Article

American Journal of Evaluation
2019, Vol. 40(2) 190-213

DOI: 10.1177/1098214018767313
journals.sagepub.com/home/aje
SAGE

The Use and Validity of Standardized Achievement Tests for Evaluating New Curricular Interventions in Mathematics and Science

Joshua Sussman[1] and Mark R. Wilson[1]

Abstract

We investigated the use and validity of standardized achievement tests for summative evaluation of 78 educational intervention projects funded by the Institute of Education Sciences (IES) math and science education program. Investigators from 46 projects evaluated curricular interventions with standardized achievement tests as outcome measures. Twenty-five of the projects had potential validity problems related to a misalignment between the achievement test and the intervention. A closer analysis of 11 of those projects flagged as high risk for validity problems showed that only 6 projects attended to the validity of the test, and only 1 project provided adequate validity evidence. We conclude that there is widespread inappropriate use of achievement tests that threatens the validity of educational evaluations. To better support innovation, evaluators must dedicate more attention to the validity of the outcome measures they use.

Keywords

evaluation, validity, assessment, outcome measures, STEM education

Leading researchers in mathematics and science education develop new classroom interventions designed to enhance student learning. The interventions include brief lesson sequences that focus on key areas of learning within a discipline, curricular supports such as computer tutors, and comprehensive year-long curricula. Researchers often use data from standardized achievement tests, such as state tests administered for accountability purposes, to evaluate the educational impact of these interventions. The central goal of this study is to investigate whether the researchers who conduct the evaluations using standardized achievement tests provide adequate evidence that the test results are valid for evaluating the interventions. We ask whether evaluations that rely on data from

[1] University of California, Berkeley, CA, USA

Corresponding Author:
Joshua Sussman, University of California, 2000 Center St #301, Berkeley, CA 94720, USA.
Email: jsussman@berkeley.edu

standardized achievement tests could misjudge the educational interventions, generating spurious results that mislead efforts for educational innovation.

The literature urges cautious use of standardized achievement tests for summative evaluation of new curricular interventions in mathematics and science. Evaluations of curricular interventions in mathematics and science have traditionally relied on data from standardized achievement tests that are not well aligned with the goals of the interventions, resulting in widespread problems with the credibility of evaluations (National Research Council [NRC], 2004, 2012; Taylor, Kowalski, Wilson, Getty, & Carlson, 2013). Historically, most achievement tests have neglected to measure important aspects of academic competence (Greeno, Pearson, & Schoenfeld, 1997; NRC, 2001) and, generally speaking, standardized achievement tests are not designed to assess the reasoning and problem-solving skills emphasized by instructional interventions in K–12 science or mathematics (Darling-Hammond et al., 2013; DeBarger, Penuel, Harris, & Kennedy, 2016; Pellegrino, Wilson, Koening, & Beatty, 2014). In spite of long-standing recognition of these issues, well-developed investigations into the validity of standardized achievement tests for evaluating new interventions constitute a relatively new area of the literature (e.g., May, Johnson, Haimson, Sattar, & Gleason, 2009; Olsen, Unlu, Price, & Jaciw, 2011; Somers, Zhu, & Wong, 2011).

The purpose of this study is to better understand the use and validity of standardized achievement tests for the summative evaluation of new mathematics and science interventions. To conduct the current investigation, we needed a sample of high-quality applied research. To achieve this, we gathered information about projects funded through the Institute for Education Sciences (IES) math and science education research program. We collected data from the IES online database, obtained reports from the principal investigators on the projects, and examined publications related to the projects. As a condition of use, we assured the principal investigators who supplied reports that this analysis would, to the extent possible, maintain the anonymity of individual projects.

Our first goal was to document the scope of the issue. The literature contains no current summaries of the prevalence of standardized achievement tests as outcome measures for evaluation of science, technology, engineering, and mathematics (STEM) interventions. Thus, we asked whether it is common for researchers to use these tests for summative evaluation of new math or science interventions. Our second goal was to investigate the potential for validity problems in the evaluation of STEM programs. To that end, we screened the studies and flagged ones that showed potential for validity problems related to using the standardized achievement test for evaluating the intervention. A team of three raters reviewed the projects and applied a straightforward rubric to determine whether each project presented potential validity problems. Then, the raters closely examined the project with potential validity problems, characterizing the nature of the validity evidence contained in the project reports or peer-reviewed research, exploring measurement issues in the data, and ultimately evaluating the adequacy of the validity evidence.

The field of evaluation can make a greater contribution to STEM education by devoting greater attention to outcome measurement. It is widely accepted that valid measurement is central to the credibility of an evaluation, and measuring the wrong outcome can produce results that fail to detect the educational benefits of an intervention (Lipsley, 1990; Rhue & Zumbo, 2008). Importantly, inaccurate results from an evaluation can have consequences, such as faulty decisions about whether or not a student is proficient in a given subject, whether to implement a program in a particular school district, or whether to consider a program evidence-based (Schoenfeld, 2006). In contrast, evaluations that use valid outcome measures and produce accurate and useful information about the educational effectiveness of STEM interventions are likely to advance beneficial programs, support students' STEM achievement, and increase access to the economic opportunities, such as job growth and income increases, associated with STEM vocations.

Standardized Achievement Tests

Characteristics of the Standardized Achievement Tests in This Study

In this study, we examine IES-funded projects, where investigators use standardized achievement tests to evaluate the impact of curricular interventions in mathematics or science. The tests have three main characteristics. First, the tests are standardized: Each test is one part of an assessment system with a defined protocol for test administration, scoring, and reporting of results. Examples of the tests represented in the data include state-developed and administered tests, commercially available measures such as the Iowa Test of Basic Skills (ITBS; Iowa Testing Programs, 2003), and subject matter tests such as the American Chemical Society General Chemistry Exam (see Brandriet, Reed, & Holme, 2015).

Second, the standardized achievement tests are *standards-based*. Developers create the tests to measure grade-level proficiency in a major subject area such as mathematics or science. A set of state or national standards, such as the Common Core State Standards in Mathematics (CCSSM; National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010), defines grade-level proficiency. The tests typically consist of several dozen items that sample from the academic content and skills taught over a year-long curriculum. The developers and/or end users conduct *alignment* research to ensure that the tests adequately cover the range, and balance, of academic content and skills within the standards (Bhola, Impara, & Buckendahl, 2003). The product is a broad measure of academic achievement in a major subject area that defines the academic skills and knowledge that students are expected to learn during a school year.

Well-validated standardized achievement tests can reasonably be considered reliable measures of grade-level academic proficiency, especially for groups of students (Popham, 1999, 2001; Schafer, Wang, & Wang, 2009). Thus, standardized achievement tests may be suitable for evaluating the impact of interventions where the goal is to increase grade-level proficiency relative to a comparison group (May et al., 2009). Although some testing systems disaggregate scores by content strand and report subscale scores, it is important to interpret these subscale scores with caution. For example, a standardized mathematics test may produce subscale scores for fractions, geometry, statistics, and so on. However, the items per subscale are generally small in number which leads to concerns about psychometric reliability and problems such as Type II error (NRC, 2004).

The third important characteristic of the standardized tests described in this study is that they are *preexisting*. The standardized tests that project investigators use to evaluate the educational impact of new interventions were originally designed for a prior purpose (i.e., assessing grade level academic achievement). Validity problems in program evaluation related to the possibility of misalignment between a preexisting test and an intervention are of major concern (American Educational Research Association[AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 2014). Investigators who use preexisting standardized achievement tests incur the responsibility to carefully examine the validity of the test for each specific purpose.

Political and Practical Factors That Influence the Use of Standardized Achievement Tests as Outcome Measures

Researchers can have varied goals for evaluation criteria (Cobb & Jackson, 2008), but IES funding for applied research usually requires summative evaluation impact of the educational program (see Campbell, 1991) for a description of this form of evaluation). This requirement, a long-standing tradition in federally sponsored research (Lagemann, 2002), generally leads investigators to experimental methods that include the use of standardized achievement tests as outcome measures (Confrey, 2006; Donaldson, Christie, & Mark, 2009; Taylor et al., 2013).

