulkadiroğlu et al., 2015).

This exercise requires three assumptions. First, we assume that there are no direct disruptive effects of school closure. This assumption seems plausible for charter conversions, which do not require students to switch schools. Any direct effects of closure are also netted out in comparisons of decisions based on hybrid and conventional estimates, the main focus of our analysis. Second, we assume that closures and

expansions have no effects on students who do not attend the closed or expansion schools. This rules out general equilibrium effects (e.g. peer effects or effort responses from other schools). Third, we assume that schools selected for expansion can be replicated with no decrease in effectiveness. This requires elastic supply of inputs such as teacher and principles. These assumptions are strong, but they allow us to construct a rough prediction of the benefits that would result from incorporating lotteries into decisionmaking.

Table 9 reports predicted impacts of school closures and expansions. Column (1) shows effects of policies based on conventional posterior means, while column (2) displays effects using hybrid posterior modes. Column (3) compares these effects to infeasible policies based on true value-added. In the first four rows we consider effects of closing the lowest-ranked school and replacing it with an average district school, an expansion of a highly-ranked school, or an average charter school. The last four rows report effects of replacing an average bottom-quintile school with each of these expansions.

The results show that decisions based on hybrid posteriors generate achievement benefits, though the magnitudes of these improvements vary. Replacing the lowest-ranked school with an average district school increases achievement by 0.26σ or 0.31σ if the closure is selected based on conventional or hybrid posteriors, respectively, implying a relatively modest improvement of 0.05σ. The added benefit of using lotteries grows to 0.11σ if the closure is replaced by expanding an average top-quintile school, and 0.13σ if the replacement is a replica of the highest-ranked district school.

The hybrid approach generates similar gains for a more general policy that closes the average failing school. Replacing a representative school in this category with an average district school leads to a gain of 0.17σ for the conventional measure and 0.21σ for the hybrid, a difference of 0.04σ. Corresponding gains generated by expanding an average successful school or the highest-ranked school are 0.10σ or 0.12σ. As in Table 7, hybrid posteriors close roughly half the gap between conventional posteriors and the gains that could be achieved with knowledge of true value-added. Both panels of Table 9 show that replacing the closed school with an average charter produces gains comparable to expanding the highest-ranked district school, again reflecting the large charter advantage documented in Table 6.

Literature estimates of relationships between achievement and long-run outcomes suggest that the improvements in Table 9 are large enough to yield substantial economic value. In a study of the long-run impacts of teacher value-added, Chetty et al. (2014a) report that a 1σ increase in test scores induced by higher teacher quality leads to a 10 percent increase in earnings. Other recent studies report similar relationships from several contexts.[19] Assuming similar relationships between score gains and earnings for schools and teachers, the net effect of 0.1σ generated by using hybrid rather than conventional posteriors to select the lowest-ranked school is expected to generate a lifetime earnings gain of roughly \$2,600 per child.[20] The

[19] Chetty et al. (2014b) report that a one standard deviation increase in teacher quality raises achievement by 0.13σ and earnings at age 28 by 1.34 percent. Chetty et al. (2011) show that a one standard deviation improvement in kindergarten class quality improves test scores by 0.024 standard deviations and earnings by 0.3 percent, implying a 12 percent earnings gain per test score standard deviation. In OLS models with controls, Heckman et al. (2006) report that a one standard deviation increase in test scores is associated with 12 percent and 17 percent higher wages for males and females in the National Longitudinal Study of Youth. Similarly, in OLS regressions for Swedish men, Lindqvist and Vestman (2011) report estimates of 13.6 percent and 10.4 percent for earnings and wages.

[20] Chetty et al. (2014b) calculate that the average present discounted value of earnings in the United States ate age 12

average 6th grade class in Boston includes 78 students, implying an aggregate gain of more than $204,000 for each cohort and closure decision. A similar calculation predicts that using hybrid rather than conventional posteriors to replace a single failing school with a successful one would produce a net benefit of $400,000 per affected cohort. These rough calculations demonstrate that the stakes for school accountability decisions are substantial, and that the improvements generated by hybrid posteriors are large enough to generate meaningful benefits for students.

8 Conclusions and Next Steps

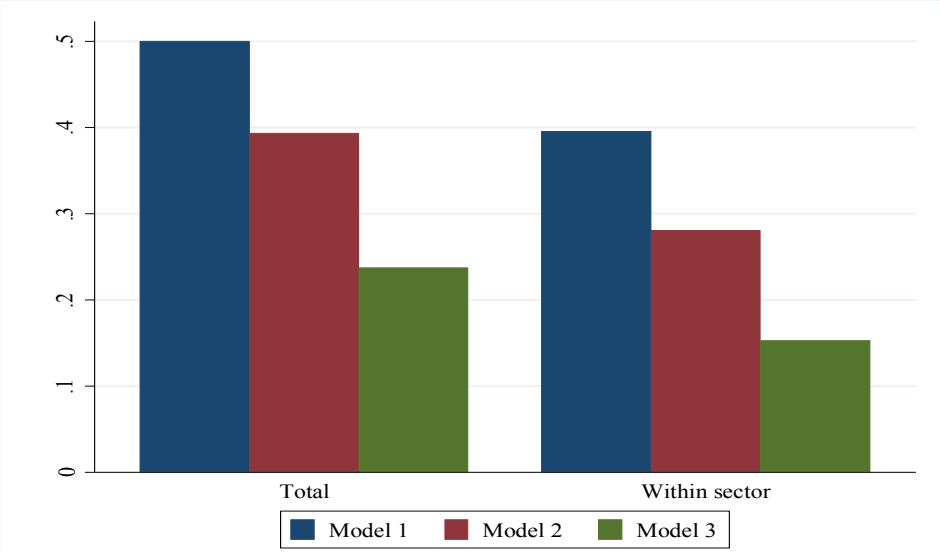
Test score value-added models are central components of school accountability policies in many districts. This paper uses randomized school admission lotteries from Boston, Massachusetts to test and improve these models. We reject the null hypothesis that conventional OLS value-added models predict admission effects in a large set of charter, pilot, and traditional public school lotteries. This implies bias in conventional models. Estimates of a random coefficients model of school effects reveal substantial variation in both true value-added and bias, which leads to errors in policies that use conventional estimates to identify effective and ineffective schools. We develop a new hybrid measure of school quality that combines lottery and OLS estimates to reduce the frequency of these mistakes.

Simulations show that this hybrid strategy generates reductions in statistical error and improvements in the accuracy of accountability policies. These improvements are large enough to generate important reductions in policy errors, and a simple calibration exercise suggests substantial benefits for students. Moreover, our analysis abstracts from other potential benefits of incorporating lotteries into policy decisions. For example, accountability schemes based on biased measures of value-added provide incentives for schools to strategically manipulate these measures by attracting students with high unobserved ability. Lottery estimates are unaffected by this kind of cream skimming, so policymakers may wish to dampen strategic incentives by putting more weight on lotteries than a simple statistical error criterion would dictate. We defer analysis of these strategic considerations to future work.

Our findings show that evidence from lotteries can act as a useful check on VAM-based policies, reducing the likelihood of mistakes that punish schools for the ability of their students rather than the quality of their instruction. Thanks to the rapid growth of centralized school assignment in public school districts, such evidence is increasingly available. Our approach may also be useful for testing and improving observational models in other contexts, including recent efforts to measure hospital, doctor and neighborhood effects (Finkelstein et al., 2013; Fletcher et al., 2014; Chetty and Hendren, 2015). The key requirement for our strategy is a source of quasi-experimental variation that can be used to discipline non-experimental comparisons.

