Journal of Research on Educational Effectiveness, 8: 35–59, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 1934-5747 print / 1934-5739 online
DOI: 10.1080/19345747.2014.981905

Evaluating Specification Tests in the Context of Value-Added Estimation

Cassandra M. Guarino
Indiana University, Bloomington, Indiana, USA

Mark D. Reckase, Brian W. Stacy, and Jeffrey M. Wooldridge
Michigan State University, East Lansing, Michigan, USA

Abstract: We study the properties of two specification tests that have been applied to a variety of estimators in the context of value-added measures (VAMs) of teacher and school quality: the Hausman test for choosing between student-level random and fixed effects, and a test for feedback (sometimes called a "falsification test"). We discuss theoretical properties of the tests to serve as background, and propose parsimonious one-degree-of-freedom versions. An extensive simulation study provides important further insight into the VAM setting. Unfortunately, although both the Hausman and feedback tests have good power for detecting the kinds of nonrandom assignment that can invalidate VAM estimates, they also reject in situations where estimated VAMs perform very well in terms of ranking teachers. Consequently, the tests must be used with caution when student tracking is used to form classrooms.

Keywords: Value-added, teacher quality, specification tests

INTRODUCTION

Measures of teacher and school quality based on value-added models (or VAMs) of student achievement are gaining increasing acceptance among policy makers as a tool for evaluating teaching and school effectiveness. Therefore, it is important for researchers and policy makers to understand the statistical properties of the estimates derived from VAMs, and to have some knowledge of when they can be expected to perform well—and when they do not perform well. One way to proceed is to apply statistical tests of the assumptions underlying VAMs to see if they appear justified. Rothstein (2010) and Harris, Sass, and Semykina (2014) are two examples of studies that develop and apply statistical tests of assumptions to VAMs designed to produce measures of teacher effectiveness.

In applying statistical diagnostics in VAM settings, it is imperative to be clear about the goal of the analysis. Is it to determine whether all of the assumptions underlying consistent estimation of the parameters in a structural production function hold? Or is the main goal to get reasonably accurate estimates of teacher value added? In the research literature and in policy applications, the main focus appears to be on getting reliable estimates of value added. Structural models are often used to motivate estimation procedures, but the performance of the value-added estimates is of primary interest. If we assume that providing relatively accurate performance estimates is the primary purpose of the VAM literature, then

Address correspondence to Cassandra M. Guarino, School of Education, Educational Leadership and Policy Studies, Indiana University, 107 S. Indiana Avenue, Bloomington, IN 47405-7000, USA. E-mail: guarino@indiana.edu

it is critical to understand the difference between rejecting assumptions of an underlying structural model and concluding that a particular procedure likely produces poor estimates of value added.

In earlier work (Guarino, Reckase, & Wooldridge, 2015), we provide a summary of the known theoretical properties of various approaches to estimating VAMs designed to yield teacher performance estimates. More important, we provide extensive simulation evidence showing how six of the most commonly used estimators behave under different mechanisms used to match teachers and students. One of the key findings is that certain estimators that are not technically consistent can perform well in estimating teacher effects for the purpose of ranking. One of the estimators—ordinary least squares (OLS) applied to a dynamic test-score equation, which we dubbed "dynamic OLS," or "DOLS"[1]—performs best across many scenarios, although other estimators are slightly better under certain assignment mechanisms.

Given the several choices among estimators in the VAM context it would be helpful to have tools for choosing among different estimators, especially when they produce very different estimated effects. More fundamentally, can we determine whether any estimator does a good enough job of estimating value added to use the estimates for policy purposes, such as rewarding or sanctioning teachers based on estimated performance? It is natural to turn to existing statistical tests to help diagnose whether or not underlying assumptions are met. Several general tests have been developed in the panel data literature, and some variants have been proposed by education researchers for the express purpose of evaluating VAMs (e.g., those in Rothstein [2010] and Harris, Sass, and Semykina [2014]).

The main purpose of this article is to determine the efficacy of specification tests in determining how well a VAM accomplishes its task, which we take to be estimating teacher effects primarily for ranking purposes. To do so, we use simulations in which we originate student test score data based on known teacher effects but then act as if we do not know the true effects and estimate them. We then assess both the degree to which a model and estimation approach yield accurate teacher effect estimates and the behavior of statistical tests of underlying assumptions. Our goal is to determine the usefulness of statistical tests in revealing the quality of specific models for estimating teacher effects.

Generally, we focus on two kinds of tests that are designed to detect nonrandom teacher assignment, that is, where teachers are assigned at least partly on the basis of observed or unobserved student characteristics. The first test—a robust version of the Hausman (1978) test comparing the student-level random and fixed-effects estimators—is primarily intended to uncover situations where teacher assignment is based on unobserved, time-constant student heterogeneity. The test has the power to detect other kinds of nonrandom assignment mechanisms but its main purpose is to determine whether teacher assignment is correlated with unobserved student heterogeneity.

The second kind of test is actually a class of tests whose purpose is to detect dynamic teacher assignment mechanisms that might cause bias in VAM estimates. Such tests were popularized in the VAM context by Rothstein (2010), who called them "falsification tests." Generally, a falsification test tries to determine whether future teacher assignments are predictive of pass scores or gain scores. Rothstein's falsification tests are closely related to tests of strict exogeneity in the panel data literature. In the VAM context, a test of strict exogeneity looks for feedback from shocks to student performance today to future teacher assignment, typically allowing for student heterogeneity that is accounted for by

[1] DOLS essentially regresses the end-of-year score on prior test scores, teacher indicators, and student characteristics.

student-level fixed-effects (FE) estimation. Rothstein proposes falsification tests in contexts with and without unobserved student heterogeneity.

Ideally, we could propose a strategy for applying diagnostic tests that would reject various estimation methods when they produce poor value-added measures and leave us with estimators that perform well. Unfortunately, our findings are not very positive from a practitioner's perspective. Although in many cases the tests properly reject when the estimation method produces poor value-added estimates, in other cases the tests strongly reject when the underlying estimation method produces very good estimates of the VAMs, at least for the purpose of ranking teachers. The particular situation that causes problems for both the Hausman and feedback tests is when students are tracked based on some observable or unobservable factor, but the classrooms are randomly assigned to teachers. In such cases a variety of estimation methods produce reliable VAMs, depending on the nature of the tracking. Yet, as we show in the section Simulation Results, the specification tests strongly reject many of the best estimators.

An important consequence of our findings is that criticisms of VAMs on the basis of evidence provided by Hausman or feedback tests are likely to be unjustified. The bottom line is that, applied in the VAM context, the tests have the power to detect nonrandom assignment schemes that have nothing to do with whether popular estimators are doing their main job: providing good VAM estimates.

Other authors have recently analyzed the performance of diagnostic tests in the context of VAM estimation, although we seem to be the first to evaluate the Hausman test for comparing the RE and FE estimators. Kinsler (2012) studies a particular version of Rothstein's (2010) falsification test in the presence of student heterogeneity based on Chamberlain's (1984) correlated random effects (CRE) framework.[2] As discussed by Kinsler (2012), Rothstein's falsification test is a test of the null hypothesis that teacher assignment is strictly exogenous with respect to time-varying student unobservables. In this article we study the robust regression-based test suggested in Wooldridge (2010, Section 10.7), which can be computed from simple fixed effects estimation (at the student level). For gain-score equations with student heterogeneity, our approach has several advantages over Rothstein's (2010) approach. First, the regression-based test is computationally (and conceptually) much simpler while still robust to student-level serial correlation and heteroskedasticity of unknown form. Second, the regression-based test can be applied to unbalanced panels without change; the CRE approach is not well suited for unbalanced panels. Third, it is straightforward to obtain a one-degree-of-freedom regression-based test, thereby conserving degrees of freedom and possibly improving finite-sample statistical properties.

Goldhaber and Chaplin (2015) independently provide an evaluation of Rothstein's regression-based falsification test in settings that do not allow for student-level heterogeneity. Goldhaber and Chaplin first determine whether Rothstein's statistic properly detects omitted variable bias. They conclude that it is possible to have data-generating mechanisms and estimators that produce unbiased VAMs but where the Rothstein test will reject the specification. We come to a similar conclusion, but our main focus is on the reliability of VAMs for ranking teachers and not on bias per se.[3] In addition, we use a version of the

[2] Kinsler (2012) shows using a simulation that Rothstein's chi-square test based on the CRE approach performs poorly when the number of student observations per teacher is small.