Many politicians, the general public, and some researchers consider standardized achievement tests to be rigorous and objective measures of academic competence independent of how they are used (Crocker, 2003; NRC, 1999; Office of Technology Assessment, 1992). The What Works Clearinghouse (WWC), created by IES to evaluate research evidence on the effectiveness of educational interventions, does not require additional evidence of reliability or validity when using standardized achievement tests (Song & Herman, 2010). Some scholars argue that standardized tests are curriculum-neutral outcome measures, which avoid the bias (i.e., inflated treatment effects) that can occur when investigators use outcome measures that are aligned with particular curricular interventions (e.g., Slavin & Madden, 2011). Others point to the importance of test scores to state and local administrators who are accountable for improving student proficiency as measured by state-administered tests. Finally, researchers may use standardized tests as outcome measures because the data are inexpensive and easy to obtain compared to the time, effort, and skill required to develop, validate, and administer researcher-developed tests (May et al., 2009; Somers et al., 2011). This confluence of factors explains, in part, the widespread use of standardized tests, in conflict with good measurement in applied research.

Problems With the Use of Standardized Achievement Tests for Evaluating New Curricular Interventions in Mathematics and Science

Standardized achievement tests may not be suitable outcome measures for evaluation of new curricular interventions because of a basic mismatch between the knowledge and the skills that the test measures with the knowledge and skills that the intervention teaches. The past several decades of research led to new learning goals in STEM education, and typical standardized achievement tests no longer adequately measure the range of skills and knowledge that are considered important in modern definitions of proficiency in K–12 science or mathematics (Darling-Hammond et al., 2013; DeBarger, Penuel, & Harris, 2013; NRC, 2001). The goal of an intervention is typically to change or modify students' behavior relating to the acquisition of specific new academic skills or new ways of thinking (Tyler, 1942). The prevailing view is that evaluators should appraise the intervention by determining the degree to which the goals of the program are being realized in students (Baker, Chung, & Cai, 2016). In the following paragraphs, we explore the disparities between learning and assessment in math and science education and the educational measurement literature.

Outcome Measurement in Mathematics Education

In mathematics education, the NRC (2004) examined 192 evaluations of the effectiveness of 19 mathematics curricula and identified widespread problems with use of outcome measures for evaluating curricular interventions. The committee found extensive use of standardized achievement tests, emphasizing skills such as computation and procedural fluency, to evaluate the impact of curricular interventions that emphasized specific learning goals such as mathematical reasoning, conceptual understanding, and problem-solving. The NRC recommended that the credibility of an evaluation of curricular effectiveness should depend on a strong match, or alignment, between what the test measures and what the intervention teaches.

The critical perspectives expressed in the NRC's (2004) report immediately preceded major assessment reforms in mathematics education. The CCSSM is the latest effort to define research-based learning goals in K–12 mathematics education that guide curriculum, instruction, and assessment. The CCSSM called for efforts to develop new assessments that measure higher order thinking skills, such as mathematical sense-making, that are emphasized in the new standards. Groups such as the Smarter Balanced Assessment Consortium (SBAC; 2016) and Partnership for Assessment of Readiness for College and Careers (PARCC; Pearson Corporation, 2017) have developed, and continue to develop, popular CCSSM-aligned tests.

At present, little information exists in the literature about whether these new tests meet the vision of mathematical proficiency defined by the CCSSM. Doorey and Polikoff (2016) reviewed tests from four assessment systems (SBAC, PARCC, ACT Aspire, and the Massachusetts Comprehensive Assessment System) and concluded that the new mathematics tests placed greater emphasis on higher order thinking than prior tests, but the relative emphasis on items that measure higher order skills varied widely across grades and testing systems. In addition, two teacher panels ($N = 12$ and $N = 13$) evaluated the quality of the SBAC assessments compared to prior state tests in mathematics and offered mixed reviews of the ability of the tests to measure the more complex thinking skills that are part of high-quality instruction (McClellan, Joe, & Bassett, 2017a, 2017b). In sum, the early evidence suggests that new standardized tests place greater emphasis on higher order thinking, but some tests have succeeded more than others. With respect to the use of the new tests as outcome measures, the differences between tests reinforce the need to carefully examine the match between the type of learning one wishes to assess with the types of knowledge measured by the test.

Outcome Measurement in Science Education

Scholars of science education point to a "disjuncture" between the skills and the knowledge measured by typical science tests with the skills and knowledge that define proficiency in science (Pellegrino, 2013). The Next Generation Science Standards (NGSS) draw from the latest research in science education to define modern learning goals in K–12 science. NGSS articulates a framework of learning along three interrelated dimensions: integrating knowledge of core ideas within a scientific discipline, engagement with scientific practices, and building connections across ideas (NRC, 2012). In contrast, most standardized tests in science align with science standards that emphasize building content knowledge over scientific practices and cross-cutting concepts (Pellegrino, 2015), overemphasizing recall of facts and recognition of correct answers (Porter, Polikoff, Barghaus, & Yang, 2013; Quellmalz et al., 2013). A systematic review of existing science tests reported that although some items, within some assessments, measure some of the learning goals within NGSS, most existing assessments are not well aligned with NGSS (Wertheim et al., 2016). Indeed, outcome measures for evaluating NGSS-inspired interventions are not yet a reality.

The high probability of misalignment between a new, research-based curricular intervention and a standardized achievement test suggests that investigators must be very careful to examine the validity of outcome measures. Pertinent to the current study, the curricular interventions in the IES database represent cutting-edge approaches to science education, which are likely to focus on many of the learning goals articulated in NGSS. We expect to identify validity problems with the outcome measures used to evaluate the effectiveness of curricular interventions in science, similar to the NRC's (2004) analysis of mathematics interventions.

Validity of Outcome Measures for Evaluating the Impact of New Curricular Interventions

AERA, APA, and NCME (2014) document the widely agreed upon professional standard that the use of test scores—for any purpose—must be supported by evidence that the test is *valid* for its intended purpose. The consensus in the field is that validity is a "necessary condition for the justifiable use of a test" (p. 11). Importantly, validity is not a property of a test; it is a property of a test for a certain purpose (Messick, 1995).

The literature in mathematics education and in science education uses different concepts to examine validity problems and define the desired qualities of outcome measures. The mathematics education literature refers to the *curricular validity of measures*: the idea that an outcome measure must accurately and comprehensively assess the curriculum's ability to meet the designer's

intended objectives (Confrey, 2006). In contrast, the science education literature uses the concept of *instructional sensitivity* to describe a test that detects the influence of instruction rather than the influence of other factors, such as general ability (Ruiz-Primo et al., 2012; Wiliam, 2008). An important area of consensus is that the tests must measure student progress in the objectives as intended by the designers.

In this article, we use an alignment framework to analyze specific problems with standardized achievement tests as outcome measures. The goal of alignment is to determine the congruence between a test and a set of learning goals—typically educational standards. In a typical process of alignment, test developers or end users match the items on an assessment with the content domains and thinking skills that define grade-level academic proficiency in a subject area, using either the standards or the enacted intervention (Bhola et al., 2003; Herman, Webb, & Zuniga, 2007; Martone & Sireci, 2009; Porter, 2002). The ultimate purpose is to ensure that the final version of a test measures the knowledge and skills of the intended, or enacted, curriculum (Porter, Smithson, Blank, & Zeidner, 2007). In particular, May et al. (2009) discussed the theoretical importance of alignment to the validity of state-administered standardized tests as outcome measures for evaluating educational interventions.

We explore the alignment between tests and curricular interventions within two domains: content and cognitive processes. Analysis within these two domains provides a basis for evaluating whether or not a test is likely to be valid for evaluating a particular intervention. A test may be misaligned with an intervention if the test measures different amounts of academic content than the intervention teaches and/or if the test measures different types of thinking skills than the intervention teaches. In other words, the test's target of *inference* must match with the intervention's target of *influence*.

Alignment problems can exist in the content domain because the developers of new curricular interventions as opposed to the developers of standardized tests often focus on different grain sizes of academic content (Wertheim et al., 2016). Both CCSSM and NGSS emphasize core ideas that are thought to be especially generative for long-term learning: Interventions inspired by these standards will often focus on fewer content areas in greater depth rather than covering a larger number of areas more superficially. Therefore, a particular standardized test (that is aligned with the grade-level standards) may contain many items that have no relevance to the academic content taught by a new intervention. The problem is further exacerbated if the intervention aims for advanced learning not covered in the grade-level standards (e.g., Confrey & Scarano, 1995). If the content is misaligned, it is likely that the test scores will not provide accurate and useful information about the intervention.

