Title
Standardized tests as outcome measures for evaluating instructional interventions in mathematics and science

Author
Sussman, Joshua Michael

Publication Date
2016

Peer reviewed|Thesis/dissertation

Standardized tests as outcome measures for evaluating instructional interventions in mathematics and science

By

Joshua Michael Sussman

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

In

Education

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark Wilson, Chair
Professor Frank Worrell
Professor Geoffrey Saxe
Professor Silvia Bunge

Fall 2016

Standardized tests as outcome measures for evaluating instructional interventions in mathematics and science

Abstract

Standardized tests as outcome measures for evaluating instructional interventions in mathematics and science

by

Joshua Michael Sussman

Doctor of Philosophy in Education

University of California, Berkeley

Professor Mark Wilson, Chair

This three-paper dissertation explores problems with the use of standardized tests as outcome measures for the evaluation of instructional interventions in mathematics and science. Investigators commonly use students' scores on standardized tests to evaluate the impact of instructional programs designed to improve student achievement. However, evidence suggests that the standardized tests may not measure, or may not measure well, the student learning caused by the interventions. This problem is special case of a basic problem in applied measurement related to understanding whether a particular test provides accurate and useful information about the impact of an educational intervention. The three papers explore different aspects of the issue and highlight the potential benefits of (a) using particular research methods and of (b) implementing changes to educational policy that would strengthen efforts to reform instructional intervention in mathematics and science.

The first paper investigates measurement problems related to the use of standardized tests in applied educational research. Analysis of the research projects funded by the Institute of Education Sciences (IES) Mathematics and Science Education Program permitted me to address three main research questions. One, how often are standardized tests used to evaluate new educational interventions? Two, do the tests appear to measure the same thing that the intervention teaches? Three, do investigators establish validity evidence for the specific uses of the test? The research documents potential problems and actual problems related to the use of standardized tests in leading applied research, and suggests changes to policy that would address measurement issues and improve the rigor of applied educational research.

The second paper explores the practical consequences of misalignment between an outcome measure and an educational intervention in the context of summative evaluation. Simulated evaluation data and a psychometric model of alignment grounded in item response modeling generate the results that address the following research question: how

do differences between what a test measures and what an intervention teaches influence the results of an evaluation? The simulation derives a functional relationship between alignment, defined as the match between the test and the intervention, and treatment sensitivity, defined as the statistical power for detecting the impact of an intervention. The paper presents a new model of the effect of misalignment on the results of an evaluation and recommendations for outcome measure selection.

The third paper documents the educational effectiveness of the Learning Mathematics through Representations (LMR) lesson sequence for students classified as English Learners (ELs). LMR is a research-based curricular unit designed to support upper elementary students' understandings of integers and fractions, areas considered foundational for the development of higher mathematics. The experimental evaluation contains a multilevel analysis of achievement data from two assessments: a standardized test and a researcher-developed assessment. The study coordinates the two sources of research data with a theoretical mechanism of action in order to rigorously document the effectiveness and educational equity of LMR for ELs using multiple sources of information.

To Marne

Contents

Acknowledgements

I would like to express my gratitude to the following people who made this dissertation possible. First, I want to thank Frank Worrell for the opportunity to attend a graduate program that helped me to grow as a researcher, a professional school psychologist, and a person. I also thank him for the opportunities to collaborate, to learn academic writing, and for the discussions that helped me, among other things, learn to appreciate the complexities of social systems. Second, I wish to thank Mark Wilson for the research opportunities and for helping me develop skills as a methodologist. I thank him for his candor that helped me simultaneously learn about and dig deeply into important issues in education. I also thank him for his guidance in pushing this research forward. I wish to thank Geoff Saxe for pushing me to think more deeply about education, cognition, and development. I would like to thank Karen Draney for helping me to climb the learning curve of Rasch modeling and for her consistent support. I aim to pay it forward with new Rasch modelers. I also wish to thank the students and professional faculty in the school psychology program for creating a learning environment that is simultaneously rigorous, supportive, and enjoyable. Thanks to my cohort in particular for inspiring me to be better both academically and professionally. My thanks also to the students in the Quantitative Methods and Evaluation Program for the productive collaborations and for treating me like one of their own. I wish to thank my mother, Joan Sussman, for her tireless dedication to her family and for instilling her strength in her children. Last but not least, thanks to Marne Sussman for her encouragement to pursue my goals, her self-sacrifice in helping me to complete this degree, and her editing.

Chapter 1

Introduction

1.1 Standardized tests in United States education

Since the middle of the 19th century, scores from standardized tests have played an increasingly important role in decision making at all levels of education in the United States (Hubert & Hauser, 1999; Office of Technology Assessment [OTA], 1992; Perie, Park, & Klau, 2007). In recent history, educational leaders have used data from standardized tests as evidence to support an array of decisions about students, teachers, schools, and educational programs (Abrams, Pedulla, & Madaus, 2003; Linn, 2001; Slavin, 2008; Wang, Beckett, & Brown, 2006). The uses span student-level decisions about admissions, tracking, promotion, graduation, and eligibility for special education; teacher-level decisions regarding quality and awards for high quality teaching; school-level decisions about autonomy and accreditation; and macro-level decisions about funding and implementing new educational programs. In addition, educators make curricular and pedagogical decisions based on test scores with the goal of improving students' test scores (Shephard, 2010). Finally, parents select schools for their children based on rankings derived in part from test scores (Billingham & Hunt, 2016). Thus, a variety of different stakeholders use data from standardized tests to support various types of decision-making.

This dissertation focuses on a specific use of standardized tests as outcome measures for evaluating the impact of instructional interventions in mathematics and science. Historically, Title I of the 1965 Elementary and Secondary Education Act (ESEA; H.R. 2632, 1965) advanced the use of standardized tests as outcome measures for evaluating new educational programs (Lagemann, 2000). ESEA provided funding for programs that aimed to improve education for "deprived" students. The legislation also required participating schools to test students using "objective measurements of educational achievement…for evaluating at least annually the effectiveness of the programs…" (p. 31). Administrators typically interpreted this requirement as a directive to assess students using standardized tests (OTA, 1992).

Current education policy encourages the continued use of scores from standardized tests for evaluating the effectiveness of educational interventions. For example, the What Works Clearinghouse, a federal agency that generates research syntheses, considers scores on standardized tests to be valid and does not require investigators to establish validity evidence for the specific use of the test (Song & Herman, 2010). Other forms of assessment, such as researcher-developed measures, must be validated—ostensibly because of the potential for bias in the measure that leads to results that overestimate treatment effects (i.e., Slavin & Madden 2011).

1.2 Problems with the use of standardized tests

Problems with the use of standardized tests as outcome measures have become increasingly salient in the literature. One important problem is the mismatch between the measurement targets of standardized tests and the goals of new instructional interventions (Pellegrino, Chudowsky, & Glaser, 2001). Standardized tests in mathematics and science contain an excess of items that measure relatively basic skills such as computation and rote recall (Greeno, Pearson, & Schoenfeld, 1996; Porter, Polikoff, Barghaus, & Yang 2013). Evaluating students' growth in these basic skills was more appropriate in the past, when instructional interventions focused on piecemeal approaches to knowledge building and emphasized direct instruction to teach content and procedures (Greeno, Collins, & Resnick, 1996). In contrast, new interventions in mathematics and science typically emphasize core generative ideas and engage students in problem-solving activities that develop their higher-order cognitive skills (Lehrer, 2009; National Research Council, 1989; Quinn, Schweingruber, & Keller, 2012). Standardized tests do not adequately measure students' higher-order (e.g., reasoning and inquiry) skills that represent key learning targets for many new instructional interventions (Quellmalz et al., 2013). Therefore, it is likely that a new instructional intervention will be badly misjudged when it is assessed using a standardized test. At present, it is unclear whether new standardized tests (e.g., Partnership for Assessment of Readiness for College and Careers; Smarter Balanced Assessment Consortium) address any of these concerns.

