2010 SREE Conference Abstract

Title:

Using State Tests *vs.* Study-Administered Tests to Measure Student Achievement: An Empirical Assessment Based on Four Recent Randomized Evaluations of Educational Interventions

Author(s):

Pei Zhu, Marie-Andrée Somers, and Edmond Wong (MDRC)

ABSTRACT

BACKGROUND/CONTEXT: The past six years have seen a large and growing number of high-quality randomized field trials of the effects of educational interventions, based in large part on sponsorship from the Institute of Education Sciences (IES) (Spybrook, 2007). Most of this research has focused on the impact of educational interventions on students' academic outcomes. An important question for these studies, therefore, has been how best to measure students' academic performance. Typically, academic achievement is measured in two ways:

- *Study-administered test:* The first approach is to administer a standardized test to the sample of students in the study. This allows the study to choose a test that is closely aligned with the intervention and suitable to the study population. It also allows the same test to be used across the full study sample, yielding consistent measures for student achievement. On the other hand, the cost of administering these tests can be high – sometimes even prohibitive – and these tests impose additional burden on students and school staff. In addition, since these tests are not linked to school accountability systems and student progression, a lack of incentives for students to perform well might lead to biased results if treatment and control students do not exert the same level of effort on the test.
- *State test*: The second approach is to collect students' test scores on state/district tests from school records. In recent years, in response to NCLB requirements, nearly all states test students yearly in grades 3-8 and one grade level in high school in both reading and mathematics, making state tests an increasingly feasible source of information on student achievement. For this reason, using state tests as part of an evaluation can reduce the cost of the study, and cam also reduce the burden of additional testing on students, teachers, and school staff. Existing state/district test scores are also considered more "policy relevant" because they are used by districts/states to make decisions about individual students and schools. However, there is substantial variation in state/district assessments, in terms of the content focus of the assessments and their scale. The format of the assessments and the "stakes" attached to them also vary across states/districts. These factors make it challenging to combine results from multiple states. In addition, state assessments usually cover a broad range of content areas which may not be the focus of the intervention, posing question about whether specific state/district assessments are suitable for research purposes.

The trade-offs between these two types of approach are discussed in greater detail in a recent concept paper commissioned by IES (May et al., 2009). The paper focuses in large part on the conceptual and technical complications associated with using state test data from different jurisdictions (rescaling, pooling, etc.). Because many questions are left unanswered, May *et. al.* (2009) conclude the paper by proposing several empirically-focused studies that would contribute to a better understanding of the issues and assumptions pertaining to the use of state tests in RCT evaluations. Of these, they suggest conducting a within-study comparison of the estimated impact of an educational intervention on state tests *vs.* a study-administered test, using data from an RCT where both types of test data are available for students. MDRC was commissioned by Mathematica Policy Research and IES to conduct such a study.

PURPOSE/OBJECTIVE/RESEARCH QUESTIONS/FOCUS OF STUDY: For this project, we use data from four IES-sponsored randomized studies to examine some of the key issues identified in May *et. al.* (2009). The first set of questions focuses on issues related to using state tests:

1. ***Do studies meet the assumptions needed for combining impacts on state tests across grades and/or states?*** In order to decide how best to pool impacts on state assessments across different states and grades, it is important to start by assessing the structure of the data, that is: 1) whether the sample from each grade and state look similar, i.e., it represents a similar student population that is targeted by the study, and 2) whether the test score distribution is similar across grades and/or states. These factors can affect the decision on how best to rescale test scores and aggregate them (see question #2).

2. ***How sensitive is the estimated impact on state tests to different strategies for rescaling test scores and/or aggregating results across states?*** State test scores must be converted to a common metric in order to be able to compare and pool findings across different grades and states. This is typically done by converting raw test scores to z-scores, or by using a nonlinear transformation such as equipercentile equating. Once test scores have been rescaled to a common metric, impacts can be estimated for each state and then aggregated based on either a weighted average or a meta-regression model. The performance of these different rescaling/aggregation approaches depends on the underlying data structure; therefore, it is important to study how these different strategies affect impact results, especially in conjunction with the assessment of data structure addressed in #1 above.

The next set of questions compares the performance of state tests and study-administered tests:

3. ***Do estimated impacts on state tests differ from estimated impacts on the study-administered test? Under what conditions is this most likely to happen?*** The extent to which results based on state tests and study-administered tests differ may depend on the subject area being tested (math, reading); on the age of the study population (elementary, middle school, secondary); and/or the degree of overlap between the content of the study test and that of the state test. It may also depend on the nature of the outcome that is targeted by the intervention.[1]

4. ***What is the reliability of study-administered versus state tests, especially for the study-targeted population?*** It is important to assess the reliability of alternative outcome measures, because reliability affects the statistical power of impact estimation. Reliability decreases as scores move away from the average (Hambleton, Swaminathan, and Rogers, 1991), hence for studies that target low-performing students (like most IES studies including the ones that will be used in this project), the reliability of outcome measure can be low, which can in turn reduce the study's power to detect program effects.