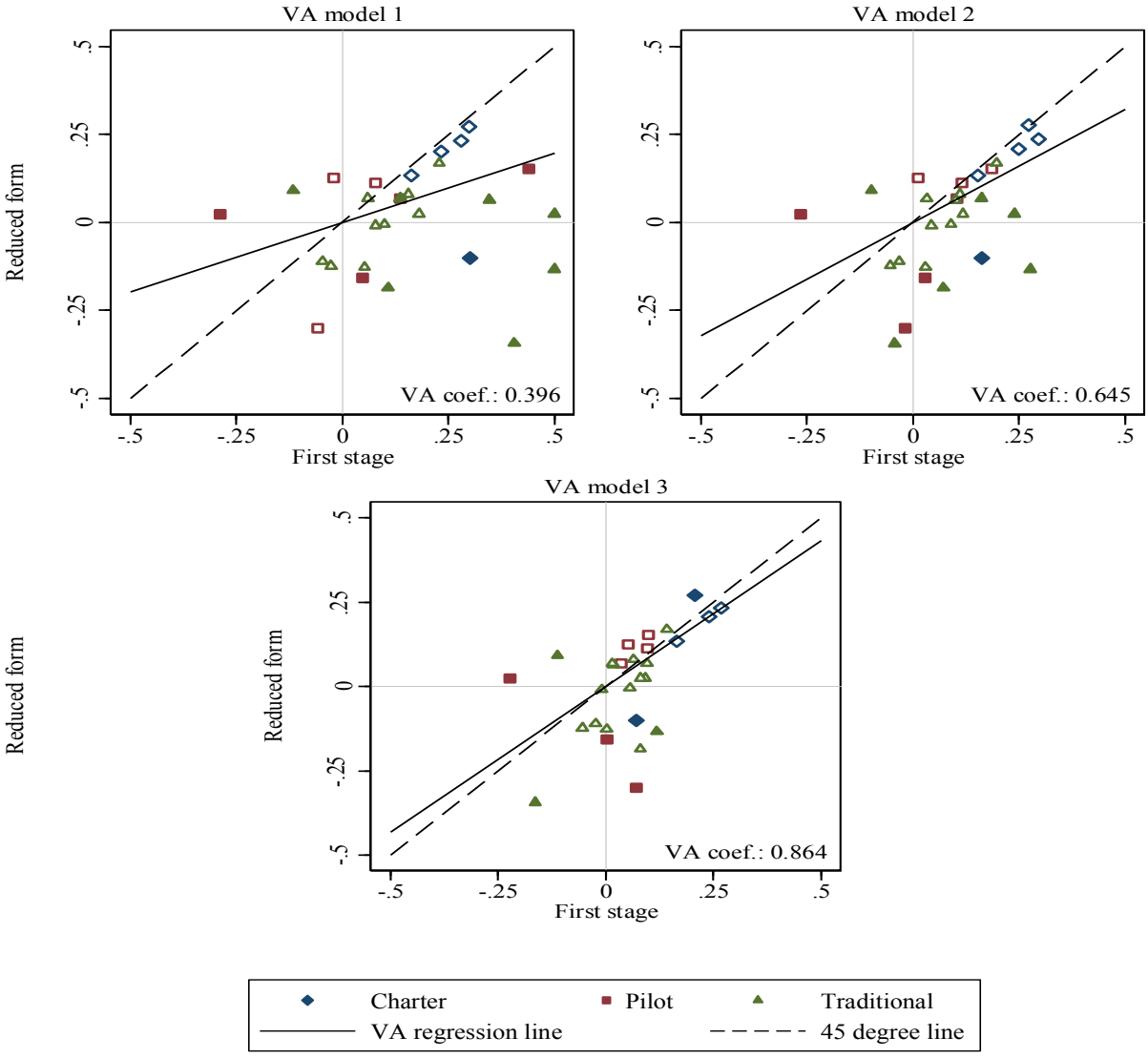
is $522,000 in 2010 dollars. On this basis, they project an increase of $52,200 in present value for each test score standard deviation.

Figure 1: Standard deviations of school effects from value-added models



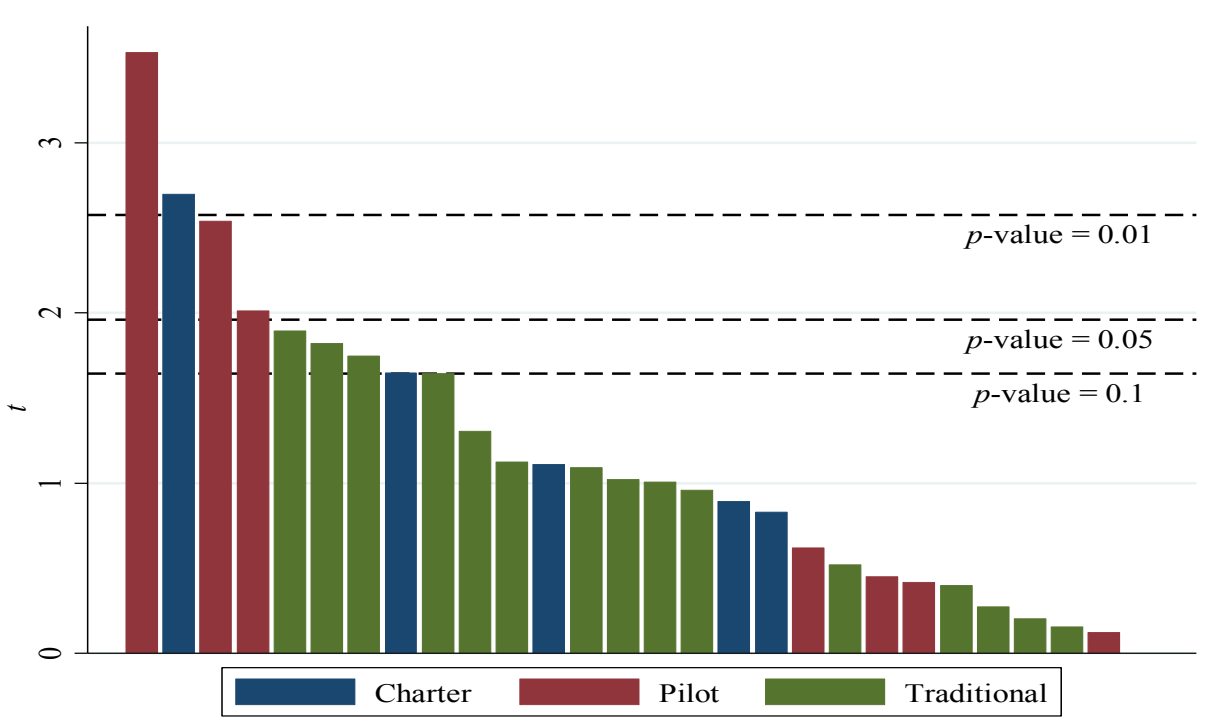
Notes: This figure compares standard deviations of school effects from three value-added models; see Table 3's notes for a description of the models. For each model, the total variance of school effects is obtained by subtracting the average squared standard error from the sample variance of value-added estimates, then taking the square root. Within sector variances are obtained by first regressing value-added estimates on charter and pilot dummies, then subtracting the average squared standard error from the sample variance of residuals and taking the square root.

Figure 2: Visual instrumental variables tests for bias



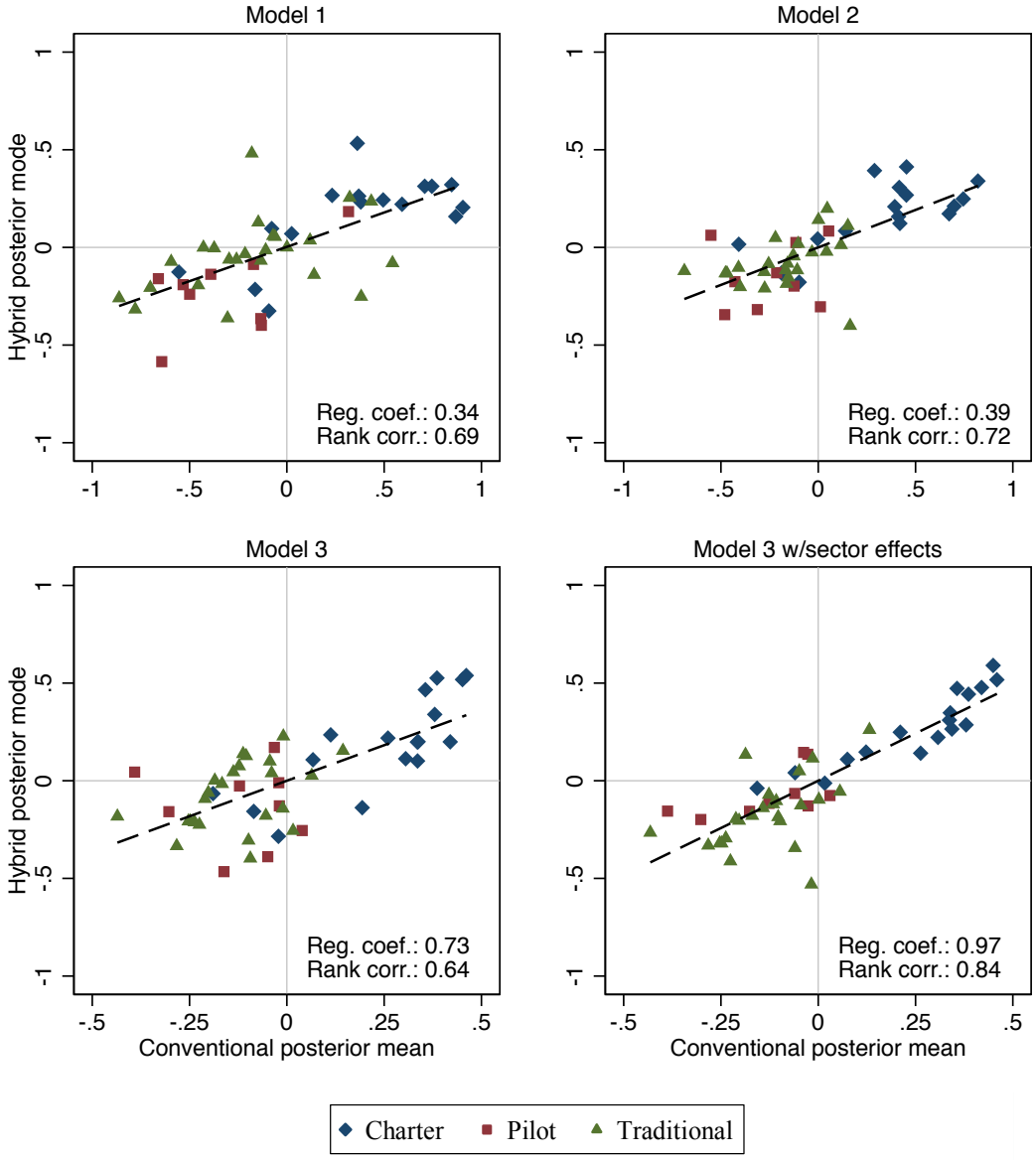
Notes: This figure plots quasi-experimental reduced form effects against value-added first stages from each of the 28 quasi-experimental samples. See the notes for Table 3 for a description of the value-added models and quasi-experimental specification. Filled markers indicate estimates that are significant at the 10% level. Slopes of solid lines corresponds to the IV value-added coefficients from Table 3, while dashed lines indicate the 45-degree line.