1.3 The use of standardized tests as outcome measures

Scholars are only beginning to directly investigate the use and validity of standardized tests for evaluation of new educational interventions (May, Johnson, Haimson, Sattar & Gleason, 2009; Olsen, Unlu, Price, & Jaciw, 2011; Somers, Zhu, & Wong, 2011). The handful of studies in the literature suggest that researchers commonly use standardized tests for evaluating the impact of educational interventions. The studies also support the concern in the prior literature about the disconnect between what the tests measure and what the interventions teach. If the wrong outcome is measured, useful interventions may be judged to be ineffective.

The purpose of this dissertation is to advance the study of standardized tests as outcome measures for evaluation of new instructional interventions. The first paper gathered data to (a) determine how often investigators used standardized tests as key outcome measures and (b) to examine whether investigators established validity evidence

for the specific use of the test per recommendations in the literature (American Educational Research Association, American Psychological Association, National Council of Measurement in Education [AERA, APA, NCME], 2014). In the second paper, I explored the consequences of misalignment between an outcome measure and an educational intervention. The third paper contains a rigorous evaluation of an equitable mathematics intervention that coordinated data from a standardized test, data from a researcher developed test, and a theoretical mechanism of action for the intervention. Together the three papers explore different aspects of the problems with using standardized tests in applied research and orient the reader to potential solutions moving forward.

Chapter 2

The use and validity of standardized tests for evaluating new curricular interventions in mathematics and science

2.1 Abstract

Scholars are beginning to conduct well-developed investigations into specific problems with the use of scores from standardized tests as dependent variables in educational research. In this paper, I investigated the use and validity of standardized tests for summative evaluation of educational interventions funded by the Institute of Education Sciences (IES) math and science education program. I examined information from 78 projects in the IES online database, final reports from 33 of the projects, and the published literature associated with selected projects. The results show that 46 projects (59%) evaluated, or planned to evaluate, a new educational intervention using data from a standardized test. I closely examined the 11 projects for which I had an adequate corpus of data and found that only 6 of the 11 projects (54.5%) considered the validity of the standardized test for the evaluation. Only 5 of the 11 projects (45.5%) conducted validation activities aimed to support the interpretation and use of the standardized test. The evidence suggests that it is common practice to use standardized tests for summative evaluation of new interventions in math and science without establishing validity evidence for the specific use of the test. I discuss the implications for applied educational research and its role in supporting educational reform.

2.2 Introduction

Leading researchers in education develop new classroom interventions designed to influence student learning. The researchers often use data from the same standardized tests, administered for accountability purposes (i.e., Linn, 2000), to evaluate the impact of the interventions. The goal of this study is to investigate whether the researchers who conduct these evaluations provide adequate evidence that the test scores are valid for the specific purpose. I ask whether evaluations that rely on data from standardized tests could misjudge new educational interventions and negatively impact educational reform.

The literature urges cautious use of scores from standardized tests. Scholars of various areas of education have argued that standardized tests neglect to measure important aspects of academic competence (Frederiksen & Collins, 1989; Greeno, Pearson, & Schoenfeld, 1996; Pellegrino, Chudowsky, & Glaser, 2001; Thier, 2004; Wiggins, 1993). However, well-developed investigations into the use of standardized tests for evaluation of educational interventions constitute a relatively new area of the literature (May, Johnson, Haimson, Sattar, & Gleason, 2009; Olsen, Unlu, Price, & Jaciw, 2011; Somers, Zhu, & Wong, 2011). It is widely accepted that the validity of an outcome measure impacts the credibility of an evaluation (Lipsley, 1990; Rhue & Zumbo, 2009), but the literature lacks information about how investigators validate scores on standardized tests to position the scores as evidence to support the efficacy of new educational interventions. The newly developed standardized tests in mathematics and science used in K-12 settings appear to complicate the situation (Pellegrino & Quellmalz, 2010) rather than resolve enduring issues in applied educational measurement related to test validity (e.g., Polikoff, 2012).

The purpose of this study was to generate information about the use and the validity of scores on standardized tests for the summative evaluation of new math interventions and science interventions. To conduct the investigation, I needed a sample of high quality applied research. I gathered information about projects funded through the Institute for Education Sciences (IES) math and science education program. I collected data from their online database, obtained reports from the principal investigators on the projects, and examined publications related to the projects. Importantly, as a condition of use, I assured the principal investigators who supplied reports that this analysis would, to the extent possible, maintain the anonymity of individual projects.

The first research goal was to document the scope of the issue. I asked how common is it for researchers to use standardized tests for summative evaluation of new math or science interventions? The second goal was to screen the studies and flag those with reasonable potential for problems related to the validity of the standardized test for evaluating the impact of the intervention. A team of raters reviewed the projects and applied a straightforward rubric to determine whether each project presented a potential validity problem. Third, the raters closely examined the flagged projects for validity evidence. If investigators presented evidence of validity in reports or peer-reviewed research, I characterized the nature of the evidence. In total, I interpreted the analysis as a snapshot of applied measurement in mathematics and science education related to the use and validity of standardized tests.

2.2.1 Problems with the use of standardized tests for evaluating new educational interventions

Standardized tests may not be suitable outcome measures for evaluation of educational interventions because of a basic mismatch between the knowledge and skills that the test measures with the knowledge and skills that the intervention teaches. In this article, I define an educational intervention as a program or activity that schools and teachers use as an organizing framework for teaching and learning. An intervention "provides structured or sequenced activities designed to influence student learning in some intended way" (Halverson, 2010, p. 136). Examples of interventions include brief lesson sequences that focus on key areas of learning within a discipline, computer tutors, and comprehensive, year-long curricula for both supplemental and replacement purposes.

Different amounts of academic content

The mismatch between a standardized test and an intervention can occur when the content measured by the test differs from the content taught by the intervention. The developers of standardized tests and the developers of academic interventions often focus on different grain sizes of academic content. Test developers typically create standardized tests for reliable measurement of a broad range of academic content and skills (i.e., a year long curriculum sequence) within a major subject area such as English Language Arts or mathematics (Popham, 1999, 2001). The test developers and/or the end users conduct *alignment* research (Bhola, Impara, & Buckendahl, 2003) to ensure that the tests adequately cover the full range of content and skills that students are expected to master over the course of an entire school year.

Standardized tests often measure more academic content than new interventions teach. The interventions focus on fewer content areas rather than covering a larger number of areas more superficially. A standardized test (that is aligned with the grade-level standards) may contain many items that have little relevance to the intervention, especially if the intervention aims for advanced learning not covered in the standards (e.g., Confrey & Scarano, 1995). In sum, the difference in specifications between standardized tests and intervention content precludes standardized tests from providing accurate and useful information about many educational interventions (Somers et al., 2011). Borrowing from Kane's (2013) vernacular, I point to the possibility of a misalignment between the *target domain* of the standardized test and the goals of the educational intervention.

Different emphasis on cognitive skills

Second, standardized tests measure only a subset of the academic skills currently considered important in mathematics education (Schoenfeld, 2006) and science education (Taylor, Kowalski, Wilson, Getty, & Carlson, 2013). For example, the mathematics education literature conceptualizes mathematical competence using five "strands" of mathematical proficiency (National Research Council, 2001) and some authors posit that standardized tests, including most of the tests that states use to evaluate educational progress, do not measure all the strands (Confrey, 2006).

Moreover, among the skills considered important in learning mathematics and science, standardized tests tend to measure lower level skills. Many tests rely on item formats with relatively simple question and answer stems that measure basic thinking skills such as computation, declarative knowledge, and recall of information in memory (Madaus, West, Harmon, Lomax, & Viator, 1992; Noble et al., 2012; Quellmalz et al., 2013). Indeed, Porter, Polikoff, Barghaus, and Yang (2013) demonstrated that items from leading standardized tests in science mainly tapped students' memorization ability. Relatively few items required students to analyze information or apply knowledge. Finally, standardized tests produce little to no information on other research based indicators of academic success, such as what a student can produce over an extended period of time using good research skills, organization, and reflective self-assessment (Fredericksen & White, 2004).

