CALDER

NATIONAL CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

Tracking Every Student's Learning Every Year

Urban Institute

a program of research from
The Urban Institute with:
Duke University
Stanford University
University of Florida
University of Missouri-Columbia
University of Texas at Dallas
University of Washington

BRIEF 2
NOVEMBER 2008

Overview of Measuring Effect Sizes: The Effect of Measurement Error

DON BOYD, PAM GROSSMAN, HAMP LANKFORD, SUSANNA LOEB, AND JIM WYCKOFF

The use of value-added models in education research has expanded rapidly. These models allow researchers to explore how a wide variety of policies and measured school inputs affect the academic performance of students. An important question is whether such effects are sufficiently large to achieve various policy goals. For example, would hiring teachers having stronger academic backgrounds sufficiently increase test scores for traditionally low-performing students to warrant the increased cost of doing so? Judging whether a change in student achievement is important requires some meaningful point of reference. In certain cases a grade-equivalence scale or some other intuitive and policy relevant metric of educational achievement can be used. However, this is not the case with item response theory (IRT) scale-score measures common to the tests usually employed in value-added analyses. In such cases, researchers typically describe the impacts of various interventions in terms of *effect sizes*, although conveying the intuition of such a measure to policymakers often is a challenge.

The *effect size* of an independent variable is measured as the estimated effect of a one standard deviation change in the variable divided by the standard deviation of test scores in the relevant population of students. Intuitively, an effect size represents the magnitude of change in a variable of interest, e.g., student achievement, resulting from a one standard deviation, or rather large, change in another variable, e.g., class-size. Effect size estimates derived from value-added models employing administrative databases typically are quite small. For example, in several recent papers the average effect size of being in the second year of teaching relative to the first year, other things equal, is about 0.04 standard deviations for math achievement and 0.025 standard deviations for reading achievement, with variation no more than 0.02. Additional research examines the effect sizes of a variety of other teacher attributes: alternative certification compared to traditional certification (Boyd et al. 2006; Kane et al. in press); passing state certification exams (Boyd et al. 2008; Clotfelter et al. 2007; Goldhaber 2007); National Board Certification (Clotfelter et al. 2007; Goldhaber and Anthony 2007; Harris and Sass 2007); and ranking of undergraduate college (Boyd et al. 2008; Clotfelter et al. 2007).

As one example, consider results from a recent paper analyzing how various attributes of teachers affect the test-score gains of their students (Boyd et al. 2008). Parameter estimates reflecting the effects of a subset of the teacher attributes included in the analysis are shown in the first column of table 1. These estimated effects, measured relative to the standard deviation of observed student achievement scores, indicate that none of the estimated effect sizes are large by standards often employed by educational researchers in other contexts (see Hill et al. 2007). However, most observers believe that the difference between first- and second-year teachers is meaningful. The effect of not being certified, and the effect of a one standard deviation increase in math SAT scores, are comparable to about two-thirds of the gain that accrues to the first year of teaching experience.

While specific attributes of teachers are estimated to have small effects, researchers and policymakers agree that high-quality teachers have large effects on student learning so that effectively choosing teachers can make an important difference

Table 1. Estimated Effect Sizes for Teacher Attributes Model for Math Grades 4 & 5, NYC 2000-2005

	Estimated effects relative to			
	S.D. of observed score	S.D. of observed score gain	S.D. of universe score	S.D. of universe score gain
First year of experience	0.065	0.103	0.072	0.253
Not certified	-0.042	-0.067	-0.046	-0.162
Attended competitive college	0.014	0.022	0.016	0.054
One S.D. increase in math SAT score	0.041	0.065	0.045	0.158
All observable attributes of teachers	0.162	0.256	0.179	0.631

in student outcomes (Sanders and Rivers 1996; Aaronson, Barrow, and Sander, 2003; Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Kane, Rockoff, and Staiger in press). The findings that teachers greatly influence student outcomes but that measures of teacher qualifications seem to matter little, taken together, have led some observers to conclude that attempting to differentiate teachers on their pre-employment credentials is of little value. Rather, they argue, education policymakers would be better served by reducing educational and credential barriers to enter teaching in favor of more rigorous performance-based evaluations of teachers. Indeed, this perspective appears to be gaining some momentum.[1] Thus, the perception that many educational interventions have small effect sizes, as traditionally measured, is having important consequences for policy.

