Abstract Title Page
Not included in page count.

Title: Using State or Study-Administered Achievement Tests in Impact Evaluations

Author(s): Robert B. Olsen, Fatih Unlu, and Andrew P. Jaciw

Abstract Body

Limit 5 pages single spaced.

Background/context:

Many educational evaluations estimate an intervention's impacts on student achievement. A key question in designing \these evaluations is how to measure achievement outcomes. Historically, most evaluations of interventions that have been developed to boost student achievement have administered tests, such as"off-the-shelf" standardized tests (e.g., Stanford 9 or Terra Nova) to students in the study sample. However, testing students is costly and burdensome.

Under No Child Left Behind, public schools are required to administer state reading and mathematics tests to students in selected grades. Scores from these tests can be used to measure achievement in educational evaluations. A small number of IES-funded evaluations rely on state tests, including an ongoing evaluation of charter schools, an evaluation of teacher induction programs (Glazerman et al., 2008), an evaluation of teacher professional development in early reading (Garet et al., 2008), and the recently completed evaluation of the impacts of the Student Mentoring Program (Bernstein et al., 2009). However, most of the 25 or so IES-funded impact evaluations have administered standardized achievement tests as part of the evaluation.

There are at least four reasons for evaluators to consider relying less on study-administered tests and more on state tests to measure student achievement. First, it would reduce the burden on students who already face substantial testing requirements. Second, it could substantially reduce the costs of conducting evaluations: testing students is expensive, while collecting state test scores is relatively inexpensive by comparison. Third, scores on state tests can have consequences for the students who take them, so they may elicit a higher level of effort than the "no-stakes" tests that studies administer. Fourth, studies based on state tests may yield results that resonate with policymakers who are focused on the state standards to which state tests have been aligned.

At the same time, there are still unanswered questions about using state tests to measure student achievement in education evaluations. One question is whether we should expect the impact estimates to be larger or smaller if studies rely on state tests instead of study-administered standardized tests. Presumably, the answer will vary on a case by case basis, and will depend largely on which test is more closely aligned to the intervention being evaluated. In general, we would expect larger impacts from measures that are more closely aligned to the treatment than from measures that are less closely aligned to the treatment.

Another unanswered question about using state tests involves statistical power—or, more fundamentally, the sample size required to achieve a pre-specified Minimum Detectable Effect Size (MDES). Commonly cited papers that help evaluators to determine minimum sample size requirements tend to rely either on standardized tests that are not part of state accountability systems (e.g., Hedges and Hedberg, 2007) or on pre-NCLB district-required tests (e.g., Bloom, Richburg-Hayes, and Black 2007). Furthermore, key parameters that determine the minimum sample size requirements, such as intra-class correlations and R-squares from regression models, can vary across measures, and there is no guarantee that these parameter estimates provide accurate guidance on the sample size requirements of studies that rely on state tests.

Therefore, the possibility of using state tests in education evaluations provides both opportunities for substantial cost savings along with some challenges and unknowns about the implications of relying on state tests.

Purpose / objective / research question / focus of study:
Description of what the research focused on and why.

This report, which has been prepared by Abt Associates for the Institute of Education Sciences' National Center for Education Evaluation and Regional Assistance, takes an important first step in sorting out the implications of relying on state tests for general, student-level measures of reading and math achievement in evaluations of educational effectiveness. More specifically, this report provides empirical evidence that may help evaluation designers decide whether to rely on state tests to measure student achievement0 in studies of educational effectiveness.

This study is designed to inform some of the choices that evaluation designers face with respect to state and study-administered standardized tests for a broad but well-defined class of study designs. This class, which covers many IES-sponsored evaluations, has two key defining features: (1) randomization (at either the student or cluster level) and (2) measurement of student achievement at two points in time—at "baseline" (prior to the implementation of the intervention) and (2) at "follow-up" (at least one point after the intervention has been implemented).

This study is designed to address three questions:

Question 1: Will impact evaluations in education yield different impact estimates and standard errors if they rely entirely on state tests—that is, they use state tests to measure achievement at both baseline and follow-up—instead of administering standardized tests at both points in time? If multiple studies find systematically larger or smaller impacts from state test than from study-administered tests, this would warrant a serious examination of the coverage of the two types of tests to try to explain the differences. Regardless of the explanation, an empirical regularity that impacts on state test scores are smaller or larger than impacts on study-administered test scores would affect our interpretation of impact estimates from education studies (e.g., a 0.10 standard deviation impact on state reading scores might be considered "larger" or "smaller" than a 0.10 standard deviation impact on a study-administered test).