[3] Guarino et al. (2015) show that systematic bias in VAMs need not cause problems for ranking teachers. A constant additive bias clearly has no effect on rankings. In some cases, the bias is in the form of amplifying the estimated effects, and this actually helps with ranking, even though the magnitudes of the estimated VAMs cannot be trusted.

feedback test that is suggested by the panel data literature, and we also propose a convenient one-degree-of-freedom version of the test. Further, the tracking mechanisms we use are different from those in Goldhaber and Chaplin, who explicitly introduce nonlinearities into their tracking mechanisms. The nonlinearities in our tracking devices are more subtle in that they are generated by fixed class sizes and assignment probabilities that are not linear functions of past performance or unobserved heterogeneity.

The rest of the article is organized as follows. We discuss the value-added modeling framework in the next section. Following that, A Discussion of the Tests in the VAM Setting describes the statistical tests that we study—both those that have been applied to VAMs by other researchers and some that have not—and discusses their theoretical properties. We discuss the kinds of nonrandom grouping and assignment mechanisms that seem particularly relevant in the section Student Grouping, Teacher Assignment, and Behavior of the Tests. In the last three sections we discuss our simulation design and the simulation results, and then provide some concluding remarks.

CONCEPTUAL FRAMEWORK FOR TESTING VALUE-ADDED MODELS

It is helpful to begin with a fairly general value-added equation and a brief discussion of the assumptions embedded in it. Assume that the achievement score, A_{it}, is generated as

$$A_{it} = \lambda A_{i,t-1} + E_{it}\beta_0 + c_i + u_{it} - \lambda u_{i,t-1} \tag{1}$$

$$u_{it} = \rho u_{i,t-1} + r_{it}, t = 1, 2, \ldots, T, \tag{2}$$

where A_{it} is a measure of achievement for student i in grade (or year) t and E_{it} is the (row) vector of educational inputs whose coefficients, β_0, are of greatest interest. Generally, E_{it} can include inputs at the school, classroom, or even individual level. In the present article, E_{it} is a vector of teacher assignment indicators. The variable c_i is unobserved, student-level heterogeneity. We assume $\{r_{it}\}$ is a sequence of independent, identically distributed random variables with mean zero so that $\{u_{it}\}$ follows an AR(1) model.

As discussed in Todd and Wolpin (2003) and Guarino et al. (2015), equation (1) can be derived from a general cumulative effects model (CEM) under certain fairly restrictive assumptions, and (2) adds the assumption that the errors $\{u_{it}\}$ have a particular pattern of serial correlation. The parameter λ is the decay parameter in the CEM. Even this restricted version of the CEM is never estimated, because accounting for the combined issues of heterogeneity, ρ differing from λ, and the lagged dependent variable is a challenging econometric problem.

Sometimes $A_{i,t-1}$ is subtracted from both sides of (1) to obtain an equation for the gain score, ΔA_{it}:

$$\Delta A_{it} = \tau_t + \alpha A_{i,t-1} + E_{it}\beta_0 + c_i + u_{it} - \lambda u_{i,t-1}, \tag{3}$$

where $\alpha = \lambda - 1$. The model with no decay in learning is given by $\alpha = 0$. Although this model alleviates some endogeneity concerns regarding the presence of a lagged dependent variable, the "no decay" assumption is quite restrictive.

There are several types of misspecifications that can cause difficulty for standard estimators of β_0. One is failure of the so-called "common factor restriction" (CFR), $\lambda = \rho$. Under the CFR the errors $u_{it} - \lambda u_{i,t-1}$in (1) have no serial correlation, but if $\lambda \neq \rho$

then (1) contains serial correlation. In general, serial correlation in the presence of a lagged dependent variable causes inconsistent estimation for many estimation procedures, including OLS applied to (1) (where we ignore both the presence of c_i and serial correlation in $u_{it} - \lambda u_{i,t-1}$). The Arellano and Bond (1991) instrumental variables procedure, which removes c_i, relies on no serial correlation in the errors in (1).

McClain and Wooldridge (1995) propose a simple test of the null hypothesis that the CFR holds in the context of time series regression models. The test is easily adapted to the panel data case when there is no heterogeneity, that is, when c_i is not in (1). Unfortunately, it is not clear how to extend the test to allow for heterogeneity. Any neglected serial correlation, due to violation of the CFR, higher-order autoregressive properties, or the presence of c_i will cause a rejection of the CFR restriction. However, in the sensitivity analysis in Guarino et al. (2015) we found that violation of the CFR did not appreciably affect the dynamic OLS estimator in the sense that DOLS still provided estimates of the teacher effects that produced reliable rankings among teachers. For these reasons we do not study the CFR test further in this article.

A second kind of misspecification arises in setting λ in equation (1) equal to unity (which is the same as dropping $A_{i,t-1}$ in equation [3]). Including the lagged achievement in equation (3) is a simple, effective way to detect dynamic misspecification. Harris et al. (2014) apply tests for dynamic misspecification by including lagged teacher assignment using data from Florida. We do not study the properties of dynamic misspecification tests in the current article because they are standard tests of omitted variables (and are confounded by the presence of c_i, in any case)[4]. Rather, we are interested in the behavior of exogeneity tests when λ is incorrectly set to unity.

Finally—and most importantly for this article—we are interested in what happens when teachers were assigned to students in such a way to make E_{it} endogenous in an estimating equation. In this third kind of "misspecification" it is not necessarily true that (1) is an incorrect equation; it is simply that inputs have been chosen in a way to violate certain exogeneity requirements, resulting in inconsistent estimators.

A DISCUSSION OF THE TESTS IN THE VAM SETTING

The general purpose of the specification tests we study is to detect nonrandom assignment of students to teachers. We consider tests of both static and dynamic assignment. One form of static assignment occurs when teachers are (partly) assigned on the basis of unobserved student heterogeneity, that is, students with fixed but unobserved characteristics are matched with particular teacher effectiveness levels. Another form of static assignment is based on an initial, or base, test score in an early grade. Dynamic assignment occurs when the prior test scores of students are matched to particular teacher effectiveness levels.

In what follows it is important to distinguish between two mechanisms that can be used for generating classrooms of students taught by particular teachers. Students may be first grouped on the basis of unobserved or observed characteristics, a process often referred to as "tracking." This kind of grouping might be done even if teachers are randomly assigned to classrooms. Nonrandom teacher assignment occurs when classrooms with different average

[4] In the current testing setting, omitting $A_{i,t-1}$ from (1) and failure of the CFR restriction can be expected to have similar consequences because, in effect, a variable that can predict the gain score is omitted from the equation, and teacher assignment might be correlated with that variable.

levels of ability or achievement are systematically assigned to teachers with different levels of competence.

Although little research exists on evaluating specification tests within the VAM setting, certain tests are well known in the panel data literature. For example, Wooldridge (2010, Section 10.7.3) discusses different versions of the Hausman test used to compare the RE and FE estimators. In the VAM context, the Hausman test is primarily a test of static assignment mechanisms because it is intended mainly to pick up any correlation between student heterogeneity and the observed inputs, in this case, teacher assignment. As discussed in Wooldridge (2010), it is generally important to use a version of the Hausman test that is robust to violations of assumptions that are not required for consistently estimating the parameters, in this case, the teacher value-added measures. In particular, we prefer tests that are robust to serial correlation and heteroskedasticity in the students' idiosyncratic shocks. A regression-based test, which we review in the section The Hausman Test Comparing RE (or POLS) to FE, provides a straightforward method for obtaining a fully robust Hausman test.

To detect dynamic forms of teacher assignment, Rothstein (2010) uses a regression-based falsification test, which is in the same spirit as tests for strict exogeneity in the panel data literature. In particular, Rothstein includes future teacher assignments in a current gain-score equation estimated by OLS. Importantly, with the simple regression-based version of the test, Rothstein does not allow for student fixed effects. Without fixed effects, standard estimators, such as OLS, do not require strict exogeneity for consistent estimation. Thus, it is not clear why one wants to test for the presence of dynamic assignment in such cases. By contrast, because failure of strict exogeneity does result in inconsistency when student-level fixed effects are included, Wooldridge (2010, Section 10.7.1) shows how feedback effects are easily tested in the context of fixed effects estimation. The most straightforward way to test for feedback effects is to include future values of the explanatory variables—usually one period ahead—and test their significance using a robust Wald test after FE estimation. We discuss this variant of Rothstein's falsification test in the section Testing Strict Exogeneity in Equations with Student Fixed Effects.

In addition to Rothstein (2010), some other recent empirical papers have applied falsification tests in the VAM context. Koedel and Betts (2009) applied falsification tests in the context of VAM estimation using data from the San Diego school district. Harris et al. (2014) apply a battery of tests (several of which we do not study here) in an empirical context. In their application to data from Florida, they generally find evidence against random assignment of students to teachers and find estimated VAMs that vary widely across procedures.

