Abstract Title Page

Title: Test Score Measurement Error, Short-term Knowledge, and Lagged Dependent Variables in Models of the Education Production Function

Authors and Affiliations:

Brian Stacy, Michigan State University, RAND Corporation

J.R. Lockwood, RAND Corporation

Daniel McCaffrey, RAND Corportation

Background

Researchers and policymakers are interested in the causal effects of educational inputs on student achievement. Unfortunately, it is not possible to directly observe student learning, so test score data is often used as an approximate measure. To measure their achievement at a given point in time (e.g., in the spring of the school year) students typically complete standardized tests composed of around forty to fifty questions per subject over one or two days of the school year. Given the small number of items, the test is an incomplete measure of students' achievement. In addition, students can get sick during testing, be distracted, or can cram, either on their own or through their teachers, all of which will affect their scores. Guessing is also an issue due to the small number of items on the tests. Combined, these factors mean that test scores are a noisy measure of a student's true level of knowledge, and so estimation of causal effects may be affected.

As discussed by Todd and Wolpin (2003), one class of models that are particularly vulnerable to measurement error are those that include lagged test scores as regressors. Under classical measurement error assumptions, the estimate of the coefficient on the lagged test score will suffer from attenuation bias. Even if this coefficient is not of interest, the bias can be transmitted to all other estimates, including estimates of the effects of teachers or other educational inputs, with the magnitude of the bias depending on how strongly correlated these variables are with the lagged test score. Estimates of the effects of schooling inputs are then rendered inconsistent.

Many solutions to the errors in variables problem have been proposed in the econometrics and statistics literature. Todd and Wolpin (2003) propose using further lagged test scores as instrumental variables for the once lagged score in the model. Andrabi, Das, Khwaja, and Zajonc (2009) make use of other subject scores as IVs. Another approach is to use known test score measurement error variances, computed by testing companies, to correct the attenuation, as implemented by the Value-Added Research Center at the University of Wisconsin-Madison (Value-Added Research Center, 2010).

It is also quite common to ignore the measurement error issue in estimation. This may not be problematic as long as the measurement error bias is small and may be a reasonable course of action if there is little evidence that the estimates of parameters of interest are noticeably affected.

It is theoretically possible for measurement error to be completely harmless to estimation, even in models with lagged test scores. With the right serial correlation structure or if there are dynamic effects of measurement error of exactly the right size, then the measurement error terms can cancel in the structural equation and estimates will become bias free but these conditions have not been empirically studied.

Objective

This paper investigates the impact of measurement error on the estimation of parameters in the education production function. The primary research questions are:

1. Are there conditions under which measurement error is harmless and do those conditions hold?
2. Does the bias introduced produce statistically and practically significant differences in estimates?
3. What solutions are available and are they valid and practical?
4. What implications are there for estimates of other parameters in the education production function, including teacher effect estimates?
5. How can measurement error affect some specification tests of the education production function?

Setting

The data used in this paper are a panel dataset from a countywide school system in an anonymous southern state. They include student-teacher links, achievement test scores from mathematics, reading, langauge arts, science and social studies, and student-level characteristics such as gifted or special education status, free and reduced price lunch status, race and gender. In addition, we have detailed information on test score measurement error in the form of standard errors of measure, which determines the variability of measurement errors attributable to the small sample of items and guessing on the tests, and which can be used to remove bias in the estimates.

Population

The primary sample includes 33,522 observations covering grades four through eight for three cohorts of students linked to 1,311 teachers. The data cover school years 2007-2010. We focus on math scores, although we repeat the analysis for English language arts. The data includes all students with three lagged math scores as well as a current and past year teacher link.

(Insert Table 1 Here)

Significance and Novelty of study

The paper adds to the literature in four primary ways:

1. A distinction is made between measurement error caused by the limited number of item or guessing during testing, and short-term knowledge fluctuations due to sickness, distractions, or cramming, which has implications for the different estimators studied

2. A discussion of dynamic effects or serial correlation in the measurement error and measurement error spillovers to other subjects

3. A novel estimator for the education literature that involves removing bias using standard errors of measure in a first difference with instrumental variables estimator

4. A discussion and empirical example of how measurement error can make testing the

education production function more difficult.

Econometric Model

This paper focuses its attention on two commonly used estimators of the parameters in the following model:

$$A_{it} = \tau_t + \lambda A_{i,t-1} + X_{it}\beta + c_i + e_{it} \qquad (1)$$

where A_{it} is achievement of student i in year t, X_{it} is a vector of current year inputs including both schooling and non-schooling inputs, c_i is unobserved student ability, and e_{it} is an error term.