Figure 3: *t*-statistics for tests of bias



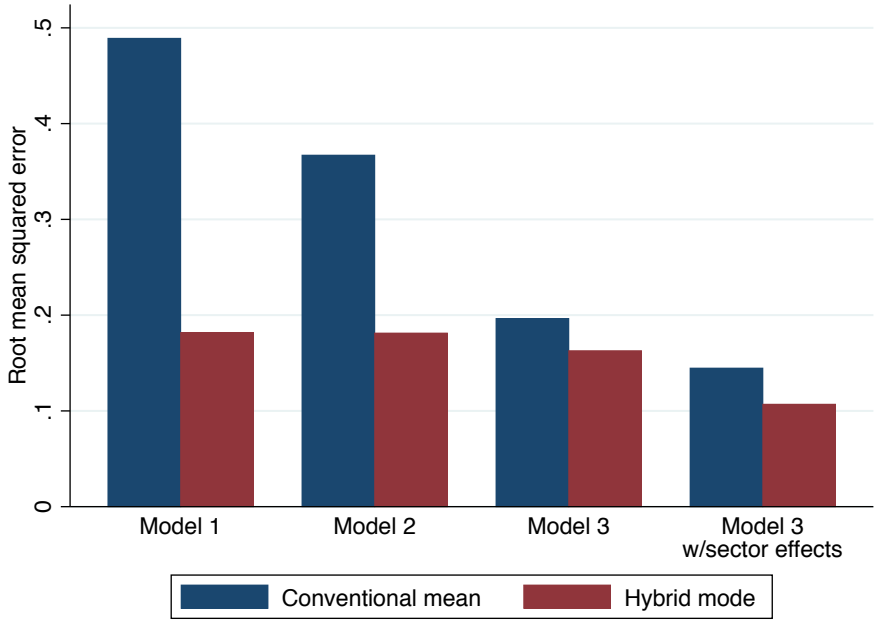
Notes: This figure plots *t*-statistic absolute values for school-specific forecast bias in conventional value-added. See the notes for Table 3 for a description of the value-added model (Model 3) and quasi-experimental specification. The horizontal lines indicate *t*-statistics significant at the 10% and 5% level.

Figure 4: Empirical Bayes posterior predictions of school value-added



Notes: This figure plots empirical Bayes posterior mode predictions of value-added from the random coefficients model against posterior means based on OLS value-added. Posterior modes are computed by maximizing the sum of the log-likelihood of the OLS, reduced form, and first stage estimates conditional on all school-specific parameters plus the log-likelihood of these parameters given the estimated random coefficient distribution. Conventional posteriors shrink observational estimates towards the mean in proportion to one minus the signal-to-noise ratio.

Figure 5: Root mean squared error for value-added posterior predictions



Notes: This figure plots root mean squared error for posterior predictions of school value-added. Conventional predictions are posterior means constructed from OLS value-added estimates. Hybrid predictions are posterior modes constructed from OLS and lottery estimates. Root mean squared error is calculated from 100 simuated samples drawn from the data genterating processes implied by the estimates in tables 5 and 6. The random coefficients model is re-estimated in each simulated sample.

Table 1: Boston students and schools

School	Enrollment		Lottery school?	School	Enrollment		Lottery school?
	OLS sample	Lottery sample			OLS sample	Lottery sample	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A. Charters				C. Traditional publics		
1	738	406	Y	1	1,095	79	Y
2	361	23		2	1,025	445	Y
3	357	215		3	1,713	1,084	Y
4	393	332	Y	4	547	218	Y
5	338	16		5	217	46	
6	511	115	Y	6	1,354	581	Y
7	71	8		7	263	44	
8	300	23		8	1,637	492	Y
9	389	342	Y	9	472	104	
10	654	34		10	1,238	591	Y
11	45	3		11	537	11	
12	53	2		12	331	35	Y
13	415	305	Y	13	335	82	
14	70	6		14	952	232	Y
15	104	23		15	294	71	Y
16	701	92		16	333	90	
17	85	37		17	766	243	Y
				18	372	47	Y
	B. Pilots			19	137	14	
1	538	310	Y	20	1,091	225	Y
2	1,260	433	Y	21	1,086	127	Y
3	585	296	Y	22	577	104	Y
4	78	5		23	622	61	
5	453	46	Y	24	906	270	Y
6	380	67	Y	25 *(Ref.)*	267	19	
7	242	179	Y				
8	558	73	Y	Total:	27,864	8,718	28
9	18	12					

Notes: This table counts the students and schools included in the observational (OLS) and quasi-experimental (lottery) samples. The sample covers cohorts attending 6th grade in Boston between the 2006-2007 and 2013-2014 school years. Traditional public school #25 is the designated omitted enrollment category for value-added estimation. Columns (4) and (8) indicate whether the school has enough students subject to lottery offer variation to be included in the quasi-experimental sample.

Table 2: Descriptive statistics

	Means		Offer instrument balance			
	OLS sample	Lottery sample	All lotteries	Charter	Pilot	Traditional
Baseline covariate	(1)	(2)	(3)	(4)	(5)	(6)
Hispanic	0.345	0.354	-0.017	-0.006	0.003	-0.007
			(0.013)	(0.018)	(0.033)	(0.017)
Black	0.410	0.485	-0.011	-0.009	-0.052	-0.005
			(0.014)	(0.020)	(0.034)	(0.018)
White	0.122	0.072	0.010	0.009	0.005	0.006
			(0.007)	(0.010)	(0.015)	(0.008)
Female	0.490	0.504	0.017	-0.025	-0.013	0.034*
			(0.014)	(0.020)	(0.037)	(0.019)
Subsidized lunch	0.806	0.830	0.020*	-0.005	0.006	0.020
			(0.010)	(0.016)	(0.026)	(0.013)
Special education	0.208	0.195	0.006	0.015	-0.022	-0.003
			(0.011)	(0.016)	(0.030)	(0.013)
English-language learner	0.205	0.206	0.006	0.004	0.018	-0.001
			(0.011)	(0.016)	(0.027)	(0.014)
Suspensions	0.093	0.076	-0.025	-0.016	0.009	-0.025
			(0.016)	(0.017)	(0.025)	(0.023)
Absences	1.710	1.534	-0.087	0.110	-0.092	-0.138*
			(0.095)	(0.167)	(0.260)	(0.080)
Math score	0.058	0.004	0.022	0.036	0.080	-0.026
			(0.024)	(0.035)	(0.061)	(0.030)
ELA score	0.030	0.013	0.035	0.013	0.060	0.045
			(0.025)	(0.036)	(0.061)	(0.030)
N	27,864	8,718	8,718	3,655	1,303	4,849

Notes: This table reports sample mean characteristics and investigates balance of random lottery offers. Column (1) shows mean characteristics for all Boston students 6th graders enrolled between the 2006-2007 and 2013-2014 school years, and column (2) shows mean characteristics for randomized lottery applicants. Columns (3)-(6) report coefficients from regressions of baseline characteristics on lottery offers, controlling for lottery strata. Robust standard errors, clustered by student, are reported in parenthenses.
*significant at 10%; **significant at 5%; ***significant at 1%

Table 3: Tests for bias in school value-added models

	Model 1 (1)	Model 2 (2)	Model 3 (3)	Model 3 without charter lotteries (4)
Value-added coefficient	0.396	0.645	0.864	0.549
	(0.056)	(0.065)	(0.075)	(0.164)
First stage *F* -statistic	45.6	36.1	29.6	11.2
p -value: VA coef. equals 1	<0.001	<0.001	0.071	0.006
p -value: overid. restrictions	<0.001	<0.001	0.003	0.043
Joint *p* -value: VA coef. equals 1 and overid. restrictions	<0.001	<0.001	<0.001	0.003
Bootstrap-refined joint *p* -value	<0.001	0.010	0.030	0.054
VAM controls:				
Year effects	X	X	X	X
Demographics		X	X	X
Baseline scores			X	X

Notes: This table reports lottery-based tests for bias in school value-added models. Value-added coefficients come from IV (two-step GMM) regressions of 6th grade math scores on estimated value-added at a student's attended school, instrumented by the set of offer dummies for all lotteries in the quasi-experimental sample. Joint *p* -values come from OLS regressions of value-added residuals on offer dummies. Value-added model 1 includes only year-of-test indicators as controls. Model 2 adds indicators for student sex, race, subsidized lunch, special education, limited-English proficiency, and counts of baseline absences and suspensions. Model 3 adds cubic polynomials in baseline math and ELA scores. Column (4) excludes charter school lotteries from the quasi-experimental sample in testing Model 3. All regressions control for lottery strata fixed effects and the baseline covariates from Model 3. Inference is robust to heteroskedasticity and accounts for first-step estimation error in school value-added. Bootstrap refinements to first-order asymptotics are based on 500 replication samples, with students drawn with replacement.