Mismatch may also occur between the thinking skills that develop through participation in the intervention compared to the thinking skills required for success on the test. New educational interventions inspired by CCSSM and NGSS aim to support students' higher level thinking skills, such as reasoning and problem solving, and will engage students in complex forms of thinking over time (Greeno, Collins, & Resnick, 1996; Lehrer, 2009). For example, a natural science intervention might aim to support students' ability to construct scientific models of natural phenomena in order to explain the nature of a scientific mechanism, requiring students to use evidence in well-defended arguments and to revise arguments in the face of new information (Berland et al., 2015). As discussed in the previous section, typical achievement tests would not produce accurate information about the impact of this sort of intervention. Although researchers are actively developing standardized tests that measure higher level thinking skills (e.g., De Barger et al., 2016), to our knowledge no literature claims that an existing standardized achievement test is a *good* measure of the types of learning outcomes targeted by many new educational interventions. In fact, scholars posit that assessing cognitively complex learning requires different approaches to assessment that are not represented in conventional standardized tests (Brown & Wilson, 2011; Catley, Lehrer, & Reiser, 2005; Frederiksen & White, 2004).

In conclusion, when a content or cognitive skills mismatch occurs, one must question whether test scores accurately capture the impact of the intervention and hence question whether these scores are valid for summative evaluation of program impact.

Establishing the validity of the test for each specific purpose is essential and evaluators should not rely on prior validation activities when using a standardized test as an outcome measure (AERA, APA, & NCME, 2014). The current study examined the extent to which investigators in the field have followed this advice.

The Current Study

In this study, we investigated the use and validity of standardized achievement tests for evaluating new educational interventions in math and science.

This study addressed five research questions

1. **Research Question 1:** What proportion of projects in the database evaluated, or planned to evaluate, a curricular intervention using a standardized achievement test?
2. **Research Question 2:** How many curricular evaluations showed validity problems, where the standardized achievement test did not appear to measure what the intervention attempted to teach?
3. **Research Question 3:** For the projects with validity problems, how many of the projects presented validity evidence to support the use of the standardized achievement test for evaluating the impact of the intervention? How many projects specifically referenced the concept of alignment?
4. **Research Question 4:** How many of the projects in Research Question 3 conducted test validation research?
5. **Research Question 5:** How many of the standardized tests were valid for their intended purpose?

The purpose of Research Question 1 was to document how often investigators funded through the IES mathematics and science education program use standardized achievement tests to evaluate new interventions. To our knowledge, no empirical data exist on the prevalence of standardized achievement tests as outcome measures in applied educational research.

The goal of Research Question 2 was to identify projects with potential validity problems, where the test did not appear to measure what the intervention attempted to teach. Three raters reviewed the information contained in the IES database and flagged the projects where the test and the intervention differed in terms of either content or cognitive process.

For Research Question 3, we assessed the validity evidence presented by investigators. We wrote to principal investigators and asked for the most recent project report to IES, which became part of the research data (e.g., Spybrook & Raudenbush, 2009). Next, we examined the reports and published literature associated with each project for evidence that the investigators considered the validity of the scores from standardized achievement tests. We searched the reports for validity evidence, described as the rhetorical use of fact or theory to (a) develop an evidence-based rationale for the meaning of standardized test scores and (b) support the interpretation of standardized test scores for summative evaluation of program impact. We also searched for specific evidence that investigators considered the alignment between the standardized test and the intervention, as recommended by the literature (May et al., 2009).

For Research Question 4, we searched the IES reports and published literature for evidence that investigators carried out *test validation* activities. Test validation is a more active process of generating validity evidence. We defined test validation as research activities that intended to produce evidence that the test scores provide accurate and useful information for a particular purpose. Examples of validation activities include pilot administrations of a test with analyses of

psychometric properties and participant interviews (i.e., think-aloud or cognitive labs) to understand how test takers approach the items.

We concluded the study with a qualitative analysis that judged the adequacy of the corpus of validity information within each project (Research Question 5). For each project, we aimed to address the overarching question "is the standardized achievement test a valid outcome measure?" Relatedly, we explore measurement issues contained in the data in order to draw more general conclusions about test validity in applied STEM education research.

Method

Data

The data for this study come from four sources. The first is the IES online database of funded research grants and contracts (termed *projects* hereafter), found at: http://ies.ed.gov/funding/ grantsearch/index.asp. At the time of writing, the database contained descriptions of 85 projects funded through the IES math and science education program awarded between 2003 and 2015. From this database, we gathered information about the purpose of each project, the goals for student learning, and the key outcome measures used to evaluate the project.

The second data source consists of reports that principal investigators submitted to IES. We contacted the principal investigator on each project via e-mail and requested copies of the final reports to IES submitted at the close of the grant. If the final reports were not available (e.g., for open grants), we requested interim reports or proposals. We contacted the 68 unique principal investigators for the 78 projects that met our criterion for inclusion in the study (see Procedure section). We e-mailed principal investigators up to 3 times. Forty-eight (70.6%) principal investigators responded to our e-mails, and investigators from 33 (48.5%) projects provided one or more documents.

Documents from principal investigators. We received a variety of documents from the projects. Most of the documents we received from principal investigators were final reports. Some were interim reports, and a small number were the research proposals that garnered IES funding. The amount of information contained in the documents was inconsistent—even across documents of the same type. The page range was between 5 and 355 pages. Thus, for some projects, we had hundreds of pages of information, and for others we had only a few pages. The diversity of information complicated the analysis. However, we applied a consistent procedure to all the documents and based the analysis on the available information.

The third data source consisted of peer-reviewed articles that contain research funded through the grants. We identified the articles using two methods. First, we gathered references from the project page in the IES database. Second, we searched *Proquest*, *Google*, and *Google Scholar* databases for articles connected to the grant. Appendix A contains more information about the search procedure.

The fourth source of data was the result of an Internet search to gather information about the achievement tests used as outcome measures. We entered the name of a test into the Google search engine and examined the results for technical information. Interested readers are referred to Appendix A for additional information.

Importantly, we assured the principal investigators that this manuscript would maintain the anonymity of individual projects. We expected to generate critical analyses that could possibly reveal specific weaknesses in the research projects. In order to encourage investigators to provide access to the reports that might contain detailed evidence of the weaknesses, we informed the investigators that, to the best of our ability, the analysis would not help readers trace measurement problems back to the original projects.

