Assessing Rothstein's critique of teacher value-added models

Josh Kinsler
Department of Economics, University of Rochester

Value-added models of teacher effectiveness yield consistent estimates of teacher quality under the assumption that students are randomly assigned to classrooms conditional on ability. Rothstein (2010) tested and rejected this underlying sorting assumption, casting doubt on the usefulness of the value-added framework. In this paper, I illustrate that the falsification tests employed by Rothstein perform quite poorly in small samples and I propose an alternative testing strategy that is less sensitive to sample size. I also show that none of the proposed falsification tests works well when the achievement production function is misspecified. Finally, I return to the same North Carolina sample employed by Rothstein and retest the assumption of conditional random assignment. Once I account for the "smallness" of the data and allow teacher inputs to persist at reasonable rates, I fail to reject conditional random assignment.

Keywords. Teacher value-added, model testing.

JEL classification. I20, C10.

1. Introduction

Teacher quality is widely recognized as the most significant institutional determinant of academic success. However, because teacher quality cannot be directly observed, schools have largely relied on a set of subjective measures to estimate teacher effectiveness. With the proliferation of standardized testing, an arguably more objective measure of teacher performance is available, student test scores.[1] The predominant methodology for mapping student test scores into estimates of teacher effectiveness is the value-added model (VAM).[2] While there are various flavors of value-added modeling, the basic strategy is to utilize changes in student test scores over time to isolate the value added by each individual teacher.

Utilizing student test scores to evaluate the effectiveness of individual teachers has been roundly criticized, both by scholars and teachers.[3] At the most basic level, critics

Josh Kinsler: jkinsler@rochester.edu

I would like to thank Peter Arcidiacono, Jesse Rothstein, and conference participants at the Duke Young Economists Jamboree for their helpful comments. All remaining errors are my own.

[1] Recent papers such as Harris and Sass (2010) and Jacob and Lefgren (2008) have investigated how well subjective principal evaluations match up with teacher value-added estimates.

[2] Prominent examples in the literature include Rivkin, Hanushek, and Kain (2005), Aaronson, Barrow, and Sander (2007), and Rockoff (2004).

[3] See, for example, Kane and Staiger (2002), Amrein-Beardsley (2008), Harris and Sass (2011), Martineau (2006), Koedel and Betts (2010b), and Koretz (2002).

DOI: 10.3982/QE132

argue that standardized tests simply cannot measure all the knowledge and skills teachers impart to their students. In addition, there are concerns that value-added estimates of teacher quality are unstable over time, and sensitive to test reliability, scaling, timing, and content selection. While these are all valid critiques, perhaps most troubling is the fact that value-added models are typically implemented using large administrative data on students and teachers who are not randomly assigned either within or across schools.[4] Consider a setting in which students are assigned to classrooms based on unobserved factors that are grade relevant, such as ability or home inputs. Estimates of teacher quality will then confound true teacher effectiveness with the distribution of unobserved student factors in the classroom. Thus, even if a perfect test were available, VAMs would yield biased and inconsistent estimates of teacher quality.

In a recent article, Rothstein (2010) (henceforth Rothstein) reviewed the most common VAMs and tests the sorting assumptions that are necessary for these methods to yield causal estimates of teacher quality. Rothstein employed a novel testing strategy that consists of investigating whether future teachers have an impact on today's test score conditional on all current and past observed inputs. If future teachers impact today's scores, this suggests that students are sorted into future classrooms based on variables the econometrician cannot account for and that may be grade relevant. The key finding of the paper is that the basic assumptions underlying the most prominent VAMs fail to hold in a sample of public-school students in North Carolina. As a result, estimates of teacher effectiveness produced by any of these VAMs will be biased, tainting any teacher personnel decisions based in part on student testing results.

However, just as the consistency of the estimated teacher effects produced by VAMs relies on a number of assumptions, so does the accuracy of the falsification tests employed by Rothstein. In particular, the tests are developed under the assumption that arbitrarily large samples are available, when in fact the estimates of each teacher effect are based on just a handful of student test score observations. The rationale for employing tests that rely on large sample theory is that unless the number of students tends to infinity as the number of teachers is held fixed, teacher effectiveness estimates are never consistent, even if students are randomly assigned to teachers. In practice, infinite amounts of data are never available and, as Rothstein pointed out, small sample corrections need to be made to the teacher effect estimates even when the assumptions regarding the asymptotic nature of the data are maintained. In a similar vein, I would argue that the falsification tests proposed by Rothstein also need to account for the small sample nature of the data.

In this paper, I illustrate the small sample bias in a key falsification test proposed by Rothstein and provide a new falsification test that is robust to small samples. The proposed test relies on an estimation technique that allows researchers to easily incorporate multiple high-dimensional fixed effect parameter vectors into their models. This methodology allows the researcher to control for student, teacher, and school effects directly. To test whether future teachers have an impact on today's scores is then a simple

[4] See Clotfelter, Ladd, and Vigdor (2006) and Jackson (2009) for empirical evidence of the sorting of teachers and students.

F-test that requires estimating an unrestricted model, where future teachers are allowed to affect current outcomes, and a restricted model, where the effect of future teachers is constrained to zero. The F statistic can then be computed using the R^2 of the two regressions. I present simulation results that indicate that the proposed F-test significantly outperforms Rothstein's falsification test in small samples for a wide range of data generating processes.

One additional finding from the simulation exercises is that none of the proposed falsification tests performs well when the underlying achievement production function is misspecified. This result turns out to be critical when I reevaluate the assumption of conditional random assignment for the same cohort of North Carolina students and teachers initially examined by Rothstein. Two of the value-added models tested by Rothstein assume that teacher inputs persist indefinitely. When I apply the proposed F-test to either of these models using the North Carolina schooling data, I continue to reject the conditional random assignment assumption. If instead I estimate a model that allows teacher inputs to persist at a more reasonable rate, the F-test fails to reject conditional random assignment. Rothstein's falsification test, however, still rejects conditional random assignment. Thus, the reversal is a function of both the change to the underlying production function and the alternative testing approach.

The remainder of the paper is as follows. Section 2 reviews Rothstein's falsification test and presents an alternative test that is likely to perform better in small samples. Section 3 investigates the performance of each test under various sorting and sample size assumptions, and assesses the robustness of the proposed tests to model misspecification. Section 4 implements the proposed test using the same cohort of North Carolina students utilized by Rothstein, and Section 5 concludes.

2. Testing the validity of VAMs

2.1 *Rothstein's approach*

I begin with a brief review of the econometric model and testing approaches employed by Rothstein. Rothstein assumed that the test score of student i at the end of grade g, A_{ig}, can be written as

$$A_{ig} = \alpha_g + \sum_{h=1}^{g} \beta_{hgc(i,h)} + \mu_i \tau_g + \sum_{h=1}^{g} \varepsilon_{ih} \phi_{hg} + \nu_{ig}, \tag{1}$$

where β_{hgc} is the causal effect of being in classroom c in grade h on the grade g test score, and $c(i,h) \in \{1, \dots, J_h\}$ indexes the classroom to which student i is assigned in grade h. Classroom and teacher effects are synonymous since, in the empirical application, only one cohort of students will be observed with each teacher. The effect of individual ability, μ_i, is allowed to vary across grades according to τ_g, and ε_{ih} captures all other inputs in grade h, such as family or community inputs; ν_{ig} is classical measurement error. Equation (1) is a relatively general form for an achievement production function in that past inputs continue to impact student performance in future grades. This is true for both past teacher inputs and all other inputs, ε. Estimation of Equation (1) is generally not possible since ε_{ih} is not observed. In addition, the large number of teacher and student fixed effects would make estimation quite difficult, even if the unobserved inputs were excluded.

As a result of these econometric difficulties, researchers typically estimate simplified versions of Equation (1). Three of the most common specifications are a simple regression of gain scores on grade and contemporaneous classroom indicators,

$$\text{VAM1:}\quad \Delta A_{ig} = \alpha_g + \beta_{ggc(i,g)} + e_{1ig}, \tag{2}$$

a regression of score levels on classroom indicators and the lagged score,

$$\text{VAM2:}\quad A_{ig} = \alpha_g + A_{ig-1}\lambda + \beta_{ggc(i,g)} + e_{2ig}, \tag{3}$$

and a regression that stacks the gain scores from several grades and adds student fixed effects,

$$\text{VAM3:}\quad \Delta A_{ig} = \alpha_g + \beta_{ggc(i,g)} + \mu_i + e_{3ig}. \tag{4}$$

Each of the above specifications can be derived from Equation (1) by subsuming all the inputs not explicitly accounted for, such as past classroom assignments or unobservable inputs, into e_{kig}.

The primary purpose of the above specifications is to produce consistent estimates of the causal impact of individual teachers. Notice that in each of the above equations, $\beta_{ggc(i,g)}$—the effect of the grade g teacher on the grade g test score—can be estimated directly. However, just as in any standard ordinary least squares (OLS) regression, $\hat{\beta}_{ggc(i,g)}$ will only have a causal interpretation under the assumption that the grade g classroom assignment is uncorrelated with e_{kig}. Rothstein developed a novel strategy for testing this assumption for each of the VAM specifications.

In this paper, I focus on the test for whether VAM3 yields causal estimates of teacher effectiveness. The consistency of the teacher quality estimates produced by VAM3 relies on the assumption that students are sorted into classrooms based on a permanent factor, μ_i, that is fixed over the student's lifetime and is observable to whomever makes the classroom assignment. If this sorting assumption is true, all the components in e_{3ig}—past teachers and past and current unobserved inputs—will be uncorrelated with the classroom assignment conditional on μ_i. The advantage of VAM3 relative to VAM1 is that it allows for nonrandom teacher assignment based on permanent unobserved heterogeneity in student test score growth.

I concentrate on the statistical test for VAM3 for a number of reasons. First, VAM3 has been widely utilized in the literature and thus it is important to be able to accurately evaluate the validity of this benchmark model.[5] Second, permanent heterogeneity in student test score growth will lead to a rejection of VAM1 if students are sorted into classrooms based on this unobserved factor. Thus, implementing the proposed test for VAM3 is the natural next step after a rejection of VAM1. Rothstein viewed the additional

[5] Boardman and Murnane (1979) initially developed VAM3. Recent papers that utilize this framework include Rivkin, Hanushek, and Kain (2005), Harris and Sass (2011), and Jacob and Lefgren (2008).

complexity of VAM3 as unnecessary, since test score growth in third grade is largely uncorrelated with test score growth in fifth grade for the North Carolina data utilized in his analysis. However, the small observed correlation does not eliminate the possibility that permanent heterogeneity in test score growth exists.[6] In addition, student heterogeneity in test score growth may play a larger role in other settings.[7] Finally, Rothstein's falsification test for VAM3 works particularly poorly in small samples. VAM1 and VAM2 also have the potential to perform poorly in small samples, an issue I discuss further in Sections 2.2.1 and 3.2.