Why might the effect sizes of teacher attributes computed from administrative databases appear so small? A variety of factors could cause estimates of the effects of teacher attributes to appear to have little or no effect on student achievement gains, even if in reality they do. These include: measures of teacher attributes available to researchers are probably weak proxies for the underlying teacher characteristics that influence student achievement; measures of teacher attributes often are made many years before we measure the link between teachers and student achievement gains; high-stakes achievement tests may not be sensitive to differences in student learning resulting from teacher attributes; and multicolinearity resulting from the similarity of many of the commonly employed teacher attributes. We believe that each of these contributes to a diminished perceived importance of measured teacher attributes on student learning. In this paper, we focus on two additional issues pertaining to how effect sizes are measured, which we believe are especially important. First, we argue that estimated model coefficients should be compared to the standard deviation of gain scores, not the standard deviation of scores. Second, it is important to account for test measurement error when calculating effect sizes.

MEASURING EFFECTS RELATIVE TO THE STANDARD DEVIATION OF GAIN SCORES

At a point in time, a student's academic achievement will reflect the history of all those factors affecting the student's cumulative, retained learning. This includes early childhood events, the history of family and other environmental factors, the historical flow of school inputs, etc. The dispersion (e.g., standard deviation) in the skills and knowledge of students at a point in time reflects the causal linkages between all such factors and how these varied and long-run factors differ across students. From this perspective, it is not surprising that estimated effect sizes are small, as almost any short-run intervention—say a particular feature of a child's education during one grade—is likely to move a student by only a modest amount up or down in the distribution of cumulative student achievement. Of course, in part this depends upon the extent to which the test focuses on current topics covered, or draws upon prior knowledge and skills.

The nature of the relevant comparison depends upon the question. For example, if policymakers want to invest in policies that provide at least a minimum year-to-year student achievement growth, for example to comply with NCLB in a growth context, or if policymakers wanted a benchmark of progress relative to improvements common in a year, then the relevant metric is the standard deviation of the gain in achievement. However, if policymakers are interested in the extent to which an intervention may close the achievement gap, then comparing the effect of that intervention to the standard deviation of achievement provides a better metric of improvement. Even in the latter case, it is important to keep in mind that interventions often are short lived when compared to the period over which the full set of factors affect cumulative achievement.

The effect of employing the standard deviation in test score gains rather than the standard deviation of test scores can be seen by comparing column 2 to column 1 in table 1. Estimated effect sizes measured relative to the standard deviation of score gains are 59 percent larger than those based on the standard deviation of observed scores.

ACCOUNTING FOR TEST MEASUREMENT ERROR

The distribution of observed scores from a test of student achievement will differ from the distribution of true student learning because of the errors in measurement inherent in testing.[2] In particular, the variance in test scores in the population of students of interest can be shown to equal $\sigma_S^2 = \sigma_\tau^2 + \sigma_\eta^2$ where σ_τ^2 is the variance measuring the dispersion in true achievement and σ_η^2 is the variance in scores attributable to test measurement error. Psychometricians long have worried how such measurement error impedes the ability of educators to assess the academic achievement, or growth in achievement, of individual students and groups of students. This measurement error is less problematic for researchers carrying value-added analyses where test scores, or gain scores, are the outcomes of interest, as the measurement error will only affect the precision of estimates, a loss in precision (but not consistency) which can be overcome with sufficiently large numbers of observations.[3]

Even though test measurement error does not complicate the estimation of how a range of factors affect student learning, such errors in measurement do have important implications when judging the sizes of those estimated effects. As noted above, the sizes of estimated effects typically are judged relative to the standard deviation of observed scores, σ_S, or the standard deviation of observed gain scores. From the perspective that the estimated effects shed light on the extent to which various factors can explain systematic differences in student learning, not test measurement error, the sizes of those effects should be judged relative to the standard deviation of true achievement, σ_τ, or the standard deviation of gains in true achievement. (As argued above, in many cases it is the latter that is pertinent.) It is the size of an estimated effect relative to the dispersion in true achievement or the gain in true achievement that is of interest. From this perspective, effect sizes as traditionally measured have led analysts to understate the magnitudes of effects because the standard deviation of observed scores overstates the dispersion of true achievement in the student population.