A Basic Gain-Score Equation and Exogeneity Assumptions

In describing tests for endogeneity of teacher assignment we start with a gain-score equation because both the Hausman test and the falsification test (test of strict exogeneity) can be reasonably applied. A standard gain-score equation is

$$\Delta A_{it} =_{t} + E_{it}\beta_0 + X_{it}\gamma_0 + c_i + e_{it}, t = 1, \ldots, T, \tag{4}$$

which maintains the "no decay" assumption but where the errors, $\{e_{it}\}$, may be serially correlated because (4) no longer includes a lagged dependent variable. The vector X_{it} includes controls, many of which may be constant, that are often included in empirical

VAM studies, such as gender, race/ethnicity, disability status, and free-and-reduced-lunch eligibility. In this article we do not include extra controls in our simulation study, but for a general discussion of how to apply tests, it is useful to explicitly include X_{it}.

The constants τ_t allow for different intercepts for different grades (or, with many cohorts, allows for cohort effects). Because we have few grades (time periods), these can be estimated precisely with a large number of students. The c_i are the time-constant student unobserved effects, sometimes called “student heterogeneity.” The presence of c_i causes the composite error, $v_{it} = c_i + e_{it}$, to be serially correlated. More important, if c_i is correlated with the inputs E_{it}, leaving c_i in the error term can cause inconsistency in estimating β_0.

The idiosyncratic errors, $\{e_{it}\}$, are time-varying unobserved factors that affect gain scores. Generally, these can be serially correlated or heteroskedastic, or both. In the context of an underlying cumulative effects model, e_{it} is a linear combination of the errors appearing in the structural production function; see, for example, Guarino et al. (2015) for a more extensive discussion of the structural model.

An important assumption required of the most common panel data estimators that recognize the presence of c_i—RE and FE—is strict exogeneity of the inputs conditional on the student heterogeneity, namely,

$$E\left(e_{it}|E_{iT}, E_{i,T-1}, \ldots, E_{i1}, X_{iT}, X_{i,T-1}, \ldots, X_{i1}, c_i\right) = 0, t = 1, \ldots, T. \tag{5}$$

Note that in (5) the expectation of the error term, conditional on heterogeneity and all current, past, and *future* inputs, is zero.

To interpret the strict exogeneity assumption, drop the $\{X_{is} : s = 1, \ldots, T\}$ for simplicity. Then, when we combine (5) with equation (1), we have

$$E\left(\Delta A_{it}|E_{iT}, \ldots, E_{i1}, c_i\right) =_t +E_{it}\beta_0 + c_i. \tag{6}$$

Equation (6) means that, once we control for student heterogeneity, only inputs at time t, E_{it}, appear in the gain-score equation at time t. This restriction implies that past inputs—in this case, previous teachers—have no effect on the current gain score, once current teacher and student heterogeneity have been accounted for. As discussed in Harris et al. (2014), it is simple to test such an assumption: simply include, say, $E_{i,t-1}$ and test for joint significance (using whatever estimation method one settles on). Of course, including lagged inputs costs us a year of data, but such tests typically can be carried out for the kinds of data sets available.

For our purposes assumption (5) has another important implication: *future* values, such as $E_{i,t+1}$, do not appear on the right-hand side of (6). Generally, if teacher assignment at time $t + 1$ depends on ΔA_{it} or A_{it}, $E_{i,t+1}$will be (partially) correlated with e_{it}. Shortly, we use this observation to obtain a test of (5).

In addition to strict exogeneity of the inputs conditional on c_i, another important assumption in the panel data literature is

$$E\left(c_i|E_{iT}, E_{i,T-1}, \ldots, E_{i1}\right) = E\left(c_i\right) = 0, \tag{7}$$

where we have again dropped the $\{X_{is} : s = 1, \ldots, T\}$. The assumption that $E\left(c_i\right) = 0$ is without loss of generality when the gain-score equation has an intercept (or a full set of time intercepts). When we combine assumptions (6) and (7), the inputs $\{E_{it}\}$ are strictly exogenous with respect to the composite error:

$$E(v_{it}|E_{iT}, E_{i,T-1}, \ldots, E_{i1}) = 0, t = 1, \ldots, T. \quad (8)$$

Assumption (8) is important, because it justifies generalized least squares estimation (GLS), including the popular RE estimator, applied to

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + v_{it}, t = 1, \ldots, T. \quad (9)$$

A special case of GLS is pooled OLS (POLS), where any serial correlation in v_{it} due to the presence of c_i—in fact, any serial correlation—is ignored. Inference is handled by using a robust variance matrix estimator (which is also robust to heteroskedasticity of arbitrary form). From (9) it is easily seen that consistency of POLS only requires that v_{it} and E_{it} are uncorrelated; the strict exogeneity assumption in (8) is not needed. However, because v_{it} includes c_i, POLS requires that the inputs are uncorrelated with the student-specific heterogeneity.

The Hausman Test Comparing RE (or POLS) to FE

In empirical panel data applications, including VAM estimation, one could estimate the gain-score equation (4) by both RE and FE. Both estimators require the strict exogeneity assumption stated in (5) for consistency. In addition, RE uses the heterogeneity exogeneity condition in (7). Therefore, it is common to compare the RE and FE estimates as a test of (7). However, it is important to remember that RE and FE—and, for that matter, POLS—will typically have different probability limits if the strict exogeneity assumption (1) is violated. Therefore, any test that explicitly or implicitly compares the RE and FE estimators (or the POLS and FE estimators) generally has power against violation of (5) or (7), and one cannot use the outcome of the Hausman test to conclude which assumption fails, or whether both fail.

The traditional form of the Hausman (1978) statistic uses a quadratic form based on the differences between the RE and FE estimators. A critical point in applying the traditional form is that it assumes that the RE estimator is (asymptotically) efficient: the variance-covariance matrix appearing in the quadratic form is valid only when RE is asymptotically efficient. As discussed by Wooldridge (2010, Section 10.7.3), the relative efficiency of the RE estimator holds only when the idiosyncratic errors $\{e_{it}\}$ are serially uncorrelated and homoscedastic, both conditional on the covariates and c_i. (See Wooldridge, 2010, 10.7.3 for a formal statement of the assumptions.) Yet the Hausman test has no power for detecting serial correlation or heteroskedasticity in $\{e_{it}\}$ because these problems do not cause inconsistency in either the RE or FE estimator. In the language of Wooldridge (1990), the traditional form of the Hausman test adds "auxiliary assumptions," which are used to get a standard null distribution even though the test has no power for detecting failure of the assumptions.

It is possible but computationally cumbersome to modify the usual Hausman statistic to be robust to arbitrary serial correlation and heteroskedasticity in $\{e_{it}\}$. One problem is that the variance-covariance matrix is singular when the estimated equation includes time effects, which is very common. A much more straightforward approach is to use a robust, regression-based test.

The regression-based Hausman test is based on the correlated random effects specification

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + \bar{E}_i\xi + a_i + e_{it} \tag{10}$$

where $\bar{E}_i = T^{-1}\sum_{r=1}^{T} E_{ir}$ is the time average and $c_i = \bar{E}_i\xi + a_i$. In this formulation, we explicitly model the heterogeneity c_i as a linear function of the time average of the inputs (which is where the name "correlated random effects" comes from). Equation (10) still contains unobserved heterogeneity, a_i, but it is uncorrelated with the entire history of inputs, $\{E_{it}\}$. If we maintain strict exogeneity conditional on c_i, then strict exogeneity holds conditional on a_i in (10). Therefore, equation (10) can be estimated by POLS or random effects.

A well-known algebraic result (e.g., Wooldridge, 2010, Section 10.7.3) is that when POLS or RE is applied to (10), the resulting estimate of β_0 is the fixed effects estimator that uses deviations from time averages to remove c_i from the equation

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + c_i + e_{it} \tag{11}$$

Therefore, equation (10) is very useful for presenting a unified setting for RE and FE estimation. In particular, if (10) is estimated by random effects (i.e., feasible GLS using the RE structure), it is straightforward to construct a robust Wald test of $H_0 : \xi = 0$, which has as many degrees of freedom as there are inputs E_{it}[5]. Obtaining a Wald test that is robust to arbitrary serial correlation or heteroskedasticity in $\{e_{it}\}$, while remaining asymptotically efficient under the traditional RE assumptions, is straightforward using popular packages that support RE and FE estimation.[6] The regression-based test is asymptotically equivalent (against local alternatives) to the traditional Hausman test when the $\{e_{it}\}$ are serially uncorrelated and homoscedastic.

A rejection of $H_0 : \xi = 0$ is typically taken to mean that c_i is correlated with $\bar{E}_i$ but, as mentioned earlier, this interpretation is based on maintaining assumption (5). If we reject $H_0 : \xi = 0$ then we have found that $\bar{E}_i$ is correlated with the composite error, $c_i + e_{it}$, which warrants a statistical rejection of the RE estimator. However, as we will see in our simulations in the section on simulation results, in the context of estimating VAMs one must be cautious in using the Hausman test in this way. It could be that RE is statistically rejected but provides better estimates of the VAM coefficients than its natural alternative, FE. Even though the RE estimates of VAMs might be systematically biased, they typically have less sampling variation—sometimes much less—and the bias may be such that the estimated VAMs do a good job of ranking teachers. We will have more to say on this in the section on simulation results.

