American Economic Review: Papers & Proceedings 2016, 106(5): 388–392
http://dx.doi.org/10.1257/aer.p20161080

Interpreting Tests of School VAM Validity†

By Joshua Angrist, Peter Hull, Parag Pathak, and Christopher Walters*

Public school districts increasingly look to the estimates generated by value-added models (VAMs) to measure school and teacher quality. A typical VAM compares average student test scores across teachers or schools while using regression models to control for students' past scores and demographic characteristics. The resulting estimates serve as inputs into teacher retention and promotion policies, school report card systems, and decisions about which schools to close, restructure, or expand.

The VAM framework relies on a selection-on-observables assumption: teachers and schools must be as good as randomly assigned conditional on previous test scores and other observed characteristics. The high stakes attached to VAM estimates have motivated research on the predictive validity of VAMs.[1] A closely related line of inquiry, pioneered by Deutsch (2012) and Deming (2014), uses randomized school admission lotteries to test VAM validity. These tests are motivated by intuitive arguments, but their formal statistical properties have yet to be fully developed.

This paper lays out the econometric theory behind lottery-based tests of VAM validity. Our working paper (Angrist et al. 2015) derives the testable implications generated by school VAMs, introduces a new test of the restrictions implied by these models, and develops an empirical Bayes strategy that uses lotteries to improve estimates of school quality. Our focus here is on the link between the test in Angrist et al. (2015) and the classical overidentification tests introduced by Anderson and Rubin (1949) and Sargan (1958). We use the general theory of specification testing presented in Newey (1985) and Newey and West (1987) to make this link. We also discuss finite-sample concerns raised by the many-weak instrument nature of empirical lottery scenarios. The theory is applied to data from the Charlotte-Mecklenburg School (CMS) district first analyzed by Deming (2014).

I. Value-Added Framework

Our analysis of VAM specification testing starts with a constant-effects causal model:

$$Y_i = D_i'\beta + X_i'\gamma + \epsilon_i, \tag{1}$$

where Y_i is a test score for student i, D_i is a $J \times 1$ vector of mutually exclusive indicators for attendance at one of J schools, X_i is a vector of control variables including past achievement and a constant, and ϵ_i is a random error that satisfies $E[X_i\epsilon_i] = 0$ by definition of γ. The $J \times 1$ vector β captures the causal effects of school attendance relative to an omitted school. In other words, β measures school value-added.

Conventional value-added models use ordinary least squares (OLS) regression coefficients to measure school effectiveness. The OLS regression of Y_i on D_i and X_i is

$$Y_i = D_i'\alpha + X_i'\Gamma + v_i. \tag{2}$$

*Angrist: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, and NBER (e-mail: angrist@mit.edu); Pathak: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, and NBER (e-mail: ppathak@mit.edu); Hull: Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139 (e-mail: hull@mit.edu); Walters: University of California, Berkeley, 530 Evans Hall, Berkeley, CA 94720, and NBER (e-mail: crwalters@econ.berkeley.edu). We gratefully acknowledge financial support from the National Science Foundation, the Laura and John Arnold Foundation, and the Spencer Foundation. Thanks also go to our discussants, Larry Katz and Guido Imbens, to Gary Chamberlain, Raj Chetty, John Friedman, Tom Kane, Jesse Rothstein, and Doug Staiger for helpful comments, and to Dave Deming for providing the CMS data.

†Go to http://dx.doi.org/10.1257/aer.p20161080 to visit the article page for additional materials and author disclosure statement(s).

[1] See, e.g., Rothstein (2010); Kane and Staiger (2008); Chetty, Friedman, and Rockoff (2014); and Rothstein (2014).

The fact that α and Γ are population regression coefficients, insures that $E[D_i v_i] = E[X_i v_i] = 0$. If the controls in X_i are sufficient to eliminate omitted variables bias, we also have that $E[D_i \epsilon_i] = 0$ and the parameters of equations (1) and (2) coincide. This scenario is described by the null hypothesis

$$H_0 : \alpha = \beta.$$

When H_0 is false, OLS and Causal parameters differ for some or all schools and conventional VAM estimates are biased.