Table 4: Robustness of bias tests to effect heterogeneity

		VAM estimated by subgroup				
	Pooled VAM estimates	Baseline year	Subsidized lunch	Special education	Baseline score tercile	Interacted groups
	(1)	(2)	(3)	(4)	(5)	(6)
	A. VAM estimated on the OLS sample					
Value-added coefficient	0.864	0.916	0.849	0.863	0.866	0.930
	(0.075)	(0.072)	(0.075)	(0.074)	(0.075)	(0.061)
Bootstrap-refined joint *p*-value	0.030	0.038	0.022	0.034	0.016	0.036
	B. VAM estimated on the quasi-experimental sample					
Value-added coefficient	0.868	0.962	0.851	0.872	0.873	0.934
	(0.070)	(0.068)	(0.069)	(0.070)	(0.070)	(0.052)
Bootstrap-refined joint *p*-value	0.036	0.044	0.020	0.034	0.018	0.040

Notes: This table reports lottery-based tests for bias in school value-added models that allow for treatment effect heterogeneity by baseline characteristics. Value-added coefficients come from IV (GMM) regressions of 6th grade math scores on estimated value-added at a student's attended school, instrumented by the set of offer dummies for all lotteries in the quasi-experimental sample. Joint p-values come from OLS regressions of value-added residuals on offer dummies. The observational value-added specification is Model 3, described in the notes to table 3. Panel A estimates value-added in the full observational sample, while Panel B restricts estimation to the quasi-experimental subsample. Column (1) repeats the estimates from Table 3, while columns (2)-(6) implement the covariate-weighted version of these estimates described in the text. The covariates used to define subgroups in column (6) are hispanic, black, and female indicators, dummies for subsidized lunch, special education, and english language learner status, and indicators for baseline score terciles, based on average 5th grade math and ELA test scores in the observational sample. All regressions control for lottery strata fixed effects and the baseline covariates from Model 3. Inference is robust to heteroskedasticity and accounts for first-step VAM estimation error. Bootstrap refinements to first-order asymptotics are based on 500 replication samples, with students drawn with replacement.

Table 5: Joint distribution of causal value-added and OLS bias

Parameter	Description	Model 1 (1)	Model 2 (2)	Model 3 (3)
σ_β	Std. dev. of causal value-added	0.237 (0.064)	0.191 (0.072)	0.211 (0.063)
σ_b	Std. dev. of bias in OLS value-added	0.438 (0.076)	0.338 (0.102)	0.158 (0.067)
ρ	Correlation of value-added and bias	0.008 (0.321)	0.069 (0.479)	-0.227 (0.455)
σ_α	Std. dev. of OLS value-added	0.499 (0.051)	0.399 (0.052)	0.232 (0.025)
$\sigma_{\alpha\beta}/\sigma_\alpha^2$	Coefficient from regression of causal VA on OLS VA	0.228 (0.134)	0.258 (0.180)	0.680 (0.249)
	J-statistic (d.f.):	5.76 (4)	7.35 (4)	5.32 (4)
	Overid. *p*-value:	0.218	0.119	0.256

Notes: This table reports simulated minimum distance estimates of parameters governing the joint distribution of causal school value-added and OLS bias. The moments used in estimation are functions of the observed OLS, reduced form, and first stage estimates, as described in Appendix B. Simulated moments are computed from 500 samples constructed by drawing estimation errors from the asymptotic covariance matrix of the observed estimates, along with school-specific parameters drawn from the random coefficient distribution. Moments are weighted by an estimate of the inverse covariance matrix of the moment conditions, calculated from a first-step estimate using an identity weighting matrix. The weighting matrix is produced using 10,000 simulations, drawn independently from the samples used to compute the estimator. See notes to Table 3 for a description of the control variables included in each value-added model.

Table 6: Random coefficient estimates with sector effects

Mean shifters	β_j (1)	b_j (2)	δ_j (3)	ξ_j (4)
Charter	0.376	0.012	0.444	-2.003
	(0.122)	(0.118)	(0.492)	(0.482)
Pilot	0.062	-0.055	0.205	-0.556
	(0.141)	(0.136)	(0.444)	(0.541)
Within-sector variances				
Std. dev.	0.163	0.129	0.978	0.960
	(0.065)	(0.056)	(0.150)	(0.278)
Correlation w/β_j	1	-0.454	0.241	0.220
	-	(0.423)	(0.437)	(0.556)
		J-statistic (d.f.):	1.92 (4)	
		Overid. *p*-value:	0.750	

Notes: This table reports simulated minimum distance estimates of parameters governing the distribution of value-added, bias, and first-stage compliance across schools. The model allows the means of the random coefficients to depend on school sector (charter, pilot or traditional). See notes to Table 5 for a description of the estimation procedure. Observational value-added estimates come from Model 3, which is described in the notes to Table 3.

Table 7: Accuracy of value-added posterior predictions

	All schools			District schools		
	Conventional	Hybrid	True VA	Conventional	Hybrid	True VA
	(1)	(2)	(3)	(4)	(5)	(6)
Value-added of highest-ranked school	0.462	0.505	0.565	0.098	0.176	0.240
Average value-added in top quintile	0.329	0.345	0.376	0.003	0.061	0.103
Value-added of lowest-ranked school	-0.379	-0.430	-0.497	-0.379	-0.430	-0.497
Average value-added in bottom quintile	-0.264	-0.299	-0.338	-0.287	-0.327	-0.373
Regression of value-added on school percentile rank	0.732	0.807	0.878	0.357	0.486	0.590

Notes: This table describes the accuracy of empirical Bayes posterior predictions of value-added. Conventional empirical Bayes posteriors are means conditional on OLS estimates only, while hybrid posteriors are modes conditional on OLS and lottery estimates Statistics are means from 100 simulated samples drawn from the data genterating processes implied by estimates of value-added Model 3 with sector effects, listed in Table 6. The random coefficients model is re-estimated in each sample. Columns (1) through (3) show results for all Boston public schools, while columns (4) through (6) show results for traditional and pilot schools.

Table 8: Misclassification rates for policies based on value-added predictions

Value-added model	Posterior method	Fraction of failing schools not classified as failing (1)	Fraction of non-failing schools classified as failing (2)	Fraction of successful schools not classified as successful (3)	Fraction of unsuccessful schools classified as successful (4)
Model 1	Conventional	0.692	0.173	0.740	0.185
	Hybrid	0.537	0.134	0.494	0.124
Model 2	Conventional	0.629	0.157	0.640	0.160
	Hybrid	0.504	0.126	0.459	0.115
Model 3	Conventional	0.478	0.120	0.461	0.115
	Hybrid	0.417	0.104	0.393	0.098
Model 3 w/sector effects	Conventional	0.402	0.101	0.287	0.072
	Hybrid	0.283	0.071	0.222	0.055
Model 3 (district schools only)	Conventional	0.458	0.115	0.522	0.131
	Hybrid	0.337	0.084	0.317	0.079

Notes: This table reports misclassification rates for policies based on empirical Bayes posterior predictions of value-added. Value-added models 1-4 are defined in the notes to Table 3. Conventional empirical Bayes posteriors are means conditional on OLS estimates only, while hybrid posteriors are modes conditional on OLS and lottery estimates. Statistics are computed using 100 simulated samples drawn from the data genterating processes implied by the estimates in tables 5 and 6. The random coefficients model is re-estimated in each sample. Column (1) shows fractions of failing schools, defined as those in the bottom value-added quintile, that are not classified as failing according to each measure of value-added. Column (2) shows fractions of non-failing schools classified as failing. Column (3) shows fractions of successful schools, defined as those in the top value-added quintile, that are not classified as successful according to each measure of value-added. Column (4) shows fractions of unsuccessful schools classified as successful. Districts schools include traditional and pilot schools and exclude charter schools.