To test the strict exogeneity assumption required by VAM3, Rothstein relied on a framework originally developed by Chamberlain (1984). Consider the projection of student ability, μ_i, on the full sequence of classroom assignments in grades 1 through G,

$$\mu_i = \xi_{1c(i,1)} + \cdots + \xi_{Gc(i,G)} + \eta_i,$$

where ξ_{hc} is the incremental information about μ_i provided by the knowledge that a student was in classroom c in grade h, conditional on all the other classroom assignments. Assuming that teacher effects do not decay over time, substituting the above expression into the first-differenced production function yields

$$\Delta A_{ig} = \Delta\alpha_g + \sum_{h=1}^{G} \pi_{hgc(i,h)} + \eta_i + e_{3ig}, \tag{5}$$

where $\pi_{ggc} = \xi_{gc}\Delta\tau_g + \beta_{ggc}$ and $\pi_{hgc} = \xi_{hc}\Delta\tau_g$ for $h \neq g$. While the notation can be a bit cumbersome, the key to the equation is that under the assumption of strict exogeneity, the effect of any teacher from grade $h \neq g$ on the grade g test score should be proportional to the incremental information that the grade h teacher provides regarding μ_i. As an example, if $\Delta\tau_g$ is equal to a constant for all g, the effect of the fifth grade teacher on a student's third and fourth grade score should be identical, since it provides the same information in either regression. If the effect of the fifth grade teacher varies across the third and fourth grade scores, it suggests that students are sorted into fifth grade based on unobserved factors that affect the third and fourth grade scores differently. This is a violation of conditional strict exogeneity, since students should be sorted into classrooms based on an unobserved fixed factor.

To test whether the restrictions imposed by strict exogeneity hold, Rothstein utilized a minimum chi-square estimator, also known as an optimal minimum distance estima-

[6] It is quite simple to generate test score data that incoporate minimal permanent heterogeneity in test score growth such that (i) there is essentially no correlation in test score growth over time and (ii) the proposed test for VAM1 fails as a result of sorting on the unobserved heterogeneity.

[7] There exists considerable debate about whether standardized tests are well designed to evaluate test score growth. See Martineau (2006) as an example. Thus, as test instruments improve, heterogeneity in growth may become more prevalent. In addition, the recent literature on the development of cognitive and noncognitive skills suggests that there are important dynamic complementarities in the production function. See Cunha, Heckman, and Schennach (2010) as an example. This is also consistent with permanent heterogeneity in skill growth.

tor. The test statistic is calculated by estimating the vector Ξ_h and the ratio $r = \frac{\Delta\tau_g}{\Delta\tau_{g-1}}$ that minimize

$$D = \left(\begin{pmatrix}\hat{\Pi}_{hg-1}\\ \hat{\Pi}_{hg}\end{pmatrix} - \begin{pmatrix}\Xi_h\\ \Xi_h * r\end{pmatrix}\right)' \hat{W}^{-1}\left(\begin{pmatrix}\hat{\Pi}_{hg-1}\\ \hat{\Pi}_{hg}\end{pmatrix} - \begin{pmatrix}\Xi_h\\ \Xi_h * r\end{pmatrix}\right),$$

where $\hat{\Pi}_{hg}$ and $\hat{\Pi}_{hg-1}$ are the stacked OLS estimates of the π_{hg} and π_{hg-1} from Equation (5), and $\hat{W}$ is the estimated sampling variance of $(\hat{\Pi}'_{hg-1}\hat{\Pi}'_{hg})'$. Under the null hypothesis of strict exogeneity, the minimized value of D is distributed χ^2 with $J_h - 1$ degrees of freedom, where J_h is the number of grade h teacher effects estimated in Equation (5). Intuitively, D is a weighted measure of the distance between the effect of the grade $h = g + 1$ teacher on student test scores in grades g and $g - 1$. Under the assumptions of conditional strict exogeneity, this difference should be close to zero.

Once $\hat{\Pi}_{hg-1}$ and $\hat{\Pi}_{hg}$ are estimated, computing the test statistic is rather straightforward. However, directly estimating Equation (5) by OLS is infeasible since it contains multiple high-dimensional fixed effect vectors. In addition, to compute D, Equation (5) must be estimated for multiple grades to enable comparison of the effects of grade h teachers on the test scores in the two previous grades. Rothstein's empirical strategy, which I follow in Section 3, is to estimate Equation (5) school-by-school, thus avoiding the need to invert a matrix that will contain thousands of columns, one for each teacher effect to be estimated.

The limitation of this empirical approach is that for the strategy to be valid, students must have remained in the same school from grade $g - 3$ to g. This sample restriction significantly reduces the number of observations per teacher, generating two problems for the proposed test. First, with few observations per teacher, it is more likely that the estimated teacher effects will be biased. In other words, even if students are randomly assigned to fifth grade classrooms, with so few test score observations per teacher, it is unlikely that the conditional mean of the unobserved test score shocks in grades 3 and 4 will equal zero across the fifth grade classrooms. Second, it is well known that the optimal minimum distance estimator performs quite poorly in small samples, potentially exacerbating the first issue.[8] Thus, the validity of the proposed test for VAM3 may be compromised by the nature of the available data. The Monte Carlo exercises in Section 3 illustrate that this is indeed a serious problem for a sample similar in magnitude to the one utilized by Rothstein.

2.2 *Alternative falsification tests for VAM3*

2.2.1 *An extension of Rothstein's approach* Using the minimum chi-square estimator to test for conditional strict exogeneity is indirect in the sense that future teachers will have an effect on current scores only if students are sorted on unobserved ability and unobserved ability is not accounted for in the achievement regression. Conditional strict

[8] For evidence regarding the poor small sample performance of the optimal minimum distance estimator, see Altonji and Segal (1996), Burnside and Eichenbaum (1996), or Horowitz (1998).

exogeneity then implies that the estimated effect of a fifth grade teacher should be identical in all previous grades. A more direct test for conditional strict exogeneity would instead estimate student achievement gains, controlling for student ability in addition to all past, current, and future teachers, and then examine whether future teachers have any explanatory power.[9] The novel testing strategy proposed by Rothstein of exploring how future teachers impact current performance remains at the heart of this alternative testing procedure.

One way to implement the direct test is simply to extend the school-by-school regressions in the first step of Rothstein's proposed test to include unobserved student ability. These regressions remain tractable since there are a limited number of fixed effects to estimate within each school. The school level regressions then take the form

$$\Delta A_{ig} = \mu_i + \sum_{h=g}^{g+1} \beta_{hgc(i,h)} + e_{ig}, \tag{6}$$

where I assume that the current teacher and the one period ahead teacher enter the gain score equation.[10] I follow the literature and assume that lagged teachers do not enter the gain score equation, though in practice, this assumption can be relaxed. The test for conditional strict exogeneity then boils down to a test of whether the future teacher effects are jointly equal to zero. However, constructing the appropriate test statistic for the null hypothesis that future teachers have no effect is not trivial, since the test statistic must combine results across numerous independent regressions. Ideally, one would estimate a single regression with all the student and teacher effects included, and construct a Wald statistic using a cluster-robust estimate of the asymptotic variance of the estimated teacher effects. Recall that since we are examining changes in test scores, the errors are likely to be correlated at the student level. Any shock to test scores in third grade will affect not only the gain in third grade, but also the gain in fourth grade.

The Wald statistic for the regression combining data across all schools is given by

$$W = (R\hat{\beta})'[R\hat{\mathrm{Var}}[\hat{\beta}]R']^{-1}(R\hat{\beta}),$$

where $\hat{\beta}$ is a $k \times 1$ vector of student and teacher effect estimates, and R is a $q \times k$ matrix such that $R\hat{\beta}$ yields a vector comprised only of the q estimated future teacher effects. Piecemeal estimation of the model does not affect our ability to construct the outer matrices, since they are essentially just the stacked vectors of the school-by-school estimates of the future teacher effects. The challenge is in generating an estimate of the inner matrix, the inverse of the asymptotic variance of the future teacher effects.

[9] Rather than proposing a new testing strategy, there are methods to correct for the small sample bias in the minimum chi-square estimator. The most common approach is to bootstrap the critical value of the test following Hall and Horowitz (1996). However, bootstrapping is not feasible in this setting, since estimating the test statistic D is computationally intensive.

[10] Note that adding a school effect here is innocuous, since it requires normalizing one of the student effects. Consistent with the standard VAM3 model, I assume that the coefficient on unobserved student heterogeneity is 1 for all grades.

For a sample containing J students and a total of n test score gain observations, the cluster-robust variance estimator, adjusted for the number of clusters and sample size, is given by

$$\hat{\text{Var}}[\hat{\beta}] = \left(\frac{J}{J-1} \times \frac{n}{n-k}\right)\left[(X'X)^{-1} \times \sum_{j=1}^{J}(u_j' \times u_j) \times (X'X)^{-1}\right],$$

where $u_j = \sum_g e_g * x_g$, e_g are the residuals for student j across multiple grades g, and X is an $n \times k$ matrix of dummy variables for all of the student and teacher effects. The difficulty in generating both the asymptotic covariance matrix and its inverse is the sheer size of X, which will necessarily contain thousands of columns, one for each teacher and student in the sample. However, limiting the sample to those students who remain in the same school makes it possible to generate an estimate of the asymptotic variance and its inverse rather simply.

Because each school is self-contained, the estimated teacher and student effects are independent across schools, implying that the asymptotic variance matrix will be block diagonal, with each block representing a separate school. Thus, I can construct an estimate of the overall asymptotic variance by combining each school-specific asymptotic variance matrix.[11] The inverse of the components related to the future teacher effects can be constructed in a similar vein, since the inverse of a block diagonal matrix is also a block diagonal matrix. Note that when constructing the asymptotic variance matrix at the school level, it is necessary to adjust using the total number of clusters (J) and the total number of observations and parameters (n and k) across all schools.