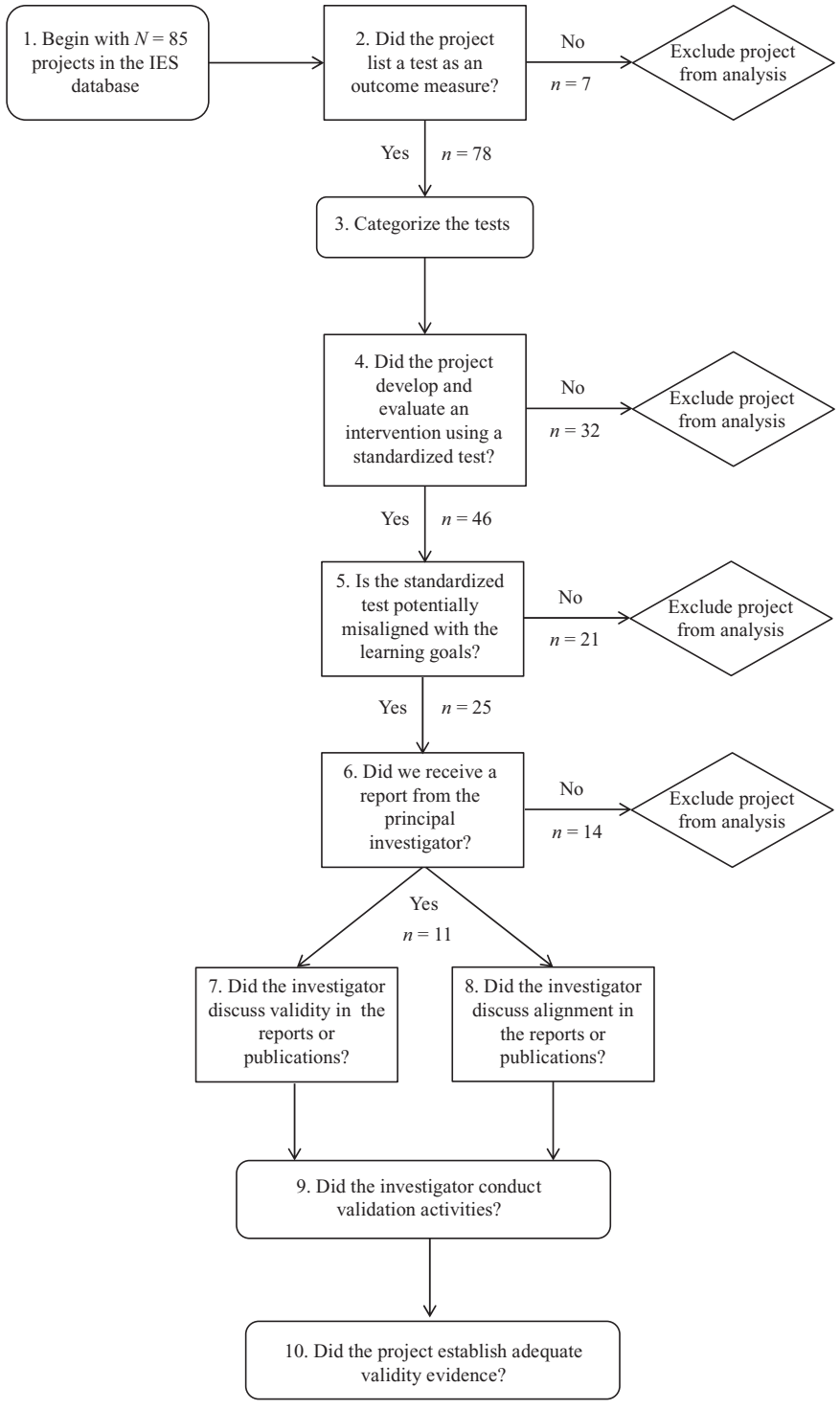

Figure 1. Logic model for research method.

Procedure

Coding the studies. Figure 1 contains a diagram of the analysis plan. The diagram demonstrates the logic of the procedure. Our research questions are nested in particular steps. The rounded rectangle on the left, Step 1, represents the start of the procedure with the initial sample of 85 projects in the IES database.

Projects that met the inclusion criterion. For Step 2, we excluded projects that did not list a key measure. We started with 85 projects in the IES math and science education database with award dates from 2003 to 2015. Seventy-eight projects listed at least one measure in the Key Measures section of the database and thus remained in the pool of research. Some projects did not list a quantitative outcome measure because they did not plan an evaluation.

Classifying the outcome measures. The rounded rectangle in Step 3 summarizes a multistep procedure for classifying the outcome measures used in each project. Our goal was to generate a high-level description of the types of outcome measures used in the studies. For each project, we examined the section in the IES database entitled Key Measures that contains descriptions of the sources of outcome data, including test scores.

The Key Measures section of each project in the IES database contained descriptions of the sources of outcome data, including tests, audio, video, and work samples. The amount of information contained in this section varied across projects: Some investigators provided detailed descriptions of specific instruments and how researchers planned to use and interpret the measures, whereas other investigators provided vague descriptions of instruments such as "math assessment," with no information about how the tests would be used.

We developed a three-category rubric to distinguish qualitatively different types of measures. The three categories were *standardized achievement tests, study-developed tests,* and *other tests.* We defined a standardized achievement test as a previously developed measure of grade-level achievement in a broad area of mathematics or science. A study-developed test included any assessment developed by the investigator for research or evaluation purposes. The category of other tests included (a) standardized measures of academic achievement that had been previously developed and validated to measure more specific domains of math or science achievement than standardized tests (e.g., fractions) and (b) measures with details that were too vague to classify into one of the other two categories.

Classifying the intervention projects that used standardized achievement tests. In Step 4, we identified the projects that met two criteria. First, the project had to be classified as an intervention project, which means it had to (a) develop and (b) evaluate a new educational intervention. Development and evaluation did not have to be the primary purpose of the project; the proposal only needed to state the intent to evaluate the educational effectiveness of a new educational intervention. For example, some projects used rigorous experimental or quasi-experimental designs, whereas other projects used less intensive nonexperimental designs that lacked control groups. Of the 78 projects in the IES database that listed an outcome measure, 64 (82.1%) evaluated the educational effectiveness of an educational intervention. Second, we asked how many intervention projects used a standardized achievement test as an outcome measure. Of the 64 intervention projects, 46 (71.9%) also used a standardized achievement test as an outcome measure. We explore these findings further in the Results section.

Screening projects for potential validity problems. In Step 5, the first author and two raters identified projects with the greatest potential for validity problems related to the use of a standardized achievement test for summative evaluation of an intervention. The crux of the analysis involved assessing the match between the intervention's learning goals and the test's target of measurement. The goal was to flag the studies with the greatest potential for problems due to mismatch.

The raters were advanced doctoral students in a graduate school of education with research training in educational measurement. The raters worked through four stages: rating the studies independently, meeting to compare notes and adjust the coding procedure, rerating the studies independently, and meeting again to compare ratings and resolve disagreements. For each project that developed an intervention and evaluated it using a standardized achievement test, the raters answered the following question: Did the test appear to measure the same content and skills that the intervention taught?

For the analysis, the raters cross-referenced the goals of each intervention with the information from the IES database and the Internet. Each entry in the IES database contained a section for investigators to articulate the learning goals of their intervention. The first author searched the Internet and located documents from test publishers that contained information about test validity and, for state-developed tests, test blueprints from state education agencies. Some of the state agencies provided detailed information: the standards assessed, descriptions of the underlying theoretical constructs that the test was developed to measure, and samples of many items or complete tests. Other state agencies provided sparse information, such as the standards assessed by the test and a small sample of released items. The first author shared the documents from the Internet with the other two raters.

The raters compared the goals of the intervention with the purpose of the test to make a binary decision about whether a test appeared to align with an intervention or additional evidence was needed to support the claim that the test was aligned with the intervention. Initially, the interrater reliability was inadequate ($\kappa^1 = 0.14$) due to the difficulty in articulating operational definitions of test validity. The raters were using conceptually different thresholds to mark a lack of validity. In response, the first author revised the coding manual to clarify the intended dimensions of validity analysis (i.e., alignment between test and intervention for content and cognitive process). Next, the raters reviewed the manual and independently recoded the studies. Finally, the raters reconvened to compare notes and were in greater agreement. In the same meeting, the raters resolved all disagreements through discussion and reaching complete agreement on the coding ($\kappa = 1.0$). The raters flagged 25 (54.3%) of the 46 intervention projects for potential validity problems. We discuss the findings further in the results section.

Documents received from principal investigators. We asked each of the principal investigators for their final reports to IES (or other reports to IES if final reports were unavailable) and received documents from 33 (42.3%) individuals. The format and length of documents varied between projects, but the data permitted us to investigate whether investigators included validity information.

Of the 33 sets of documents, only 11 matched the 25 projects that the team of raters flagged as potentially problematic (Step 6 of Figure 1). The other 22 sets of documents related to studies were eliminated from our analysis or did not flag. Although we reviewed all the documents, we limited the remainder of the analyses to the 11 studies that had been flagged as potentially problematic. We reasoned that discussions about test validity were more likely to be within the documents than other sources of information, and analysis of a project without a report would be lacking the best source of evidence. We view this as a conservative choice: Although we limited the number of projects in the sample, we also ensured that we did not include incomprehensive analyses of projects, thereby underestimating the attention to validity in applied STEM education research.

In-depth analysis of validity evidence. In Step 6, the same three raters reviewed the corpus of information that supported the validity of the standardized achievement test for evaluating the intervention. The raters worked independently, and then met to compare notes. The raters determined that it was necessary to exclude the studies for which the investigator did not provide a report. Without the report, the amount of information varied too widely between studies.

In Steps 7 and 8 of Figure 1, the raters determined whether investigators discussed the validity and the alignment of the standardized achievement test for evaluating the intervention. Three raters searched the data corpus that included the IES database, reports provided by each principal investigator, and the literature associated with each of studies. The first step was to generate a binary answer to the following question: Do the investigators or authors discuss the validity or the alignment of the test in the context of the evaluation? For the alignment question, the raters searched for literal discussions about alignment as well as more general discussions about the overlap between the test and the intervention at the item or the construct level.

In Step 9, the raters independently coded whether the investigators conducted research for the purpose of validating the standardized achievement test. The raters searched the documents for evidence of research activities, where investigators positioned the results as evidence that warranted the suitability of the test for evaluating the impact of the intervention. In Step 10, the raters attempted to address the question of whether the corpus of validity information was adequate for supporting the proposed use of the test for evaluating the intervention.