Table 9: Test score impacts of policies based on value-added predictions

Closed school	Replacement school	Conventional (1)	Hybrid (2)	True VA (3)
Lowest-ranked district school	Average district school	0.261	0.313	0.379
	Average successful district school	0.382	0.492	0.600
	Highest-ranked district school	0.477	0.606	0.736
	Average charter school	0.621	0.672	0.739
Average failing district school	Average district school	0.169	0.209	0.255
	Average successful district school	0.290	0.388	0.476
	Highest-ranked district school	0.385	0.502	0.613
	Average charter school	0.529	0.569	0.615

Notes: This table dispays predicted effects of accountability policies on test scores. Column (1) shows effects of school closure decisions based on posterior means from conventional OLS value-added models. Column (2) shows effects of decisions based on posterior modes that incorporate lottery estimates. Column (3) shows effects of decisions based on true school value-added. Calculations are based on simulations of Model 3 with sector effects, as described in Table 7. Failing schools are those ranked in the bottom quintile, and successful schools are those ranked in the top quintile.

References

Abdulkadiroğlu, A., J. D. Angrist, S. Dynarski, T. J. Kane, and P. A. Pathak (2011): "Accountability and flexibility in public schools: Evidence from Boston's charters and pilots," *Quarterly Journal of Economics*, 126(2), 699–748.

Abdulkadiroğlu, A., J. D. Angrist, P. D. Hull, and P. A. Pathak (2015): "Charters without lotteries: Testing takeovers in New Orleans and Boston," IZA discussipn paper no. 8985.

Abdulkadiroğlu, A., P. A. Pathak, A. E. Roth, and T. Sönmez (2006): "Changing the Boston school choice mechanism," NBER working paper no. 11965.

Altonji, J. G. and L. M. Segal (1996): "Small-sample bias in GMM estimation of covariance structures." *Journal of Business and Economic Statistics*, 14, 353–366.

Angrist, J. D. (1998): "Estimating the labor market impact of voluntary military service using Social Security data on military applicants," *Econometrica*, 66, 249–288.

Angrist, J. D., S. R. Cohodes, S. M. Dynarski, P. A. Pathak, and C. R. Walters (forthcoming): "Stand and deliver: Effects of Boston's charter high schools on college preparation, entry and choice," *Journal of Labor Economics*.

Angrist, J. D. and I. Fernandez-Val (2013): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE framework," in *Advances in Economics and Econometrics*, ed. by D. Acemoglu, M. Arellano, and E. Dekel, Cambridge University Press, vol. III of *Econometric Society Monographs, Econometrics*, chap. 11, 401–434.

Angrist, J. D., G. W. Imbens, and D. B. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91, 444–455.

Angrist, J. D., P. A. Pathak., and C. R. Walters (2013): "Explaining charter school effectiveness," *American Economic Journal: Applied Economics*, 5, 1–27.

Angrist, J. D. and J.-S. Pischke (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Baicher-Hicks, A., T. J. Kane, and D. O. Staiger (2014): "Validating teacher effect estimates using changes in teacher assignments in Los Angeles," Mimeo, Harvard University.

Bloom, H. S. and R. Unterman (2014): "Can small high schools of choice improve educational prospects for disadvantaged students?" *Journal of Policy Analysis and Management*, 33, 290–319.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak," *Journal of the American Statistical Association*, 90, 443–450.

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

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the impact of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 104, 2593–2563.

——— (2014b): "Measuring the impact of teachers II: Teacher value-added and student outcomes in adulthood," *American Economic Review*, 104, 2633–2679.

Chetty, R. and N. Hendren (2015): "The impacts of neighborhoods on intergenerational mobility: childhood exposure effects and county-level estimates," Mimeo, Harvard University.

Condie, S., L. Lefgren, and D. Sims (2014): "Heterogeneous match quality and teacher value-added: theory and empirics," *Economics of Education Review*, 40, 76–92.

Cullen, J. B., B. A. Jacob, and S. D. Levitt (2006): "The effect of school choice on participants: evidence from randomized lotteries," *Econometrica*, 74, 1191–1230.

Deming, D. (2014): "Using school choice lotteries to test measures of school effectiveness," *American Economic Review: Papers & Proceedings*, 104, 406–411.

Deming, D. J., J. S. Hastings, T. J. Kane, and D. O. Staiger (2014): "School choice, school quality, and postsecondary attainment," *American Economic Review*, 104, 991–1013.

Deutsch, J. (2012): "Using school lotteries to evaluate the value-added model," Mimeo, University of Chicago.

Dobbie, W. and R. G. Fryer (2013): "Getting beneath the veil of effective schools: evidence from New York City," *American Economic Journal: Applied Economics*, 5, 28–60.

——— (forthcoming): "The medium-term impacts of high-achieving charter schools," *Journal of Political Economy*.

Finkelstein, A., A. Sacarny, and C. Syverson (2013): "Healthcare exceptionalism? Productivity and allocation in the US healthcare sector," NBER working paper no. 19200.

Fletcher, J. M., L. I. Horwitz, and E. Bradley (2014): "Estimating the value added of attending physicians on patient outcomes," NBER working paper no. 20534.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013): *Bayesian Data Analysis*, Chapman and Hall/CRC, third ed.

Hahn, J. and J. Hausman (2002): "A new specification test for the validity of instrumental variables," *Econometrica*, 70, 163–189.

Hall, P. and J. L. Horowitz (1996): "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrica*, 64, 891–916.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

Hausman, J. A. (1983): "Specification and estimation of simultaneous equation models," in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator, North-Holland, vol. 1, 391–448.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor Economics*, 24, 411–482.

Imbens, G. W. and J. D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467–475.

Jacob, B. A. and L. Lefgren (2008): "Principals as agents: subjective performance assessment in education," *Journal of Labor Economics*, 26, 101–136.

Judge, G. G. and R. C. Mittlehammer (2004): "A semiparametric basis for combining estimation problems under quadratic loss," *Journal of the American Statistical Association*, 99, 479–487.

——— (2005): "Combing estimators to improve structural model estimation and inference under quadratic loss," *Journal of Econometrics*, 128, 1–29.

——— (2007): "Estimation and inference in the case of competing sets of estimating equations," *Journal of Econometrics*, 138, 513–531.

Kane, T. J., D. F. McCaffrey, and D. O. Staiger (2013): "Have we identified effective teachers? Validating measures of effective teaching using random assignment," *Gates Foundation Report*.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (2008): "What does certification tell us about teacher effectiveness? Evidence from New York City," *Economics of Education Review*, 27, 615–631.

Kane, T. J. and D. O. Staiger (2008): "Estimating teacher impacts on student achievement: An experimental evaluation," NBER working paper no. 14607.

Kinsler, J. (2012): "Assessing Rothstein's critique of teacher value-added models," *Quantitative Economics*, 3, 333–362.

Koedel, C. and J. R. Betts (2011): "Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique," *Education Finance and Policy*, 6, 18–42.

Lindqvist, E. and R. Vestman (2011): "The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment," *American Economic Journal: Applied Economics*, 3, 101–128.

McFadden, D. (1989): "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 57, 995–1026.

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

Rothstein, J. (2010): "Teacher quality in educational production: Tracking, decay, and student achievement," *Quarterly Journal of Economics*, 125, 175–214.

——— (2014): "Revisiting the impacts of teachers," Mimeo, University of California, Berkeley.

Sargan, J. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica*, 26, 393–415.

Staiger, D. and J. H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557–586.

Stock, J. H. and M. Yogo (2005): "Testing for weak instruments in linear IV regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews and J. H. Stock, Cambridge University Press, 80–108.

White, H. (1982): "Instrumental variables regression with independent observations," *Econometrica*, 50, 483–499.

Appendix A: Data

The data used for this project come from student demographic and attendance information in the Massachusetts Student Information Management System (SIMS), standardized student test scores from the Massachusetts Comprehensive Assessment System (MCAS) database, Boston charter school admission lottery records, and information from the centralized BPS student assignment system. We describe each data source and our cleaning and matching process in detail below; the construction of our main analysis file closely follows that of previous studies, in particular Abdulkadiroğlu et al. (2011).

A.1 Student enrollment, demographics, and test scores

The Massachusetts SIMS contains snapshots of all students in a public school in Massachusetts in October and at the end of each school year. These records contain demographic information on students, their current schools, their residence, and their attendance. We work with SIMS files for the 2005-2006 through the 2013-2014 school years and limit the sample to students enrolled in a Boston school over this period. Schools are classified as charters by the Massachusetts Department of Elementary and Secondary Education website (`http://www.profiles.doe.mass.edu`), and as pilots by the Boston pilot school network website (`http://www.ccebos.org/pilotschools/schools.html`). All remaining Boston schools are considered traditional public schools for the purposes of this study.