While the above approach avoids utilization of the optimal minimum distance estimator, there are still concerns regarding construction of the cluster-robust Wald statistic when the effective sample size is small. There exists ample Monte Carlo evidence to suggest that robust estimators of the variance matrix perform poorly in small samples.[12] In the data Rothstein utilized for estimation, teachers have on average only 11 associated test score observations. Thus, there is again concern that the tests based on the cluster-robust Wald statistic will be poorly sized. Utilizing a Wald statistic based on the assumption of homoskedasticity is also likely to yield poorly sized tests, since the test score gain residuals are likely to be correlated at the student level.

Although the focus of this paper is on VAM3, it is worth pointing out that the inaccuracy of robust variance estimators when working with small samples will also impact Rothstein's proposed tests for VAM1 and VAM2. Take, for example, the falsification test for VAM1, which is built on the same idea as the test for VAM3. Conditional on the fourth grade teacher assignment, the fifth grade teacher should have no impact on the fourth grade gain. Implementing the test is straightforward, since VAM1 assumes that there is no unobserved student heterogeneity. Simply regress fourth grade gains on indicators for the fourth grade teacher and the future fifth grade teacher, and test whether the fifth

[11] This is quite similar to methodology employed by Rothstein when generating the variance–covariance matrix for the estimates of the fifth grade teachers on the third and fourth grade score. See Appendix B.3 in Rothstein (2010) for further details.

[12] See MacKinnon and White (1985) and Kezdi (2004) as examples.

grade teacher effects are jointly equal to zero. The question is, which test statistic is appropriate? With small samples, there is a concern that a heteroskedastic robust test will perform poorly.[13]

2.2.2 *Examining test score levels* As Section 3 will show, the test for conditional strict exogeneity based on the cluster-robust Wald statistic significantly outperforms the test originally proposed by Rothstein. However, for sample sizes likely to be encountered in practice, the test remains undersized, likely a result of the inaccuracy in the cluster-robust estimate of the variance matrix. In this section, I propose a test that has the potential to perform well regardless of sample size by eliminating the most obvious source of heteroskedasticity: the correlation in gain score residuals at the student level.

Consider the gain score equation of VAM3, as illustrated in Equation (6). Underlying this model of test score gains is the equation for test score levels,

$$A_{ig} = \tau_g \mu_i + \sum_{h=1}^{g} \beta_{hgc(i,h)} + \epsilon_{ig}, \tag{7}$$

where I exclude the potential effect of future teachers. Notice that ϵ_{ig} now only appears in the grade g outcome, eliminating the most obvious source of clustering from the model. As a result, it may now be possible to construct a falsification test for VAM3 that relies on the assumption of homoskedasticity, avoiding the need to estimate a robust variance matrix.[14]

However, it is not obvious how to estimate the parameters of Equation (7) or the corresponding homoskedastic variance matrix. The school-by-school approach is no longer applicable, since there now exist parameters that are common across schools, τ_g.[15] A direct approach is also infeasible, since there are an enormous number of fixed effects to estimate. Arcidiacono, Foster, Goodpaster, and Kinsler (2012) demonstrated that there is a rather simple estimation technique that can easily deal with the computational complexity inherent in Equation (7).[16] The basic idea is to estimate the

[13] In the online Appendix to his paper, Rothstein performed a sequence of Monte Carlo experiments that show that VAM1 and VAM2 are in fact missized in small samples. While the bias in his Monte Carlo exercises is mild, I find that it can be quite large if the data generating process is altered slightly. I discuss the performance of VAM1 and VAM2 further in Section 3.2.

[14] Of course, it is still possible that the variance of ϵ_{ig} varies in the population and/or that ϵ_{ig} and ϵ_{ig-1} are correlated conditional on the student fixed effects. I investigate the sensitivity of the proposed test in the next section.

[15] The grade-specific coefficients on student ability are not all identified. However, for the proposed levels model to be consistent with a growth model that contains student fixed effects, the impact of student ability must vary across grades. In practice, I capture this by normalizing $\tau_2 = 1$ and setting τ_g for $g > 2$ such that $\tau_g - \tau_{g-1} = \gamma - 1$, where γ is estimated within the model. This essentially assumes that student ability has a constant effect on test score growth. This is the standard assumption in growth models that allow for unobserved student heterogeneity.

[16] Similar estimation techniques have been proposed in Abowd, Creecy, and Kramarz (2002), Guimaraes and Portugal (2010), and Kramarz, Machin, and Ouazad (2008). See Smyth (1996) for a review of common iterative estimation methods.

model iteratively, where each step estimates a portion of the parameters, conditional on the current values of all the other parameters.[17]

To illustrate the estimation method, consider the achievement level formulation given by Equation (7). Estimation would start with an initial guess of the parameters, μ_i^0, $\beta_{ggc(i,g)}^0$, and τ_g^0, with the qth iteration consisting of the following steps:

Step 1. Update μ_i^q conditional on $\beta_{ggc(i,g)}^{q-1}$ and τ_g^{q-1}.

Step 2. Update $\beta_{ggc(i,g)}^q$ conditional on μ_i^q and τ_g^{q-1}.

Step 3. Estimate τ_g^q conditional on μ_i^q and $\beta_{ggc(i,g)}^q$.

The updating equations for each vector of parameters are derived from the first order conditions of the least squares problem. Precise updating equations are included in Appendix A for the case where test scores are available for grades 2–5. At each step of the estimation procedure, the sum of the squared residuals is minimized conditional on the other parameter values. Iteration continues until the parameters converge, at which point the parameters will be identical to those that would be obtained by minimizing the least squares problem over the entire parameter space in one step.

Testing for strict exogeneity using the above technique is rather straightforward. First, estimate Equation (7) using the iterative methodology, restricting the impact of future teachers to zero. Then estimate an unrestricted version of Equation (7) that allows future teacher effects to enter the achievement equation. The unrestricted model is given by

$$A_{ig} = \tau_g \mu_i + \sum_{h=1}^{g+1} \beta_{hgc(i,h)} + \epsilon_{ig}. \tag{8}$$

The iterative estimation methodology remains the same, except that additional steps need to be added so as to update the effect of the grade $g+1$ teacher on the grade g score. Finally, perform a simple test using the F statistic given by

$$F = \frac{(R_u^2 - R_r^2)(n-k)}{(1-R_u^2)J}, \tag{9}$$

where R_u^2 and R_r^2 denote the fits of the unrestricted and restricted model, respectively, $n-k$ is the model degrees of freedom, and J is the number of restrictions.[18]

In addition to the potential for better performance in small samples, there are a number of advantages to the testing approach utilizing the iterative estimation strategy. First, the estimation sample does not need to be limited to students who remain in

[17] Recently, a number of canned STATA routines have been developed for estimating linear models that contain two high-dimensional vectors of fixed effects (see McCaffrey, Lockwood, Mihaly, and Sass (2010)). I do not pursue these methods here, since in some of the models to follow, there will be as many as four high-dimensional fixed effect vectors, as well as interactions between the fixed effects and other parameters of interest.

[18] For the proposed F-test to work well, ϵ_{ig} must be both homoskedastic and normally distributed. The distribution of standardized student test scores in North Carolina is well approximated by a normal distribution.

the same school. If school switching is prominent and these students tend to be sorted differently, then excluding them can potentially bias the results of the test.[19] Second, it is much simpler to incorporate additional regressors to capture other observed student or classroom factors that are changing over time. For example, a student's free lunch status might change, affecting both academic performance and sorting across classrooms. If the effect of these observed attributes is common across schools as is typically assumed, then estimating school-by-school is not possible.

3. Monte Carlo simulations of VAM3 falsification tests

Section 2 presents three alternative falsification tests for VAM3, a model that is only valid under the assumption that students are randomly assigned to classrooms conditional on unobserved student ability. The first test, originally proposed by Rothstein, utilizes an optimal minimum distance estimator to test the restriction that future teachers have identical effects across earlier grades when unobserved student ability is excluded. The second test utilizes a cluster-robust Wald statistic to test the restriction that future teachers have no effect conditional on unobserved student ability. There is a concern that both tests will perform poorly when teachers are observed with only a handful of students, a feature that is common when evaluating teacher effectiveness using student test scores. The final statistical test considered will likely perform better in small samples; however, it requires additional assumptions. The goal of this section is to investigate how well Rothstein's test and the alternative tests proposed in the previous section detect violations of conditional strict exogeneity under various sample sizes and sorting assumptions.

Prior to discussing how sample size and selection vary across the simulations, it is useful to consider what is common across the various Monte Carlo exercises. For each simulation, I assume that student test scores and classroom assignments are available from second grade through fifth grade. In keeping with the sample restrictions necessary to estimate the model, students remain in the same school across the four observations. Student test scores are determined according to the simple production function

$$A_{igs} = \tau_g \mu_i + \sum_{h=2}^{g} \beta_{hgc(i,h)} + \rho_g \kappa_s + \epsilon_{ig} \quad \text{for } g = 2, 3, 4, 5, \tag{10}$$

where κ_s is the school-specific contribution to test scores, which is allowed to vary by grade.[20] Because I will eventually estimate gains in student test scores, I make a number of simplifying assumptions. First, I restrict $\beta_{hgc(i,h)} = \beta_{hhc(i,h)}$, requiring the effect of being in classroom c in grade h on the grade g test score to equal the effect of being in classroom c in grade h on the grade h test score. In other words, teacher effects persist

[19] Allowing for switching across schools complicates identification, since it is no longer necessary to normalize one teacher effect per school grade. Thus, calculating the degrees of freedom and the number of restrictions can be cumbersome.

[20] I include a school effect since Rothstein included one in his original model.

with no decay.[21] Second, I restrict $\tau_{g+1} = \tau_g + 1$ and $\rho_{g+1} = \rho_g + 1$. First-differencing the production function outlined in Equation (10), imposing the aforementioned restrictions, yields

$$\Delta A_{igs} = \mu_i + \beta_{ggc(i,g)} + \kappa_s + \epsilon_{ig} - \epsilon_{ig-1} \quad \text{for } g = 3, 4, 5. \tag{11}$$

While the restrictions on τ and $\beta_{hgc(i,h)}$ make data generation and estimation slightly easier, the production function illustrated in Equation (11) is nested within the original production function hypothesized by Rothstein. Thus, the proposed tests for strict exogeneity should not be affected by these restrictions.