A practical problem with using equation (10) as the basis for the Hausman test is that, with many teachers, (10) contains many regressors: the original teacher dummies and then the proportion of times the student sees that teacher over a student's entire observed history

[5] The POLS and RE estimates of β_0 are equal to the FE estimator. Further, with a balanced panel the POLS and RE estimates of ξ are the same. (They generally differ with an unbalanced panel, in which case RE will be more efficient under the standard RE assumptions.) Whether POLS or RE is used, the test should be made fully robust to serial correlation and heteroskedasticity in $\{e_{it}\}$.

[6] In Stata, a fully robust Wald test is easily obtained using the "cluster" option, which is how we carry out the test in our simulations.

(i.e., the time average). Computationally, many regressors are not too difficult to handle with modern computers and statistical packages. A more pressing concern is potential finite-sample distortions in using large-sample critical values (which is what the Hausman approach necessarily uses). The proper asymptotics rely on the number of students per teacher getting "large." In practice, we may not have many student outcomes associated with some teachers, in which case a one-degree-of-freedom test may have better size properties.

Rather than including the entire vector $\bar{E}_i$ and testing joint significance, which necessitates a separate variable for each teacher in the data set, we propose a new test. This consists of substituting, for each student, the estimated *average teacher effect* across all years for the vector $\bar{E}_i$. This is identical to estimating the equation

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + \alpha\left(\bar{E}_i\hat{\beta}_0\right) + error_{it} \tag{12}$$

and performing a t test of $\alpha = 0$. The estimate $\hat{\beta}_0$ is the RE estimate from the equation (11) obtained in a first stage. One could also apply pooled OLS to (11), but the error term in (11) contains the student heterogeneity term, and under standard assumptions RE is more efficient than POLS. For the same reason we apply RE to equation (12), but we use a fully robust t statistic for $\hat{\alpha}$ to allow for arbitrary serial correlation and heteroskedasticity.

A test using (12) rather than (10) conserves on degrees of freedom, but it may not detect certain kinds of teacher assignment mechanisms. In our simulation we study the properties of both tests and find that our new "one-degree-of-freedom Hausman test" has substantial power against nonrandom assignment alternatives.

In many applications of RE estimation in the VAM context, other explanatory variables are included as controls. Often such controls are student characteristics, such as family background, socioeconomic status, or baseline test scores that do not vary over time. (Test scores lagged one or more period are not allowed in RE estimation because lagged dependent variables always violate the strict exogeneity assumption.) When available, it is important to include such controls in equation (10), leading to an equation such as

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + \bar{E}_i\xi + Z_i\gamma + a_i + e_{it}, \tag{13}$$

where Z_i is the vector of time-constant controls. With good controls, it is more plausible that the (remaining) unobserved heterogeneity is uncorrelated with $\{E_{it}\}$. One can also include time-varying, strictly exogenous controls, say $\{X_{it}\}$, and then (10) becomes

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + \bar{E}_i\xi + Z_i\gamma + X_{it}\eta + \bar{X}_i\lambda + a_i + e_{it}, \tag{14}$$

where we also include the time averages of $\{X_{it}\}$. To test whether the inputs are partially correlated with heterogeneity we would still test $H_0 : \xi = 0$; failing to reject means we can drop $\bar{E}_i$ from (11) and estimate the equation by RE, typically obtaining a more precise estimator of β_0.[7] The test described in equation (12) can be also applied when additional covariates are included in the model. In our simulations, we only consider an equation with teacher dummies and no other inputs.

[7] Guggenberger (2010) warns of the problems of using the Hausman test as a pretest for choosing between RE and FE. The regression-based version of the Hausman test makes it clear that the Hausman pretesting problem is essentially the same as the problem of pretesting whether a set of regressors belongs in an equation and then using an F or Wald test to determine whether those regressors appear in the final model.

Wooldridge (2009) shows that equation (13) can be used as the basis for a Hausman test even in the case of unbalanced panels, provided the reason the panel is unbalanced is appropriately exogenous. One subtle point is that a time period should be used in constructing the time averages only when observations on all variables are available. In the simulations later we only consider balanced panels but most panel data sets are, at least initially, unbalanced.

Testing Strict Exogeneity in Equations with Student Fixed Effects

If the RE estimator is rejected using the regression-based Hausman statistic from the previous section, a natural step is to use the FE estimator so that arbitrary correlation is allowed between c_i and $\{E_{it}\}$. Because consistency of the FE estimator relies on strict exogeneity, it is (potentially) important to test that assumption. Here we are interested in testing for feedback, assuming under the null that only current inputs appear in the gain-score equation at time t.

An auxiliary equation that leads to a simple test is

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + E_{i,t+1}\delta + c_i + r_{it}, t = 1, \ldots, T - 1, \tag{15}$$

where we lose the last time period (grade) by putting the future inputs into the equation at time t. Equation (15) should be estimated by fixed effects in order to allow the heterogeneity and inputs E_{it} to be correlated under the null, making a test of $H_0 : \delta = 0$a pure test of strict exogeneity (feedback in this case). Naturally, the test should be made robust to arbitrary serial correlation and heteroskedasticity in $\{r_{it}\}$ (through what are commonly called "cluster robust" test statistics).

We can add additional time-varying covariates to equation (15) and we may or may not include their lead values. As in any testing context, including a lot of irrelevant variables (lead values in this case) tends to reduce the power of the test. In our simulation study we do not have extra covariates.

Rothstein (2010) uses a version of the test from equation (15) but he applies the test one grade at a time. By using deviations from school means, Rothstein allows school fixed effects, but he does not allow unobserved student effects that are correlated with teacher assignment. In effect, Rothstein imposes the restriction $c_i = 0$, something that is important to recognize in interpreting the outcome of the test. Rothstein effectively applies the test to a cross-sectional regression with school fixed effects. Importantly, strict exogeneity of teacher assignment is *not* required for OLS with school dummies to consistently estimate teacher effects, provided there are many children per school (which is true in Rothstein's setting and reasonable in general). In other words, if Rothstein thinks it is sufficient to control for school but not student effects, then he is testing an assumption that is not needed for consistent estimation of teacher effects.[8] It is only when student fixed effects are allowed in panel data that feedback necessarily causes inconsistent estimation of the teacher effects.

Rothstein also applies a version of the feedback test that is equivalent to testing the coefficients on $E_{i,t+1}$ in the following equation (which does not include student

[8] Rothstein (2010) proposes two versions of the test, one that excludes current teacher assignment and one that includes it. In practice, one should include current teacher assignment because it may be correlated with the next grade's teacher assignment.

heterogeneity):[9]

$$A_{i,t-1} = \tau_t + \alpha A_{it} + E_{i,t+1}\delta + r_{it}, t = 1, \ldots, T - 1; \tag{16}$$

See also Goldhaber and Chaplin (2015), who focus on this particular test among those proposed by Rothstein. This test can be interpreted as checking whether future teacher assignment is related to the test score two years prior after controlling for the previous year's score. For example, it checks whether fifth grade teacher assignment depends on the third grade score once the fourth grade score has been partialled out. Consequently, this test should have power for detecting dynamic assignment mechanisms that depend on multiple lagged test scores, but it has little to do with whether standard VAM estimators are consistent. Essentially, the Rothstein test is irrelevant for evaluating VAM estimators provided we are willing to use dynamic regression with multiple lags of student achievement to control for nonrandom assignment. Goldhaber and Chaplin (2015) make a similar argument and obtain bias formulas for the estimated teacher effects under some simple scenarios. But it is easier, and more general, to simply understand that the regression in (16) is just one way of testing whether $E_{i,t+1}$ and $A_{i,t-1}$are correlated after partialling out A_{it}. The absence of partial correlation is neither necessary nor sufficient for dynamic VAM estimators to consistently estimate teacher VAMS, let alone provide good rankings of teachers.[10]

In the context of dynamic regression where, for simplicity, we include only a single lagged test score, a more natural test comes from the equation

$$A_{it} = \tau + \lambda A_{i,t-1} + E_{it}\beta_0 + E_{i,t+1}\delta + r_{it}. \tag{17}$$

The test of strict exogeneity of teacher assignment is that all elements of δ are zero. Unlike the Rothstein approach, (17) properly controls for current teacher assignment, and answers the question: Is future assignment correlated with current test scores after we partial out lagged test scores *and* current teacher assignment? Even though we prefer (17) to Rothstein's approach, we must emphasize again that dynamic OLS does *not* require strict exogeneity of teacher assignment to consistently estimate the teacher VAMs. Consequently, it is not clear what we can learn, in general, from such tests. Nevertheless, because Rothstein-type tests are popular, we will evaluate the tests in a simulation study in the chance that the tests provide useful information.