Enrollment in the SIMS is grade-specific. When a student repeats grades, we retain the first school a student attended in that grade. We then record students attending multiple schools in a given school year as enrolled in the school for which the attendance duration is longest, with duration ties broken randomly. This results in a unique student panel across grades; for the purposes of this study we restrict focus to 6th grade students enrolled from 2006-2007 to 2013-2014, using their 5th grade information for baseline controls. These controls include indicators for student race (Hispanic, black, white, Asian, and other race), sex, free- or reduced-price lunch eligibility, special education status, and English-language learner status, as well as counts of the number of days a student was suspended or truant over the school year. Suspension data are unavailable in the SIMS starting in the 2012-2013 school year; we include an indicator for students missing this baseline information whenever suspensions are used.

Our primary outcome for measuring school value-added are 6th grade standardized test scores from the Massachusetts Comprehensive Assessment System (MCAS) database. We normalize MCAS math and ELA scores by grade and year to be mean-zero and have standard deviation one within a combined BPS and Boston charter school reference population. MCAS scores are merged to SIMS data via a state-assigned unique student identifier. We also merge baseline (5th grade) math and ELA test scores for each student in our sample (5th grade MCAS information is available starting in the 2005-2006 school year).

A.2 Charter school lotteries

We use annual lottery records for five of the six Boston middle school charters with 6th grade admission for the 2006-2007 through the 2013-2014 academic year. These schools are Academy of the Pacific Rim, Boston Preparatory, MATCH Charter Public Middle School, Roxbury Preparatory, and UP Academy Charter School of Boston. The remaining school, Smith Leadership Academy, has declined to participate in our studies. For each school and each oversubscribed year we obtain a list of names of students eligible for entry by lottery, as well as information on whether each student was offered a seat on lottery night. Students are marked as ineligible if they submit an incomplete or late application; we also exclude students with a sibling currently enrolled in the school, as they are guaranteed admission. For UP Boston, which is an in-district charter school, students applying from outside of BPS are placed in a lower lottery priority group.

A student is coded as receiving a charter admission offer if she is offered a seat on lottery night. These offers are randomly assigned within strata defined by school, application year, and, in the case of UP Boston, BPS priority group. Students are retained the first year they apply to a charter school. We match the set of charter offers and randomization strata to state data by student name, grade, and application year; 97% of charter lottery applicants are successfully matched.

A.3 The BPS mechanism

We obtain a complete record of student-submitted preferences, school priorities, random tiebreaking sequence numbers, and assignments from the BPS deferred-acceptance mechanism, 2006-2007 though 2013-2014. For each year, we identify groups of students subject to the same priorities (given by whether a student has an enrolled sibling and whether she resides in a school's walk-zone, a 1.5 mile radius) at schools that they rank first. In forming these groups we exclude students that are guaranteed admission by virtue of being currently enrolled in the school, as well as certain other students with guaranteed or nonstandard priorities (see Abdulkadiroğlu et al. (2006) for a complete description of priorities in BPS). Within groups we construct indicators for whether an applying student has a random sequence number that is better than the worst number belonging to a student in the group that is assigned to each school. A student qualified in such a way is assigned to it by the mechanism, and such offers are randomly assigned within strata defined by school, application year, and priority group. We drop all schools with fewer than 50 students subject to conditionally-random admission, and match offers and randomization strata to state data via a BPS unique student identifier. Students are retained the first year they enter the BPS mechanism for 6th grade entry.

A.4 Sample Selection

We restrict attention to Boston public schools with at least 25 6th grade students enrolled in each year of operation from 2006-2007 to 2013-2014. In our merged analysis file this leaves 51 schools (see Table 1). Students enrolled at these schools are retained if they were enrolled in Boston in both 5th and 6th grade, if

their baseline demographic, attendance, and test score information is available, and if we observe their 6th grade MCAS test scores. These restrictions leave a total of 27,864 Boston students, summarized in detail in Table 2. Of these, 8,718 students are subject to quasi-experimental variation in 6th grade admission at 28 schools, either from a charter school lottery or from assignment by the BPS mechanism.

Appendix B: Heterogeneous School Effects

This appendix generalizes our lottery-based tests for bias to a model that allows school effects to vary across students. For a set of mutually exclusive and exhaustive student "types" t, we write potential achievement as

$$Y_{ij} = \sum_t \mu_{jt} T_{it} + a_i + m_{ij}, \tag{20}$$

where T_{it} is an indicator for belonging to type t, μ_{jt} is average potential achievement at school j for individuals of type t, and a_i is student ability, which is again orthogonal to student type by definition. The decomposition in equation (20) also allows for an unrestricted match component in achievement, m_{ij}, and is therefore fully general.

Equation (20) implies that the observed outcome for student i can be written

$$Y_i = \sum_t \mu_{0t} T_{it} + \sum_{j=1}^{J} \sum_t \beta_{jt} D_{ij} T_{it} + X_i'\gamma + \epsilon_i, \tag{21}$$

where $\beta_{jt} = \mu_{jt} - \mu_{j0}$ is the value-added of school j for type t and the error term is $\epsilon_i = \sum_{j=1}^{J} \sum_t (m_{ij} - m_{i0}) D_{ij} T_{iw} + \epsilon_{i0}$, with ϵ_{i0} the residual from a projection of $a_i + m_{i0}$ on X_i.

The following assumptions extends selection-on-observables to the case with effect heterogeneity:

$$E\left[\epsilon_{i0} | D_i\right] = 0, \tag{22}$$

$$E\left[m_{ij} - m_{i0} | D_i, T_i\right] = 0, \tag{23}$$

where T_i is the vector of all type dummies. Assumption (22) requires that general student ability is unrelated to school choices after controlling for X_i, similar to (6). The additional assumption (23) requires that idiosyncratic match effects are also independent of school choices conditional on T_i. In other words, any relationship between school effects and school choices occurs through the type-specific effects β_{jt}, not through sorting on gains within type. Assumption (23) is similar to the "conditional effect ignorability" assumption described by Angrist and Fernandez-Val (2013), which rules out sorting on unobserved treatment gains within covariate cells.

The OLS regression corresponding to the causal model (21) is

$$Y_i = \sum_t \alpha_{0t} T_{it} + \sum_{j=1}^{J} \sum_t \alpha_{jt} D_{ij} T_{it} + X_i'\Gamma + v_i, \tag{24}$$

with v_i orthogonal to T_{it}, D_{ij} and X_i by definition. Together with the maintained exclusion restriction (7), assumptions (22) and (23) imply that residuals from model (24) should be orthogonal to lottery offers after controlling for randomization strata. If rejection of the more restrictive model (4) is caused by heterogeneity

in school effects across student types rather than bias, tests based on (24) will not reject. In section 4.4 we conduct tests with student types defined in a variety of ways.

Appendix C: Simulated Minimum Distance

Overview

This appendix describes the simulated minimum distanced (SMD) approach used to compute estimates of the random coefficients model. The vector of parameters to be estimated is

$$\theta = \left(\alpha_0, \beta_0, \delta_0, \xi_0, \Sigma, \sigma_\nu^2\right)'.$$

We estimate these parameters by fitting means, variances, and covariances of OLS value-added, lottery reduced form, and first stage estimates. The complete vector of observed estimates is

$$\hat{\Omega} = \left(\hat{\alpha}_1, ..., \hat{\alpha}_J, \hat{\rho}_1, ..., \hat{\rho}_L, \hat{\pi}_1^1, ..., \hat{\pi}_1^L, ..., \hat{\pi}_J^L\right)'.$$

Let $\Omega = (\alpha_1, ..., \pi_J^L)'$ denote the probability limits of these estimates. We assume that the sampling distribution of $\hat{\Omega}$ is well approximated by asymptotic theory, so that

$$\hat{\Omega} \sim N\left(\Omega, V\right),$$

where V is a covariance matrix derived from conventional asymptotics. This assumption requires within-school and within-lottery samples to be large enough for asymptotic approximations to be accurate.

Under this assumption and the distributional assumptions in equations (14) through (17), we can simulate values of Ω and $\hat{\Omega}$ for any value of θ. We use this procedure to generate simulated data sets, and estimate θ by minimizing the distance between simulated and observed moments.

Moments

We select 19 moments to match. We next list these moments, and provide intuition for the information they contain about the parameters to be estimated.