I choose the standard deviation of each of the model components in an effort to yield results that mimic those obtained when using the estimation sample from North Carolina. In particular, I want the estimated standard deviation of teacher quality, the fit of the achievement regressions, and the correlations in test score gains from the Monte Carlo experiments to be similar in magnitude to those found in the data. In that vein, I assume that

$$\begin{aligned} &\mu_i \sim N(0, 0.15^2), \\ &\beta_{ggc(i,g)} \sim N(0, 0.15^2), \\ &\kappa_s \sim N(0, 0.25^2), \end{aligned}$$

and that $\epsilon_{ig} \sim N(0, 0.5^2)$ and $E(\epsilon_{ig}\epsilon_{ig'}) = 0$ for $g' \neq g$.[22] I examine the robustness of the results to these assumptions and the assumptions regarding the sorting of students to classrooms (discussed below) at the end of Section 3 and in Appendix B.

To close the model, I need to specify how students are assigned to classrooms, since the sorting method will determine whether the strict exogeneity test should reject or fail to reject the null hypothesis. I consider two assignment rules: sorting on permanent ability (μ_i), and sorting on ability and the lagged test score shock (μ_i and ϵ_{ig-1}). In each case, the sorting is not perfect, as there is also a random sorting shock in each period. This ensures that there will be significant classroom switching across grades, a condition necessary for identification. The formulas for the two sorting scenarios are, respectively,

$$\text{rank}_{ig} = 0.6\mu_i + 0.15 * N(0, 1), \tag{12}$$

$$\text{rank}_{ig} = 0.3\mu_i + 0.3\epsilon_{ig-1} + 0.5 * N(0, 1). \tag{13}$$

Students are then assigned to classrooms within schools according to their rank, with the first 20 students slotted in the first class and so on. For the case where students are sorted based on their lagged residuals, I choose the coefficient on the lagged residual such that

[21] It is possible to relax this restriction when using Rothstein's original testing approach or the approach that utilizes level equations. See Kinsler (2012) for an example of how to incorporate varying degrees of teacher decay in the levels framework.

[22] Under this data generating process, the estimated adjusted standard deviation of the grade g teacher on the grade g gain is 0.22, the adjusted R-square is 0.16, and the 1- and 2-year correlations in test score gains are -0.30 and 0.09. The corresponding values in the North Carolina data are approximately 0.20, 0.13, -0.35, and 0.02.

the estimated standard deviations of the future teacher effects are similar in magnitude to what is estimated in the actual data. For all simulations, students and teachers are randomly assigned to schools, and teachers are randomly assigned to classrooms.[23] The first scenario, sorting students by permanent ability, clearly satisfies the strict exogeneity assumption. However, the second sorting scenario fails the strict exogeneity assumption since the fifth grade teacher assignment depends directly on the unobserved achievement shock in the previous grade.

I simulate student achievement outcomes and teacher assignments for each sorting scenario described above using four different sample sizes. In all four settings, I assume that there are 300 schools and 3 teachers per grade within each school. The crucial parameter is then the number of student observations per teacher, which will obviously also determine the overall sample size. I set the number of observations per teacher equal to 10, 20, 50, and 100 to illustrate how the test performs when the sample is similar in size to (i) the actual sample utilized by Rothstein, (ii) a single student cohort under ideal conditions, (iii) multiple student cohorts, and (iv) a sample for which the minimum chi-square estimator should perform as expected. Class sizes typically range between 20 and 30 students, thus under ideal conditions, meaning few missing test scores and school switches, a researcher might have 20 observations per teacher on average. If a researcher has access to test scores and classroom assignments across multiple cohorts, the average number of student observations per teacher could potentially climb close to 50.

3.1 *Rothstein's falsification test*

For each simulation, I estimate a version of Equation (11) that replaces the student fixed effects with teacher assignments for grades 3, 4, and 5. I estimate the gain score equations for grades 3 and 4 separately, and then calculate D, the statistic for testing whether strict exogeneity holds. I then compare D to the 95th percentile of the appropriate chi-square distribution, the critical value for a test with a significance level of 5%.[24] Table 1 reports the adjusted standard deviation of the teacher effects, fit of the regression, and rejection probability averaged across simulations. The table is split into two panels, reflecting the different sorting mechanisms.

As the first panel illustrates, the test for strict exogeneity based on the minimum chi-square estimator performs extremely poorly when the null hypothesis is in fact true. With only 10 observations per teacher, the null hypothesis is rejected in 96% of the simulations. Increasing the sample size to 20 observations per teacher improves performance significantly; however, the null hypothesis is still rejected in almost half of the simulations. At a sample size of 50 students per teacher, the size of the test approaches its asymptotic value. Finally, with 100 student observations per teacher, the rejection probability is the expected 5%. The top panel of Table 8 below contains Rothstein's results of

[23] The results are in no way sensitive to these assumptions.

[24] The degrees of freedom is 599 for all of the sample sizes since the number of teachers does not change across samples. For each Monte Carlo experiment, I estimate 600 future teacher effects and one parameter that pins down the relative importance of ability across grades 3 and 4.

Table 1. Testing for strict exogeneity using the minimum distance estimator.[a]

Obs. per Teacher	Rejection Probability	Adj. SD($\hat{\Pi}_{53}$)	Adj. SD($\hat{\Pi}_{54}$)	Adj. R^2 ΔA_{i3}	Adj. R^2 ΔA_{i4}
Sorting on Time Invariant Ability ONLY					
10	0.97	0.16	0.16	0.15	0.15
20	0.49	0.11	0.11	0.16	0.17
50	0.11	0.08	0.08	0.15	0.16
100	0.05	0.07	0.07	0.16	0.16
Sorting on Ability and Lagged Residual					
10	0.97	0.14	0.22	0.19	0.19
20	0.47	0.10	0.19	0.20	0.20
50	0.14	0.06	0.18	0.19	0.19
100	0.05	0.04	0.18	0.20	0.20

[a] Reported results are averages over 250 simulations for each sample size and sorting scenario. Each simulation contains 300 schools and 900 fifth grade teachers. Test scores are generated according to Equation (10), and the underlying distributions of teacher quality and student ability are also provided in the text. For each simulation, students are assigned to classrooms based on either their fixed ability or their fixed ability and their lagged test score residual. Only in the latter case should the test for conditional strict exogeneity fail. The rejection probability is the proportion of the 250 simulations for which conditional strict exogeneity is rejected using the minimum chi-square estimator as outlined in the text.

the falsification test for VAM3 using a cohort of North Carolina public school students. Note that the model fit and adjusted standard deviations (SD) are quite similar in magnitude to those reported in Table 1.

The poor performance of the minimum chi-square estimator in the first panel of Table 1 likely stems from two sources. The first is that the estimates of the teacher effects themselves, $\hat{\Pi}_{hg}$ and $\hat{\Pi}_{hg-1}$, are inconsistent and likely extremely noisy, making it difficult to compare them with one another. The weighting matrix attempts to account for this by weighting the estimates according to their heteroskedasticity-robust sampling variance. However, this weighting scheme further exacerbates the problem, since it has also been documented that the heteroskedasticity-robust variance matrix estimator of White (1980) performs poorly in small samples.[25] In fact, if I calculate $\hat{W}$ assuming that the errors in the third and fourth grade gain score equations are homoskedastic, then the rejection probabilities in the top panel decrease to 0.43, 0.15, 0.07, and 0.04, respectively.

Given the high rejection probability when the null hypothesis is in fact true, one would expect the test to perform well when the null hypothesis is actually false. However, the second panel of Table 1 illustrates that this is generally not the case. For small samples (10 observations per student), the test performs well, rejecting the null when it is actually false in 97% of the simulations. When the sample size increases, the performance of the test actually deteriorates.

The intuition for deterioration in performance is that r, the ratio of the importance of individual ability in grade g relative to grade $g-1$, $\frac{\Delta\tau_g}{\Delta\tau_{g-1}}$, adjusts to improbable values, making it appear as if the restrictions implied by strict exogeneity hold. When students are assigned to their classroom based on ability and the previous year's transitory

[25] Following Davidson and MacKinnon (1993), I adjust the estimated robust variance matrix by $\frac{n}{n-k}$ to minimize the small sample biases.

test score shock, the effect of the fifth grade teacher on the third grade score will be very small since the fifth grade classroom assignment contains little information about individual ability and no information about the transitory performance shock in third grade. However, the effect of the fifth grade teacher on the fourth grade score will be very large, since the fifth grade classroom assignment contains information about both ability and the fourth grade transitory shock. Given that the variance of the transitory shocks is quite high, $\hat{\Pi}_{54}$ will dwarf $\hat{\Pi}_{53}$.

The minimum chi-square estimator will account for this pattern of coefficients by setting r to unrealistically high numbers, potentially on the order of 25 or more. This would imply that student ability is 25 times more important in explaining ΔA_{i4} relative to ΔA_{i3}. When this occurs, it makes it appear as if the conditions necessary for strict exogeneity hold, since the π_{53} are all close to zero and can be easily scaled to make them appear similar to the π_{54}. The problem can be ameliorated by restricting $\frac{\Delta\tau_g}{\Delta\tau_{g-1}}$ to take on reasonable values. For example, if I restrict $\frac{\Delta\tau_g}{\Delta\tau_{g-1}} = 1$, I reject the null hypothesis of strict exogeneity with probability 1 when students are sorted based on their lagged test score shock. However, overrejection will remain an issue in small samples regardless of the restrictions imposed on r.

3.2 *Extending Rothstein's falsification test*

The concern that the test for conditional strict exogeneity based on the minimum distance estimator would perform poorly in small samples appears to be valid. However, by estimating unobserved student heterogeneity in test score gains directly, the minimum distance estimator can be completely avoided. In this case, I estimate Equation (11) school-by-school, adding an additional set of teacher effects for the one grade ahead teacher. I then construct the cluster-robust Wald statistic associated with the restriction that the future teacher effects are jointly equal to zero. Finally, I compare the test statistic to the 95th percentile of the appropriate chi-square distribution.[26]

The results of the testing procedure, along with the adjusted standard deviation of the future teacher effects, are shown in Table 2. In addition to the cluster-robust Wald statistic, I also tested the restriction that the future teacher effects are zero using a homoskedastic Wald statistic. When the null hypothesis is in fact true, both tests fail to reject in all simulations. If the test were appropriately sized, it should reject approximately 5% of the time. When the null hypothesis is in fact false, both tests significantly underreject in small samples. The bias in the test using the cluster-robust Wald statistic is the result of biases in the estimated variance matrix resulting from the small sample size. The test based on the homoskedastic Wald statistic is also biased since the homoskedastic variance matrix does not account for the correlation in residuals within students.[27]

[26] The number of restrictions in this case is 1200. There are 300 schools and 4 estimated future teacher effects in each school. Note that one of the future teachers in third and fourth grade within each school is not identified.