5. ***What are the implications for precision of using state test scores rather than a study-administered test to measure achievement in the baseline period?*** Even if a study-administered test is used to measure student outcomes, it may still be worthwhile to use state tests to measure student achievement *at baseline*, since state tests are more cost-effective than a study-administered baseline test. That said, if state tests at baseline differ greatly from the study-administered test at follow-up, this may reduce the explanatory power of baseline test scores in the impact model. This means that a larger sample would be required to maintain a given level of statistical precision, which has its own cost implications.

[1] For interventions that target a specific type of skill or knowledge (e.g., rational numbers), researchers may have to use their own test, because the coverage of state tests is too broad to detect impacts on the specific skill that is the focus of the intervention. In this situation, one would expect impacts on the study test to be larger than impacts on the state test. Conversely, for interventions that aim to improve proficiency or achievement more broadly, one would expect impacts on the two types of test to be more similar since in this case, the tests measure similar outcomes.

DATA (INTERVENTION/RESEARCH DESIGN): The research questions above are examined using data from four IES-sponsored randomized experiments. These studies were selected primarily because they use both study-administered tests and state/district tests to measure student achievement. These studies also represent useful variation that we can exploit to answer some of the research questions in the proposed analysis, with respect to grade levels (elementary school, middle school, and high school), states (22 states), and subject area (reading and math). These studies also represent a continuum in terms of the focus of the intervention being tested: two of the interventions aim to improve proficiency or achievement more broadly (e.g. math achievement), one of the interventions targets skills that are more specific yet still relatively broad (reading comprehension and vocabulary), while the last intervention targets a very specific skill (rational numbers). Table 1 summarizes the features of these studies that are most relevant to the current project. Also below is a brief description of each of the studies:

- The first study tests the impact of after-school programs with an instructional focus in mathematics or reading on students' academic achievement. Twenty-five after-school centers in the study implemented an after-school reading program and 25 centers implemented a math program, so there are actually *two studies* in this project (one for math and one for reading). Students in each center were randomly assigned to the enhanced program or to the regular after-school program offered by the center (where academic support consists largely of homework help). The target population is students in 2nd to 5th grade who are behind grade level but not by more than two years. The target sample for this project consists of students in the 3rd to 5th grade who participated in the programs during their first year of implementation (2005-2006).[2]
- The next study uses a random assignment design to test the effectiveness of a year-long math teacher professional development (PD) intervention in improving teacher knowledge of rational numbers, teacher instruction, and student math achievement in high-poverty schools. The study was implemented in 77 schools in 12 districts (a total of 195 7th grade math teachers), with approximately equal numbers of schools randomly assigned in each district to receive the PD treatment provided by the study, or a "business as usual" group, which participated only in the usual PD offered by the district. The target sample for this project consists of 7th grade students enrolled in regular math classes in the study schools in the first school year of implementation (2007-2008).
- The last study is a random assignment evaluation of two supplemental literacy programs that aim to improve the reading comprehension skills and school performance of struggling 9th grade readers. The supplemental literacy programs are full-year courses targeted to students whose reading skills are two to five years below grade level as they enter high school. These programs were implemented in 34 schools; in each school, students were randomly assigned to either enroll in the supplemental reading class or to remain in a regular ninth-grade elective class. The target sample for this project consists of the first cohort of 9th grade students to participate in the reading programs (2005-2006).

Analyses of test scores are based on students with both a state test score *and* a study-administered test score (see Table 1 for sample sizes).

[2] State test scores are not available for 2nd grade students so they are excluded from the analysis.

DATA ANALYSIS: The following analyses are conducted for each of the research questions:

1. Do studies meet the assumptions needed for combining impacts on state tests across grades and/or states?

For each of the studies, the distribution of students' demographic characteristics is compared across states (and grades, in the case of the first study) to assess the similarity of the student samples across grade-by-state cells.[3] The distribution of students' scores on state tests, by cell, are also compared to the respective statewide test score distribution, to see if the study samples in each cell are located at similar positions in their state distribution.[4] We also collect information on the content coverage of each test in the studies, in order to examine whether there is variation across states in terms of what each test emphasizes.

2. How sensitive is the estimated impact on state tests to different strategies for rescaling test scores and/or aggregating results across states?