As in the case of the Hausman test, a one-degree-of-freedom test can also be used here as an alternative to conserve on degrees of freedom. Rather than include the full set of teacher indicators, one includes the estimated teacher effect for next year's teacher. If $\hat{\beta}_0$ denotes the estimated teacher effect—using whichever method under study—then the regressor is simply $E_{i,t+1}\hat{\beta}_0$, a single linear combination of the lead teacher dummies. We call this the "one-degree-of-freedom leads test." As always, it is prudent to make the *t* statistic robust to arbitrary within-student serial correlation and heteroskedasticity.

[9] The test based on (16) is the same if $A_{i,t-1}$is replaced with the gain score, ΔA_{it} because A_{it} is included as a regressor. Therefore, the coefficients on $E_{i,t+1}$ are the same whether $A_{i,t-1}$or the gain score is used.

[10] Rothstein (2010) rejects almost all specifications that use either no lags or a single lagged test score, which is consistent with teacher assignment that may depend on more than just the most recent test score. This testing outcome likely explains why Rothstein finds that the VAM estimates differ when more flexible dynamic models are used. But this simply means one should use the dynamic models that include several lags.

STUDENT GROUPING, TEACHER ASSIGNMENT, AND BEHAVIOR OF THE TESTS

In our previous study of the properties of various estimators (Guarino et al., 2015), we used several mechanisms for grouping students into classrooms and assigning teachers to those classrooms. In this article, we study the behavior of the tests described in the section A Discussion of the Tests in the VAM Setting under the same scenarios.

As in Guarino et al. (2015), we consider grouping students—to simulate the practice of "tracking"—in four different ways. The first method of grouping students is random grouping (RG), which means there is no tracking. We then consider grouping students on the basis of their most recent test score (dynamic grouping, or DG), on the basis of their base (second-grade) test score (base grouping, or BG), and on the basis of their unobserved student heterogeneity (heterogeneity grouping, or HG). In the latter three cases noise is added to the grouping of students to reflect the reality that, even with tracking, not all of the top students will be assigned to the same class.

We consider three ways of assigning teachers to classes: random assignment (RA) and two types of nonrandom assignment (NRA)—(a) assignment where good teachers, based on their teacher effects, are assigned to better classes (positive assignment, or PA), and (b) assignment where good teachers are assigned to worse classes (negative assignment, or NA). With random grouping of students there is only random assignment of teachers, but all three kinds of teacher assignments can be applied to the three different ways of tracking students. Therefore, in total there are 10 different grouping/assignment scenarios.

It is important to keep separate the issues of tracking and teacher assignment. As discussed in Guarino et al. (2015), tracking by itself does not cause problems for VAM estimates. Even when the dynamics are misspecified in the regression analysis, most of the common estimators perform well in terms of ranking teachers. By contrast, several of the estimators perform poorly when teachers are nonrandomly assigned to groups of students.

To study the tests under dynamic misspecification we consider the case where the decay (or persistence) parameter is $\lambda = .5$ along with the baseline $\lambda = 1$ (no decay). We also considered a scenario where the student heterogeneity c_i is uncorrelated with the base score, A_{i2}, and this has implications for some of the tests in certain scenarios. But we present in our tables the more realistic case where c_i and A_{i2} are positively correlated.

Given the discussion of the various specification tests in the section A Discussion of the Tests in the VAM Setting, we can predict the outcomes of the tests across different scenarios. It is important to remember that specification tests are intended to detect inconsistent parameter estimation and not to determine when various procedures may or may not do well ranking teachers. For example, as discussed in Guarino et al. (2015), in some scenarios the VAM estimates are amplified, and this actually makes it easier to rank teachers, even though comparing the magnitudes of the estimated teacher effects could be misleading. Unfortunately, we cannot expect specification tests to distinguish between biases that help with ranking and those that hurt. Currently available tests are devised to detect inconsistent estimation of parameters.

Table 1 shows the predicted outcomes for the Hausman test based on the RE estimator. We assume that the common factor restriction holds. In constructing the tables, we show how the tests would behave if we had an infinite amount of data. In other words, we do not worry about sampling error and the fact that with any particular sample size we can always make Type I or Type II errors. Thus, the entries in the tables are "Reject" or "Accept."

Consider first the case $\lambda = 1$ and random assignment. This is a clear-cut case where the Hausman test should not reject RE estimation in favor of FE estimation: assignment of

Table 1. Predicted outcomes for Hausman test

Grouping/Assignment Mechanism	Lambda = 1	Lambda < 1
RG/RA	ACCEPT	REJECT
DG/RA	REJECT	REJECT
DG/NRA	REJECT	REJECT
BG/RA	REJECT	REJECT
BG/NRA	REJECT	REJECT
HG/RA	REJECT	REJECT
HG/NRA	REJECT	REJECT

teachers is exogenous with respect to the student heterogeneity c_i and strictly exogenous with respect to the idiosyncratic shocks e_{it} in the equation

$$\Delta A_{it} = \tau + E_{it}\beta_0 + c_i + e_{it}. \tag{18}$$

Therefore, no function of the history of teacher indicators should help to predict the gain score from grade $t-1$ to t.

Unfortunately, the conclusion for RG/RA does not carry over to other scenarios with random assignment of teachers to classrooms. Consider the DG/RA case (still with $\lambda = 1$), where students are grouped together based on past test scores but the resulting classrooms are randomly assigned to teachers. As shown in Table 2, the RE estimator works quite well in this case, producing a rank correlation between the estimated and true teacher effects of .85. Yet the Hausman test will reject because past teacher assignment contains information on the ability level of the student. For example, if students with above-average previous test scores tend to be grouped together, having had a third-grade teacher with a high estimated VAM tells us that, on average, the student has higher ability. Therefore, in a fourth-grade gain score equation the third-grade teacher assignment has some predictive

Table 2. Results from 100 replications—vertically scaled test scores

Rank Correlations	$\lambda = 1$				$\lambda = .5$			
Estimator Assignment Mechanism	POLS	DOLS	RE	FE	POLS	DOLS	RE	FE
RG-RA	0.87	0.87	0.88	0.62	0.85	0.87	0.85	0.57
DG-RA	0.80	0.87	0.85	0.58	0.76	0.87	0.76	0.48
DG-PA	0.90	0.86	0.90	−0.31	0.11	0.87	0.11	−0.44
DG-NA	0.30	0.87	0.30	0.73	0.89	0.87	0.89	0.72
BG-RA	0.83	0.87	0.85	0.62	0.83	0.87	0.83	0.57
BG-PA	0.90	0.89	0.91	0.61	0.62	0.9	0.62	0.60
BG-NA	0.53	0.84	0.64	0.60	0.86	0.77	0.86	0.47
HG-RA	0.81	0.84	0.84	0.62	0.84	0.84	0.84	0.57
HG-PA	0.90	0.91	0.90	0.62	0.87	0.91	0.87	0.51
HG-NA	0.37	0.58	0.53	0.61	0.75	0.56	0.75	0.59

Note. Correlation of student fixed effect with base score is .5. The entries are the average rank correlations between the estimated and true teacher effects.

power for the gain score because third-grade assignment is correlated with ability, c_i. A similar mechanism comes into play in the other random assignment, nonrandom grouping mechanisms.

The situation is even worse when $\lambda < 1$. In this case, the Hausman test will reject even in the random grouping, random assignment case. Rejection occurs because when $\lambda < 1$, equation (18) effectively omits the lagged dependent variable. Although it is true that under RA the assignment of a teacher in grade t does not depend on $A_{i,t-1}$, teacher assignment at time $t - 1$ is correlated with $A_{i,t-1}$ whenever teachers have an effect on achievement (which we assume here). Thus, $E_{i,t-1}$ is correlated with the error term at time t because the error term effectively includes a fraction of $A_{i,t-1}$. In other words, lagged teacher assignment helps to predict the gain score, conditional on current teacher assignment, because the lagged teacher assignment is correlated with $A_{i,t-1}$. It is important to understand that, unlike in the usual settings where the Hausman test is applied, in the current scenario neither the random or fixed-effects estimator produces consistent estimates of the teacher VAMs. Rather, the RE and FE estimators have different (incorrect) probability limits.

Rejection in the RG/RA scenario with $\lambda < 1$ is unfortunate because, as shown in Guarino et al. (2015), the RE estimator again fares well; in fact, it is scarcely worse than in the $\lambda = 1$ case.