II. Testing for Bias

School admission lotteries can be used to test for bias in OLS estimates of value-added. Let Z_i denote an $L \times 1$ vector of indicators for offers of admissions made in L random lotteries, one for each over-subscribed school. In practice offers are randomized conditional on observed stratification variables, such as indicators for having applied to a particular set of over-subscribed schools. We ignore this complication in the theoretical discussion.[2]

Assuming lottery offers affect test scores solely by changing school attendance, we have

$$E[Z_i \epsilon_i] = 0. \tag{3}$$

This vector of L restrictions provides the basis for our omnibus VAM validity test. Since $L < J$, these restrictions are insufficient to identify the coefficients on the J school indicators in equation (1). Yet they can still be used to test H_0, which implies $v_i = \epsilon_i$, and therefore that $E[Z_i v_i] = 0$.

A Lagrange multiplier (LM) test of VAM validity checks this implication directly. Let $\hat{Y}_i = D_i'\hat{\alpha} + X_i'\hat{\Gamma}$ denote the fitted values from (2), where $\hat{\alpha}$ and $\hat{\Gamma}$ are OLS estimates from a random sample of N students. Collect observations on Y_i and $\hat{Y}_i$ in the $N \times 1$ vectors Y and $\hat{Y}$, and let Z denote the $N \times L$ matrix of lottery offer data. Suppose that ϵ_i is homoskedastic, so that $E[\epsilon_i^2 | D_i, X_i, Z_i] = \sigma^2$. The LM test statistic for VAM validity is then

$$\hat{T} = \frac{(Y - \hat{Y})' P_Z (Y - \hat{Y})}{\hat{\sigma}^2}, \tag{4}$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the lottery projection matrix and $\hat{\sigma}^2 = (Y - \hat{Y})'(Y - \hat{Y})/N$ estimates σ^2. Under H_0 and appropriate regularity conditions, $\hat{T}$ has an asymptotic χ^2_L distribution.[3]

Note that we can rewrite the LM statistic as

$$\hat{T} = (\hat{\rho} - \hat{\pi})' \hat{\Sigma}^{-1} (\hat{\rho} - \hat{\pi}), \tag{5}$$

where $\hat{\rho} = (Z'Z)^{-1}Z'Y$ and $\hat{\pi} = (Z'Z)^{-1}Z'\hat{Y}$ are coefficients from regressions of Y_i and $\hat{Y}_i$ on Z_i, and $\hat{\Sigma} = \hat{\sigma}^2 (Z'Z)^{-1}$ is a restricted estimate of the asymptotic variance of $(\hat{\rho} - \hat{\pi})$ that imposes H_0. Equation (5) shows that the LM test can be interpreted as a Wald test of the hypothesis that effects of lottery offers on test scores equal effects of offers on OLS-predicted value-added. Equality of (4) and (5) is a consequence of Proposition 4 in Newey and West (1987), which shows that Wald and LM tests of linear restrictions in linear generalized method of moments (GMM) models generate identical test statistics when the same residual variance estimate is used.

We can also write the LM statistic as

$$\hat{T} = \frac{((Y - \hat{\varphi}\hat{Y}) + (\hat{\varphi} - 1)\hat{Y})' P_Z ((Y - \hat{\varphi}\hat{Y}) + (\hat{\varphi} - 1)\hat{Y})}{\hat{\sigma}^2} \tag{6}$$

$$= \frac{(\hat{\varphi} - 1)^2}{\hat{\sigma}^2 (\hat{Y}' P_Z \hat{Y})^{-1}} + \frac{(Y - \hat{\varphi}\hat{Y})' P_Z (Y - \hat{\varphi}\hat{Y})}{\hat{\sigma}^2},$$