In contrast, many new educational interventions create learning environments that aim to develop students' higher-level thinking skills. Many interventions aim to engage students in the complex forms of thinking over time (e.g., Greeno, Collins, & Resnick, 1996; Lehrer, 2009). For example, science interventions may aim to develop students' ability to construct scientific models of natural phenomena in order to explain the nature of a scientific mechanism (Berland et al., 2015). The interventions may require students to use evidence in well-defended arguments and to revise arguments in the face of new information—skills that most items on standardized tests do not measure. Although the utility of standardized tests for measuring higher-level thinking skills is an open debate in the literature (i.e., Baxter & Glaser, 1998), to my knowledge no literature claims that a standardized test is a *good* measure of the types of learning outcomes targeted by many new educational interventions. Instead, many scholars posit that assessing cognitively complex learning requires different approaches to assessment outside of conventional standardized tests (Brown & Wilson, 2011; Catley, Lehrer, & Reiser, 2005; Frederiksen & White, 2004).

Studies in the literature have discussed the misalignment between standardized tests and new educational interventions in various ways. Some scholars who seek to evaluate innovative educational interventions talk about standardized tests as *narrow* measures of academic competence (e.g., Collins, Joseph, & Bielaczyc, 200). Kennedy (1999) interpreted the data from standardized test as an "approximate" measurement of complex forms of learning that investigators truly want to assess. This paper extends related ideas with a framework based on misalignment along dimensions of content and cognitive process. When a content or cognitive skills mismatch occurs, it is reasonable to question whether the test scores accurately capture the impact of the intervention and hence whether the test scores are valid for summative evaluation of program impact. With respect to the ongoing changes in standardized assessment, it is unclear whether or how the new standardized tests (e.g., Partnership for Assessment of Readiness for College and Careers; Smarter Balanced Assessment Consortium) have advanced in these arenas.

2.2.2 Appropriate use of standardized tests as outcome measures

The recommendations in the literature that support the appropriate use of standardized tests are the same recommendations that support the appropriate use of all tests. It is

widely agreed that the use of test scores—for any purpose—must be supported by evidence that the test is valid for its intended purpose (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, & NCME], 2014). Test validity is the degree to which evidence and theory support the interpretation and use of scores. The consensus is that validity is a "necessary condition for the justifiable use of a test" (AREA, APA, & NCME, 2014, p.11). Importantly, validity is not a property of a test; it is a property of a test for a certain purpose (Messick, 1995). Evaluators should not rely on prior validation activities when using a standardized test as an outcome measure. If the evaluators do not verify the connection between the learning caused by the intervention and the performance on the test items, the test may in fact be measuring the wrong outcomes, leading to an evaluation that badly misjudges the intervention.

Alignment and the validity of standardized tests

A series of reports sponsored by the National Center on Educational Evaluation recommended that investigators should use alignment methods to determine whether state-administered standardized tests are valid for evaluating educational interventions (May et al., 2009; Olsen, et al., 2011; Somers et al., 2011). The basic goal of alignment is to determine the congruence between a test and a set of learning goals—typically educational standards. The purpose of alignment is to ensure that the test measures the knowledge and skills of the intended, or enacted, curriculum (Porter, Smithson, Blank, & Zeidner, 2007). 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, per educational standards or the enacted curriculum (Bhola et al., 2003; Herman, Webb, & Zuniga, 2007; Martone & Sireci, 2009)

Although the typical goal of an alignment method is to verify the match between an assessment and a set of educational standards in a subject area such as mathematics or science, the same ideas apply to alignment of tests with the goals of an educational intervention (Porter, 2002). According to May and colleagues (2009), alignment between the test and the intervention can be established "by determining the proportion of test items that measure skills and knowledge targeted by the intervention" (p. 7) using the alignment methods described in Porter, Polikoff, and Smithson (2009), or Webb (2007). In this study, I investigated whether researchers followed this advice and used the concept of alignment in the context of validity discussions or validation activities.

Federal leadership and the use of standardized tests

Federal policy encourages the use of standardized tests for evaluating instructional interventions. IES distributes an annual budget for research and development in the 170 million to 200 million dollar range (IES, 2015). The Education Sciences Reform Act (H.R. 3801, 2002) describes the agency's goals:

> "[The] systematic use of knowledge or understanding gained from the findings of scientifically valid research and the shaping of that knowledge or understanding into products or processes that can be applied and evaluated and may prove useful

> in areas such as the preparation of materials and new methods of instruction and practices in teaching, that lead to the improvement of the academic skills of students, and that are replicable in different educational settings" (p. 1942).

IES maintains the tradition of Federal support for educational innovation paired with the requirement for particular forms of program evaluation (i.e., Lagemann, 2000). The idea that evaluation is a necessary part of educational reform is not a controversial idea; it is widely agreed that new math or science interventions must be validated (e.g., National Mathematics Advisory Panel, 2008). However, the IES guidelines prescribe certain evaluation goals and methods. Researchers can have varied goals for evaluation, but federal support usually requires investigators to conduct a summative evaluation of the impact of the educational program (see Campbell, 1991, for a description of this form of evaluation). This evaluation model leads investigators to certain experimental methods that include random sampling and, relevant to the current investigation, the use of standardized tests as outcome measures (Donaldson, Christie & Mark, 2007; Comfrey, 2006; Taylor, et al., 2013).

Politicians, some researchers and the public have long considered standardized tests to be rigorous and objective assessments of academic competence (Office of Technology Assessment, 1992). Policymakers care about how interventions influence the outcomes of standardized tests (Somers et al., 2011). Some authors argue that different testing options, such as researcher-developed tests, lead to experimenter bias and to inflation of estimated treatment effects (Slavin & Madden, 2011). Indeed, policy from the What Works Clearinghouse (WWC), created by IES to evaluate research evidence on the effectiveness of educational interventions, codifies the perspective that researcher-developed tests have a second-class status in research syntheses. Per WWC policy, impact evaluations of new educational interventions submitted for inclusion in the WWC database need not contain evidence of reliability or validity for standardized tests (Song & Herman, 2010). Thus, evaluations based on standardized tests (a) gain an automatic level of credibility and (b) circumvent the need to gather and present validity evidence. Indeed, data from standardized tests are inexpensive and easy to obtain compared to the time, effort, and skill required to develop, validate, and administer a researcher-developed test (May et al., 2009). The convenience factor tilts the cost benefit ratio in favor of standardized tests compared to other options.

Taken together, the evidence suggests there is a tension between research and practice. The literature recommends rigorous validation of standardized tests, whereas policy paves the way for their permissive use. Research is needed to better understand whether the use of standardized tests in applied research is widespread, and whether there is in fact a gap between research and practice. The sample of projects funded through the IES mathematics and science education program offer a useful data source for investigating these research questions.

2.2.3 The current study

In this study, I investigated the use and validity of standardized tests for evaluating new educational interventions in math and science. First, I documented how often investigators funded through the IES mathematics and science education program use

standardized tests to evaluate new interventions. I calculated the proportion of projects that evaluated, or planned to evaluate, new interventions using data from standardized tests. Second, I identified the projects with potential validity problems where the standardized test did not appear to measure what the intervention attempted to teach. Multiple 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. I discuss potential validity problems rather than actual problems because the analysis did not permit us to confidently state whether a test use was valid or invalid.

Third, I assessed the validity evidence presented by investigators. I wrote to principal investigators and asked for the most recent project report to IES, which became the research data (e.g., Spybrook & Raudenbush, 2009). I examined the reports for evidence that the investigators discussed the validity of the standardized test scores. I 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 test scores and to (b) support the interpretation of standardized test scores for summative evaluation of program impact. I also searched for specific evidence that investigators considered the alignment between the test and the intervention as recommended by the literature (May et al., 2009).