The three raters worked independently and came together to compare notes. Notably, the independent ratings from the three raters were in complete agreement on whether or not the documents in the study data (a) contained validity discussions, (b) contained information about validation activities, and (c) supported the validity of the standardized achievement test for evaluating the curricular intervention. At no point did the raters disagree about coding, and the raters generally cited similar evidence from the data to support their coding. Appendix B contains additional information about how the raters scored the validity discussions.

Data Analysis Plan

First, we conducted a quantitative analysis of the binary codes represented in Steps 4, 5, and 7–9 of Figure 1. We calculated arithmetic means for each category and the conditional arithmetic means for combinations of categories. Next, we conducted a qualitative analysis of the project data to describe the observed validity problems and the approaches to validation represented in the projects. Finally, we conducted an analysis of the validity information to (a) judge the adequacy of the validity information presented within each project and (b) elucidate the measurement issues documented by the data.

Results

Projects That Used Standardized Achievement Tests as Outcome Measures

The first research question asked how many of the projects used standardized achievement tests to evaluate the impact of a curricular intervention. We coded the tests listed in the Key Measures section of the research proposals into three categories: standardized achievement tests, researcher-developed tests, and other tests. The answer to this research question was not straightforward because most studies in the IES database listed multiple outcome measures. In the sample of 64 intervention projects within the IES database, only 17 projects listed a single outcome measure: 10 listed a standardized achievement test, 4 listed a test from the other category, and 3 listed a researcher-developed test. For multiple measures, investigators listed tests from two categories in 35 (54.7%) of the projects and from all three categories in 12 (18.8%) of the projects. Some investigators listed multiple tests within the same category, but the descriptions were not sufficiently precise as to allow a more fine-grained analysis (e.g., "This study will administer standardized tests of cognitive ability").

Table 1 contains descriptive statistics showing the categories of tests that investigators most commonly listed as key outcome measures. The rows of Table 1 are nonexclusive, and the analysis

Table 1. Category and Frequency of Key Measures Listed in the IES Database.

Category	*n*	%
Standardized achievement	46	71.9
and researcher-developed	10	15.6
and other	14	21.9
Researcher-developed	36	56.3
and other	14	17.9
Other	41	64.1
Standardized achievement, researcher-developed, and other	12	18.8

Note. IES = Institute of Education Sciences.

may count the same project multiple times. For example, the 12 projects that included all three types of tests increase the number in each row of the table by one. The results show that 46 (71.9%) projects listed a standardized achievement test as an outcome measure—the most out of any category. Although investigators listed standardized achievement tests most often, investigators typically used, or planned to use, standardized achievement tests as part of a portfolio of research evidence that included both quantitative and qualitative information. Closer examination of the type of standardized achievement tests showed that 30 (46.9%) of the projects used state tests, 5 (7.8%) projects used subtests from the Woodcock Johnson-3 Test of Achievement, 3 (4.7%) projects used the American Chemical Society high school chemistry exam, and 2 (3.1%) projects used tests from the ITBS system.

In sum, the results suggest that the use of standardized achievement tests for evaluating new interventions in math and science education is widespread.

Projects with a Mismatch Between the Standardized Achievement Test and the Intervention

Research Question 2 asked whether a standardized achievement test used as an outcome measure appeared to measure the same content and skills that the intervention taught. This question helped us to flag projects for closer examination. Three raters examined the 46 intervention projects that listed a standardized achievement test as an outcome measure, and the raters agreed that 25 (54.3%) of the projects had a potentially problematic mismatch between the test and the intervention. For these projects, validity evidence is especially important if the investigators use the scores to document the educational impact of the program. For five (10.9%) of the projects, one standardized achievement test was the only test listed in the Key Measures section of the database. The investigators on the remaining 20 projects listed at least one other category of test.

We identified two main types of potential validity problems across the 25 projects we flagged. The first issue was that the academic content covered by the intervention was narrow relative to the larger amount of content measured by the achievement test. This was because many interventions were brief units that focused on core ideas within a subdomain of science or math (such as energy or fractions). The academic content of the intervention thus constituted a narrow subset of the content on the test. In these cases, the test measured more, and in many cases much more, content than the intervention taught.

Second, in other (sometimes overlapping) examples, the project listed goals for student learning that would be difficult to measure with a typical standardized achievement test. In one example, an intervention fostered students' ability to conduct scientific investigations. In another example, the goal of the intervention was to nurture student participation (e.g., debate) within a learning community. It is clear that the developers do not design standardized achievement tests to measure these learning goals. We conclude that, at best, the achievement tests would be an approximate measure of

Table 2. Results of Validity and Alignment Analyses.

Study Number	Document Provided	Scholarly Articles Located	Used Standardized Test	Validity Evidence	Alignment Evidence	Validation Activities	Adequate Validity Evidence
1.	Final report	2	Y	N	N	N	N
2.	Interim report	8	Y	Y	N	N	N
3.	Final report	4	Y[a]	N	N	N	N
4.	Final report	6	Y	Y	N	Y	N
5.	Final report	0	Y	N	N	N	N
6.	Interim report	1	Y	Y	Y	Y	N
7.	Final report	3	Y	Y	Y	Y	N
8.	Final report	16	Y[b]	Y	Y	Y	Y
9.	Final report	0	Y	N	N	N	N
10.	Final report	1	Y	N	N	N	N
11.	Final report	8	Y	Y	Y	Y	N

Note. Y = Yes; N = No.
[a]Selected items only. [b]Modified the test

the complex forms of learning that the investigators wanted to know about (Kennedy, 1999). A statistical perspective on the results suggests that a large portion of the test variance may be unrelated to evaluating the direct impact of the intervention. The raters agreed that, in some cases, it would be appropriate to consider the standardized achievement test as a *distal* indicator of student learning (see Ruiz-Primo et al., 2012 for a discussion of the term distal).

Discussions About Validity and Alignment

Research Question 3 asked how many of the projects contained validity discussions that supported the use of standardized achievement tests. We examined the validity evidence contained within the 11 sets of documents, including the report provided by the principal investigator and the published articles associated with the project. Our goal was to comment on the presence or absence of validity evidence, including evidence of alignment, in studies with the greatest potential for problems. To this end, we first evaluated the adequacy of the data for our analysis. Then, we present the results of our search for validity discussions.

Type of Information

Table 2 contains the results of our central analysis. The first column contains a number for each of the 11 projects, and the next three columns describe the evidence that supported our analysis of each project. Column 2 indicates the type of document that the principal investigator provided. Across all studies, investigators provided nine final reports and two interim reports. The interim reports are from the final year of the grant and thus are comparable to the final reports.

Column 3 contains the number of studies published in peer-reviewed journals associated with each grant. The average number of published studies per grant was 4.45 ($SD = 4.60$). The range was large: between zero and 16 studies. The group of projects that contained validity discussions had an average of 4.3 published studies (standard error [SE] = 2.73), whereas studies for which we did not find validity discussions had an average of 2.5 published studies ($SE = 1.84$). Although the validity group had more studies, the result was not statistically significant ($p = .149$).

Column 4 of Table 2 documents whether the principal investigators of all 11 projects completed the evaluation of the intervention using data from the standardized achievement test described in the

IES database. We found that all 11 investigators completed the evaluation. This is important because investigators typically discussed the outcome measure in the context of an evaluation. The superscripts in column 4 indicate that for 2 of the projects, investigators modified the assessment by changing items but otherwise conducted the planned evaluation. The two investigators used very different approaches to validating the modified outcome measure, which we discuss below in context of the validity analysis.

Validity Analysis

Three raters searched the documents and articles and coded whether they found a discussion about the validity of the standardized achievement test for evaluating the intervention. Column 5 of Table 1 contains the consensus among the raters. Six (54.5%) of the 11 projects presented validity evidence to support the use of the test in the evaluation (Projects 2, 4, 6, 7, 8, and 11). We describe the nature of the evidence below. We did not find any validity evidence in the documents and articles for five (45.5%) of the projects. Projects 3 and 8 are noteworthy because the investigators used modified standardized achievement tests. The report from Project 3 stated that the outcome measure was a subset of items from the standardized test. However, the report did not contain validity evidence to justify the choice of items. In contrast, Project 8 used a subset of items, some of which were modified, and included validity evidence to support the use of the test. In sum, the results suggest that investigators are evenly divided on whether it is necessary to provide validity evidence that justifies the use of an outcome measure.