Adjusting estimates of effect-size to account for these considerations is straightforward if one knows the extent of test measurement error. Technical reports provided by test vendors typically only provide information regarding the measurement error from a subset of possible sources. However, there are a number of other factors, including variation in scores resulting from students having particularly good or bad days, which can result in a particular test score not accurately reflecting true academic achievement.[4]

Using the covariance structure of student test scores across grades three through eight in New York City from 1999 to 2007, we estimate the overall extent of test measurement error and how measurement error varies across students. Our estimation strategy follows from two key assumptions: (1) there is no persistence (correlation) in each student's test measurement error across grades, and (2) there is at least some persistence in true achievement, with the degree of persistence constant across grades.[5]

Employing the covariance structure of test scores for NYC students and alternative models characterizing the growth in academic achievement, we find estimates of the overall extent of test measurement error to be quite robust, with our lowest estimate of the overall test measurement error variance indicating that roughly 17 percent of the variance in student test scores is attributable to test measurement error.[6] Because test score gains

Table 2. Average Qualifications of Teachers in Poorest Quartile of Schools by Math Achievement Quintiles Predicted Solely Based on Teacher Qualifications (excluding experience), 2000–20005

Value-added quintile	Mean value added	Not certified	LAST pass first	LAST score	Math SAT	Verbal SAT	College ranking competitive or higher
1	-0.068	0.731	0.46	227	355	440	0.101
2	-0.032	0.141	0.656	239	414	467	0.121
3	-0.01	0.076	0.779	245	423	462	0.224
4	0.01	0.031	0.851	252	450	470	0.352
5	0.045	0.013	0.908	254	512	474	0.494
Range	.113	-0.718	0.448	27	157	34	0.393

have measurement error in both pre- and post-tests, and gains in actual achievement are smaller than levels of achievement, measure error is a much greater proportion of the variance in test score gains. We estimate about 84 percent of the variance in gain scores is attributable to measurement error. This result underscores the problem in using test score gains for individual students, or small groups of students, as indicators of actual achievement gains—large parts of observed test-score differences frequently will merely reflect test measurement error.

In contrast, measurement error being a large portion of the total variation in gain scores across students does not create a problem in quantifying the magnitudes of estimated effects relative to the dispersion in true achievement gains. Even if test measurement error limits our ability to make inferences regarding the true achievement gains for individual students, it need not limit our ability to accurately estimate the distribution of true achievement or the distribution of true achievement gains. In fact, a central contribution of our measurement error paper is to demonstrate how credible estimates of the standard deviation in the overall dispersion in true achievement gains can be obtained, thus allowing the magnitudes of estimated effects to be judged relative to the overall dispersion in true achievement gains.

Again consider table 1. The effect of accounting for measurement error in test scores is shown in column 3 and the joint effect of employing gains scores and accounting for measurement error is shown in column 4.

Accounting for test measurement error increases the effect size estimates for teacher attributes in both instances, but the interaction of employing gain scores and accounting for measurement error increases effect sizes four-fold relative to estimates typically reported. For example, the effect of a student having a second year teacher, rather than a teacher having no prior experience, is estimated to be over a quarter of a standard deviation in the true achievement gain experienced by students. Although somewhat smaller, the effect of having an uncertified teacher, or a teacher with a one standard deviation lower math SAT, is 16 percent of the standard deviation of the gain in achievement net of measurement error.