In addition, answers to Question 1 will also influence the sample size requirements of future studies. If two tests are both considered to measure the same domain, but one test consistently yields smaller standard errors than the other test, then evaluations based on the first test would have smaller sample size requirements than evaluations based on the second test.

Question 2: Does measuring achievement using one type of test at baseline and another type of test at follow-up substantially reduce the precision of the impact estimates? The second question is motivated by the fact that some evaluations may choose to administer a standardized test at follow-up and rely on state tests to measure achievement at baseline, or vice-versa. In principle, using different tests at baseline and follow-up could reduce the R-square of the impact regression and, as a consequence, reduce the precision of the impact estimates. This

suggests that studies that rely on "mismatched tests" at baseline and follow-up may require larger samples to achieve the same precision as studies that rely on "matched tests".

Question 3: Can collecting both types of achievement measures substantially increase the precision of the impact estimates? A richer set of control variables can often increase the precision of the impact estimates. Therefore, an obvious question is whether controlling for baseline measures of achievement from *both* study-administered standardized tests and state tests yields more precise impact estimates than controlling for only one of the two achievement measures. If so, studies could be conducted with smaller samples if they would collect baseline achievement scores from both types of tests.

In addition, under some circumstances, collecting multiple outcome measures in the same domain may lead to more precise impact estimates. For multiple outcomes in the same domain, Schochet (2008b) proposes constructing composite outcomes by averaging the scores on the individual tests, and basing the "confirmatory" impact analysis on the composite measures. It is possible, though by no means certain, that composite outcome measures would yield more precise impact estimates by generating more reliable outcome measures (see forthcoming report for more details).

Setting:
Description of where the research took place.

To address the three research questions, we have re-analyzed the data from three randomized controlled trials (RCTs), as indicated in the previous section. Each of the three studies drew a sample from a single district. From this point forward, we will refer to each study by the state in which the district is located - Arizona, California, or Missouri.

Population / Participants / Subjects:
Description of participants in the study: who (or what) how many, key features (or characteristics).

All three of the studies that we have reanalyzed drew samples of elementary or middle school teachers and their students. In Arizona and California, the student sample included students in grades 3-5 (98 students in Arizona and 564 students in California). In Missouri, the student sample included students in grades 7-8 (567 students). In all three studies, students were pre-screened prior to random assignment to identify the students to which the intervention was targeted. This screening process varied from study to study. More details are provided in the forthcoming report (and more would be provided here if this study were not a methodological exercise).

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

It is important to recognize that the three experiments tested the effectiveness of three *different* interventions. The studies were not chosen because they reflect a common intervention approach: they were chosen because they were RCTs that measured student outcomes using both state

reading tests and study-administered reading tests. Two of the three interventions were clearly interventions focused on reading. While the third intervention was focused on science, it had a reading component:

- **Arizona.** The treatment was a reading intervention system that provides explicit, systematic instruction with ongoing progress monitoring, and it was designed for struggling readers in elementary schools.
- **California.** The treatment was Pearson Education's Scott Foresman Science, a year-long science curriculum for daily instruction, which is based on inquiry-rich content with a sequence of structured and supportive inquiry activities. A key feature of the curriculum is the Leveled Reader, which helps teachers differentiate instruction by reading level. Although the main purpose of the intervention is to improve science skills, the program provides reading supports to make the science content accessible.
- **Missouri.** The treatment was a middle school reading curriculum that was developed for struggling adolescent readers. It was designed to supplement the core reading program, and it provides explicit, systematic instruction with ongoing progress monitoring.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

While the data for this study come from three randomized field trials, the study itself is a methodological exercise to compare the impact estimates and standard errors across different models that use different tests to measure achievement.

In particular we estimate the models summarized in Exhibit 1 (see Appendix B). To address Question 1, we compare the impact estimates and standard errors from Model A (MAP post-test, MAP pre-test) to the impact estimates and standard errors from Model B (state post-test, state pre-test). To address Question 2, we compare the standard errors from Model A (MAP post-test, MAP pre-test) to the standard errors from Model C (MAP post-test, state or district pre-test), and we compared the standard errors from Model B (state post-test, state or district pre-test) to the standard errors from Model D (state post-test, MAP pre-test). To address Question 3, we conducted two separate analyses. To see if controlling for both types of pre-test scores substantially reduces the standard error of the impact estimate, we compared the standard errors from Model A (MAP post-test, MAP pre-test) to the standard errors from Model E (MAP post-test, both pre-tests), and we compared the standard errors from Model B (state post-test, state pre-test) to the standard errors from Model F (state post-test, both pre-tests). In addition, to see if creating a composite outcome measure substantially reduced the standard error of the impact estimate, we compared the standard errors from Model G (composite post-test, both pre-tests) to the standard errors from Model E (MAP post-test, both pre-tests) and the standard errors from Model F (state post-test, both pre-tests).