Alignment Analysis

The raters also searched for evidence of alignment as a specific form of validity. Column 6 of Table 1 shows the results of the analysis. Four of the six investigators who discussed validity also discussed alignment (Project 6, 7, 8, and 11). Interpreting the results carefully because of the few studies in our analysis, the evidence suggests that most investigators who discuss validity also reference the concept of alignment. Although alignment is a relatively new concept in test validity, it is interesting to consider the utility of alignment and whether strengthening the role of alignment within argument-based approaches to validity (Kane, 1992) may lead to improvements in applied measurement.

Validation Activities

Research Question 4 asked how many of the projects conducted test validation research. The raters searched the documents for evidence that investigators/authors conducted validation research for the purpose of generating evidence to support the validity of the standardized achievement test. Five (45.5%) of the 11 projects reported conducting validation activities (Project 4, 6, 7, 8, and 11). Four projects reported the conclusions from alignment analyses that investigators had conducted on the test and the intervention (Project 6, 7, 8, and 11). However, only one project elaborated the details of the alignment research (Project 8). The three other projects provided brief conclusions based on methods that they did not identify, so we were unable to assess the rigor or the conclusions of the alignment procedure. Other validation activities included calculating test reliability, calculating the convergent validity of the standardized achievement test with a researcher-developed measure, and analyzing the floor and ceiling effects of the standardized test.

Adequacy of the Validity Discussions

Research Question 5 asked how many of the standardized achievement tests were supported by adequate validity evidence. As expected, we found a range of approaches to validating the use of the standardized tests. However, we were surprised that most of the validity discussions were concise and,

indeed, less than one paragraph. Most of the discussions presented fragments of decontextualized evidence (e.g., reliability coefficients and correlations with other tests) rather than well-developed arguments that use specific forms of evidence to support particular interpretations of test scores. For example, 1 project contained a single sentence of psychometric information from the test publisher: The investigators did not provide validity evidence relevant to their particular investigation.

Only 1 project supplied a detailed validity discussion to support the use of the standardized test for evaluating the impact of the project. The investigators for Project 8 described a formal analysis of the alignment between the standardized test and the intervention. The project described a framework for validating the alignment of the curricular intervention with the test. The investigators conducted an item analysis: They mapped items from a standardized test to (a) the academic content covered by the intervention and (b) the cognitive processes that students might use to solve problems. The authors separated items into those that measured the same content and same processes taught by the intervention and transfer items that measured the application of knowledge to other phenomena (e.g., different contexts or relatively complex process goals such as prediction and explanation). This study was unique in our data and an example of careful measurement using item-level alignment to ensure that the items measured the same content and skills that the intervention taught. The investigators from Project 8 described the outcome measure as an adaptation of a subset of items from a standardized achievement test. Whether or not this counts as a "preexisting" test is perhaps an open question. However, we cast the results as evidence of an important connection between attention to test validity and the careful selection of items that comprise an outcome measure.

Measurement Issues

At least five principal investigators discussed measurement problems in their reports to IES. Four of the six investigators who provided validity evidence indicated that measurement issues related to the use of standardized achievement tests interfered with their planned evaluation in some way (Project 4, 6, 7, and 11). For example, the principal investigator on Project 7 stated that the standardized achievement test was invalid because it did not have enough test items that tapped the content taught by the intervention. This investigator reported that she or he learned a lesson to be more specific about the learning outcomes she or he wants to measure and to select an assessment that will be more sensitive to measuring those outcomes. The investigator of Project 11 indicated that measurement issues related to the lack of alignment between the intervention and the standardized achievement test called into question the credibility of the evaluation.

The investigator from Project 4 described a post hoc process of modifying the standardized achievement test for the purpose of creating a valid outcome measure. At the outset of the project, investigators considered the standardized achievement test a key measure for evaluating the impact of the intervention. However, after the first round of data collection, they reported that the test did not measure the same content and skills that was taught by the intervention. The investigators substituted items and conducted a program of validation research. They described the accumulation of validity evidence over successive publications in peer-reviewed journals and constructed a validity argument to explain proper interpretation of the scores from the standardized achievement test for evaluating the intervention. Ultimately, the validity argument branded the standardized achievement test a transfer test and articulated a theory about how the test measured learning transfer from the intervention to new educational problems, in new learning contexts.

The investigators from 4 projects (6, 7, 8, and 11) discussed other consequences of measurement problems. One investigator ignored the results from the standardized achievement test and used a different source of data collected during the study for a summative evaluation. A second investigator, who determined that the standardized achievement test was not well aligned with the learning goals of the intervention, unsuccessfully attempted to acquire item-level data in order to conduct an

evaluation using only relevant items. Unfortunately, she or he had not collected other outcome data and was thus unable to rigorously evaluate the impact of the intervention. A third investigator conducted the planned evaluation but suggested that the results should be interpreted with caution because of the alignment problems. Another investigator learned, after data collection, that the standardized achievement test contained many irrelevant items.

In addition, some projects that did not contain validity discussions showed evidence of measurement issues. The investigator from Project 3 altered the standardized achievement test after the first round of data collection by selecting a subset of items from the test, but we did not find the investigator's rationale for the changes. A measurement issue seems to be the most likely reason to alter a test mid-study, but without specific information we cannot evaluate the adequacy of the outcome measure and by extension the intervention. In another example, an impact evaluation of an intervention aiming to increase students' conceptual understanding in mathematics used a standardized test of mathematical fluency as an outcome measure. The investigators did not detect statistical differences between treatment and control groups, and they did not discuss the apparent mismatch between the target of influence of the intervention and the target of measurement of the test.

Summary and Discussion

Scholars are just beginning to investigate the issues surrounding the use of standardized achievement tests for evaluating the impact of new educational interventions (e.g., May et al., 2009). This study examined the use and validity of standardized achievement tests for evaluating new educational interventions in mathematics and science. We examined the projects funded through the IES mathematics and science education program. The analysis of information from the IES database, final reports from principal investigators, and published literature associated with each study afforded a unique opportunity to review the use of standardized achievement tests in leading applied research in math and science education.

First, we found that investigators commonly use standardized achievement tests as outcome measures for evaluating educational interventions. Indeed, the investigators used standardized achievement tests as outcome measures more than other forms of tests (e.g., researcher developed tests). Second, many of the projects had potential validity problems related to misalignment between the goals of the intervention and the knowledge and skills required for success on the standardized achievement test used as a key outcome measure. For some projects, the academic content covered by the intervention was narrow relative to the broad amount of academic content measured by the test. For other projects, the intervention had goals for student learning (e.g., scientific reasoning) that were difficult to measure with typical standardized achievement tests. Our analytic framework based on alignment was useful for identifying projects at risk for measurement problems related to test-intervention misalignment in the domains of content and cognitive process.

Third, we closely inspected the 11 projects that we (a) flagged for validity problems and (b) obtained project reports from principal investigators. We found that only 6 of the 11 projects presented or generated evidence that the standardized achievement test was valid for evaluating the intervention. Further, we found that only 1 of the 11 projects presented adequate validity evidence to support the use of the standardized achievement test for evaluating the impact of the intervention. The results confirm our hypothesis that investigators commonly use standardized achievement tests to evaluate new educational interventions in math and science without evidence that the test is valid for this purpose. Validity evidence is the main route through which researchers ensure that a particular test produces accurate and useful information about the impact of an educational intervention. As a matter of principle, without validity evidence, we cannot fully evaluate the adequacy of an outcome measure and, by extension, we cannot fully evaluate the adequacy of an intervention.

Therefore, the lack of validity evidence revealed in this study underlies an inattention to measurement that, we suggest, has the potential to obstruct educational innovation (i.e., Raudenbush, 2005).

Consequences of Measurement Issues

We found copious evidence that measurement issues related to the use of standardized achievement tests had real-world consequences for IES-funded projects. Multiple investigators realized, only after data collection, that the standardized test was not well aligned with the learning goals targeted by their intervention. In response, some investigators used alternate (more useful) sources of outcome data. However, at least one investigator could not evaluate the impact of their intervention for lack of an alternate (i.e., valid) outcome measure. In other situations, investigators neglected to discuss measurement problems that were apparent to the raters. We are left to wonder if the results of these evaluations would have been different under conditions of better measurement. Taken together, the results show that additional attention to measurement is needed to improve the rigor of applied research in math and science education.