[27] It is possible to generate test score gains where there is no within-student correlation in residuals. Simply replace the independent and identically distributed (i.i.d.) test score level shocks with i.i.d. test score growth shocks. The test based on the homoskedastic Wald statistic is appropriately sized in this case. Note

Table 2. Testing for strict exogeneity using the cluster-robust Wald statistic.[a]

Obs. per Teacher	Rejection Prob. Cluster Robust	Rejection Prob. Homoskedastic	Adj. SD of Future Teacher Effects	R^2
Sorting on Time Invariant Ability ONLY				
10	0	0	0.07	0.40
20	0	0	0.03	0.35
50	0	0	0.02	0.30
100	0	0	0.01	0.28
Sorting on Ability and Lagged Residual				
10	0.53	0.71	0.17	0.44
20	0.46	1	0.16	0.37
50	1	1	0.16	0.34
100	1	1	0.16	0.33

[a] Reported results are averages over 250 simulations for each sample size and sorting scenario. Each simulation contains 300 schools and 900 fifth grade teachers. Test scores are generated according to Equation (10), and the underlying distributions of teacher quality and student ability are also provided in the text. For each simulation, students are assigned to classrooms based on either their fixed ability or their fixed ability and their lagged test score residual. Only in the latter case should the test for conditional strict exogeneity fail. The two rejection probabilities are the proportion of the 250 simulations for which conditional strict exogeneity is rejected using a cluster-robust and homoskedastic Wald statistic as outlined in the text.

As noted earlier, the inaccuracies in the cluster-robust variance estimator are also likely to impact the small sample performance of Rothstein's falsification tests for VAM1 and VAM2. In fact, Rothstein showed in an online appendix that the tests for VAM1 and VAM2 over-reject slightly when teachers are observed with 20 students.[28] However, the degree to which the tests fail is quite sensitive to the choices made regarding the data generating process. Take, for example, the test for VAM1. If I simply tweak the data generating process discussed at the beginning of Section 3 such that there is no student heterogeneity in test score growth and implement Rothstein's robust score test for VAM1 that includes controls for the current teacher, I reject the null hypothesis when it is true 25% of the time. Note that if I use the more standard heteroskedasticity-robust Wald test, I reject close to 50% of the time.[29] Thus, the choice of test statistic is crucial for making correct inference when dealing with small effective samples. This is true when testing VAM1, VAM2, or VAM3.

3.3 *Falsification test using the underlying levels equation*

Extending Rothstein's original test to avoid using the minimum distance estimator yields an improvement in detecting conditional strict exogeneity, particularly with few observations per teacher. However, the test is still not appropriately sized when the number of observations per teacher is small. In this section, I examine how a test built off the underlying levels equation performs.

that the correlation in fourth and fifth grade test score gains in Rothstein's North Carolina data is −0.41, suggesting that i.i.d. test score level shocks are a more accurate depiction of the true underlying production function.

[28] The tests reject the null 8–9% of the time instead of the expected 5%.

[29] If I use a homoskedastic Wald test, I reject approximately 5% of the time.

Rather than attempt to estimate the gain score model, I instead estimate the underlying levels model given by Equation (10). However, estimating Equation (10) as it is written is not possible without further restrictions. The grade-specific coefficients on student ability and school inputs (τ_g and ρ_g) are not identified. However, to be consistent with a growth model that contains both student and school fixed effects, the impact of student and school heterogeneity must vary across grades. To capture this, I normalize $\tau_2 = 1$ and set τ_g for $g > 2$ such that $\tau_g - \tau_{g-1} = \gamma - 1$, where γ is estimated within the model.[30] This essentially assumes that student ability has a constant effect on test score growth and is identical to the implicit assumption in standard test score growth models that incorporate unobserved student heterogeneity. I restrict ρ_g similarly and, for ease of estimation, also assume that the heterogeneity in the school effect across grades follows the same pattern as that of student ability.[31] In implementation, I further assume that the second grade test score is only a function of student ability and the test score shock.[32] This is consistent with the North Carolina data since the second grade score is actually a pre-test taken at the beginning of third grade. It is possible to include school and teacher components for the first test score observation; however, additional normalizations would be required.

The results of the simulations are illustrated in Table 3. The test of conditional strict exogeneity based on the simple F statistic significantly outperforms both of the alternative tests.[33] In the levels model, the null hypothesis that future teacher assignments

TABLE 3. Testing for strict exogeneity using the underlying levels model.[a]

Obs. per Teacher	Rejection Probability	R_u^2	R_r^2
Sorting on Time Invariant Ability ONLY			
10	0.06	0.764	0.751
20	0.04	0.735	0.728
50	0.06	0.733	0.731
Sorting on Ability and Lagged Residual			
10	1	0.774	0.759
20	1	0.745	0.735
50	1	0.742	0.736

[a] Reported results are averages over 250 simulations for each sample size and sorting scenario. Each simulation contains 300 schools and 900 fifth grade teachers. Test scores are generated according to Equation (10) with $\gamma = 2$. For each simulation, students are assigned to classrooms based on either their fixed ability or their fixed ability and their lagged test score residual. Only in the latter case should the test for conditional strict exogeneity fail. The rejection probability is the proportion of the 250 simulations for which conditional strict exogeneity is rejected using the F-test as outlined in the text.

[30] In this case, γ is equal to 2 since I set $\tau_2 = 1$ and $\tau_{g+1} = \tau_g + 1$ for $g > 2$. Note that the structure of the τ_g's is the same across all testing permutations.

[31] This restriction can be relaxed in practice.

[32] The performance of the other tests is unaffected by this small change in the data generating process. Results are available on request.

[33] Appendix B contains additional comparisons of the three tests for conditional strict exogeneity. These additional comparisons allow for varying levels of dynamic sorting and alter the relative importance of student ability, teacher ability, and the idiosyncratic shock for test scores. The results in Appendix B are consistent with results obtained under the baseline data generating process.

contain no grade relevant information is rejected approximately 5% of the time, as expected. When students are sorted into classrooms based on their lagged residual, the test based on the F statistic constructed from the levels model rejects the null hypothesis 100% of the time.

The fact that the proposed F-test works well when compared with the two other approaches is not surprising, since in the data generating process I have assumed that the test score residuals are both homoskedastic and normally distributed. In addition, the tests based on the cluster-robust Wald statistic and the F statistic make additional assumptions regarding the underlying data generating process as compared to Rothstein's original test for VAM3. In particular, both tests assume that past teacher inputs persist indefinitely and that student ability has a constant effect on test score growth.[34] To investigate the sensitivity of the proposed tests to these assumptions, I alter the data generating process in a variety of ways and reassess test performance.

Table 4 contains the results of the sensitivity analysis. The first column describes how the data generating process (DGP) was altered, and the second and third columns contain the rejection rate across 250 simulations using the F-test and the cluster-robust Wald test, respectively. For all of the simulations in the sensitivity analysis I generate the data such that we observe 10 student observations per teacher and students are sorted into classrooms based solely on unobserved ability. Thus, if the test were correctly sized the rejection rate should be 5%.

The first four modification rows of Table 4 explore how sensitive the tests are to the functional specification of the levels model. In particular, the results in the first modification row illustrate how sensitive the F statistic is to the assumption that student and

TABLE 4. Sensitivity of tests to persistence and homoskedasticity.[a]

	Rejection Rates	
	F-Test	Cluster-Robust Wald Test
Baseline	0.06	0
DGP modification		
Increasing importance of student and school components	0.10	0
Differing weights on student and school component	0.07	0
Teacher inputs decay at 50%	0.11	0.02
Teacher inputs decay at 100%	0.28	0.18
Student-level heteroskedasticity	0.05	0.01
Teacher-level heteroskedasticity, balanced panel	0.05	0
Teacher-level heteroskedasticity, unbalanced panel	0.05	0
Teacher-level heteroskedasticity related to class size	0.16	0.03

[a] Reported results are averages over 250 simulations. The baseline results are taken from the first row of results in Tables 2 and 3. The first column indicates how the DGP was altered. All simulations are executed assuming each teacher is observed with 10 students and that students are assigned to classes based solely on unobserved ability. Thus, we would expect the rejection rate to equal 0.05.

[34] While Rothstein's test for VAM3 does not impose the assumption that student ability has a constant effect on test score growth, the results from Section 3.1 suggests that this is an assumption that should be imposed.

school unobserved effects have a constant impact on test score growth.[35] In the second modification row, I relax the assumption that the weights on the student and school components vary across grades in the same manner.[36] The third and fourth modification rows contain results when I relax the assumption of no decay in past teacher inputs. The rejection rate increases in all four cases, as now there are components in the residual that may be correlated with future teacher assignments. Aside from the scenario that assumes teacher effects decay completely, the F-test continues to perform rather well, with a slight tendency to overreject. The cluster-robust Wald test underrejects in all cases except when teacher inputs decay entirely. The assumption regarding the rate at which teacher inputs persist is clearly paramount, an issue I return to when testing VAMs using the North Carolina data.

The final four rows of Table 4 explore how various forms of heteroskedasticity impact the accuracy of the proposed tests for conditional strict exogeneity. I first consider student-level heteroskedasticity, allowing the variance of the test score residual to vary with student ability.[37] The overall size of each test is unaffected. I then allow for heteroskedasticity at the teacher level in both balanced and unbalanced settings. For the unbalanced data, I allow the number of student observations per teacher to vary but hold the mean at 10. Again, the overall size of each test is unaffected in the basic balanced and unbalanced cases.[38] However, when the variance of the test score residuals is related directly to class size, the test based on the F statistic significantly overrejects the null hypothesis.[39] As the final row of Table 4 shows, the rejection rate increases to 16%. While the overrejection is significant, the amount of heteroskedasticity is also significant[40] and it is directly related to the number of observations available per teacher. Note that the Wald test performs better in this extreme case.

Finally, an interesting result that emerges from the Monte Carlo exercises is that the correlation between the estimates of teacher effectiveness and the true effectiveness measures are quite similar regardless of whether students are sorted based solely on

[35] Instead, I assume that the coefficients on student ability in second, third, fourth, and fifth grade are 1, 1.5, 2.5, and 4. The coefficient on the school effects in third, fourth, and fifth grade are 1, 1.5, and 2.5. Thus I maintain the assumption that the heterogeneity in the school effect across grades follows the same pattern as that of student ability.