Table 1 shows that dynamic misspecification ($\lambda < 1$) in any grouping/assignment combination results in rejection of the RE estimator. Again, such an outcome is unfortunate because the RE estimator does well in several (but not all) of these scenarios. It is not surprising that the Hausman test detects dynamic misspecification in the RE model, but when we couple this analysis with the findings in Guarino et al. (2015), some of which are provided in Table 2, we are left to conclude that the Hausman test for choosing between RE and FE is not very informative for VAM applications.

Because the remaining $\lambda < 1$ scenarios combine various grouping/assignment mechanisms along with dynamic misspecification, we use simulations to obtain an idea of how often the Hausman test rejects.

We can create a similar table for the feedback (or falsification) test for three commonly used estimators: random effects, fixed effects, and dynamic OLS. (The entries for pooled OLS on the gain-score equation (18) are identical to RE.) It includes only the $\lambda = 1$ case. We should emphasize that Wooldridge (2010) applies the leads test only to the fixed effects estimator; not to POLS, RE, or DOLS. The test can provide useful information for RE in the sense that it can detect any correlation between E_{it} and the two sources of error in (18). But the Hausman test for choosing between RE and FE can too, and it is usually applied to the RE estimator to see whether one should use FE. If the RE estimator is rejected based on the Hausman test, the leads test is applied to the FE estimator because FE relies on the strict exogeneity assumption.

As discussed in the section A Discussion of the Tests in the VAM Setting, the case for applying the feedback test to DOLS is a priori weak. It can only tell us whether the random-grouping/random-assignment scenario holds, not whether DOLS does a good job of estimating the teacher VAMs. Because DOLS works well in many scenarios, the feedback test is likely to be very misleading.

To isolate the key problem with the feedback test, suppose that equations (1) and (2) hold with the common factor restriction and no student heterogeneity, so we can write

$$A_{it} = \tau + \lambda A_{i,t-1} + E_{it}\beta_0 + r_{it}, \tag{19}$$

where $\{r_{it}\}$ is unpredictable given past test scores and current and past inputs. This is the ideal setup for DOLS estimation of β_0 (and λ): the estimated teacher effects will be consistent and, under a homoscedasticity assumption on $\{r_{it}\}$, asymptotically efficient. This is true regardless of whether E_{it} is correlated with $A_{i,t-1}$; in fact, the main reason for including the lagged test scores is to allow this kind of nonrandom assignment. It is exactly because E_{it} and $A_{i,t-1}$ are correlated that the lead teacher assignments likely will be significant when added to (19). Thus, even though DOLS will produce good estimates of the teacher effects, the leads test will reject. Other versions of the test, such as Rothstein's (2010), potentially reject when assignment is based on the past two years of test scores. Technically, this outcome is the correct outcome, but the outcome is somewhat misleading because the misspecification is easily corrected: include additional lagged test scores in equation (19). We focus on the one-lag DOLS estimator in this study because a version of the leads test has been applied by Rothstein (2010) and Harris et al. (2014).

Table 3 contains the predicted outcomes if the leads test is applied to the three estimators. With random grouping of students and random assignment of teachers, none of the tests should reject—this is the first row of Table 3. Any deviation from random group or random assignment causes the feedback test to reject for RE and DOLS. Again, rejection is not surprising when the teacher assignment is nonrandom. For example, if teacher assignment is based on past test score, then next grade's teacher will predict the current gain score regardless of the estimation method. As with the Hausman test, the reason for rejection with random assignment but nonrandom grouping is more subtle. If, say, the students are grouped based on their unobserved heterogeneity, and the better teachers get the better students, then the estimated lead teacher effect is, on average, higher for the better students, and so is the student's gain score. It is the opposite for the worse teachers and lower-performing students. Therefore, the estimated lead teacher effects are positively correlated with the students' gain scores.

As with the Hausman test, it is unfortunate that the leads test rejects both RE and DOLS in cases where they produce very reliable teacher VAMs. In Simulation Results, we include RE and DOLS in the simulations to see how often the tests actually reject in reasonable scenarios.

The rejection scenarios for the FE estimator are more subtle. Because FE removes a time-constant student effect, grouping on the basis of time-constant variables does not cause a rejection using the leads test, provided the assignment of teachers to classrooms does not depend on time-varying factors, such as a lagged test score. Therefore, when the grouping of students is done using the base score or student heterogeneity, the FE test will not reject because it is assignment based on time-constant factors that FE is intended to be

Table 3. Predicted outcomes for feedback test: lambda = 1

Grouping/Assignment Mechanism	Random Effects	Fixed Effects	DOLS
RG/RA	ACCEPT	ACCEPT	ACCEPT
DG/RA	REJECT	REJECT	REJECT
DG/NRA	REJECT	REJECT	REJECT
BG/RA	REJECT	ACCEPT	REJECT
BG/NRA	REJECT	ACCEPT	REJECT
HG/RA	REJECT	ACCEPT	REJECT
HG/NRA	REJECT	ACCEPT	REJECT

robust against. Consequently, the leads test is informative for FE: it tests whether grouping or assignment (or both) are based on an omitted factor that varies over time, and omitting these time-varying factors generally causes poor performance by the FE estimator.

THE SIMULATION DESIGN

In the tables and surrounding discussion in the previous section, we effectively assumed that we have an infinite amount of data. In practice, we will not always reject with certainty for entries labeled "REJECT." Some estimation methods will control for more of the factors causing nonrandom assignment, in which case rejection rates will be lower. Also, when a test should not reject in theory it might reject due to poor finite-sample performance. To learn about the size and power of the tests it is very helpful to simulate the statistics in plausible scenarios to see how they perform.

Our simulation design closely follows that in Guarino et al. (2015), although we restrict our attention to the case where students and teachers are randomly assigned to schools. (In Guarino et al. [2015], where we evaluated the ability of VAM estimates to track the true teacher effects, we considered mechanisms where students and teachers sorted into schools. Such sorting had little effect on the rankings of the different estimators.) An important reason for following the Guarino et al. (2015) design is that the findings show which estimators work well across a variety of situations. We can compare those findings with the properties of the test statistics to determine when the tests provide useful information and when they do not. Along with the rejection frequencies for the tests reported in the next section on simulation results, we also report a statistic measuring how well the estimated VAMs mimic the true teacher effects, that is, the rank correlation between the true and estimated teacher effects. These rank correlations are reported in Table 2.

In generating the data, we assume that test scores are perfect reflections of the sum total of a child's learning (i.e., no measurement error) and that they are on an interval scale that remains constant across grades. We assume that teacher effects are constant over time, that teachers remain in the same grade, and that unobserved child-specific heterogeneity has a constant effect in each time period. We allow for unobserved time-varying shocks to the test scores, but we do not allow other time-varying factors (such as family effects correlated with teacher assignment). Also, we omit school effects and peer effects, and we assume that teachers have the same effect on each student in a class (so no interactions between students and teachers). We also assume a constant decay parameter, and we assume the shocks in the gain score equation are serially uncorrelated.

Our data represent three elementary grades per student in a hypothetical district. We can think of these as grades 3 through 5 over the course of three years, where we observe an initial second-grade test score. We create data sets that contain students nested within teachers nested within schools, with students followed over time. Our simple baseline data-generating process (DGP) is as follows:

$$\begin{aligned} A_{i3} &= \lambda A_{i2} + \beta_{i3} + c_i + e_{i3} \\ A_{i4} &= \lambda A_{i3} + \beta_{i4} + c_i + e_{i4} \\ A_{i5} &= \lambda A_{i4} + \beta_{i5} + c_i + e_{i5} \end{aligned} \tag{20}$$

where A_{i2} is a baseline score reflecting the subject-specific knowledge of child i entering third grade, λ is a time-constant decay parameter, β_{it} is the teacher-specific contribution (the true teacher value-added effect), c_i is a time-invariant child-specific effect, and e_{it} is a random deviation for each student. Because we assume independence of e_{it} over time, we are maintaining the common factor restriction in the underlying cumulative effects model.

The random variables A_{i2}, β_{it}, c_i, and e_{it} are drawn from normal distributions with mean zero, where we adjust the standard deviations to allow different relative contributions to the scores. We choose the same second moments as in Guarino et al. (2015); we refer the reader to that paper for a survey of the literature underlying our choices. Specifically, the standard deviation of the teacher effect is .25, while that of the student fixed effect is .5, and that of the random noise component is 1, each representing approximately 5%, 19%, and 76% of the total variance in gain scores, respectively.[11] Also, the correlation between the time-invariant child-specific heterogeneity c_i and the baseline score A_{i2} is about .5.