The first estimator simply estimates the parameters in the model using OLS, which is refered to as "DOLS" (dynamic ordinary least squares). The second uses instrumental variables after first differencing to remove the ability term, which is refered to as "IVFD" (instrumental variables after first differencing). Both use observed test scores instead of true achievement and so are potentially biased.

We offer two alternate estimators for both DOLS and IVFD that take steps to correct measurement error. The first uses other subjects and further lagged scores as IVs, which we refer to as the "Robust IV" method. The other makes use of known standard errors of measure to essentially subtract off the measurement error variance from the regression cross-product matrix in the case of DOLS, or the matrix with fitted values in the IVFD case. We refer to this as the "Corrected IV" method.

Usefulness / Applicability of Method

Our two alternate estimators apply to widely available data. The Robust IV estimator can be used as long as either another subject or further lagged scores are available for each student. The Corrected IV estimator requires an estimate of the variance of the standard error of measure, which is available from testing companies.

Results

We begin by formally testing whether the measurement error is biasing estimates. We do this by comparing a naïve estimator that ignores measurement error, to an estimator that can consistently estimate the parameters of the structural model in Equation (1) in either the case of no measurement error terms or with measurement error terms. Under a null of no measurement error, the estimates should differ only by sampling variation. We find a statistically significant difference in the estimates, so we conclude that measurement error biases estimates. The difference is also large in a practical sense.

We also compare estimates using the Robust IV and Corrected IV methods with those generated by a naïve method that completely ignores potential measurement error bias. We find a consistent pattern of larger estimates of λ using the methods that account for measurement error. This pattern holds for both the DOLS and the IVFD estimators and in both mathematics and English language arts.

(Insert Table 2 and Table 3 Here)

We conducted a “variable-addition” test of the model by adding a second lagged score to the model and testing the significance of its coefficient. It is theoretically possible that measurement error could cause a false rejection of this test, showing that a second lag is significant when in fact the coefficient is zero. We perform this test naïvely ignoring measurement error, and repeat it taking steps to correct estimation for measurement error. We find that the naïve estimates give a strongly significant rejection, while the estimates from the alternative estimator are insignificant, although the alternative estimates are much less precise.

(Insert Table 4 Here)

We compare estimates of teacher effects using the multiple methods and find some differences. In the DOLS environment, we find correlations of around .8 comparing the naïve estimator to the alternate estimators. In the IVFD, the results are more promising, with correlations between the naïve and the alternative estimators around .99.

Conclusions:

Estimation done without accounting for measurement error has the potential to be costly. We find evidence that estimates of the coefficient on the lagged achievement measure are biased, and so all estimates in the model have the potential to be biased. The bias will depend on the degree of correlation between lagged achievement and the other variables.

We study two techniques which can help reduce the bias in the estimates: an IV estimator that uses other subjects and/or further lags and an estimator that corrects the regression cross-product matrix by subtracting off known measurement error variances. Both are widely available to researchers and practitioners.

It is possible that naïve methods that ignore measurement error are still a wise choice for estimation. For estimates of teacher effects, we find moderately strong to very strong correlations among naïve and alternative estimates that correct for measurement error, but the naïve estimates are more precise. There is a tradeoff between efficiency and consistency, and further study is needed to determine which estimator is preferable. At this point, we can offer a caution while conducting estimation and recommend robustness checks using the alternative methods proposed in this paper.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Andrabi, T., Das, J., Khwaja, A.I., Zajonc, T., (2009). Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *The World Bank: Policy Research Working Paper*. Downloaded on September 30, 2011, at http://www.hks.harvard.edu/fs/akhwaja/papers/ value_added_jan27.pdf

Fuller, W. (1987). Measurement Error Models. Hoboken, NJ: John Wiley & Sons, Inc.

Harris, D., Sass, T., Semykina, A. (2010). Value-Added Models and the Measurement of Teacher Productivity. *Calder Working Paper.* 54. Downloaded on September 30, 2011 at http://www.urban.org/publications/1001508.html

Jacob, B., Lefgren, L., Sims, D. (2009) The Persistence of Teacher-Induced Learning. The *Journal of Human Resources.* 45 no. 4, 915-944. Downloaded on September 30, 2011 at http://jhr.uwpress.org/content/45/4/915.short

Todd, P., Wolpin, K. (2003) On The Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal.* 113, F3-F33. DOI: 10.1111/1468-0297.00097

Value-Added Research Center (2010). NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model. Madison Wisconsin: Wisconsin Center for Education Research. Downloaded on September 30, 2011, at http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5-43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf

Appendix B. Tables and Figures
Not included in page count.