Implications for Practice and Policy

The results of this study demonstrate the importance of attending to measurement issues during the planning phases of research to avoid unintended outcomes. First, this study shows an initial proof in concept that measurement specialists can prospectively identify measurement problems in particular evaluation studies. Consultation with measurement experts at an early stage of research design could have saved several IES-funded investigators from wasting time and effort of conducting an evaluation with an inappropriate standardized test. Such consultation could have given researchers the ability to find an appropriate test for their evaluation.

Second, policies that promote good measurement may be able to improve the validity of outcome measures used in applied educational research. IES proposals should require detailed measurement plans that (a) review the validity of each outcome measure and (b) propose analyses to ensure that outcome measures produce accurate and useful information about the educational impact of the curricular intervention. Scoring criteria should include the alignment between the learning goals of the intervention with the skills and knowledge required for success on the outcome measure. In addition, both investigators and reviewers should generally be skeptical of the validity of standardized achievement tests—unless the goals of the intervention align well with the measurement objectives of the test. Standardized achievement tests can produce useful information, but the primary outcome measure must align with the intervention (NRC, 2004; Ruiz-Primo et al., 2012; Tyler, 1942).

We recognize the long-term efforts in the measurement community to promote best practices and testing reform, including construct modeling and a reconceptualization of classroom assessment (i.e., NRC, 2001; Wilson, 2005). In spite of the new tests' improvements in measuring complex learning (Doorey & Polikoff, 2016), it is unclear whether the new tests address *any* of the concerns raised by this research. If complex thinking skills contribute more to total scores for some tests and less for other tests, it is critical that evaluators consider carefully if the particular test provides useful information for evaluating a particular intervention. Future tests in mathematics and science may measure cognitively complex learning and reduce their dependence on multiple-choice items, but other issues (i.e., reliability and adequate coverage) remain.

Limitations and Future Directions

One strength of this study was the straightforward research questions that led to empirical answers. We coded the projects in the IES database along relatively objective criteria that led to straightforward interpretations. However, the trade-off of this method was a lack of ability to conduct deeper

investigations into *why* so many investigators eschew discussions about test validity. We speculate that some investigators do not understand principles of measurement or are unwilling to invest the resources involved in good measurement. Future research with the power to examine the obstacles to good measurement may lead to more specific ways to improve measurement in practice.

A second limitation was an imperfect ability to judge the quality of the validity evidence and by extension our ability to confidently assess the adequacy of each project's validation activities. Fundamentally, we analyzed the presence or the absence of evidence. Although we desired to differentiate between disconnected bits of validity evidence and well-developed (e.g., argument-based) approaches to validity, we recognize that our assessment is subjective and susceptible to human error. A stronger study would have greater power to examine and characterize the relative strength of the evidence. Such an analysis would be a difficult undertaking, as the careful study of validity is complex (i.e., Newton, 2012), and the evidence supplied by the studies is extremely variable. However, attention to new research methods that characterize the use of validity in practice, for example, by contextualizing the scope of analysis as we did in this study, may bring clarity to the concept of validity.

The data in the form of the documents we received from investigators supplied another limitation. We requested reports from principal investigators and received documents from only 11 of the 25 projects that we flagged for potential validity problems. Our conclusions could have been stronger if we had been able to obtain documents from a greater number of projects. In addition, we received a diverse set of documents across projects, which varied in length and detail. Standardization of information would help to ensure a more common basis for comparison. For example, investigators may need to provide a validity discussion that differs according to the complexity of measuring the key outcome. Studies that examine validity in practice must find ways to extract additional, reliable, information and generate new ways of thinking about this complicated idea.

Conclusion

In this study, we examined projects funded by the IES math and science education program to address the use and validity of standardized achievement tests as outcome measures for impact evaluation of new educational interventions. The results of our research demonstrate that (a) investigators use standardized achievement tests as outcome measures more than other forms of tests (e.g., researcher-developed tests) and (b) many projects show the potential for validity problems related to a misalignment between the standardized achievement test and the intervention that it is used to evaluate. Closer examination of a subset of projects with potential validity problems showed that about half of the investigators did not present any validity evidence for the standardized achievement test, and only one investigator adequately validated the test. At the same time, many investigators discussed validity problems related to the use of a standardized achievement test. Some investigators concluded, only after conducting the evaluation, that the test did not measure the learning caused by their intervention. This study highlighted an important weakness in applied research in mathematics and science education. Without validity evidence, we cannot evaluate the adequacy of an outcome measure and, by extension, we cannot evaluate the adequacy of an intervention. The ability of educational research to support innovation depends on addressing the measurement issues highlighted in this study.

Appendix A

Additional Information about Gathering the Data

Literature search. Another data source consisted of peer-reviewed articles that contain research funded through the grants. We used two main search queries. The first was the grant number and

the second was the name of the intervention. Occasionally, we used additional queries such as the name of the principal investigator. We conducted the literature search for only the studies that we flagged as potentially problematic (see procedure).

Web search. The fourth source of data was the result of an Internet search to gather information about the standardized tests used as outcome measures. We used a set of search criteria; pairing the name of the test with the search terms *valid(ity), reliability*, *alignment, construct, and psychometric(s)* in separate searches. Generally, the most helpful search results linked to the websites of test publishers. For example, the search engine results linked to a technical manual for the ITBS (Iowa Testing Programs, 2003). For other tests, technical information was available for a fee (e.g., through the purchase of a manual), and we did not pursue this information. For state-developed and administered tests, the websites of state educational agencies contained test blueprints and released items and thought the amount of information varied between examples.

Appendix B

Analyzing the Validity Evidence within Each Project

Finding validity discussions in the project data. We used inclusive criteria for what constitutes a discussion about validity and alignment. We searched the documents for discussions about validity in accordance with mainstream views on test validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). However, it was not necessary to use the standard lexicon (jargon). The critical feature is that the discussion supports the use and/or the interpretation of the standardized test for the evaluation of program impact

Three raters coded reports and the literature associated with each study, following the same search protocol. We read each document for evidence of validity. Then, to check our work, we re-searched the document text for the following keywords: validity, test(s), assessment(s) item(s), reliability, psychometrics, outcome(s), measure(ment), and alignment. After coding independently, the raters discussed their findings and achieved consensus on the results.

Finding evidence of validation activities in the project data. We searched the documents for evidence of research activities where investigators positioned the results as evidence that warranted the suitability of the standardized test for evaluating the impact of the intervention. We searched for validation activities related to alignment, including conventional validation methods (Roach, Niebling, & Kurz, 2008) and ad hoc methods such as those that rely on examination of test blueprints (Somers et al., 2011) and other forms of item analysis.

Then, we asked whether the discussion occurred in the context of an argument-based approach to validity rather than through disconnected pieces of information. At this stage, our intent was not to judge the merit of the discussions for supporting the use of the test but to only determine whether the document contained these discussions. Finally, we judged the adequacy of the validity information contained within each project. We searched each project's corpus of data for a discussion that incorporated multiple sources of evidence into a persuasive argument that supported the particular use of the test.

Authors' Note

The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences or the U.S. Department of Education. This study is based on the dissertation of the first author.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported in part by the Institute of Education Sciences grant R305B090026 to UC Berkeley.

Note

1. Cohen's Kappa.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Berland, L. K., Schwarz, C. V., Krist, C., Kenyon, L., Lo, A. S., & Reiser, B. J. (2015). Epistemologies in practice: Making scientific practices meaningful for students. *Journal of Research in Science Teaching*, *53*, 1082–1123. doi:10.1002/tea.21257

Baker, E. L., Chung, G. K., & Cai, L. (2016). Assessment gaze, refraction, and blur: The course of achievement testing in the past 100 years. *Review of Research in Education*, *40*, 94–142.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, *22*, 21–29. doi:10.1111/j.1745-3992.2003.tb00134

Brandriet, A., Reed, J. J., & Holme, T. (2015). A historical investigation into item formats of ACS exams and their relationships to science practices. *Journal of Chemical Education*, *92*, 1798–1806.