Means

The first six moments are means of functions of the parameter estimates. The moment conditions are:

$$\hat{m}_1 = \tfrac{1}{J} \textstyle\sum_j \hat{\alpha}_j,$$

$$\hat{m}_2 = \tfrac{1}{L} \textstyle\sum_\ell \hat{\rho}_z^\ell,$$

$$\hat{m}_3 = \tfrac{1}{L} \textstyle\sum_\ell \hat{\pi}_\ell^\ell,$$

$$\hat{m}_4 = -\tfrac{1}{L} \textstyle\sum_\ell \sum_{j \neq \ell} \hat{\pi}_j^\ell,$$

$$\hat{m}_5 = -\tfrac{1}{L(J-1)} \textstyle\sum_\ell \sum_{j \neq \ell} \frac{\hat{\pi}_j^\ell}{\hat{\pi}_\ell^\ell},$$

$$\hat{m}_6 = \tfrac{1}{L} \textstyle\sum_\ell \left[\frac{\left(\hat{\pi}_\ell^\ell\right)^2}{\sum_k \left(\hat{\pi}_k^k\right)^2}\right] \cdot \left(\frac{\hat{\rho}_z^\ell}{\hat{\tau}_z^\ell}\right).$$

$\hat{m}_1$ is the mean OLS coefficient, which provides information about $\beta_0 + b_0$, the sum of mean value-added and mean bias. $\hat{m}_2$ is the mean reduced form, which provides information about β_0. $\hat{m}_3$ is the mean first stage across lotteries, which can be used to estimate δ_0. $\hat{m}_4$ is the average sum of fallback probabilities for included schools across lotteries, and $\hat{m}_5$ is the average ratio of this sum to the first stage, which gives the share of compliers drawn from included schools. These two moments help to estimate ξ_0, the mean fallback utility for included schools relative to the omitted school. $\hat{m}_6$ is the average ratio of the lottery reduced form to a "pseudo-reduced form" prediction that uses OLS value-added estimates, given by $\hat{\tau}_z^\ell = \sum_j \hat{\pi}_j^\ell \hat{\alpha}_j$. We weight this average by the squared lottery first stage to avoid unstable ratios caused by small first stages. This moment yields information about the variance of b_j, the bias in conventional value-added estimates, along with the correlation between β_j and b_j.

Variances

The next seven moments are variances of parameter estimates:

$$\hat{m}_7 = \tfrac{1}{J} \textstyle\sum_j \left(\hat{\alpha}_j - \bar{\alpha}\right)^2,$$

$$\hat{m}_8 = \tfrac{1}{L} \textstyle\sum_\ell \left(\hat{\rho}_z^\ell - \bar{\rho}\right)^2,$$

$$\hat{m}_9 = \tfrac{1}{L} \textstyle\sum_\ell (\hat{\tau}_z^\ell - \bar{\tau})^2,$$

$$\hat{m}_{10} = \tfrac{1}{L} \textstyle\sum_\ell \left(\hat{\pi}_\ell^\ell - \bar{\pi}_{own}\right)^2,$$

$$\hat{m}_{11} = \tfrac{1}{J} \textstyle\sum_j \left[\left(\tfrac{1}{L-1} \sum_{\ell \neq j} \hat{\pi}_j^\ell\right) - \bar{\pi}_{other}\right]^2,$$

$$\hat{m}_{12} = \tfrac{1}{J} \textstyle\sum_j \left[\left(\tfrac{1}{L-1} \sum_{\ell \neq j} \frac{\hat{\pi}_j^\ell}{\hat{\pi}_\ell^\ell}\right) - \bar{s}_{other}\right]^2,$$

$$\hat{m}_{13} = \tfrac{1}{J(L-1)} \textstyle\sum_j \sum_{\ell \neq j} \left(\hat{\pi}_j^\ell - \bar{\pi}_{j,other}\right)^2.$$

Here $\bar{\alpha}$ indicates the sample average of the α_j, and similarly for other variables. $\hat{m}_7$ is the variance of conventional value-added estimates across schools, which depends on the variances of value-added and bias as well as their covariance. $\hat{m}_8$ and $\hat{m}_9$ are variances of the lottery reduced form and predicted reduced form, which contain additional information about the joint distribution of value-added and bias. $\hat{m}_{10}$ is the variance of the first stage across lotteries, which helps to identify the variance of δ_j. $\hat{m}_{11}$ computes the mean share of students drawn from each school across lotteries, then takes the variance of this mean share across schools. This is the between-school variance in fallback probabilities. $\hat{m}_{12}$ is the variance of the mean share of compliers drawn from a particular school; $\bar{s}_{other}$ is the mean of this variable. These two moments yield information about the variances of ξ_j and ν_j^ℓ, which govern heterogeneity in fallback probabilities. $\hat{m}_{13}$ computes the variance of fallback shares across lotteries at every school, then averages across schools. This is the average within-school variance in fallback probabilities. This moment helps to separate the variance of ξ_j, the school-specific mean fallback utility, from σ_ν^2, the variance of idiosyncratic school-by-lottery utility shocks.

Covariances

The final six moments are covariances:

$$\hat{m}_{14} = \frac{1}{L}\sum_{\ell}\left(\hat{\rho}_z^{\ell} - \bar{\rho}\right)\left(\hat{\tau}_z^{\ell} - \bar{\tau}\right),$$

$$\hat{m}_{15} = \frac{1}{L}\sum_{\ell}\left(\hat{\rho}_z^{\ell} - \bar{\rho}\right)\left(\hat{\pi}_{\ell}^{\ell} - \bar{\pi}_{own}\right),$$

$$\hat{m}_{16} = \frac{1}{L}\sum_{\ell}\left(\hat{\alpha}_{\ell} - \bar{\alpha}\right)\left(\hat{\pi}_{\ell}^{\ell} - \bar{\pi}_{own}\right),$$

$$\hat{m}_{17} = \frac{1}{L}\sum_{\ell}\left(\hat{\rho}_z^{\ell} - \bar{\rho}\right)\left[\left(\frac{1}{L-1}\sum_{k\neq\ell}\hat{\pi}_{\ell}^{k}\right) - \bar{\pi}_{other}\right],$$

$$\hat{m}_{18} = \frac{1}{L}\sum_{\ell}\left(\hat{\alpha}_{\ell} - \bar{\alpha}\right)\left[\left(\frac{1}{L-1}\sum_{k\neq\ell}\hat{\pi}_{\ell}^{k}\right) - \bar{\pi}_{other}\right],$$

$$\hat{m}_{19} = \frac{1}{L}\sum_{\ell}\left(\hat{\pi}_{\ell}^{\ell} - \bar{\pi}_{own}\right)\left[\left(\frac{1}{L-1}\sum_{k\neq\ell}\hat{\pi}_{\ell}^{k}\right) - \bar{\pi}_{other}\right].$$

$\hat{m}_{14}$ is the covariance of the reduced form and pseudo-reduced form, which helps to identify variation in bias, as well as the covariance between bias and value-added. $\hat{m}_{15}$ is the covariance between reduced forms and first stages, which is informative about the covariance between β_j and π_j^j. $\hat{m}_{16}$ is the covariance of conventional value-added and the first stage, which helps to identify the covariance between b_j and δ_j. $\hat{m}_{17}$ is the covariance of the reduced form and average fallback probability, which helps to identify the covariance of β_j and ξ_j. $\hat{m}_{18}$ is the covariance of OLS value-added with the average fallback probability, which depends on the covariance between b_j and ξ_j. $\hat{m}_{19}$ is the covariance of a school's first stage and average fallback probability, which provides information about the covariance of ξ_j and δ_j.

There are 15 elements of θ and 19 moments, so the model has four overidentifying restrictions. Models that include charter and pilot school effects add sector-specific values of $\hat{m}_1$, $\hat{m}_2$, $\hat{m}_3$ and $\hat{m}_4$.

Estimation and Inference

Let $\hat{m} = (\hat{m}_1, ..., \hat{m}_{19})'$ be the vector of observed moments, and let $\tilde{m}(\theta)$ be the corresponding vector of simulated predictions. The simulated minimum distance estimator with weighting matrix W is

$$\hat{\theta}_{SMD}(W) = \arg\min_{\theta} \; J\left(\hat{m} - \tilde{m}(\theta)\right)' W \left(\hat{m} - \tilde{m}(\theta)\right).$$

We produce a first-step estimate of θ with an identity weighting matrix, then use this estimate to compute a model-based covariance matrix by simulation. The weighting matrix is given by

$$\hat{W} = \left[J \cdot \frac{1}{R}\sum_{r}\left(\tilde{m}^r\left(\hat{\theta}_{SMD}(I)\right) - \bar{m}\right)\left(\tilde{m}^r\left(\hat{\theta}_{SMD}(I)\right) - \bar{m}\right)'\right]^{-1},$$

where r indexes a second independent set of $R = 10,000$ simulation draws and $\bar{m}$ is the mean of the simulated moments. An efficient two-step estimate is given by $\hat{\theta}_{SMD}\left(\hat{W}\right)$.

Under the null hypothesis that the model's overidentifying restrictions hold and standard regularity conditions, the minimized SMD criterion function follows a χ^2 distribution with degrees of freedom equal to the number of overidentifying restrictions (Sargan, 1958; Hansen, 1982):

$$J\left(\hat{m}-\tilde{m}\left(\hat{\theta}_{SMD}\left(\hat{W}\right)\right)\right)'\hat{W}\left(\hat{m}-\tilde{m}\left(\hat{\theta}_{SMD}\left(\hat{W}\right)\right)\right)\sim\chi^2(4).$$

Tables 5 and 6 report overidentification tests based on this J-statistic.