Fourth, I searched the IES reports for evidence that investigators carried out *test validation* activities. Test validation is a more active process of generating validity evidence. For this study, I define test validation as research activities that produce evidence that 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 alouds or cognitive labs) to understand how test takers approach the items. Fifth, I concluded the study with a qualitative analysis that permitted general comments on the nature of test validity in applied educational research.

2.3 Method

2.3.1 Data

The data for this study comes from four sources. The first source is the Institute for Education Sciences (IES) online database of funded research grants and contracts (termed *projects* hereafter), found at the following URL: 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 (2003 – 2015). From this database, I 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. I 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), I requested interim reports or proposals. I contacted 68 principal investigators (some individuals served as principal investigators on multiple grants) up to three times if I did not receive responses. Forty-eight principal investigators responded (70.6%), and investigators from 33 (42.3%) projects provided one or more documents.

I received varied documents between projects. Most of the documents were final reports; some were interim reports, and a handful 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 five and 355 pages. Thus, for some projects, I had hundreds of pages of information. For other projects, I had a few pages. The diversity of information complicated the analysis. However, I applied a consistent procedure to all the documents and based the analysis on the information that was available.

The third data source consisted of peer-reviewed articles that contain research funded through the grants. I identified the articles using two methods. First, I gathered references from the project page in the IES database. Second, I searched Proquest, Google, and Google Scholar databases for articles connected to the grant. I used two main search queries. The first was the grant number and the second was the name of the intervention. Occasionally, I used additional queries such as the name of the principal investigator. I conducted the literature search for only the studies that I flagged as potentially problematic (see procedure).

The fourth source of data was the result of an Internet search to gather information about the standardized tests used as outcome measures. I entered the name of a test into the Google search engine and examined the results for technical information. I 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 Iowa Test of Basic Skills (ITBS; Iowa Testing Programs, 2003). For other tests, technical information was available for a fee (e.g., through the purchase of a manual) and I did not pursue this information. For state-developed and administered tests, the websites of state educational agencies contained test blueprints and released items, thought the amount of information varied between examples.

I assured the principal investigators that this manuscript would maintain the anonymity of individual projects. I expected to generate critical analyses that would reveal specific weaknesses in the research projects. In order to encourage investigators to provide access to the IES reports that might contain detailed evidence of the weaknesses, I informed the investigators that the analysis would not directly enable readers to connect the measurement problems I discuss with the individual studies that supplied evidence of these problems.

2.3.2 Procedure

Coding

Figure 2-1 contains a diagram of the analysis plan. The diagram demonstrates the logic of the procedure. The research questions are nested in particular steps. The rounded rectangle on the left represents the beginning of the procedure and the initial sample of 85 projects in the IES database. The first arrow pointing to the rectangle signifies the first step in the method. In this step, I retained 78 projects for the next step in the analysis and excluded seven projects that did not list at least one measure in the key measures section.

Classifying the outcome measures. The rounded rectangle in Step 3 summarizes a multistep procedure for classifying the outcome measures used in each project. The purpose was to generate a high-level description of the types of outcome measures used in the studies. For each project, I examined the section in the IES database entitled *Key Measures*. The Key Measures section of each project 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.

I developed a three-category classification scheme to distinguish qualitatively different types of measures. The three categories were *standardized tests, study-developed tests,* and *other tests.* I defined the three categories of tests as follows: a standardized test is an academic assessment that was designed to measure broad range of mathematics or science content. The term broad means that the content is taught and learned over a long period (months) of time. Examples of standardized tests found in the studies include state-developed and administered tests used for accountability and other commercially available assessments such as the Iowa Test of Basic Skills (ITBS), tests from the Terra-Nova series, or the American Chemical Society (ACS) high school chemistry test.

A study-developed test included any assessment developed by the investigator for research or evaluation purposes. For example, some researchers developed tests for measuring complex skills such as scientific inquiry or mathematical reasoning. The category of other tests included 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. I also placed tests into the "other" test category when the description in the key measures section was too vague to classify into one of the other two categories. Although most of the investigators included qualitative forms of data in the key measures section (e.g., observations, field notes, audio and video), I focus only on tests in this study.

Classifying the intervention studies. Step 4 of Figure 2-1 identified the studies that meet two criteria. The first criterion was use of a standardized test as described in Step 3. The second criterion was whether the study was an intervention study or not an intervention study. An intervention study developed and evaluated a new educational intervention. Development and evaluation did not have to be the primary purpose of the study; the proposal only needed to state an intent to evaluate a new educational intervention. Studies that met both criteria moved on to the next step in the analysis.

Screening studies for potential validity problems. In Step 5, the first author and two raters carried out a procedure to identify studies with the greatest potential for validity problems related to the use of the standardized test for summative evaluation of the intervention. The raters were advanced doctoral students in a graduate school of

education with research training in educational measurement. The initial inter-rater reliability was inadequate ($\kappa = 0.14$) due to the difficulty of articulating operational definitions of test validity. However, after clarifying the definition, the raters reached complete agreement on the coding.

For each project that developed an intervention and evaluated it using a standardized test, we asked the following question: did the standardized test appear to measure the same thing that the intervention taught? We addressed the question using information in the IES database and the information about each test available on the Internet. Each entry in the IES database contained a section for investigators to articulate the learning goals of their intervention. From the Internet, we located documents from test publishers that contained information about test validity and test blueprints from state education agencies. Each state agency provided information relevant to the current study. Some of the 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. In contrast, some of the agencies provided sparse information that was limited to the standards assessed by the test and a small sample of released items.

For the analysis, we cross-referenced the goals of each intervention with the information we found on the Internet. 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 only to flag the studies with the greatest potential for problems due to a mismatch. We compared the goals of the intervention with purpose of the test and decided whether the test appeared to fit the intervention, or whether we needed to see additional evidence before we would be confident stating that the tests were probably aligned.

In-depth analysis of validity evidence. In Step 6, two raters and the author reviewed the corpus of information that supported the validity of the standardized test for evaluating the intervention. We determined that it was necessary to exclude studies for which the investigator did not provide a report. Without this criterion, the amount of information would vary too widely between studies.

In Step 7 and Step 8 of Figure 2-1, we determined whether investigators discussed the validity of the standardized 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 standardized test in the context of the evaluation? Second, 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, the 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.

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 (AERA, APA, & NCME, 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. For alignment, we 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.

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.

Gathering evidence of validation activities. In Step 9, we coded whether the investigators conducted research for the purpose of validating the standardized test. 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.

2.3.3 Analysis plan

The study emphasized a quantitative analysis of the binary codes represented in Step 4, 5, and 7-9 of Figure 2-1. We calculated arithmetic means for each category and the conditional arithmetic means for combinations of categories. In addition, we supplemented the quantitative analyses with qualitative descriptions of salient features of the projects. We highlight consistencies between projects in the same category (i.e., all the projects that provided alignment evidence) that bear on the research questions and present implications for research or policy.

2.4 Results

2.4.1 Use of outcome measures

The initial step in the research was to exclude projects that did not list a key measure. I started with all 85 projects in the IES math and science education database. 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.

Projects that developed and evaluated educational interventions

The first research question asked what proportion of the projects in the IES math and science education program aimed to develop and evaluate new a classroom intervention? I examined the 78 projects in the IES database that met the inclusion criterion and determined that 64 (82.1%) evaluated an educational intervention. It is clear that,

consistent with IES' overarching mission, the development of new educational interventions is a priority for the math and science education program.