Here we focus on the rescaling and aggregation approaches described in May *et. al.* (2009):

- Rescaling methods: We rescale state test scores using three different approaches: (1) z-scores based on the *sample* mean and standard deviation for each state in the study; (2) z-scores based on the *state* mean and standard deviation; and (3) equipercentile equating.[5]
- Aggregation methods: Impacts estimates for each state/grade are aggregated using four different approaches: (1) weighting by the sample size in each state, (2) weighting by precision, (3) combining estimates using a random-effects regression model, and (4) combining estimates using a fixed-effects model.[6]

For each of the four studies, we estimate impacts on state test scores using the 12 possible combinations of these rescaling and aggregation approaches (3 rescaling * 4 aggregation approaches), and we compare the impact estimates (and standard errors) from each of these combinations. Results for each rescaling/aggregation combination are based on a consistent sample of students and modeling strategy (covariates, clustering/nesting corrections). The findings from this analytical exercised will be evaluated in conjunction with the results from the first analysis the assessment of assumptions in #1 above, since the performance of different rescaling/aggregation approaches depends in part on the underlying data structure.

3. Do estimated impacts on state tests differ from estimated impacts on the study-administered test? Under what conditions is this most likely to happen?

To answer this question, we conduct the following analyses for each of the four studies:

- We start by descriptively comparing scores for the two types of test. Specifically, the relationship between the study test and each of the state tests is examined by producing graphs (scatter plots or kernel density graphs), estimating correlation coefficients, and looking at cross-tabulations of the data. These analyses provide background information for the comparison of impact estimates using these two types of tests.

[3] These characteristics include students' test score at baseline, their gender, age, race/ethnicity, as well as their family/school information.
[4] This is done using the control group's follow-up test scores, which are unaffected by the intervention being tested.
[5] For this approach, test scores are first converted to percentile ranks by grade and state; percentile ranks are then converted to z-scores using the standard normal distribution.
[6] Methods #3 and #4 are "meta-regression" approaches (May *et al.*, 2009).

- We then conduct a systematic comparison of estimated program impacts on state tests *vs.* the study test. As shown in Table 2 (first column of results), we estimate impacts on the study-administered test using each of the aggregation methods used to pool state test impacts (i.e., sample size weighting, precision weighting, random-effects, and fixed-effects). The scores for the study-administered test do not require rescaling, because the same study test is used for all students across grades and states; however, impacts will be converted to effect sizes in order to make it possible to compare impacts on the study test to impacts on the state tests.[7]

This analysis allows us to compare the estimated impact on the study-administered test to the impact on the state tests. We can also descriptively examine the conditions under which estimated impacts on study tests and state tests are more similar. In particular, because the four studies span different age groups and subject areas, we can see whether impacts are more similar for some age groups (elementary, middle school, high school), or subject areas (math, reading). Based on content coverage information obtained in #1, we can also confirm that estimated impacts on state tests and study tests are more similar when there is more overlap in the coverage of the assessments, and when the nature of the outcome targeted by the intervention is more general ("math achievement") as opposed to specific ("rational numbers"). Finally, we can investigate whether estimated impacts are more similar based on the choice of rescaling/aggregation strategies for the analysis of state tests.

4. What is the reliability of study-administered versus state tests, especially for the study-targeted population?

To answer this question, we are collecting information on the reliability of the state tests in the four studies, for all students as well as for students who are "below basic" (since these students are the target population of the interventions). We can use these data to examine whether there is variation across states in the reliability of their tests, and also whether the state tests have lower/higher reliability than the study-administered test. We also examine the density of scores for each test in order to see whether there is ceiling or floor effect for a given test (a floor effect would reduce the reliability of the test for students at the lower end of the distribution). These analyses will help to better understand the answers to the previous research question.

5. What are the implications for precision of using state test scores rather than a study-administered test to measure achievement in the baseline period?

To answer this question, we estimate the impact of the program on the study-administered test using two different model specifications: Model 1 controls for baseline test scores from the same study-administered test, while Model 2 controls for baseline achievement as measured by state test scores. Data from two of the studies allow us to conduct this analysis.

FINDINGS/RESULTS/CONCLUSIONS: Analysis for this project is almost complete; our findings provide valuable empirical evidence of the trade-offs between using state tests and study-administered test and have implications for designing future RCT studies. However, the findings cannot be divulged until a draft report of the results is submitted to IES in December 2009. If this proposal is accepted for the conference, we will update this abstract and submit a paper prior to the conference. In the meantime, Table 2 shows the types of key findings that will be presented.

[7] Effects sizes are based on the standard deviation of the sample, but we will also calculate effect sizes based on the standard deviation of the state instead, if this information is available for the study tests.

APPENDICES

APPENDIX A. REFERENCES

Hambleton, R.K., H. Swaminathan, and H. J. Rogers. (1991) *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Press.

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., and Gleason, P. (2009). *Making Good Use of State Tests in Education Experiments* (Draft Report, February 16th 2009). Washington, DC: Mathematica Policy Research.