Boyd et al. (2008) also examine the joint effect of all observable attributes of teachers, by using the estimated model to predict the value-added for each student based only on the observed teacher attributes included in the estimated model, holding teacher experience and all non-teacher variables constant. The teachers in the poorest quartile of New York City schools are divided into quintiles based on their predicted value-added. As shown in the second column of table 2, the difference in mean estimated teacher effects between teachers in the highest and lowest quintiles is 0.11 (0.18 when experience is not held constant), measured relative to the standard deviation of observed scores. When the estimated effect is adjusted to account for test measurement error, the effect size is almost half a standard deviation of the true achievement gains. As shown in columns 3–8 of table 2, the meaningful difference in teacher value added is systematically related to teacher attributes—

attributes that many have concluded are unrelated to teacher effectiveness. We see that only one percent of the teachers in the top quintile of effectiveness are not certified, compared to 73 percent in the bottom quintile. The more effective teachers are less likely to initially have failed the general knowledge certification exam and more likely to have higher scores on this exam as well as on the SAT. Furthermore, almost half of the teachers in the most effective quintile graduated from a college ranked competitive or higher by Barron's, compared to only ten percent of the teachers in the least effective quintile. These differences in effectiveness and teacher qualifications reflect differences *within* the poorest quartile of New York City schools. Given the systematic sorting of teachers between high-poverty and other schools, the differences in teacher effects and attributes likely would be larger had we considered teachers in all NYC schools. At least in this case there are important differences in teacher effectiveness that are systematically related to observed teacher attributes.

Measuring effect sizes relative to the dispersion in gain scores net of test measurement error will result in all the estimated effect sizes implied by an estimated model being larger by the same multiplicative factor, so that the relative sizes of effects will not change. Such relative comparisons are important in cost-effectiveness comparisons where the effect of one intervention is judged relative to some other. However, even here there will be a need to account for test measurement error when the estimated effects are drawn from multiple studies employing different tests, possibly having varying degrees of measurement error. As noted in the introduction, the absolute magnitudes of effect sizes for measurable attributes of teachers are relevant in the formulation of optimal personnel (e.g., hiring) policies. More generally, the absolute magnitudes of effect sizes are relevant in cost-benefit analyses and when making comparisons across different outcome measures. In such cases, accounting for test measurement error is important.

In conclusion, this brief has shown that accounting for measurement error meaningfully increases effect size estimates associated with teacher attributes. These effects are substantially magnified when accounting for measurement error in achievement gains, as we argue is often relevant. It is equally important to account for test measurement error when estimating how other intervention affects student achievement. More generally, accounting for measurement error in the computation of effect sizes is important in non-education settings as well.

REFERENCES

Abowd, J. M., and D. Card. 1989. "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* 57(2): 411–45.

Aaronson, D., L. Barrow, and W. Sander. 2003. "Does Teacher Testing Raise Teacher Quality? Evidence from State Certification Measurements," Working Paper. Research Department Federal Reserve Bank of Chicago.

Boyd, D., H. Lankford, S. Loeb, J. Rockoff, and J. Wyckoff. 2008 "The Narrowing Gap in New York City Teacher Qualifications and its Implications for Student Achievement in High-Poverty Schools." *Journal of Policy Analysis and Management* 27(4): 793–818.

Boyd, D., P. Grossman, H. Lankford, S. Loeb, and J. Wyckoff. 2006. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Education Finance and Policy* 1(2).

Clotfelter, C., H. Ladd, and J. Vigdor. 2007. "How and Why Do Teacher Credentials Matter for Student Achievement?" CALDER working paper.

Goldhaber, D. 2007 "Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness?" *Journal of Human Resources* 42(4): 765–94.

Goldhaber, D., and E. Anthony. 2007. "Can Teacher Quality Be Effectively Assessed? National Board Certification as Signal of Effective Teaching." *The Review of Economics and Statistics* 89(1): 134–50.

Gordon, R., T. Kane, and D. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." The Hamilton Project Discussion Paper 2006-01. Brookings Institution.