Projects that used standardized tests as outcome measures

Research Question 2 asked how many of the projects used standardized tests to evaluate the impact of an educational intervention? I coded the tests listed in the Key Measures section of the research proposals into three categories: standardized tests, researcher-developed tests, and other tests. Most studies in the IES database listed multiple outcome measures. In the sample of 78 projects, investigators listed tests from one category in 17 of the projects (21.8%), from two categories in 41 of the projects (52.6%), and from all three categories in 17 of the projects (21.8%). Only 14 projects (17.9%) listed a single outcome measure. Some investigators listed multiple tests within a category but the descriptions were not sufficiently precise so to allow fine-grained analysis (e.g., " This study will administer standardized tests of cognitive ability.").

Table 2-1 contains descriptive statistics showing the types of tests that investigators listed as key outcome measures. I limited the analysis to only the 64 projects that used tests to evaluate the impact of an intervention. The rows of Table 2-1 are non-exclusive and the analysis 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 projects (71.9%) listed a standardized test as an outcome measure—the most out of any category. Although investigators listed standardized tests most often, the use of tests was balanced across categories; investigators typically used or planned to use standardized tests as part of a portfolio of research evidence that included both quantitative and qualitative information. Examination of the type of standardized tests showed that 30 of the projects (46.9%) used state tests (i.e., federal accountability). Five projects (7.9%) used subtests from the Woodcock Johnson-3 Test of Achievement, and three projects (4.7%) used the American Chemical Society high school chemistry exam. Two projects (3.1%) used tests from the ITBS system.

The results suggest that the use of standardized tests for evaluating new interventions in math and science education is widespread. Although these data are from projects funded by the IES program, it is likely that other research in mathematics and science education—and potentially other disciplines—follows suit in the common use of standardized tests for evaluating new interventions.

2.4.2 Validity of outcome measures

Mismatch between the test and the intervention

Research Question 3 asked whether the standardized test appeared to measure the same thing 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 test as an outcome measure, and the raters agreed that 25 of the projects (54.3%) had a potentially problematic mismatch between the intervention and the standardized test used as an outcome measure. For these projects, validity evidence will be especially important if the investigators position the scores as central to documenting the educational impact

of the program. For five of the projects (10.9%), one standardized 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.

In the qualitative analysis, we found two main issues across the 25 projects we flagged. The first issue was that the academic content covered by the intervention was narrow relative to the content measured by the standardized test. This is because many of the 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 constituted a narrow subset of the content on the standardized test. In such cases, the test measured more, and in many cases, much more, 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 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 tests to measure these learning goals. We conclude that, at best, standardized tests would be an approximate measure of the complex forms of the learning that the investigators really wanted to know about (Kennedy, 1999). The raters agreed that, in some cases, it would be appropriate to consider the standardized test as a distil indicator (i.e., Ruiz-Primo et al., 2012) of student learning. We hypothesized that our subsequent examination of the documents and publications associated with the projects would reveal similar interpretations within discussions about the validity of the standardized test for evaluating the intervention.

Taken together, the results suggest the possibility of a significant mismatch between the tests and the interventions. A statistical perspective on the results suggests that a meaningful portion of the test variance may be unrelated to evaluating the direct impact of the intervention. As the sensitivity of a test for detecting the impact of educational is a basic problem in measurement (Polikoff, 2010), it seems critical that investigators should explain how this content representation issue bears on the interpretation of their results and ultimately the conclusions of their evaluation.

The documents received from principal investigators

I 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 individuals (42.3%). Although the documents were diverse as described in the Method section, the data permitted us to investigate whether investigators thought the validity of the standardized test was important enough to discuss.

Out of the 33 sets of documents, only 11 matched to the 25 projects that the team of raters flagged as potentially problematic. The other 19 sets of documents related to studies that we eliminated at some point from the analysis or did not flag. Although we reviewed all the documents, we limited the current analysis to the 11 studies. We reasoned that discussions about test validity were more likely to be within the documents than other sources of information. Analysis without a report would be lacking the best source of evidence and the investigations would be more likely to turn up negative for validity discussions.

Discussions about validity and alignment

Research Question 4 asked how many of the studies contained validity discussions that supported the use of the standardized tests? To this point we have flagged 25 projects with a potentially problematic mismatch between the goal of the intervention and the standardized test used as an outcome measure. The next step was to examine the validity evidence contained within the documents provided by principal investigators and the articles associated with the projects. The primary purpose was to comment on the presence or absence of validity evidence, including evidence of alignment. A qualitative analysis described on the nature of the evidence, including whether the evidence was decontextualized or contextualized in argument-based approached to validity. To this end, I first provide a general description of the pool of reports and articles that comprise the data. Then I present the results of the search for validity discussions.

Type of information received. Table 2-2 contains the results of the central analysis. The first column contains a number for each of the 11 projects, and the next three columns characterize the evidence that supported the analysis of each study. Colum 2 indicates the type of document that the principal investigator provided. Across all studies, investigators provided nine final reports and two interim reports. The two interim reports compare with the final reports because both are from the final year of the grant.

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 (SE = 2.73), whereas studies for which the team of raters 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 = 0.149).

Column 4 of Table 2-2 documents whether the principal investigators of all 11 projects completed the evaluation of the intervention using data from the standardized test described in the IES database. The raters found that all 11 investigators completed the evaluation. This is important because investigators typically discussed the standardized test in the context of an evaluation. In addition, a closer examination of the reports showed that for two of the projects, investigators modified the assessment by changing items but otherwise conducted the planned evaluation. I discuss the implications of the modifications below.

Validity analysis. Three raters searched the documents and articles and coded whether they found a discussion about the validity of the standardized test for evaluating the intervention. Column 5 of Table 2-2 contains the consensus among the raters. Six of the 11 projects (54.5%) presented validity discussions to support the use of the standardized test in the evaluation (Projects 2, 4, 6, 7, 8, and 11). I describe the nature of these discussions below. The team of raters did not find validity discussions in the documents and articles for five of the projects (45.5%). The result suggests that it is common to use standardized tests for evaluative purposes without offering evidence that the test is valid.

Alignment analysis. The raters searched for evidence of alignment as a specific form of validity. Column 6 of Table 2-2 shows the results of the analysis. Four of the six investigators who discussed validity also discussed alignment (Project 6, 7, 8, 11). Interpreting the results carefully because of the few studies in the analysis, the evidence suggests that most investigators who discuss validity also reference the concept of alignment. It is interesting to consider the utility of alignment and whether strengthening the role of alignment within argument-based approaches to validity may lead to improvements in applied measurement.

Validation activities

The raters also searched the documents for evidence that investigators/authors conducted validation research for the purpose of generating evidence to support the validity of the standardized test. Five out of the 11 projects (45.5%) 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 standardized 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 unidentified methods, 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 test with a researcher-developed measure, and analyzing the floor and ceiling effects of the test.

Qualitative analysis of the validity discussions

The team of raters distilled common themes from the projects and characterized the nature of the validity discussions in ways that provide a snapshot of general approaches to applied measurement. The raters found a range of approaches to validity discussions, some of which may be considered argument-based approaches, some of which may not. Most of the validity discussions were concise and less than one paragraph. For example, Project 11 discussed validity *vis-à-vis* the alignment between the standardized test and the intervention in a single sentence. The sentence indicated that measurement issues related to the lack of alignment between the intervention and the standardized test called into question the credibility of the evaluation. Although a one-sentence discussion of alignment is obviously underdeveloped, it provided critical information about the interpretation of the instrument. Other brief discussions provided the sort of decontextualized evidence that would not be considered argument-based approaches. For example, one project contained one sentence of psychometric information from the test publisher. The investigators did not provide validity evidence relevant to their particular investigation.

Two projects provided more elaborate discussions of validity. The investigators on project number four used successive articles to develop a validity argument and to refine their outcome measure in ways that enhanced its validity. At the outset of the project, the investigators considered the standardized test a key measure for evaluating the impact of the intervention. However, after the first round of data collection, they stated that the test

did not measure the same thing that was taught by the intervention. The investigators substituted items and constructed a *post hoc* validity argument to explain proper interpretation of the scores from the standardized test for evaluating the intervention. The validity argument branded the standardized test a transfer test and articulated a theory about how the standardized test measured learning transfer from the intervention to new educational problems, in new learning contexts. The investigators interpreted scores on the standardized test as evidence that their intervention improved students' skills beyond the intervention's focal areas and into different educational domains.