where $\hat{\varphi} = (\hat{Y}' P_Z \hat{Y})^{-1} \hat{Y}' P_Z Y$ is the two-stage least squares (2SLS) coefficient estimate from a procedure that uses Z_i to instrument $\hat{Y}_i$ in an equation for Y_i. Previous efforts to validate VAMs have focused on testing whether forecast coefficients of this type equal 1 (Chetty, Friedman, and Rockoff 2014; Deming 2014).[4] Equation (6) shows that $\hat{T}$ is the sum of two test statistics. The first is a Wald statistic testing whether the 2SLS forecast coefficient equals 1. (The denominator in this term is the variance of $\hat{\varphi}$.)

[2] Stratification can be accommodated by projecting structural and OLS residuals on stratifying variables and working with the residuals from this projection.

[3] As in Hausman (1983), $\hat{T} = NR^2$, where R^2 is the R-squared from a regression of $Y_i - \hat{Y}_i$ on Z_i.

[4] These applications involve more elaborate multi-step computations of the forecast coefficient that use transformations of lottery instruments and conventional VAM fitted values. The motivation for these procedures nevertheless appears to be the set of restrictions described by (3).

The second term is the Sargan (1958) statistic for an LM test of 2SLS overidentifying restrictions. The decomposition in (6) reveals that $\hat{T}$ combines a test of forecast bias, which checks the predictive accuracy of a particular weighted average of lottery-specific forecasts, with an overidentification test, which checks whether VAM estimates are equally predictive within every lottery.[5] Under H_0, the forecast coefficient test statistic has a limiting χ^2_1 distribution and the Sargan statistic has a limiting χ^2_{L-1} distribution.[6]

Tests based on the 2SLS forecast coefficient may be misleading when lotteries shift students across schools with similar value-added predictions. When a large collection of lottery dummies generate only small shifts in OLS value-added, the resulting many-weak instrument scenario produces a 2SLS estimate that is biased toward the corresponding OLS estimate. In this case, the OLS regression of Y_i on $\hat{Y}_i$ necessarily yields a coefficient of one. This suggests that a test based on the forecast coefficient alone may be biased against rejecting invalid VAMs when lottery offers have little effect on value-added. By contrast, $\hat{T}$ has the form of an Anderson and Rubin (1949) statistic, which has better finite-sample performance with many weak instruments (Stock and Wright 2000).

Finally, note that with $L < J$ (fewer lotteries than schools) there are necessarily alternatives to H_0 against which lottery-based tests of VAM validity have no power. This is a consequence of Proposition 1 in Newey (1985), which shows the inconsistency of GMM-based tests against general misspecification. In particular, for alternatives with $E[Z_i Y_i] = E[Z_i \hat{Y}_i]$ (in other words, when $E[Z_i D_i'(\beta - \alpha)] = 0$), the non-centrality parameter that determines the distribution of the test statistic under the alternative will be zero even while estimated VAM is biased.[7] This scenario arises, for example, when lotteries fail to change patterns of school enrollment, or when they only move students across sets of schools within which both causal and OLS value-added are constant.

[5] See Angrist and Pischke (2009) for the weighting formula implicit in overidentified 2SLS models.

[6] In practice, the Sargan and Wald statistics typically use the unrestricted variance estimate $\hat{\sigma}^2_U = (Y - \hat{\varphi}\hat{Y})'(Y - \hat{\varphi}\hat{Y})/N$ in place of $\hat{\sigma}^2$.

[7] Note that $E[Z_i Y_i] = E[Z_i(D_i'\beta + X_i'\gamma + \epsilon_i)] = E[Z_i D_i'\beta]$ since random offers are orthogonal to X_i and ϵ_i. Similarly, $E[Z_i \hat{Y}_i] = E[Z_i(D_i'\alpha + X_i'\Gamma)] = E[Z_i D_i'\alpha]$. Thus $E[Z_i Y_i] = E[Z_i \hat{Y}_i]$ is equivalent to $E[Z_i D_i'(\beta - \alpha)] = 0$.