Harris, D., and T. Sass. 2007. "The Effects of NBPTS-Certified Teachers on Student Achievement." CALDER working paper.

Hill, C., H. Bloom, A. Black, and M. Lipsey. 2007. “Empirical Benchmarks for Interpreting Effect Sizes in Research.” MDRC Working Paper.

Kane, T., J. Rockoff, and D. Staiger. In press. “What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City.” *Economics of Education Review*.

Rivkin, S., E. Hanushek, and J. Kain. 2005. “Teachers, Schools, and Academic Achievement,” *Econometrica* 73(2): 417–58.

Rockoff, J. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review* 94(2): 247–52.

Sanders, W. and J. Rivers. 1996. “Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.” Working paper, University of Tennessee Value-Added Research and Assessment Center.

NOTES

[1] See, for example, Gordon, Kane, and Staiger (2006).

[2] From the perspective of classical test theory, an individual’s observed test score is the sum of two components, the first being the *true score* representing the expected value of test scores over some set of test replications. The second component is the residual difference, or random error, associated with test measurement error. Generalizability theory, which we draw upon in the paper, extends test theory to explicitly account for multiple sources of measurement error. The *universe score* in generalizability theory is similar to the true score defined above. To avoid technical matters, this policy brief employs phrases such as “true achievement” and “actual achievement”. These terms should be interpreted as being equivalent to the “universe score.”

[3] It is important to note that the two central issues discussed in the paper—measuring effects relative to the standard deviation of gain scores and accounting for test measurement error—are relevant regardless of how one estimates causal effects (e.g., randomized trials, quasi-experiments, or other regression based methods). The central issue is how one should judge whether estimated effects are large or small.

[4] Thorndike (1951) provides a useful, detailed classification of factors that contribute to test measurement error.

[5] Our estimation strategy draws upon an approach developed by Abowd and Card (1989) to study the covariance structure of individual- and household-level earnings over time, accounting for both permanent and transitory components in earnings where the latter includes error in the measurement of earnings.)

[6] In contrast, information in the technical reports for the New York tests imply that approximately 10 percent of the variation in test scores reflects measurement error from the sources analyzed. Both our empirical results and theoretical considerations support the proposition that the overall extent of test measurement error is substantially larger than that reported in the technical reports provided by test providers.

ABOUT THE AUTHORS

Don Boyd is Deputy Director of the Center for Policy Research at SUNY, Albany, where his research work focuses on teacher labor markets. Dr. Boyd is a part of the CALDER New York team.

Pam Grossman is Professor and Chair of Curriculum and Teacher Education in the School of Education at Stanford University. Her research focuses on teacher professional education.

Hamp Lankford is Professor of Educational Administration, Policy & Economics at SUNY, Albany, and part of the CALDER New York team.

Susanna Loeb is Professor of Education at Stanford University, specializing in the economics of education. Dr. Loeb oversees the CALDER New York team.

Jim Wyckoff is Professor of Education at the Curry School of Education at the University of Virginia. He is a senior researcher with CALDER.

THE URBAN INSTITUTE
2100 M Street, N.W.
Washington, D.C. 20037

Phone: 202-833-7200
Fax: 202-467-5775
http://www.urban.org

National Center for Analysis of Longitudinal Data in Education Research

IN THIS ISSUE

Overview of Measuring Effect Sizes:
The Effect of Measurement Error

For more information, call Public Affairs at 202-261-5709 or visit our web site, http://www.urban.org.

To order additional copies of this publication, call 202-261-5687 or 877-UIPRESS, or visit our online bookstore, http://www.uipress.org.

National Center for Analysis of Longitudinal Data in Education Research

This research is part of the activities of the National Center for the Analysis of Longitudinal Data in Education Research (CALDER). CALDER is supported by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education. More information on CALDER is available at http://www.caldercenter.org.

The views expressed are those of the authors and do not necessarily reflect the views of the Urban Institute, its board, its funders, or other authors in this series. Permission is granted for reproduction of this file, with attribution to the Urban Institute.