For the other project that supplied a detailed validity discussion, the investigators completed a more formal alignment analysis. Project 8 described a framework for validating the alignment of a curricular intervention with a standardized test. The investigators conducted an item analysis: They mapped test items 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 the data and an example of careful measurement using item-level alignment to ensure that the items measured the same thing that the intervention taught.

2.4.3 Measurement issues

Many of the 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 tests interfered with their planned evaluation in some way (Project 4, 6, 7, 11). For example, the principal investigator on project 7 declared that the standardized test was invalid because it did not have enough test items that tapped the content taught by the intervention. This investigator reported that s/he learned a lesson to be more specific about the learning outcomes s/he wants to measure and to select an assessment that will be more sensitive to measuring those outcomes.

Projects that did not contain validity discussions also showed evidence of measurement issues. One investigator altered the standardized test after the first round of data collection by selecting a subset of items from the test. We did not find a rationale for the changes but a measurement issue seems to be the most likely reason to alter a test mid-study. In another example, an intervention aimed to increase students' conceptual understanding but the impact evaluation 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. Investigators must address these issues through an explicit focus on measurement.

I asked what was the real world impact of measurement issues? The investigators from four projects (6, 7, 8, and 11) administered standardized tests and noticed misalignment (e.g., that some items on the test measured what their intervention taught and some did not). The reports show that each investigator had different experiences. One

investigator simply ignored the data from the standardized test and used a different source of data collected during the study for a summative evaluation.

A second investigator determined that the standardized test was not well aligned with the learning goals of the intervention. This investigator attempted to acquire item-level data from the test administrator in order to conduct an evaluation using only relevant items. Unfortunately, the investigator was unsuccessful. The investigator had not collected other outcome data and was thus unable evaluate the impact of the intervention. A third investigator attempted to address the misalignment by dedicating considerable resources towards studying the validity of the standardized test. A fourth investigator conducted the planned evaluation but suggested that the results should be interpreted with caution because of the alignment problems. The array of measurement issues all point to the importance of emphasizing measurement during the planning phases of research to avoid unproductive outcomes.

Can Researchers Predict Measurement Problems?

This study permitted us to investigate whether, and how accurately, raters predicted measurement issues in the projects that comprised the data. In the first part of this study, a team of raters flagged projects with potential measurement problems using key pieces of information from the IES database. Importantly, the raters were effectively blind to the results of the studies.

The investigators from four projects explicitly discussed measurement issues and investigators from two additional projects described a situation that constituted a measurement issue. Thus, six of the 11 studies flagged by the raters (54.5%) showed evidence of measurement problems. I interpret this success rate as evidence of the ability of raters to detect measurement problems. The results are an initial proof in concept that prospective, qualitative analyses of validity by researchers with measurement training may add value to research by identifying measurement problems that might be avoided through modifications of the research method. Alternatively, we might interpret the results as evidence that programs that train educational researchers should place additional emphasis on educational measurement.

2.5 Summary and discussion

This study examined the use and validity of standardized tests for evaluation of new educational interventions in mathematics and science. Scholars are only beginning to investigate the issues surrounding the use of standardized tests for evaluating the impact of new educational interventions (May et al., 2009). We examined the projects funded through the Institute of Education Sciences (IES) mathematics and science education program and found that researchers commonly use standardized tests as outcome measures for evaluating educational interventions. Further, three raters agreed that many projects showed potential validity problems related to a basic misalignment between the goals of the intervention and the standardized test used as an outcome measure. Close inspection of 11 projects found that over half used standardized tests to evaluate the impact of an intervention without presenting or generating evidence that the test was valid. Indeed, researchers who neglected to discuss the validity of the standardized test

also reported measurement problems related to the misalignment of the standardized tests with the interventions. Taken together, the results point to a serious problem with applied measurement that should be addressed through changes to policy and practice.

Key findings and interpretations

This study aimed to investigate the use and the validity of standardized tests for summative evaluation of new math education and science education programs. I utilized a sample of applied research that demonstrated how investigators typically use and validate standardized tests in practice. The information from projects funded by the IES mathematics and science education program afforded a unique opportunity to review the use of standardized tests in leading applied research in math education and science education.

The first goal was to learn more about the scope of the problem: how many projects developed new interventions and how many projects used data from standardized tests to evaluate these interventions? I examined 78 projects in the IES database and found that 64 (82.1%) developed and evaluated a new educational intervention. Next, I asked how many of these intervention projects used, or planned to use, standardized tests for evaluating the interventions? I coded the tests used by researchers into one of three categories: standardized test, researcher-developed test, or other test. The results of the analysis showed that principal investigators used, or planned to use, standardized tests more often than tests belonging to the other two categories. Although most of the intervention projects (78.1%) listed multiple outcome measures belonging to more than one category, investigators used standardized tests more often than other forms of testing.

Taken together, the results demonstrate that it is common practice to evaluate new educational interventions in mathematics education and science education using data from standardized tests. The data comes from leading research; therefore, I interpret the results as evidence that the fields of mathematics education and science education possess a similar applied focus. Second, the use of standardized tests as outcome measures is also common. The results suggest that students' scores on standardized tests play a significant role in educational research and reform.

The validity of standardized tests in practice. Second, a team of raters investigated whether investigators documented the validity of standardized tests used as outcome measures. A broad consensus in the measurement community is that tests must be validated for each specific purpose. We asked whether investigators discussed the validity of the standardized test or conducted research activities aimed to generate information that supported the validity of the test.

Potential validity problems. The author and two raters screened the entries in the IES database in order to flag the projects that appeared to be especially in need of validity evidence. We reviewed each intervention project that listed a standardized test as a key measure and flagged 25 of the 46 intervention projects (54.3%) as having a potentially problematic mismatch between the intervention and the standardized test used as an outcome measure. In some instances, the standardized test was valid by design because the goal of the intervention was to increase students' scores on the standardized test. In

other instances, the test appeared to measure something different than what was taught by the intervention.

A subsequent qualitative analysis revealed that the raters flagged projects for two general reasons. The first reason was that the academic content covered by the intervention was narrow relative to the broad amount of academic content measured by the standardized test. The second reason was that the intervention had goals for student learning (e.g., scientific reasoning) that were difficult to measure with typical standardized tests. Both examples are different instances of construct underrepresentation: In the first example, the test items inadequately emphasize certain academic content and in the second example the cognitive skills of interest are not necessary for success on most of the items.

In-depth examination of validity. Next, we closely examined the validity evidence contained within the studies that we flagged. We asked whether each investigator discussed the validity of the standardized test for the summative evaluation of the intervention or conducted research activities for validating the test. We searched for particular evidence related to the alignment (Porter, 2002) between the test and the intervention because some scholars posit that alignment is a critical aspect of determining whether a test is suitable for evaluating a new educational intervention (May et al., 2009).

We conducted the in-depth validity analysis for 11 projects that we had an adequate corpus of information. The results showed that six of the projects (54.5%) discussed validity related to the use of the standardized test for the evaluation. Regarding the use of alignment, four of the six projects that discussed validity also discussed alignment. Additionally, five out of the 11 projects (45.5%) reported conducting validation activities for the standardized test. Roughly half of the investigators showed concern for the validity of the standardized test but it is also common for investigators to use data from standardized tests without providing any evidence that the test is valid.

Most of the investigators provided very brief discussions about validity, contained within one or two sentences. Two investigators provided more robust discussions of validity. One investigator described an analysis of the alignment between test items and the skills and knowledge taught by the intervention. Although we can find criticisms (as in any research), we believe that this approach is an exemplar of good measurement that should be emulated by others.