III. Validating VAM in CMS

Our investigation of bias in VAM estimates for CMS is based on the Deming (2014) sample of 87,351 fourth to eighth grade students attending CMS schools between 1996 and 2004. We use this sample to estimate three OLS value-added models for the average of math and reading test scores: an "uncontrolled" model that adjusts only for year-of-test effects, a "lagged score" model that adds cubic polynomials in math and reading test scores from the previous grade, and a "gains" model that replaces the outcome variable with grade-to-grade test score changes in the uncontrolled model. Seats at CMS schools are assigned via a centralized matching mechanism that randomly breaks ties between students with the same preferences and priorities at oversubscribed schools, inducing a set of stratified admission lotteries. Our lottery sample is restricted to schools with at least 25 students subject to random assignment in the 2002–2003 school year. The resulting sample includes 2,213 students, each of whom picked 1 of 24 over-subscribed schools as a first choice.

Application of our test to CMS data reveals that, except for the most naïve VAM model, failures of over-identifying restrictions are a more important specification error than forecast bias. This can be seen in Table 1, the first three columns of which report results of the tests developed in Section II. As shown in column 1, the 2SLS forecast coefficient equals 0.109 for the uncontrolled model, an estimate that is statistically different from one. The overidentifying restrictions for this model are also rejected, and the joint test of all restrictions generates a stronger rejection than tests of either forecast bias or overidentification alone. At 136.58, the sum of forecast bias and overidentification test statistics (which use the unrestricted estimate of σ^2) generates a result qualitatively similar to that generated by the joint test statistic of 111.86 (computed using the restricted estimate).

Forecast coefficient estimates for the lagged score and gains models equal 0.848 and 0.960, estimates that are not statistically different from 1. These estimates are imprecise, however, with

Table 1—Tests for Bias in Charlotte-Mecklenburg Value-Added Models

	Assuming homoskedasticity			Allowing for heteroskedasticity		
	Uncontrolled (1)	Lagged score (2)	Gains (3)	Uncontrolled (4)	Lagged score (5)	Gains (6)
Forecast coefficient	0.109	0.848	0.960	0.119	0.969	1.074
	(0.088)	(0.576)	(0.792)	(0.075)	(0.543)	(0.791)
First stage *F*-stat.	13.46	11.75	11.50	19.46	11.87	9.09
Bias tests:						
Forecast bias (1 d.f.)	102.71	0.07	<0.01	137.07	<0.01	<0.01
	[<0.001]	[0.792]	[0.960]	[<0.001]	[0.955]	[0.925]
Overidentification (23 d.f.)	33.87	33.64	34.01	39.49	39.29	40.18
	[0.067]	[0.071]	[0.065]	[0.018]	[0.018]	[0.015]
All restrictions (24 d.f.)	111.86	34.33	34.56	147.03	39.32	40.17
	[<0.001]	[0.079]	[0.075]	[<0.001]	[0.025]	[0.021]
Sum of FC bias and overid.	136.58	33.71	34.02	176.57	39.30	40.19

Notes: This table reports estimates of the VAM forecast coefficient and the results of test for bias in value-added models for 4th–8th graders in Charlotte-Mecklenburg. The lottery sample includes 2,213 students applying to oversubscribed schools. The uncontrolled model includes only year-of-test indicators as controls, while the lagged score model adds cubic polynomials in baseline math and reading test scores. The gains model controls for year-of-test indicators and uses score gains from baseline as the outcome. Forecast coefficients are estimated by 2SLS (assuming homoskedasticity) or by two-step optimal IV (allowing for heteroskedasticity) using lottery offers as instruments. The full set of restrictions is tested using an unrestricted residual variance estimate. Standard errors are reported in parentheses; *p*-values are reported in brackets.