Our data structure has the following characteristics that do not vary across simulation scenarios:

- 10 schools
- 3 grades (3rd, 4th, and 5th) of scores and teacher assignments, with a base score in 2nd grade
- 4 teachers per grade (thus 120 teachers overall) who remain in the same grade
- 20 students per classroom
- 4 cohorts of students
- No crossover of students to other schools

To create different scenarios, we vary certain key features: the grouping of students into classes, the assignment of classes of students to teachers, and the amount of decay in prior learning from one period to the next.

Given the 10 different ways of grouping students and assigning them to teachers and two specifications pertaining to the amount of decay (1 and .5), we have 20 different scenarios per estimator. We use 100 replications per simulation.[12]

SIMULATION RESULTS

We begin with the Hausman test for comparing the RE estimator to the FE estimator. For completeness, we also include the POLS estimator. As discussed earlier in the section on tests in the VAM setting, in practice one should allow for general serial correlation and heteroskedasticity in the composite error term, and so we report findings for the test robust to cluster correlation at the student level. (The findings are similar when we use the nonrobust tests, provided the nonrobust tests are asymptotically valid.)

Because the grouping mechanisms generate within-classroom correlation, we also, when it is mechanically possible, cluster at the school level. We cannot cluster at the classroom level because the students change classrooms over time. Besides, grouping different students into different classrooms in different grades generally creates correlation across all students at a school within a grade. Thus, if one is to cluster at a level higher than the individual student then the school level is natural. We use school-level clustering even though we only have 10 schools, which makes applying the asymptotic theory where the number of schools is large suspect. Nevertheless, some simulation studies have shown

[11] Guarino et al. (2015) also explores the sensitivity of results to teacher effects that are larger relative to the other factors in the data-generating process.

[12] However, we have tested the sensitivity of our results to much higher numbers of replications and found no substantive difference in the results.

Table 4. Hausman test rejection rates: Results from 100 replications—vertically scaled test scores: $\lambda = 1$

Hausman Test $\lambda = 1$	Cluster at Student Level		Cluster at School Level	
Estimator Assignment Mechanism	POLS	RE	POLS	RE
	0.19	0.06	0.18	0.08
RG-RA	0.12	0.13		
	1	0.95	0.99	0.67
DG-RA	1	1		
	1	1	1	1
DG-PA	1	1		
	1	1	0.75	0.83
DG-NA	0.91	0.6		
	0.99	0.99	0.71	0.45
BG-RA	1	1		
	1	1	1	1
BG-PA	0.25	0.89		
	1	1	0.07	0.78
BG-NA	0.99	0.83		
	1	1	0.86	0.6
HG-RA	1	1		
	1	1	1	1
HG-PA	0.76	0.67		
	1	1	0.24	0.27
HG-NA	0.19	0.06		

Note. Correlation of student fixed effect with scorebase is .5. Row 1: Rejection rate of test with estimated mean teacher effect. Row 2: Rejection rate of test with mean teacher indicators.

clustering with as few as 10 clusters can work reasonably well, and much better than doing nothing. We do not have enough schools to cluster when using the test that includes the average of all teacher indicators.

Table 4 contains the results for the Hausman test with $\lambda = 1$. The first panel considers clustering at the student level. Each scenario and estimator has two entries. The first row contains the rejection frequencies of the one-degree-of-freedom test that includes the estimated teacher effect across all three grades. The second row has the rejection rates for the test that includes the full set of teacher dummies.

We focus on the RE results because POLS seems to have more small sample bias. Under the RG/RA scenario, the one-degree-of-freedom test rejects about 6% of the time, which is close to the expected 5% rate predicted for this scenario in Table 1. The full test somewhat over-rejects (13% rejection rate). Clustering at the school level leads to a test with fairly good size (8%), even though we only have 10 schools.

The remaining predictions from the section Student Grouping, Teacher Assignment, and Behavior of the Tests are born out as well. The test that includes the estimated teacher effect almost always rejects in every other kind of grouping/assignment scenario. To see the practical problems this causes, consider the HG-RA scenario in Table 4. The test clustered at the student level rejects 100% of the time, and the lowest rejection rate is 60% (RE with clustering at the school level). This means that we would traditionally reject the RE estimator in favor of FE. Yet in this simulation the rank correlation of the estimated VAMs, reported in Table 2, for the RE estimator is .84 compared with .62 for FE. In fact, of the four estimators—POLS, DOLS, RE, and FE—RE (tied with DOLS) works the best.

As mentioned earlier, the rejection of RE using the Hausman test in the HG-RA scenario is essentially mechanical due to the fact that good students are grouped with other good students. But this grouping has no effect on the quality of RE as an estimator of teacher VAMs. We are forced to conclude that the Hausman test is very misleading in this case.

In other scenarios the situation is even worse. For example, in the HG-PA setting, where we fully expect RE to be rejected, and it is, the RE estimator actually does even better in ranking teachers. The rank correlation jumps to .90; because FE removes the heterogeneity when estimating the teacher effects, we expect the FE rank correlation to remain about the same, and it does to two decimal places.

As if things were not bad enough, against the one alternative where some versions of the Hausman test do not detect nonrandom assignment, HG-NA, the FE estimator outperforms the RE estimator (with rank correlations of .61 and .53, respectively). In other words, when we want to reject RE in favor of FE the Hausman test has the lowest power. To be fair, the version of the test that uses the estimated teacher effects has unit power when we do not cluster, but this is not the standard form of the Hausman test.

Table 5 contains the simulation rejection rates when $\lambda = 1/2$. With a handful of exceptions, the test rejects the RE estimator 100% of the time.

Table 5. Hausman test rejection rates: Results from 100 replications—vertically scaled test scores: $\lambda = .5$

Hausman Test $\lambda = .5$	Cluster at Student Level		Cluster at School Level	
Estimator Assignment Mechanism	POLS	RE	POLS	RE
	1	1	1	1
RG-RA	0.94	0.94		
	1	1	1	1
DG-RA	1	1		
	1	1	1	1
DG-PA	1	1		
	1	1	1	1
DG-NA	1	1		
	1	1	1	1
BG-RA	1	1		
	1	1	1	1
BG-PA	1	1		
	0.53	0.53	0.32	0.41
BG-NA	1	1		
	1	1	1	1
HG-RA	0.98	0.98		
	0.63	0.63	0.47	0.53
HG-PA	0.97	0.97		
	1	1	1	1
HG-NA	1	1		

Note. Correlation of student fixed effect with scorebase is .5. Row 1: Rejection rate of test with estimated mean teacher effect. Row 2: Rejection rate of test with mean teacher indicators.

Table 6. Leads test rejection rates: Results from 100 replications—vertically scaled test scores: $\lambda = 1$

Leads Test $\lambda = 1$	Cluster at Student Level				Cluster at School Level			
Estimator Assignment Mechanism	POLS	DOLS	RE	FE	POLS	DOLS	RE	FE
	0.1	0.03	0.01	0.07	0.15	0.03	0.05	0.07
RG-RA	0.05	0.04	0.05	0.03				
	1	0.36	0.95	0.16	0.97	0.02	0.57	0.16
DG-RA	1	1	1	0.18				
	1	1	1	1	1	1	1	1
DG-PA	1	1	1	1				
	1	1	1	1	0.74	1	0.8	1
DG-NA	1	1	1	1				
	0.65	0.08	0.35	0.09	0.55	0.03	0.31	0.06
BG-RA	0.68	0.09	0.61	0.09				
	1	0.14	1	0.06	1	0.15	1	0.04
BG-PA	0.97	0.11	0.97	0.1				
	0.13	0.22	0.6	0.06	0.06	0.22	0.49	0.13
BG-NA	0.97	0.1	0.97	0.1				
	0.86	0.36	0.54	0.04	0.7	0.25	0.43	0.1
HG-RA	0.94	0.73	0.88	0.07				
	1	1	1	0.06	1	1	1	0.06
HG-PA	1	1	1	0.09				
	0.59	0.76	0.56	0.08	0.19	0.51	0.2	0.06
HG-NA	1	1	1	0.09				

Note. Correlation of student fixed effect with scorebase is .5. Row 1: Rejection rate of test with estimated lead teacher effect. Row 2: Rejection rate of test with future teacher indicators.

The situation is somewhat improved for the leads test in the sense that it has roughly size 5% for the fixed effects estimator under static assignment mechanisms and it detects dynamic forms of teacher assignment. Table 6 contains the rejection frequencies when $\lambda = 1$ and Table 7 contains the frequencies when $\lambda = 1/2$. The discussion below pertains to the results where $\lambda = 1$.

In the RG-RA case, the test has size roughly 5% for all estimators with the exception of pooled OLS, where the rejection rates are somewhat high. For FE under base score and heterogeneity grouping, the feedback test using the estimated lead teacher effect rejects between 4% and 9% of the time, reasonably close to a 5% significance level. Clustering by school causes some distortions, but the rates are acceptable with only 10 schools.