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review*, *23*, 221–234. doi:10.1007/s10648-011-9161-z

Campbell, D. T. (1991). Methods for the experimenting society. *American Journal of Evaluation*, *12*, 223–260. doi:10.1177/109821409101200304

Catley, K., Lehrer, R., & Reiser, B. (2005). *Tracing a prospective learning progression for developing understanding of evolution*. Retrieved from https://www.researchgate.net/profile/Richard_Lehrer/publication/253384971_Tracing_a_Prospective_Learning_Progression_for_Developing_Understanding_of_Evolution/links/541ec24e0cf241a65a1a90ca.pdfss

Cobb, P., & Jackson, K. (2008). The consequences of experimentalism in formulating recommendations for policy and practice in mathematics education. *Educational Researcher*, *37*, 573–581. doi:10.3102/0013189X08327826

Confrey, J. (2006). Comparing and contrasting the National Research Council report on evaluating curricular effectiveness with the What Works Clearinghouse approach. *Educational Evaluation and Policy Analysis*, *28*, 195–213. doi:10.3102/01623737028003195

Confrey, J., & Scarano, G. H. (1995, October). *Splitting reexamined: Results from a three-year longitudinal study of children in grades three to five*. Paper presented at the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Columbus, OH.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, *22*, 5–11. doi:10.1111/j.1745-3992.2003.tb00132.x

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., … Steele, C. M. (2013). *Criteria for high-quality assessment*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Retrieved from https://edpolicy.stanford.edu/sites/default/files/publications/criteria-higher-qualityassessment_0.pdf

DeBarger, A. H., Penuel, W. R., & Harris, C. J. (2013). *Designing NGSS assessments to evaluate the efficacy of curriculum interventions*. Paper presented at the K-12 center at ETS invitational research symposium on science assessment, Washington, DC. Retrieved from https://www.ets.org/Media/Research/pdf/debarger-penuel-harris.pdf

DeBarger, A. H., Penuel, W. R., Harris, C. J., & Kennedy, C. A. (2016). Building an assessment argument to design and use next generation science assessments in efficacy studies of curriculum interventions. *American Journal of Evaluation*, *37*, 174–192. doi:10.1177/1098214015581707

Donaldson, S. I., Christie, C. A., & Mark, M. M. (2009). *What counts as credible evidence in applied research and evaluation practice?* Thousand Oaks, CA: Sage. doi:10.4135/9781412995634

Doorey, N., & Polikoff, M. (2016). *Evaluating the content and quality of next generation assessments*. Washington, DC: Thomas B. Fordham Institute.

Frederiksen, J. R., & White, B. Y. (2004). Designing assessments for instruction and accountability: An application of validity theory to assessing scientific inquiry. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 74–104). Chicago, IL: University of Chicago Press.

Greeno, J., Collins, A., & Resnick, L. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology (pp. 15–46)*. New York, NY: Macmillan.

Greeno, J., Pearson, P., & Schoenfeld, A. (1997). Implications for the National Assessment of Educational Progress of research on learning and cognition. In R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Assessment in transition: Monitoring the nation's educational progress, background studies (pp. 151–215)*. Stanford, CA: National Academy of Education.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments. *Applied Measurement in Education*, *20*, 101–126. doi:10.1080/08957340709336732

Iowa Testing Programs. (2003). *Iowa test of basic skills research guide*. Retrieved from https://itp.education.uiowa.edu/ia/documents/ITBS-Research-Guide.pdf

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535. doi:10.1037/0033-2909.112.3.527

Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis*, *21*, 345–363.

Lagemann, E. C. (2002). *An elusive science: The troubling history of education research*. Chicago, IL: University of Chicago Press. doi:10.1086/386402

Lehrer, R. (2009). Designing to develop disciplinary disposition: Modeling natural systems. *American Psychologist*, *64*, 759–771. doi:10.1037/0003-066x.64.8.759

Lipsley, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. New York, NY: Sage.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*, 1332–1361. doi:10.3102/0034654309341375

May, H., Johnson, I. P., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009–013). Washington, DC: National Center for Education Evaluation and Regional Assistance.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. doi:10.1037/0003-066x.50.9.741

McClellan, C., Joe, J., & Bassett, K. (2017a). *Still on the right trajectory: State teachers of the year compare former and new state assessments*. Retrieved from National Network of State Teachers of the Year website: http://www.nnstoy.org/wp-content/uploads/2017/04/Still-on-the-Right-Trajectory.pdf

McClellan, C., Joe, J., & Bassett, K. (2017b). *Beginning a higher trajectory: Grade 11 study. State teachers of the year compare former and new state assessments*. Retrieved from National Network of State Teachers of the Year website: http://www.nnstoy.org/wp-content/uploads/2017/04/NNSYOY-11th-Grade-Report.pdf

National Research Council. (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press. doi:10.17226/6336

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: The National Academies Press. doi:10.1037/e302582005-001

National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: Author. doi:10.17226/11025

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Committee on Conceptual Framework for the New K-12 Science Education Standards. Board on Science Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. doi:10.17226/13165

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, *10*, 1–29. doi:10.1080/15366367.2012.669666

Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. Washington, DC: United States Government Printing Office.

Olsen, R., Unlu, F., Price, C., & Jaciw, A. (2011). *Estimating the impacts of educational interventions using state tests or study-administered tests* (NCEE 2012–4016). Washington, DC: National Center for Educational Evaluation and Regional Assistance.

Pearson Corporation. (2017). *PARCC: Final technical report for 2016 administration*. Retrieved from https://www.isbe.net/Documents/PARCC%202016%20Tech%20Report.pdf

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, *340*, 320–323.

Pellegrino, J. (2015). *Session L: Measuring what matters: Challenges and opportunities in assessing science proficiency*. Paper presented at the 2015 ACER Research Conference. Retrieved from http://research.acer.edu.au/research_conference/ RC2015/17august/15

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the next generation science standards*. Washington, DC: National Academies Press.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, *56*, 8–16.

Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*, 3–14. doi:10.3102/0013189x031007003

Porter, A. C., Polikoff, M. S., Barghaus, K. M., & Yang, R. (2013). Constructing aligned assessments using automated test construction. *Educational Researcher*, *42*, 415–423. doi:10.3102/0013189X13503038

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education*, *20*, 27–51. doi:10.1080/08957340709336729

Quellmalz, E. S., Davenport, J. L., Timms, M. J., DeBoer, G. E., Jordan, K. A., Huang, C.-W., & Buckley, B. C. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology*, *105*, 1100–1114. doi:10.1037/a0032220

Raudenbush, S. W. (2005). Learning from attempts to improve schooling: The contribution of methodological diversity. *Educational Researcher*, *34*, 25–31. doi:10.3102/0013189X034005025

Rhue, V., & Zumbo, B. D. (2008). *Evaluation in distance education and e-learning: The unfolding model*. New York, NY: Guilford Press.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools*, *45*, 158–176. doi:10.1002/pits.20282

Ruiz-Primo, M. A., Li, M., Wills, K., Giamellaro, M., Lan, M.-C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching*, *49*, 691–712. doi:10.1002/tea.21030

Schafer, W. D., Wang, J., & Wang, V. (2009). Validity in action: State assessment validity evidence for compliance with NCLB. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 195–212). Charlotte, NC: Information Age.

Schoenfeld, A. H. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher*, *35*, 13–21. doi: 10.3102/0013189X035002013

Slavin, R., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, *4*, 370–380. doi:10.1080/19345747.2011.558986

Smarter Balanced Assessment Consortium. (2016). *2014-15 technical report*. Retrieved from https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf

Somers, M., Zhu, P., & Wong, E. (2011). *Whether and how to use state tests to measure student achievement in a multi-state randomized experiment: An empirical assessment based on four recent evaluations* (NCEE 2012-015). Washington, DC: National Center for Educational Evaluation and Regional Assistance.

Song, M., & Herman, R. (2010). *A practical guide on designing and conducting impact studies in education: Lessons learned from the What Works Clearinghouse*. Retrieved from http://www.air.org/sites/default/files/downloads/report/Song__Herman_WWC_Lessons_Learned_2010_0.pdf

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*, 298–318.

Taylor, J., Kowalski, S., Wilson, C., Getty, S., & Carlson, J. (2013). Conducting causal effects studies in science education: Considering methodological trade-offs in the context of policies affecting research in schools. *Journal of Research in Science Teaching*, *50*, 1127–1141. doi:10.1002/tea.21110

Tyler, R. W. (1942). General statement on evaluation. *The Journal of Educational Research*, *35*, 492–501.

Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). *An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS*. Stanford, CA: Stanford NGSS Assessment Project Team. Retrieved from https://snapgse.stanford.edu/snap-reports/snap-reports

Wiliam, D. (2008). International comparisons and sensitivity to instruction. *Assessment in Education: Principles, Policy & Practice*, *15*, 253–257.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.