Another investigator, responding to initial measurement problems related to the use of the standardized test, added a measurement strand of research to their project. This investigator realized only after data collection that the standardized test did not measure what he expected it to measure—that is, the same knowledge and skills that were taught by the intervention. This investigator reconceptualized the standardized test as a test of learning transfer and studied the connections between their intervention and performance on the transfer test and other tests.

Investigator-reported measurement issues. Indeed, we found evidence that measurement issues related to the use standardized tests impacted other projects. Some investigators clearly articulated their mistakes: one investigator stated that s/he realized, only after data collection, that the standardized test did not measure the learning outcomes targeted by the intervention. This investigator learned a lesson to be more careful about measurement and more specific about the outcomes one wants to measure.

Other investigators omitted discussions about measurement where they were clearly indicated. One investigator evaluated a conceptual learning intervention with a test of computational fluency and, not surprisingly, observed no treatment effect. The investigator chose not to comment on the validity of the test that was used as the outcome, or on the apparent difference between the intervention's target of influence (concept development) and the test's target of inference (quick computation). We are left to speculate as to the reasons why.

2.5.1 Implications for practice

It appears all too common to use standardized tests to evaluate new educational interventions in math and science without evidence that the test is valid. I suggest that it is a serious problem when only half of investigators provide any validity evidence at all. This research noted multiple consequences of inattention to measurement issues, including the inability to conduct an evaluation for sudden lack of an outcome measure. This lack of validity evidence underlies an inattention to measurement that, we suggest, has the potential to obstruct educational innovation (i.e., Raudenbush, 2004).

Consider the situation where a new educational intervention teaches students valuable new skills, but the impact evaluation does not detect the learning caused by the intervention because the outcome measure was a standardized test that measured a different set of skills. Consequently, stakeholders jettison the intervention, to the detriment of educational reform. This is a situation we aim to prevent through this research that shows both potential and actual measurement problems in practice.

I believe that emphasizing good measurement practice can forestall some of these difficulties. I recognize the long-term efforts in the measurement community to promote best practice and testing reform, including a reconceptualization of standardized testing (i.e., Pellegrino et al., 2001). This study provided evidence that it is possible for measurement specialists to prospectively identify measurement problems in particular evaluation studies. It seems that consultation with the team of raters at an early stage of research design could have saved several investigators the time and effort that it took them to conduct an evaluation with a standardized test.

It is up to scholars and policymakers to sustain focus on this issue and to promote research methods that include good measurement practices. One solution is to develop additional policy around measurement. Proposals should have more robust measurement plans that contain validity information. IES could prioritize efforts to develop and validate appropriate measures. In addition, I repeat the call for wholesale reconceptualization of the place and purpose of standardized tests—and careful consideration of whether the information they provide really is the information that we need. At present, it is unclear whether new standardized tests (e.g., Partnership for Assessment of Readiness for College and Careers; Smarter Balanced Assessment Consortium) adequately address *any* of the concerns raised by this research. Future tests in science education may reduce their dependence on multiple-choice items, but other issues (i.e., reliability) emerge.

2.5.2 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. I 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 imply specific ways to improve measurement in practice.

A second limitation was the inability of the raters to judge the quality of the validity evidence. Fundamentally, we analyzed the presence or the absence of evidence. Although we wished to differentiate between disconnected bits of validity evidence and argument-based approaches to validity, we observed no clear way to differentiate between the two categories. 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). However, research that characterizes the use of validity in practice may bring clarity to the concept of validity as the practitioners show scholars what in the concept of validity is possible and useful.

The data in the form of the documents I received from investigators supplied another limitation. I requested reports from principal investigators and received a diverse set of documents in response. The documents supplied more information for some projects and less information for others. Standardization of information would be desirable to ensure a 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. Future studies that examine validity in practice should consider new methods for extracting more information about test validity. For example, through surveys, interviews, or by policies that require scholars to comment on test validity in research proposals and reports.

2.5.3 Conclusion

In this study, a team of raters examined projects funded by the Institute of Education Sciences (IES) math and science education program to address the use and validity of standardized tests as outcome measures for impact evaluation of new educational interventions. The results of the research suggest that the primary goal of the IES mathematics and science education program is the development of new interventions. Further, investigators use standardized tests as outcome measures more than other forms of tests (i.e., researcher-developed tests). The analysis suggests that many projects show the potential for validity problems related to a misalignment between the standardized test and the intervention that it is used to evaluate. Closer examination of the projects with potential validity problems showed that many investigators did not present validity evidence for the standardized test. At the same time, many investigators discussed measurement problems related to the use of the standardized test. Some investigators concluded, only after conducting the evaluation, that the standardized test did not measure the learning caused by their intervention. This study highlighted an important

weakness leading applied research in mathematics education and science education. The ability of educational research to support beneficial reform depends on addressing the measurement issues highlighted in this study.

2.6 Tables and figure

Table 2-1

Category and Frequency of Key Measures Listed in the IES Database

Category	n	%
Standardized	46	71.8
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, Researcher Developed, And Other	12	18.8

Table 2-2

Results of Validity and Alignment Analyses

Study Number	Document provided	Scholarly articles located	Used ST?	Validity Evidence?	Alignment Evidence?	Validation Activities?
1.	Final report	2	Y	N	N	N
2.	Interim Report	8	Y	Y	N	N
3.	Final report	4	Y*	N	N	N
4.	Final report	6	Y	Y	N	Y
5.	Final report	0	Y	N	N	N
6.	Interim (yearly) report	1	Y	Y	Y	Y
7.	Final report	3	Y	Y	Y	Y
8.	Final Report	16	Y**	Y	Y	Y
9.	Final report	0	Y	N	N	N
10.	Final report	1	Y	N	N	N
11.	Final report	8	Y	Y	Y	Y