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

The three studies that we selected for this exercise were chosen because they *measured achievement using both state assessment tests and study-administered standardized tests*.[†] In all three studies, the study team collected baseline and follow-up measures of reading achievement from state reading tests. In addition, students in all three studies were administered tests offered by the Northwest Evaluation Association (NWEA). In Arizona and Missouri, the study administered the Measures of Academic Progress (MAP) reading test. In California, the study administered the Achievement Level Test Series (ALT).

To conduct the analysis, we estimated standard multi-level models. These models regressed the post-test on an indicator of treatment status, the pre-test variable or variables, and a set of demographic covariates. We accounted for clustering in estimating standard errors. The goal was to estimate impacts and standard errors for each of the models specified in Exhibit 1 to help us in addressing the study's three research questions.

Findings / Results:
Description of main findings with specific details.

These findings cannot be reported at this time because the work is in progress. However, IES has encouraged us to submit the report for the SREE conference, and we expect to be allowed to present interim findings if the abstract is accepted.

Conclusions:
Description of conclusions and recommendations based on findings and overall study.

To be determined.

[†] The three evaluations collected student-level scale scores, not just their proficiency levels as defined by state accountability standards.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Black, A.R., Doolittle, F., Zhu, P., Unterman, R., and Grossman, J. B. (2008). The Evaluation of Enhanced Academic Instruction in After-School Programs: Findings After the First Year of Implementation (NCEE 2008-4021). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2007). Using Covariates to Improve Precision for Studies that Randomize Schools to Evaluate Educational Interventions. Educational Evaluation and Policy Analysis, 29(1), 30 - 59.

Brown, R.S. and Coughlin, E. The Predictive Validity of Selected Benchmark Assessments Used in the Mid-Atlantic Region (Issues & Answers Report, REL 2007-No. 017). Washington, DC: Regional Educational Laboratory Mid-Atlantic , National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Cronin, J. (2004). Aligning the NWEA RIT Scale with the South Carolina High School Assessment Program. Northwest Evaluation Association.

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., Means, M., Murphy, R., Penuel, W., Javitz, H., Emery, D., and Sussex, W. (2007). Effectiveness of Reading and Mathematics Software Products: Findings from the First Student Cohort (NCEE 2007-4005), Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hedges, L. V., and Hedberg, E. C. (2007). Intraclass Correlation Values for Planning Group-randomized Trials in Education. Educational Evaluation and Policy Analysis, 29(1), 60 - 87.

Institute of Education Sciences, U.S. Department of Education (2008). Rigor and Relevance Redux: Director's Biennial Report to Congress (IES 2009-6010). Washington D.C.

Kingsbury, G.G. (2001). A Comparison of MAP and ALT Scores. Northwest Evaluation Association.

Linn, R.L. (2005). Fixing the NCLB Accountability System. Policy Brief on the National Center for Research on Evaluation, Standards and Student Testing.

NWEA (2004). A Few Notes about Reliability and Validity as They Are Reported in "NWEA Reliability and Validity Estimates: Achievement Level Tests and Measures of Academic Progress." Northwest Evaluation Association.

Schochet, P. Z. (2008a). Statistical Power for Random Assignment Evaluations of Education Programs. Journal of Educational and Behavioral Statistics, 33(1), 62 - 87.

Schochet, P. Z. (2008b) Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations. Washington D.C.: U. S. Department of Education, Institute of Education Sciences.

Shields, J. (2008). A Comparison of the NWEA Measures of Academic Progress and the Missouri Assessment Program. Doctoral Dissertation, University of Missouri.

Torgesen, J., Schirm, A., Castner, L., Vartivarian, S., Mansfield, W., Myers, D., Stancavage, F., Durno, D., Javorsky, R., and Haan, C. (2007). National Assessment of Title I, Final Report: Volume II: Closing the Reading Gap, Findings from a Randomized Trial of Four Reading Interventions for Striving Readers (NCEE 2008-4013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Appendix B. Tables and Figures

Not included in page count.

Exhibit 1: Analysis Models to Address the Research Questions

Research Question	Model	Outcome Variable		Pre-test Covariates	
		Study Test	State Test	Study Test	State Test
1	A	√		√	
	B		√		√
2	C	√			√
	D		√	√	
3	E	√		√	√
	F		√	√	√
	G	Average of two test scores		√	√