Appendix D: Empirical Bayes Posteriors

We use estimates of the random coefficients model to construct hybrid posterior predictions of school value-added that condition on lottery and OLS estimates. The vector of school-specific parameters is given by:

$$\Theta = \left(\beta_1, b_1, \delta_1, \xi_1, ..., \beta_J, b_J, \delta_J, \xi_J, \nu_1^1, ..., \nu_J^L\right)'.$$

The posterior density of Θ conditional on the observed estimates $\hat{\Omega}$ and the prior parameters θ is

$$f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\theta\right) = \frac{f_{\hat{\Omega}|\Theta}\left(\hat{\Omega}|\Theta\right) f_{\Theta}\left(\Theta;\theta\right)}{f_{\hat{\Omega}}\left(\hat{\Omega};\theta\right)}. \tag{25}$$

Here $f_{\hat{\Omega}|\Theta}(\cdot)$ is the density of observed estimates conditional on the school-specific parameters, $f_{\Theta}(\cdot)$ is the prior density of these parameters and $f_{\hat{\Omega}}(\cdot)$ is the marginal density of the observed estimates. The denominator of (25) does not depend on Θ, so we can write

$$f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\theta\right) \propto \phi_m\left(\hat{\Omega} - \Omega(\Theta); V\right) \phi_m\left(\Theta - \bar{\Theta}(\theta); \Gamma(\theta)\right),$$

where

$$\bar{\Theta}(\theta) = (\beta_0, b_0, \delta_0, \xi_0, ...\beta_0, b_0, \delta_0, \xi_0, 0, ...0)'$$

is the prior mean of Θ, $\phi_m(x; v)$ is the multivariate normal density function with mean zero and covariance matrix v, and

$$\Gamma(\theta) = \begin{bmatrix} I_J \otimes \Sigma & 0 \\ 0 & \sigma_\nu^2 I_{LJ} \end{bmatrix}.$$

Here I_K is the K-dimensional identity matrix. Note that the probability limit of the vector of observed estimates, Ω, is a function of Θ, so we write $\Omega(\Theta)$.

We form an empirical Bayes posterior density by plugging $\hat{\theta}_{SMD}$ into (25). The empirical Bayes posterior mean is

$$\Theta^*_{mean} = \int \Theta f_{\Theta|\hat{\Omega}}\left(\Theta|\hat{\Omega};\hat{\theta}_{SMD}\right) d\Theta.$$

This integral does not have a closed form and it is not possible to sample directly from the posterior distribution. To avoid integration we instead work with the posterior mode:

$$\Theta^*_{mode} = \arg\max_{\Theta} \ \log\phi_m\left(\hat{\Omega} - \Omega(\Theta); V\right) + \log\phi_m\left(\Theta - \bar{\Theta}\left(\hat{\theta}_{SMD}\right); \Gamma\left(\hat{\theta}_{SMD}\right)\right).$$

In Table 7 we compare posterior modes for the β_j with conventional empirical Bayes posterior means based on OLS estimates of value-added. The conventional predictions are given by

$$\alpha^*_{j.} = \hat{\lambda}_j \hat{\alpha}_j + (1 - \hat{\lambda}_j)\hat{\mu}_\alpha, \tag{26}$$

where

$$\hat{\lambda}_j = \frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_\alpha^2 + Var\left(e_j^\alpha\right)},$$

$$\hat{\mu}_\alpha = \frac{1}{J}\sum_j \hat{\alpha}_j,$$

$$\hat{\sigma}_\alpha^2 = \frac{1}{J}\sum_j \left[\left(\hat{\alpha}_j - \hat{\mu}_\alpha\right)^2 - Var\left(e_j^\alpha\right)\right].$$

Models with sector effects replace $\hat{\mu}_\alpha$ in equation (26) with the regression predictions

$$\hat{\mu}_{\alpha j} = W_j' \left[\frac{1}{J}\sum_k W_k W_k'\right]^{-1} \left[\frac{1}{J}\sum_k W_k \hat{\alpha}_k\right],$$

where W_j is a vector including a constant and charter and pilot school indicators.

Table A1: Tests for bias in school value-added models for ELA scores

	Model 1 (1)	Model 2 (2)	Model 3 (3)	Model 3 without charter lotteries (4)
Value-added coefficient	0.358	0.660	0.864	0.423
	(0.087)	(0.130)	(0.167)	(0.310)
First stage *F* -statistic	33.1	27.0	26.8	14.0
p -value: VA coef. equals 1	<0.001	0.009	0.416	0.063
p -value: overid. restrictions	0.011	0.057	0.039	0.157
Joint *p* -value: VA coef. equals 1 and overid. restrictions	<0.001	0.004	0.010	0.050
Bootstrap-refined joint *p* -value	0.050	0.108	0.118	0.120
VAM controls:				
Year effects	X	X	X	X
Demographics		X	X	X
Baseline scores			X	X

Notes: This table reports lottery-based tests for bias in school value-added models. Value-added coefficients come from IV (two-step GMM) regressions of 6th grade ELA scores on estimated ELA value-added at a student's attended school, instrumented by the set of offer dummies for all lotteries in the quasi-experimental sample. Joint p-values come from OLS regressions of value-added residuals on offer dummies. Value-added model 1 includes only year-of-test indicators as controls. Model 2 adds indicators for student sex, race, subsidized lunch, special education, limited-English proficiency, and counts of baseline absences and suspensions. Model 3 adds cubic polynomials in baseline math and ELA scores. Column (4) excludes charter school lotteries from the quasi-experimental sample in testing Model 3. All regressions control for lottery strata fixed effects and the baseline covariates from Model 3. Inference is robust to heteroskedasticity and accounts for first-step estimation error in school value-added. Bootstrap refinements to first-order asymptotics are based on 500 replication samples, with students drawn with replacement.

Table A2: Tests for bias using school mean residuals

	Model 1 (1)	Model 2 (2)	Model 3 (3)	Model 4 (4)	Model 4 without charter lotteries (5)
Value-added coefficient	0.412	0.675	0.889	1.003	0.475
	(0.055)	(0.066)	(0.081)	(0.098)	(0.217)
First stage *F* -statistic	45.1	35.4	29.4	29.1	12.0
p -value: VA coef. equals 1	<0.001	<0.001	0.169	0.979	0.015
p -value: overid. restrictions	<0.001	<0.001	0.002	<0.001	0.026
Joint *p* -value: VA coef. equals 1 and overid. restrictions	<0.001	<0.001	<0.001	<0.001	0.001
Bootstrap-refined joint *p* -value	<0.001	0.012	0.030	0.030	0.044
VAM controls:					
Year effects	X	X	X	X	X
Demographics		X	X	X	X
Baseline scores			X	X	X
School-level averages				X	X

Notes: This table reports lottery-based tests for bias in school value-added models using school mean residuals as estimates of value-added. Value-added coefficients come from IV (two-step GMM) regressions of 6th grade ELA scores on estimated ELA value-added at a student's attended school, instrumented by the set of offer dummies for all lotteries in the quasi-experimental sample. Joint p-values come from OLS regressions of value-added residuals on offer dummies. Value-added model 1 includes only year-of-test indicators as controls. Model 2 adds indicators for student sex, race, subsidized lunch, special education, limited-English proficiency, and counts of baseline absences and suspensions. Model 3 adds cubic polynomials in baseline math and ELA scores, while Model 4 adds school-level average demographics and baseline achievement as controls. Column (5) excludes charter school lotteries from the quasi-experimental sample in testing Model 4. All regressions control for lottery strata fixed effects and the baseline covariates from Model 3. Inference is robust to heteroskedasticity and accounts for first-step estimation error in school value-added. Bootstrap refinements to first-order asymptotics are based on 500 replication samples, with students drawn with replacement.

Table A3: Random coefficient distributions

	β_j	b_j	δ_j	ξ_j
	(1)	(2)	(3)	(4)
Standard deviation	0.166	0.110	0.955	0.927
	(0.059)	(0.047)	(0.103)	(0.246)
Correlation w/β_j	1			
	-			
Correlation w/b_j	-0.439	1		
	(0.450)	-		
Correlation w/δ_j	0.189	0.396	1	
	(0.415)	(0.518)	-	
Correlation w/ξ_j	0.317	-0.564	-0.642	1
	(0.474)	(0.582)	(0.190)	-
Std. dev. of ν_{lj}		1.325		
		(0.155)		

Notes: This table reports simulated minimum distance estimates of parameters governing the distribution of value-added, bias, and first-stage compliance across schools. See notes to Table 5 for a description of the estimation procedure. OLS estimates come from Model 3 and the model includes sector effects.