95 percent confidence intervals including values below zero and close to two. Moreover, the first stage *F*-statistics for these models equal 11.75 and 11.50, close to the rule-of-thumb threshold of 10 often used to diagnose weak instruments (Staiger and Stock 1997). This suggests that forecast bias tests for the lagged score and gains models may not be reliable. The overidentification and joint tests of all restrictions reject VAM validity for these models despite imprecision of the forecast coefficient. This highlights the value of looking at the full set of VAM restrictions as well as forecast bias.

Columns 4–6 of Table 1 report test results allowing for heteroskedasticity in ϵ_i. The forecast coefficients in these models are estimated using the efficient IV estimator introduced in White (1982). The estimated forecast coefficients in columns 4–6 are similar to those in columns 1–3, with results close to 1 in the lagged score and gains models. On the other hand, heteroskedasticity-robust overidentification test statistics are mostly larger than those imposing homoskedasticity. Robust standard errors for estimates of reduced form parameters (not reported in the table) are also smaller than those computed assuming homoskedasticity. These findings suggest the robust variance estimates are biased downwards.

IV. Summary and Conclusions

School admission lotteries offer the opportunity to validate school value-added models. The restrictions in the VAM framework can be checked by specification tests of the sort traditionally associated with simultaneous equation models. We show here that an omnibus test of the restrictions generated by admissions lotteries combines a test of forecast bias with a Sargan-style overidentification test. Applied to data from the Charlotte-Mecklenburg school district, our test rejects conventional value-added models, mostly because of a failure of the over-identifying restrictions implicit in the VAM framework.

VAMs that fail to pass an omnibus specification test may nevertheless be useful. In Angrist et al. (2015), we use a random coefficients model to quantify the joint distribution of causal value-added and OLS bias, and show how this model can be used to generate

improved value-added predictions that partially correct for bias. Estimates from Boston suggest that policies based on VAMs can generate large achievement gains even when the underlying estimates are biased. Hybrid value-added predictions that incorporate lottery information generate further gains. At the same time, rejections of the assumptions underlying conventional VAMs offer an important caution, and highlight the value of model assessment procedures that go beyond conventional specification testing.

REFERENCES

Anderson, T. W., and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics* 20 (1): 46–63.

Angrist, Joshua, Peter Hull, Parag A. Pathak, and Christopher Walters. 2015. "Leveraging Lotteries for School Value-Added: Testing and Estimation." National Bureau of Economic Research Working Paper 21748.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104 (9): 2593–2663.

Deming, David J. 2014. "Using School Choice Lotteries to Test Measures of School Effectiveness." *American Economic Review* 104 (5): 406–11.

Deutsch, Jonah. 2012. "Using School Lotteries to Evaluate the Value-Added Model." Unpublished.

Hausman, Jerry A. 1983. "Specification and Estimation of Simultaneous Equation Models." In *Handbook of Econometrics*, Vol. 1, edited by Zvi Griliches and Michael D. Intriligator, 391–448. Amsterdam: North-Holland.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." National Bureau of Economic Research Working Paper 14607.

Newey, Whitney K. 1985. "Generalized Method of Moments Specification Testing." *Journal of Econometrics* 29 (3): 229–56.

Newey, Whitney K., and Kenneth D. West. 1987. "Hypothesis Testing with Efficient Method of Moments Estimation." *International Economic Review* 28 (3): 777–87.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1): 175–214.

Rothstein, Jesse. 2014. "Revisiting the Impacts of Teachers." Unpublished.

Sargan, J. D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica* 26 (3): 393–415.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65 (3): 557–86.

Stock, James H., and Jonathan H. Wright. 2000. "GMM with Weak Identification." *Econometrica* 68 (5): 1055–96.

White, Halbert. 1982. "Instrumental Variables Regression with Independent Observations." *Econometrica* 50 (2): 483–99.