[36] The coefficients on student ability in second, third, fourth, and fifth grade are 1, 2, 3, and 4. The coefficients on the school effects in third, fourth, and fifth grade are 1, 1.25, and 1.5.

[37] In the original Monte Carlo experiments, ϵ_{ig} was normally distributed with mean 0 and variance equal to 0.5^2. Here, the variance of ϵ_{ig} is given by $(0.5 * (1 - \mu_i))^2$, where μ_i is student ability. Recall that $\mu_i \sim \mathcal{N}(0, 0.15^2)$.

[38] In both cases, the variance of the test score residual is related to "current" teacher ability according to $(0.5 * (1 + \frac{\beta_{ggc(i,g)}}{2}))^2$. Recall that $\beta_{ggc(i,g)} \sim \mathcal{N}(0, 0.15^2)$.

[39] The variance of ϵ_{ig} is now given by $(0.5 * (1 - \frac{\text{class size}-10}{22.85}))^2$, where the mean class size is 10 with a standard deviation equal to approximately 2. The overrejection is a well known result in the analysis of variance literature. Heteroskedastic robust tests have been developed for unbalanced one- and two-way tests; however, the achievement model considered here is a nested, three-way unbalanced model. To my knowledge, no generalized correction is available.

[40] The variance of the residuals in the smallest classes is four times as large as the variance of the residuals in the largest classes.

ability or sorted on ability and the lagged residual.[41] With 10 observations per teacher, the correlations are 0.48 and 0.47, respectively, and with 20 observations per teacher, the correlations are 0.61 and 0.57, respectively. Thus, the rejected models have almost as much information about underlying teacher effectiveness as the nonrejected models. It is possible to weaken the correlations in the dynamic sorting model by ratcheting up the amount of sorting on the lagged residual. However, an interesting by-product of this is that both the standard deviation of future teacher effects and the fit of the model increase drastically. In general, these numbers would not be in line with the results obtained using the North Carolina sample.

4. Reassessing the validity of VAMs in North Carolina

Since Rothstein's test for VAM3 is likely to reject regardless of the underlying data generating process, I return to the same cohort of North Carolina students to implement a more accurate test for conditional strict exogeneity. I utilize test score and classroom assignment data for the cohort of North Carolina public school students who were in fifth grade in 2001. I omit the details governing the construction of the estimation samples since Rothstein lays out rather clearly in both the text and the appendix the steps taken in cleaning the data. However, it is important to note that my estimation samples are significantly larger despite following the text as precisely as possible.

Table 5 presents the summary statistics from Rothstein's estimation samples and the summary statistics for my estimation samples. Despite the significant sample size differences, the summary statistics are extremely similar across the two data sets. In particular, the test score distributions appear almost identical across the two samples.

As an additional check on the similarity of the two samples, I replicate Rothstein's tests for VAM1 and VAM3, the results of which are shown in Tables 6 and 7. Table 6 presents both Rothstein's results (top panel) and my results (bottom panel) for the specification outlined in VAM1. The various columns reflect estimates of the impact of the fourth and fifth grade teachers on the fourth and fifth grade gain scores. Notice that my estimates of the adjusted standard deviations of the teacher fixed effects and R^2 measures are almost identical to Rothstein's results.[42] Similar to Rothstein, I find that fifth grade teachers appear to have a significant impact on test score gains in fourth grade.

The results of VAM3 and the accompanying strict exogeneity test are illustrated in Table 7. The top panel again lists Rothstein's findings, while the bottom panel illustrates my estimates and tests of the model. Again, the standard deviation of the estimated teacher

[41] Rothstein also found that the bias in the estimated teacher effects may be quite small when students are dynamically sorted. He examined the bias in teacher estimates further in Rothstein (2009).

[42] For each set of estimated teacher effects, I report an unadjusted and an adjusted standard deviation. The unadjusted measure is derived from a simple weighted variance of the teacher effect estimates, where the weights are determined by class size. When calculating this variance, an adjustment is made for the fact that one teacher in each school–grade combination must be normalized to zero. The adjusted standard deviation accounts for the sampling error in each of the teacher effect estimates by subtracting the average variance of the teacher effect estimates, again weighting by class size. See Appendix B.2 of Rothstein for further details.

Table 5. Summary statistics.[a]

	Rothstein Table		Replication Table	
	Base (1)	Restricted (2)	Base (3)	Restricted (4)
No. of students	60,740	23,415	64,367	29,490
No. of schools	868	598	898	684
1 fifth grade teacher	0	0	0	0
2 fifth grade teachers	207	122	199	134
3–5 fifth grade teachers	602	440	638	506
>5 fifth grade teachers	59	36	61	44
No. of fifth grade classrooms w/ valid teacher	3040	2116	3170	2447
Complete test score record: G4–5	99%	100%	100%	100%
G3–5	91%	100%	93%	100%
G2–5	80%	100%	82%	100%
Changed schools between G3 and G5	27%	0%	26%	0%
Valid teacher assignments in grade 3	78%	100%	85%	100%
grade 4	86%	100%	87%	100%
grade 5	100%	100%	100%	100%
Math scores: third grade (beginning of year)	0.14	0.20	0.10	0.16
	(0.96)	(0.96)	(0.95)	(0.96)
third grade (end of year)	0.11	0.19	0.12	0.20
	(0.94)	(0.91)	(0.93)	(0.91)
fourth grade (end of year)	0.07	0.20	0.09	0.21
	(0.97)	(0.93)	(0.96)	(0.93)
fifth grade (end of year)	0.09	0.20	0.08	0.19
	(0.98)	(0.94)	(0.99)	(0.98)
third grade gain	−0.02	0.00	0.02	0.04
	(0.69)	(0.69)	(0.69)	(0.69)
fourth grade gain	−0.01	0.01	−0.01	0.01
	(0.58)	(0.56)	(0.57)	(0.56)
fifth grade gain	0.01	−0.01	−0.01	−0.02
	(0.55)	(0.53)	(0.55)	(0.54)

[a] Unit of observation is a North Carolina public school student who attended fifth grade in 2001. Columns 1 and 2 are taken from Table 1 in Rothstein (2010). Columns 3 and 4 are constructed by the author. Valid teachers are those identified in the data as teaching a self-contained class for the relevant grade in the relevant year. All scores are standardized at the grade–year level. A complete test score record indicates no missing test scores. The restricted sample includes only those students with a complete test score record, with a valid teacher in every grade, and who remained in the same school between third and fifth grade.

fixed effects and the R^2 measures are extremely similar across the two samples. In addition, the outcomes of the strict exogeneity tests are almost identical. I easily reject the strict exogeneity assumption for both math and reading test scores, and find ratios for the effect of ability in grade 4 relative to grade 3 similar to Rothstein.

I now reevaluate the validity of VAM3 on the same estimation sample utilizing the F-test proposed in Section 2.[43] The results of the F-test and the adjusted standard deviation of the estimated teacher effects are listed in Table 8. The p-values for the test of

[43] Rothstein's original approach requires that students remain in the same school and have complete test score records for grades 2–5. These restrictions are not necessary for the F-test. However, to make direct comparisons, I maintain these assumptions.

Table 6. Regression of gain scores on teacher indicators, VAM1.[a]

	Fifth Grade Gain		Fourth Grade Gain		Fifth Grade Gain		Fourth Grade Gain	
	Math (1)	Reading (2)	Math (3)	Reading (4)	Math (5)	Reading (6)	Math (7)	Reading (8)
Teacher Coefficients—Rothstein								
Fifth grade teachers								
Unadjusted SD	0.179	0.160	0.134	0.142	0.197	0.181	0.151	0.168
Adjusted SD	0.149	0.113	0.077	0.084	0.163	0.126	0.090	0.105
p-value	<0.001	<0.001	0.016	0.002	<0.001	<0.001	0.035	<0.001
Fourth grade teachers								
Unadjusted SD					0.188	0.181	0.220	0.193
Adjusted SD					0.150	0.125	0.182	0.140
p-value					<0.001	<0.001	<0.001	<0.001
R^2	0.195	0.100	0.132	0.086	0.297	0.176	0.254	0.174
Adjusted R^2	0.148	0.047	0.081	0.033	0.203	0.066	0.154	0.064
Teacher Coefficients—Replication								
Fifth grade teachers								
Unadjusted SD	0.179	0.156	0.132	0.137	0.198	0.174	0.149	0.156
Adjusted SD	0.150	0.111	0.077	0.078	0.163	0.118	0.085	0.084
p-value	<0.001	<0.001	<0.001	0.025	<0.001	<0.001	<0.001	0.03
Fourth grade teachers								
Unadjusted SD					0.186	0.175	0.218	0.188
Adjusted SD					0.149	0.118	0.181	0.133
p-value					<0.001	<0.001	<0.001	<0.001
R^2	0.203	0.100	0.130	0.083	0.300	0.172	0.252	0.167
Adjusted R^2	0.158	0.050	0.081	0.031	0.211	0.067	0.156	0.061

[a] Unit of observation is student gain scores in fourth and fifth grade for the cohort of North Carolina public school students who attended fifth grade in 2001. The top panel is taken directly from Table 3 in Rothstein (2010). The bottom panel is results produced by the author. Dependent variables are as indicated at the top of each column. Regressions include school indicators, fifth grade teacher indicators, and (in columns 5–8) fourth grade teacher indicators, with one teacher per school per grade excluded. The p-values are for the test of the hypothesis that all teacher coefficients equal zero, using the heteroskedasticity-robust score test proposed by Wooldridge (2010). Standard deviations are of teacher coefficients, normalized to have mean zero at each school and weighted by the number of students taught. Adjusted standard deviations are computed to account for sampling error in the teacher effect estimates. The sample for columns 1–4 includes students from the base sample with nonmissing scores in each subject in grades 3–5. Columns 5–8 exclude students without valid fourth grade teacher matches and those who switched schools between fourth and fifth grade.

whether the future teacher effects are jointly equal to zero are well below 0.05 for both math and reading test score gains. Implementing the alternative falsification test alone does not alter the basic finding that students appear to be sorted into classrooms based on unobserved, time-varying inputs.