Table 1: Summary Statistics from Primary Sample

VARIABLES	N	mean	sd	min	max
Limited English Proficiency	33,705	0.0166	0.128	0	1
Special Education	33,705	0.101	0.302	0	1
Free and Reduced Priced Lunch	33,705	0.260	0.439	0	1
Gifted	33,705	0.0296	0.169	0	1
Asian	33,705	0.103	0.304	0	1
Black	33,705	0.260	0.438	0	1
Hisp	33,705	0.0867	0.281	0	1
White	33,705	0.515	0.500	0	1
Male	33,705	0.499	0.500	0	1
Math Score	33,705	0.315	0.956	-2.701	4.084
Language Arts Score	33,696	0.314	0.942	-3.575	3.329
Reading Score	33,704	0.293	0.941	-3.237	4.002
Science Score	33,662	0.316	0.944	-4.141	4.992
Social Studies Score	18,367	0.295	0.960	-2.465	4.428
Math Std Error of Measure	33,705	0.316	0.170	0.183	1.624
Language Arts Std Error of Measure	33,695	0.412	0.230	0.220	1.834
Reading Std Error of Measure	33,704	0.444	0.230	0.246	2.297
Science Std Error of Measure	33,658	0.322	0.122	0.202	1.638
Social Studies Std Error of Measure					
Class Size		25.709	33.726	1	229

Table 2: Results for IVFD: Comparing Naïve to Alternative Estimators

				IVFD w/	
	IVFD w/			Larger	
Estimator	Smaller		Estimator	Dataset	
Naive	0.039	Obs=33522	Naive	0.068	Obs=47919
	(.013)			(.011)	
Corrected	0.219		Corrected	0.35	
	(.105)			(.090)	
Robust	0.181		Robust	0.193	
	(.091)			(.051)	
Robust-Naive	0.143		Robust-Naive	0.125	
	(.087)	P-value=.078		(.044)	P-value=.004

Uses Math achievement scores. Estimates of lambda shown from model above. Std Errors computed using panel bootstrapping procedure. Naive estimator refers to IVFD using Math lag2 as IV. Corrected uses Math lag2 as IV, but fixes estimates using known Std Errors of Measure. Robust in smaller dataset uses Math lag3 as well as Reading and Language Arts lag2 and lag3. Robust in larger dataset uses Reading and Language Arts lag2 as IVs.

Table 3: Results for DOLS: Comparing DOLS to Alternative Estimators

	Levels w/ Lag 3			Levels w/ out	
Estimator	Data			Lag 3	
DOLS	0.606	Obs=33522	DOLS	0.602	Obs=47919
	(.005)			(.004)	
Corrected	0.893		Corrected	0.891	
	(.008)			(.007)	
Robust	0.946		Robust	0.915	
	(.006)			(.005)	
Robust-Naive	0.34		Robust-Naive	0.313	
	(.007)	P-value=.000		(.006)	P-value=.000

Uses Math achievement scores. Estimates of lambda shown from model above. Std Errors computed using panel bootstrapping procedure. DOLS estimator refers to OLS on the above model. Corrected fixes DOLS estimates using known Std Errors of Measure. Robust in smaller dataset uses Math lag2 and lag3, Reading, and Language Arts lag1, lag2, and lag3. Robust in larger dataset uses Math lag2 as well as Reading, and Language Arts lag1 and lag2 as IVs.

Table 4: Test of Further Lags: Naive vs Robust Estimation

Math Results		
Variable	Naive Estimate	Robust Estimate
$\Delta A_{i,t-1}$	.0563*** (.019)	.245*** (.093)
$\Delta A_{i,t-2}$	.017* (.010)	.028 (.035)
Language Arts Results		
$\Delta A_{i,t-1}$	.123*** (.019)	.268*** (.082)
$\Delta A_{i,t-2}$	.048*** (.011)	.038 (.036)

Cluster robust std errors in parenthesis. Naive IVFD uses Math lag2 as IV for $\Delta A_{i,t-1}$. Robust IVFD uses Reading and Language lag2 and lag3 as IVs for both $\Delta A_{i,t-1}$ and $\Delta A_{i,t-2}$.

Testing: $H_o : \gamma = 0$ for
$\Delta A_{i,t} = \Delta\alpha_t + \lambda\Delta A_{i,t-1} + \gamma\Delta A_{i,t-2} + \Delta X_{i,t}\ \beta + \Delta\eta_{it}$