The FE estimator is strongly rejected when dynamic grouping is coupled with nonrandom teacher assignment, either positive or negative: the rejection rates are all 100%. This shows that the test works as it is supposed to in detecting a failure of strict exogeneity when using FE estimation. Under dynamic grouping but random teacher assignment, the test rejects about 16% of the time. What appears to be happening is that removing a student fixed effect largely, but not entirely, accounts for the grouping by past test scores.

Table 6 also shows that POLS and RE are strongly rejected in most scenarios, even though in some of these RE is the best estimator for ranking the teacher effects. We already

Table 7. Leads test rejection rates: Results from 100 replications—vertically scaled test scores: $\lambda = .5$

Leads Test $\lambda = .5$	Cluster at Student Level				Cluster at School Level			
Estimator Assignment Mechanism	POLS	DOLS	RE	FE	POLS	DOLS	RE	FE
	0.13	0.03	0.13	0.06	0.11	0.02	0.16	0.09
RG-RA	0.02	0.04	0.02	0.03				
	1	0.38	1	0.62	0.99	0.02	0.99	0.55
DG-RA	1	1	1	0.38				
	1	1	1	1	1	1	1	1
DG-PA	1	1	1	1				
	1	1	1	1	1	1	1	1
DG-NA	1	1	1	1				
	0.1	0.09	0.1	0.09	0.11	0.03	0.14	0.05
BG-RA	0.52	0.08	0.52	0.13				
	0.98	0.28	0.98	0.07	0.88	0.26	0.95	0.07
BG-PA	0.9	0.15	0.91	0.14				
	0.78	0.32	0.78	0.17	0.65	0.25	0.71	0.14
BG-NA	0.78	0.23	0.78	0.14				
	0.1	0.35	0.1	0.04	0.06	0.25	0.1	0.06
HG-RA	0.23	0.73	0.23	0.05				
	0.71	1	0.71	0.12	0.56	1	0.65	0.15
HG-PA	0.7	1	0.71	0.12				
	0.99	0.63	0.99	0.07	0.97	0.32	0.97	0.12
HG-NA	0.76	1	0.77	0.12				

Note. Correlation of student fixed effect with scorebase is .5. Row 1: Rejection rate of test with estimated lead teacher effect. Row 2: Rejection rate of test with future teacher indicators.

discussed a similar situation for the Hausman test. DOLS is rejected much less often with base score grouping than are POLS and RE. One way to understand why this happens is that DOLS is "almost" controlling for the right variable that determines teacher assignment: the most recent test score rather than the base score. The feedback test applied to DOLS is strongly rejected in the HG case: grouping is based directly on a factor affecting test scores (the student effect) and controlling for the lagged test score is not sufficient when assignment is based on c_i. To see why applying the leads test to DOLS is problematical, we again turn to the rank correlations between the DOLS VAM estimates and the true teacher effects. In the HG-PA scenario, the leads test rejects 100% of the time, yet the rank correlation is about .91. Therefore, DOLS is doing a good job of ranking teachers even though the falsification test virtually always rejects.

The feedback test also rejects DOLS 100% of the time in the DG-PA and DG-NA cases, even though these are cases where DOLS does well in ranking the teachers (rank correlation .86 and .87). Interestingly, both POLS and RE do notably better, with both having rank correlations of about .90 in the case of DG-PA. Of course, the leads test strongly rejects when they are used, too.

The results in Table 7, where $\lambda = 1/2$, similarly show 100% rejection rates for the test when applied to DOLS in the DG-PA and DG-NA scenarios. Again, these are cases where DOLS performs well at ranking the teachers (rank correlations of .87).

If we relied on the rejection by this "falsification test" to determine whether the estimates are doing a good job of estimating the teacher effects, we would be led badly astray; we would conclude none of the estimates can be trusted. In effect, Rothstein's (2010) conclusion was that the VAM estimates could not be trusted because his version of the falsification test always rejected. Our simulations show that this test actually has very little to say about when POLS, RE, and DOLS are working well or not.

CONCLUDING REMARKS

In this article we have discussed two specification tests that are applied in the literature. The first test is a robust, regression-based version of the Hausman test that compares the random effects and fixed effects estimators (at the student level). The second test is a feedback or leads test that was originally designed to test for violation of the strict exogeneity assumption in the context of fixed effects estimation. Versions of this test were used in an influential paper by Rothstein (2010) to detect nonrandom teacher assignment in the context of several regression equations, including dynamic equations.

The most important takeaway from this article is that neither the Hausman test nor the feedback test is very helpful for choosing among estimators or for determining whether a particular estimation method is providing good estimates of teacher VAMs. The Hausman test rejects RE in favor of FE in many cases where the RE estimator is clearly superior for ranking teachers based on estimated VAMs. The source of the problem with the Hausman test is that nonrandom grouping of students—often called tracking—leads to rejection even though teacher assignment to classrooms is random. Under random teacher assignment, RE does well for estimating teacher value added.

The feedback test is a little more successful, but only when it is applied to the fixed effects estimator, the original application of the test described in Wooldridge (2010). The test has good-size properties under static assignment mechanisms and detects dynamic assignment, which has deleterious effects on the FE VAM estimates, with high probability. Nevertheless, we must emphasize that the falsification test applied to pooled OLS, random effects, and dynamic regression produces misleading results. Often the test rejects even though the estimation method is working well. Conversely, sometimes the test fails to reject when the estimated VAMs are poor.

The findings in this article can be combined with those in Guarino et al. (2015) to provide some practical advice to those wanting to estimate teacher VAMs. Guarino et al. (2015) found that, generally, dynamic regression methods provide the best and most robust estimates, although there are exceptions. The current article shows that applying a falsification test to dynamic regression—whether it is the simple form studied here, with just a single lagged score, or more sophisticated methods with multiple lags—is a poor idea. A rejection has very little to do with whether dynamic regression produces good VAM estimates. A similar comment holds for RE estimation, whether one applies the Hausman test or the falsification test: the outcome of the test is practically useless for the main aim of estimating VAMs.

The inappropriateness of applying the falsification test to dynamic regression methods for estimating VAMs can be further understood by viewing dynamic regression through the lens of estimating average treatment effects, where being assigned a particular teacher is the "treatment," rather than thinking of dynamic regression as estimating a structural cumulative effects model. From the modern treatment effects perspective, controlling for lagged test scores and perhaps other observables is intended to make teacher assignment random conditional on the observables. This "unconfoundedness of treatment assignment" assumption is at the heart of regression, propensity score, and matching methods for estimating treatment effects; see, for example, Imbens and Wooldridge (2009). As is well known, the unconfoundedness assumption is not testable: it exactly identifies the teacher effects. Moreover, the treatment is assumed, or at least allowed, to be correlated with the conditioning variables—usually the past test scores. Generally, testing whether teacher assignment is correlated with past test scores is not a test of the unconfoundedness assumption unless some strong assumptions are imposed about the nature of any nonrandom assignment. Imbens and Wooldridge (2009) discuss a set of sufficient conditions, but the spirit of them can be easily described. In effect, in order to construct a falsification test one must assume that unconfoundedness holds conditional on a short history of test scores, with more lags excluded from the conditioning set. We see no reason to think such assumptions are plausible when trying to estimate teacher effectiveness.

Given the frailty of the cumulative effects model as a description of educational production, viewing dynamic regression methods as flexible ways to estimate VAMs, without worrying about "structural" parameters, appears to be the most promising way forward.

FUNDING

This work was supported by grant number R305D10002 from the Institute of Education Sciences in the U.S. Department of Education.

REFERENCES

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277–297.

Chamberlain, G. (1984). Panel data. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics* (Vol. II). Amsterdam, the Netherlands: Elsevier North-Holland.

Goldhaber, D., & Chaplin, D. (2015). Assessing the "Rothstein Falsification Test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, *8*, 8–34.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1).

Guggenberger, P. (2010). The Impact of a Hausman pretest on the size of a hypothesis test: The panel data case. *Journal of Econometrics*, *156*(2), 337–343.

Harris, D., Sass, T., & Semykina, A. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, *38*, 9–23.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, *46*, 1251–1271.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics*, *3*(2), 333–362.

Koedel, C., & Betts, J. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique* (Working Paper). Retrieved from http://econpapers.repec.org/paper/umcwpaper/0902.htm

McClain, K. T., & Wooldridge, J. M. (1995). A simple test of the consistency of dynamic linear regression in rational distributed lag models. *Economic Letters*, *48*(3), 235–240.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Todd, P., & Wolpin, K. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, F3–F33.

Wooldridge, J. M. (1990). A unified approach to robust, regression-based specification tests. *Econometric Theory*, *6*(1), 17–43.

Wooldridge, J. M. (2009), *Correlated random effects models with unbalanced panels* (Working Paper). Michigan State University Department of Economics.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.