At this point, should we simply concede on trying to measure teacher quality using observed student test scores? I think the answer is no for two reasons. First, despite the statistical rejection of VAM3, the extent of the bias may still be small. As discussed in the previous section, when students were sorted based on their lag test scores, the correlation between the estimated teacher effects and the truth was only slightly smaller than when students were sorted based strictly on ability. So while our estimates

TABLE 7. Gain score specification with student fixed effects, VAM3.[a]

	Math		Reading	
	Third grade (1)	Fourth grade (2)	Third grade (3)	Fourth grade (4)
Unrestricted Model—Rothstein				
Standard deviation of teacher effects, adjusted				
Fifth grade teacher	0.135	0.099	0.144	0.123
Fourth grade teacher	0.136	0.193	0.160	0.163
Third grade teacher	0.228	0.166	0.183	0.145
Fit statistics				
R^2	0.314	0.376	0.245	0.284
Adjusted R^2	0.129	0.209	0.042	0.092
Restricted Model—Optimal Minimum Distance				
Ratio, effect on G4/effect on G3	0.14		1.17	
Objective function	2136		2174	
95% critical value	1684		1684	
p-value	<0.001		<0.001	
Unrestricted Model—Replication				
Standard deviation of teacher effects, adjusted				
Fifth grade teacher	0.127	0.098	0.133	0.113
Fourth grade teacher	0.121	0.189	0.145	0.154
Third grade teacher	0.227	0.163	0.171	0.140
Fit Statistics				
R^2	0.305	0.371	0.237	0.273
Adjusted R^2	0.128	0.211	0.042	0.089
Restricted Model—Optimal Minimum Distance				
Ratio, effect on G4/effect on G3	0.23		1.11	
Objective function	2233		2178	
95% critical value	1813		1813	
p-value	<0.001		<0.001	

[a] Unit of observation is student gain scores in third and fourth grade for the cohort of North Carolina public school students who attended fifth grade in 2001. The top panel is taken directly from Table 5 in Rothstein (2010). The bottom panel is results produced by the author. Students who switched schools between third and fifth grade, who are missing test scores in third or fourth grade (or on the third grade beginning-of-year tests), or who lack valid teacher assignments in any grade 3–5 are excluded. Schools with only one included teacher per grade or where teacher indicators are collinear across grades are also excluded. Unrestricted Model reports estimates from a specification with school indicators and indicators for classrooms in grades 3, 4, and 5. Restricted Model reports optimal minimum distance estimates obtained from the coefficients from the unrestricted models for the third and fourth grade gains, excluding the largest class in each grade in each school. The restriction is that the fourth grade effects are a scalar multiple of the third grade effects. The weighting matrix is the inverse of the robust sampling variance–covariance matrix for the unrestricted estimates, allowing for cross-grade covariances.

of teacher value-added may be slightly flawed, they still likely contain important information about teacher performance.

More important is the fact that VAM3 may be misspecified. A strong assumption implicit in VAM3 is that past teacher inputs persist in perpetuity.[44] As Table 4 indicates, if this assumption is incorrect, the proposed F-test has a tendency to overreject the null

[44] Recent papers, such as Kinsler (2012) and Jacob, Lefgren, and Sims (2010), found strong evidence that teacher inputs decay quite rapidly. In fact, Rothstein also found support for this in the penultimate section of his paper.

Table 8. Testing for strict exogeneity in the NC sample using the F-test.[a]

	Math		Reading	
Teacher Coefficients	Restricted	Unrestricted	Restricted	Unrestricted
Third grade teachers on *3rd* grade scores				
Unadjusted SD	0.211	0.220	0.200	0.209
Adjusted SD	0.137	0.133	0.098	0.087
Fourth grade teachers on *3rd* grade scores				
Unadjusted SD		0.194		0.217
Adjusted SD		0.060		0.078
Fourth grade teachers on *4th* grade scores				
Unadjusted SD	0.171	0.228	0.162	0.232
Adjusted SD	0.117	0.120	0.095	0.104
Fifth grade teachers on *4th* grade scores				
Unadjusted SD		0.163		0.172
Adjusted SD		0.077		0.081
Fifth grade teachers on *5th* grade scores				
Unadjusted SD	0.194	0.218	0.176	0.208
Adjusted SD	0.152	0.157	0.116	0.128
R-squared	0.864	0.871	0.831	0.839
Degrees of freedom	75,959		75,959	
Restrictions	3439		3439	
F statistic	1.08		1.07	
p-value	<0.001		0.003	

[a] Results correspond to the estimation of a levels model of student achievement where public student test scores from North Carolina are observed in grades 2–5. Equation (12) in the text corresponds to the restricted model, while the unrestricted model adds fourth and fifth grade teacher assignments to the third and fourth grade level outcomes, respectively. The F-test determines the significance of these additional regressors and provides a test of conditional strict exogeneity. The estimation sample is identical to the sample utilized in Table 3. Students who switched schools between third and fifth grade, who are missing test scores in third or fourth grade (or on the third grade beginning-of-year tests), or who lack valid teacher assignments in any grade 3–5 are excluded. Schools with only one included teacher per grade or where teacher indicators are collinear across grades are also excluded.

hypothesis. To investigate the importance of this assumption, I estimate and test the levels production function

$$A_{igs} = \mu_i + \beta_{ggc(i,g)} + \kappa_s + \epsilon_{ig} \quad \text{for } g = 2, 3, 4, 5, \tag{14}$$

where I restrict the teacher and school effects for the second grade outcome to be zero since the second grade score is actually a pre-test taken in third grade. I also assume that there is no unobserved student heterogeneity in test score growth, since the estimates from VAM3 indicate that the coefficient on μ_i is close to 1 across g. In contrast to the growth specification of VAM3, however, this levels equation assumes that teacher inputs do not persist at all.

To test whether students are randomly assigned to teachers conditional on μ_i, I estimate an unrestricted version of Equation (14) that allows the one grade ahead teacher to enter into the grade g equation. I then test whether these effects are jointly significant. Using the iterative methodology and accompanying F-test, I fail to reject the null hypothesis that the future teacher effects are jointly equal to zero for both math and reading. The respective p-values are 0.99 and 1.00. Note that if I employ the school-by-school approach and construct either a cluster-robust Wald statistic, a heteroskedastic Wald statistic, or a robust score statistic, I continue to reject the null hypothesis that the future teacher effects are jointly equal to zero. The associated p-values are all very close to zero. If I allow the variance of ϵ_{ig} to vary across schools but maintain the restriction of homoskedasticity within schools, I am unable to reject the null hypothesis that the future teacher effects are zero. Clearly the assumptions regarding the error structure are paramount. Given the tendency for robust variance estimators to perform poorly in small samples, I am partial to the results under either homoskedasticity or a very mild form of heteroskedasticity.[45]

Of course, the assumption that teacher effects decay completely from one grade to the next is also extreme. As a compromise, I also estimate the levels specification assuming that teacher inputs persist at a constant geometric rate equal to 0.35.[46] I again fail to reject the null hypothesis that the future teacher effects are jointly equal to zero. The p-values decline slightly to 0.87 and 0.99, but remain well above the 5% critical value.

5. Conclusion

Two trends in the field of education—the explosion of standardized testing and a recognition of the importance of teacher quality in developing students—have recently been married in proposals that would link teacher tenure and salary decisions directly to student test score performance.[47] There are wide-ranging criticisms of such plans, but one fundamental issue is the extent to which teacher quality can be causally identified using student outcomes. The primary threat to identification is the fact that students and teachers are not randomly assigned within or across schools, making it difficult to separate teacher quality from unobserved student inputs.

In a recent article, Rothstein (2010) illustrated the difficulty in identifying teacher effectiveness using standard value-added modeling techniques. Rothstein's (2010) key contribution is a methodology for testing whether the assumptions necessary for identifying causal estimates of teacher quality actually hold in the data. However, as the current paper illustrates, the proposed tests in Rothstein (2010) perform quite poorly in empirical settings likely to be encountered by researchers. I then develop an alternative test that not only performs significantly better in small samples, but is also less computationally intensive. However, the improved small sample performance does come at a cost—the assumption that test score residuals are homoskedastic.

[45] I perform an additional set of Monte Carlo exercises using the levels model outlined in Equation (14). In small samples, all of the heteroskedasticity robust test statistics overreject at a high rate. When I allow for heteroskedasticity at either the student level or the teacher level, the falsification test based on the F-statistic mildly overrejects. Thus, if there were heteroskedasticity in the underlying data, it would make the F-test more likely to reject. Results from the additional Monte Carlo exercises are available on request.

[46] This number is similar to persistent rates estimated in Kinsler (2012), Rothstein (2010), and Jacob, Lefgren, and Sims (2010). For details on how to estimate this model, see Kinsler (2012). Note that this production function is similar to the approach taken by Koedel and Betts (2010a), who estimated a hybrid version of VAM2 and VAM3, incorporating both student fixed effects and lag test scores.

[47] See, for example, the recent Race to the Top federal grants program.

In practice, this last assumption cannot be easily dispensed with in small samples. The standard cluster-robust and heteroskedasticity-robust variance estimators tend to perform quite poorly when teachers are observed with only a handful of student test scores. As a result, VAM falsification tests and estimates of the dispersion in teacher quality that rely on these estimators will likely be inaccurate. Obviously, as the number of observations per teacher increase, it becomes possible to allow for more flexible error structures. This would require either grouping test score outcomes across multiple years or testing students in one cohort more often. Alternatively, it may be possible to improve on the standard robust variance estimators using bootstrap methods.

Appendix A: Updating equations for levels model with student heterogeneity in test score growth

Consider the achievement levels formulation given by

$$\begin{aligned}
A_{i2s} &= \mu_i + \epsilon_{i2},\\
A_{i3s} &= \gamma\mu_i + \beta_{33c(i,3)} + \epsilon_{i3},\\
A_{i4s} &= (2\gamma-1)\mu_i + \beta_{44c(i,4)} + \beta_{33c(i,3)} + \epsilon_{i4},\\
A_{i5s} &= (3\gamma-2)\mu_i + \beta_{55c(i,5)} + \beta_{44c(i,4)} + \beta_{33c(i,3)} + \epsilon_{i5}.
\end{aligned}$$

This formulation is equivalent to Equation (7) under the assumption that there is no teacher associated with the second grade score, and that $\tau_2 = 1$ and $\tau_g - \tau_{g-1} = \gamma - 1$ for $g > 2$. Estimation would start with an initial guess of the parameters μ_i^0, $\beta_{ggc(i,g)}^0$, and γ^0, with the qth iteration consisting of the following steps:

Step 1. Update μ_i^q according to

$$\begin{aligned}
\mu_i^q = \Biggl(& A_{i2s} + \gamma(A_{i3s} - \beta_{33c(i,3)}) + (2\gamma-1)\Biggl(A_{i4s} - \sum_{g=3}^{4}\beta_{ggc(i,g)}\Biggr)\\
&+ (3\gamma-2)\Biggl(A_{i5s} - \sum_{g=3}^{5}\beta_{ggc(i,g)}\Biggr)\Biggr)\Big/\bigl(1+\gamma^2+(2\gamma-1)^2+(3\gamma-2)^2\bigr),
\end{aligned}$$

where all other parameters are evaluated at their $q-1$ iteration values.[48]

Step 2. Update $\beta_{ggc(i,g)}^q$ according to

$$\beta_{33c(i,3)}^q = \frac{1}{3 * N_{\beta_{33c(i,3)}}} \sum_{i=1}^{N_{\beta_{33c(i,3)}}} \Biggl(\sum_{g=3}^{5} A_{igs} - (6\gamma-3)\mu_i - 2\beta_{44c(i,4)} - \beta_{55c(i,5)}\Biggr),$$

$$\beta_{44c(i,4)}^q = \frac{1}{2 * N_{\beta_{44c(i,4)}}} \sum_{i=1}^{N_{\beta_{44c(i,4)}}} \Biggl(\sum_{g=4}^{5} A_{igs} - (5\gamma-3)\mu_i - 2\beta_{33c(i,3)} - \beta_{55c(i,5)}\Biggr),$$

[48] I suppress the iteration notation in the formulas for ease of presentation.

$$\beta^q_{55c(i,5)} = \frac{1}{N_{\beta_{55c(i,5)}}} \sum_{i=1}^{N_{\beta_{55c(i,5)}}} \bigl(A_{i5s} - (3\gamma - 2)\mu_i - \beta_{33c(i,3)} - \beta_{44c(i,4)}\bigr),$$

where $N_{\beta_{ggc(i,g)}}$ is the number of students assigned to classroom c in grade g, μ_i takes on its qth iteration value, and γ is set at its $q-1$ iteration value. When updating $\beta^q_{33c(i,3)}$, the fourth and fifth grade teacher effects are set at their $q-1$ values. Then, when updating $\beta^q_{44c(i,4)}$, I set the fifth grade teacher effects at their $q-1$ value, but use the qth iteration value for the third grade teacher parameters. Both the third and fourth grade teacher parameters take their qth iteration values when updating $\beta^q_{55c(i,5)}$.

Table 9. Sorting on time invariant ability ONLY.

			Ability Weight in	Rejection Rates		
SD(μ_i)	SD($\beta_{ggc(i,g)}$)	SD(ϵ_{ig})	Sorting Formula	Rothstein	Wald Test	F-Test
0.15	0.15	0.5	0.1	0.92	0	0.06
0.15	0.15	0.5	0.4	0.92	0	0.06
0.15	0.15	0.5	0.7	0.94	0	0.02
0.15	0.3	0.5	0.1	0.9	0	0.06
0.15	0.3	0.5	0.4	0.92	0	0.04
0.15	0.3	0.5	0.7	0.98	0.02	0.02
0.3	0.15	0.5	0.1	0.94	0	0.06
0.3	0.15	0.5	0.4	0.98	0	0.04
0.3	0.15	0.5	0.7	1	0.88	0.08
0.15	0.15	1	0.1	0.78	0	0.02
0.15	0.15	1	0.4	0.86	0	0.04
0.15	0.15	1	0.7	1	0	0.04

Table 10. Sorting on lagged residual.

			Lagged Residual Weight	Rejection Rates		
SD(μ_i)	SD($\beta_{ggc(i,g)}$)	SD(ϵ_{ig})	in Sorting Formula	Rothstein	Wald Test	F-Test
0.15	0.15	0.5	0.1	0.95	0	0.16
0.15	0.15	0.5	0.4	0.97	1	1
0.15	0.15	0.5	0.7	0.96	1	1
0.15	0.3	0.5	0.1	0.97	0	0.16
0.15	0.3	0.5	0.4	0.98	1	1
0.15	0.3	0.5	0.7	0.95	1	1
0.3	0.15	0.5	0.1	1	0	0.12
0.3	0.15	0.5	0.4	0.99	1	1
0.3	0.15	0.5	0.7	0.96	1	1
0.15	0.15	1	0.1	0.98	0.06	0.96
0.15	0.15	1	0.4	0.98	1	1
0.15	0.15	1	0.7	0.99	1	1

Step 3. Finally, update γ^q using OLS, fixing all the other parameters at their qth iteration values and treating them as known. Only third, fourth, and fifth grade outcomes are used in this regression, since γ does not appear in the second grade test score.

Appendix B: Sensitivity of VAM3 tests to sorting and scale assumptions

The results reported in Tables 9 and 10 are averages over 50 simulations. The standard deviations and sorting assumptions utilized in the baseline framework are provided in Section 3 of the text. All simulations are executed assuming each teacher is observed with 10 students.

References

Aaronson, D., L. Barrow, and W. Sander (2007), "Teachers and student achievement in the Chicago public high schools." *Journal of Labor Economics*, 25 (1), 95–135. [333]

Abowd, J., R. H. Creecy, and F. Kramarz (2002), "Computing person and firm effects using linked longitudinal employer–employee dataset." Technical report, US Census Bureau. [341]

Altonji, J. and L. Segal (1996), "Small sample bias in GMM estimation of covariance structures." *Journal of Business & Economic Statistics*, 14 (3), 353–366. [338]

Amrein-Beardsley, A. (2008), "Methodological concerns about the education value-added assessment system." *Educational Researcher*, 37 (2), 65–75. [333]

Arcidiacono, P., G. Foster, N. Goodpaster, and J. Kinsler (2012), "Estimating spillovers using panel data, with an application to the classroom." *Quantitative Economics* (forthcoming). [341]

Boardman, A. and R. Murnane (1979), "Using panel data to improve estimates of the determinants of educational achievement." *Sociology of Education*, 52, 113–121. [336]

Burnside, C. and M. Eichenbaum (1996), "Small-sample properties of GMM-based Wald tests." *Journal of Business & Economic Statistics*, 14 (3), 294–308. [338]

Chamberlain, G. (1984), *Handbook of Econometrics*. Elsevier, Amsterdam. [337]

Clotfelter, C., H. Ladd, and J. Vigdor (2006), "Teacher–student matching and the assessment of teacher effectiveness." *Journal of Human Resources*, 41 (4), 778–820. [334]

Cunha, F., J. Heckman, and S. Schennach (2010), "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78 (3), 883–931. [337]

Davidson, R. and J. MacKinnon (1993), *Estimation and Inference in Econometrics*. Oxford University Press, New York. [346]

Guimaraes, P. and P. Portugal (2010), "A simple feasible procedure to fit models with high-dimensional fixed effects." *The Stata Journal*, 10 (4), 628–649. [341]

Hall, P. and J. Horowitz (1996), "Bootstrap critical values for tests based on generalized-method-of-moments estimators." *Econometrica*, 64 (4), 891–916. [339]

Harris, D. and T. Sass (2010), "What makes for a good teacher and who can tell?" Working paper, Georgia State University. [333]

Harris, D. and T. Sass (2011), "Value-added models and the measurement of teacher productivity." Working paper, Georgia State University. [333, 336]

Horowitz, J. (1998), "Bootstrap methods for covariance structures." *Journal of Human Resources*, 33 (1), 39–61. [338]

Jackson, K. (2009), "Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation." *Journal of Labor Economics*, 27 (2), 213–256. [334]

Jacob, B. and L. Lefgren (2008), "Can principals identify effective teachers? Evidence on subjective performance evaluation in education." *Journal of Labor Economics*, 26, 101–136. [333, 336]

Jacob, B., L. Lefgren, and D. Sims (2010), "The persistence of teacher-induced learning gains." *Journal of Human Resources*, 45, 915–943. [355, 357]

Kane, T. and D. Staiger (2002), "The promises and pitfalls of using imprecise school accountability measures." *Journal of Economic Perspectives*, 16 (4), 91–114. [333]

Kezdi, G. (2004), "Robust standard error estimation in fixed-effects panel models." *Hungarian Statistical Review*, 9, 95–116. [340]

Kinsler, J. (2012), "Beyond levels and growth: Estimating teacher value-added and its persistence." *Journal of Human Resources* (forthcoming). [344, 355, 357]

Koedel, C. and J. R. Betts (2010a), "Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique." *Education Finance and Policy*, 6 (1), 18–42. [357]

Koedel, C. and J. R. Betts (2010b), "Value-added to what? How a ceiling in the testing instrument influences value-added estimation." *Education Finance and Policy*, 5, 54–81. [333]

Koretz, D. (2002), "Limitations in the use of achievement tests as measures of educators' productivity." *Journal of Human Resources*, 37 (4), 752–777. [333]

Kramarz, F., S. Machin, and A. Ouazad (2008), "What makes a test score? The respective contributions of pupils, schools, and peers in achievement in English primary education." Discussion paper, London School of Economics and Political Science. [341]

MacKinnon, J. and H. White (1985), "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties." *Journal of Econometrics*, 29, 305–325. [340]

Martineau, J. (2006), "Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability." *Journal of Educational and Behavioral Studies*, 31 (1), 35–62. [333, 337]

McCaffrey, D., J. R. Lockwood, K. Mihaly, and T. Sass (2010), "A review of Stata routines for fixed effects estimation in normal linear models." Working paper, Georgia State University. [342]

Rivkin, S., E. Hanushek, and J. Kain (2005), "Teachers, schools, and academic achievement." *Econometrica*, 73 (2), 417–458. [333, 336]

Rockoff, J. (2004), "The impact of individual teachers on student achievement: Evidence from panel data." *American Economic Review*, 94, 247–252. [333]

Rothstein, J. (2009), "Student sorting and bias in value-added estimation: Selection on observables and unobservables." *Education Finance and Policy*, 4 (4), 537–571. [352]

Rothstein, J. (2010), "Teacher quality in educational production: Tracking, decay, and student achievement." *Quarterly Journal of Economics*, 125 (1), 175–214. [333, 334, 340, 353, 354, 355, 357]

Smyth, G. K. (1996), "Partitioned algorithms for maximum likelihood and other nonlinear estimation." *Statistics and Computing*, 6, 201–216. [341]

White, H. (1980), "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica*, 48, 817–838. [346]

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, second edition. MIT Press, Cambridge. [354]

Submitted January, 2011. Final version accepted March, 2012.