Note. Y=Yes; N=No. * Selected items only. **Modified the test

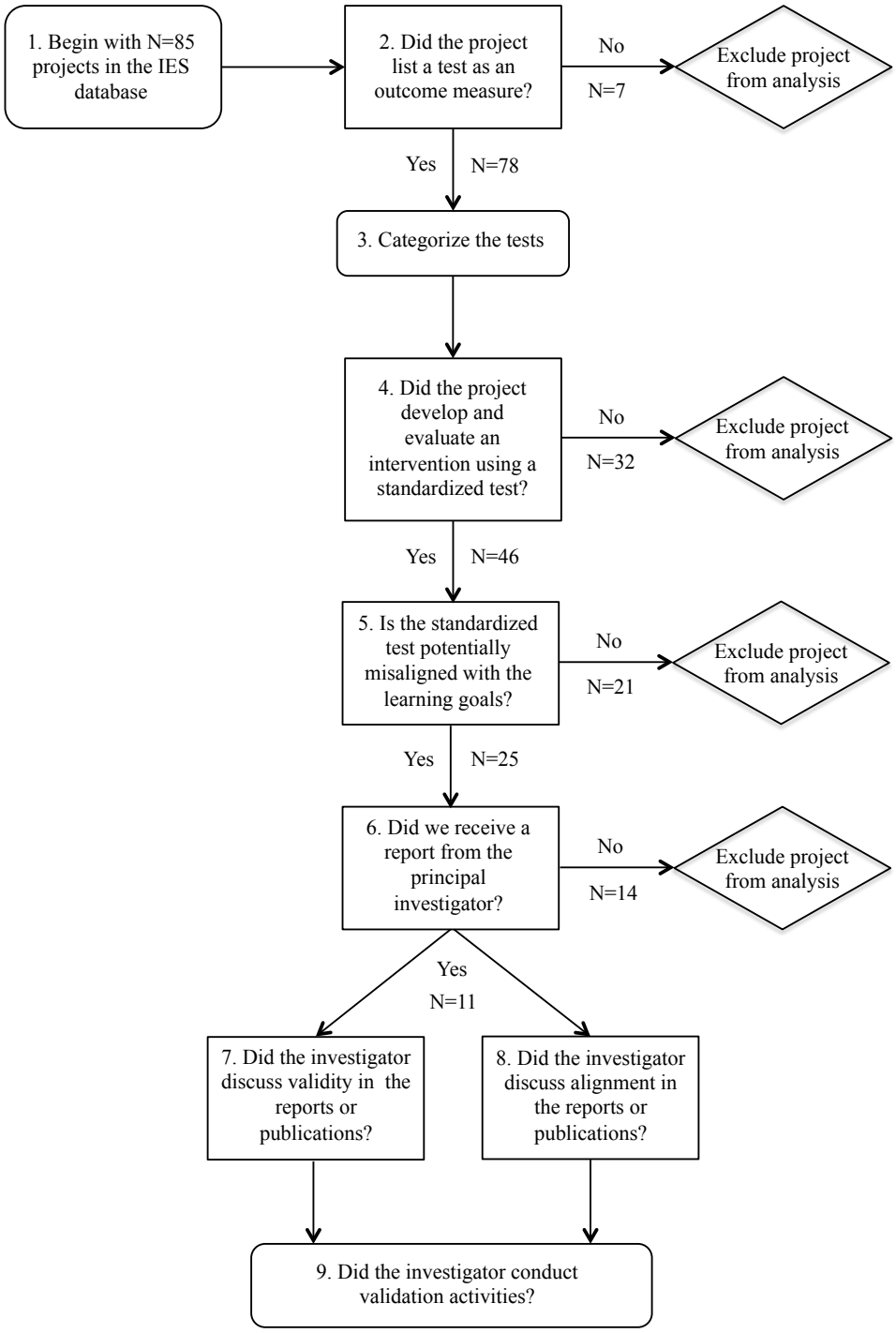

Figure 2-1. Logic model for research method.

Chapter 3

Standardized tests as outcome measures in applied research: A psychometric simulation of the relationship between alignment and treatment sensitivity

3.1 Abstract

In this research, I developed a method for modeling the impact of degree of alignment (Porter, 2002) between an outcome measure and an educational treatment on the results of an evaluation. The purpose was to investigate alignment problems related to the use of standardized tests for evaluating the educational impact of instructional interventions. I used data simulation to assess the effect of alignment (2% – 100%) on simulated experimental evaluations with different effect sizes (0.1 – 2.0). I studied three outcomes of alignment: statistical power, the estimated treatment effect, and the reliability of the test. The results showed that close alignment is required for adequate statistical power to detect smaller effect sizes (0.1 – 0.3 SDs). When effect sizes are larger (> 0.4), statistical power remains adequate under greater misalignment. However, decreases in alignment lead to an attenuation of the estimated treatment effect. Test reliability was unaffected by changes in alignment. An empirical test of the model provided evidence for the practical usefulness of the results. I discuss implications of this research for the use of standardized tests in applied educational research and for the concept of alignment.

3.2 Introduction

In this study, I explored measurement problems related to the use of standardized tests as outcome measures for evaluating the impact of instructional interventions. Researchers commonly use scores on standardized tests, such as those used for accountability purposes (i.e., Linn, 2000), as data for evaluating the effectiveness of new interventions such as curricular units or lesson sequences. Unfortunately, the standardized tests may

not measure, or may not measure well, the knowledge and skills that are the focus of new instructional interventions. Researchers commonly use standardized tests as outcome measures and sometimes learn only after data collection that the tests do not measure the same skills and knowledge that were taught by the intervention (Sussman, 2016).

This paper conceptualizes the problem as poor *alignment* (Porter, 2002) between the test and the goals of the intervention. To gain an empirical understanding of the problem, I developed a psychometric model of the relationship between alignment and the *treatment sensitivity* of an evaluation, defined as the ability of an evaluation to detect the effect of an educational intervention (e.g., Lipsley, 1990). The practical goal of this paper is to develop and explore a method, akin to power analysis, that helps researchers account for misalignment when they design evaluations.

Although authors have pointed out inconsistencies between standardized tests and the goals of educational instruction (e.g., Pelligrino, Chudowsky, & Glaser, 2001; Schoenfeld, 2006; Wiggins, 1993), the literature contains few well-developed investigations into the problems related to the use of standardized tests as outcome measures for impact evaluation of new educational interventions (May, Johnson, Haimson, Sattar & Gleason, 2009; Olsen, Unlu, Price, & Jaciw, 2011; Somers, Zhu, & Wong, 2011). Among these few studies, a consensus exists that weaknesses in the alignment between the outcome measure and the intervention threaten the validity of an evaluation. The researchers also agree that alignment impacts the sensitivity of an evaluation for detecting the intended effect of the intervention. However, the literature lacks empirical evidence to support the putative relationship between alignment and treatment sensitivity.

The purpose of this study is to demonstrate empirically how alignment influences the sensitivity of an evaluation for detecting the impact of an educational intervention. The research method simulates experimental evaluations with standardized test-like outcome measures that are more or less aligned with a theoretical intervention. I assume a functional relationship between alignment and the estimated treatment effect, where decreases in alignment reduce the observed impact of the intervention. The research method includes psychometric models for data generation and for data analysis, from the Rasch family of item response models (Rasch, 1960/1980; Adams, Wilson, & Wang, 1997). Monte Carlo simulation method allows simulation of treatment sensitivity for multiple conditions of alignment and for different effect sizes. The results provide estimates of three evaluation outcomes (mean group differences, statistical power, and implied reliability of the outcome measure) that lead to different representations of the impact of alignment. Each representation helps to demonstrate the importance of alignment for rigorous evaluation.

The remainder of this introduction contains four sections. In the first section, I discuss the problems related to the use of standardized tests as outcome measures in experimental evaluations of new educational interventions. The second section contains a theoretical perspective on the relationship between alignment and the sensitivity of an evaluation for detecting the impact of an intervention. The third section contains an operationalization of the theory in measurement and statistical models. The final section contains a review of the research questions and hypotheses.

3.2.1 Using standardized tests to evaluate instructional interventions

May et al. (2009) explored issues related to the suitability of standardized tests as outcome measures for evaluation of new educational programs. The authors reviewed 58 studies funded by the National Center of Educational Evaluation that either used or planned to use a state-administered standardized test to evaluate the impact of an educational program. The review concluded that (a) investigators commonly use standardized tests as outcome measures for experimental evaluations of new educational interventions, and (b) questions exist about the suitability of the standardized tests for research purposes. May and colleagues discussed specific problems related to the alignment of the test with the goals of the intervention. They defined alignment as the degree to which a particular standardized test measures the same knowledge and skills that are taught by an intervention and suggested that weaknesses in alignment could lead to an invalid evaluation.

Sussman (2016) explored the use and alignment of standardized tests for applied educational research in greater detail using a sample of investigator-initiated studies funded through IES research grants. The author examined 78 projects funded through the IES math and science education program. The research data consisted of information from the IES online database, published articles related to the projects, and reports to IES. First, Sussman found that researchers commonly use standardized tests as outcome measures in investigator-initiated studies. Fifty-nine percent of the projects used, or planned to use, a standardized test as an outcome measure for evaluating a new educational intervention.

Second, many of the studies demonstrated potential and actual problems related to a lack of alignment between the knowledge and skills required by the test and the knowledge and skills taught by the intervention. Sussman (2016) cross-referenced test blueprints with the goals of the intervention and found cases where the standardized tests measured much more academic content than the intervention taught. This discrepancy occurred when investigators used standardized tests to evaluate brief (e.g., three week long) curriculum units. The standardized tests measured grade-level achievement across a year's worth of academic content, whereas the brief interventions focused on teaching and learning within a small subset of the content areas.

Another type of potential misalignment occurred when the intervention focused on developing certain academic skills that were not well measured by the standardized test. As a straightforward example, consider a hypothetical evaluation that measures the impact of a reading fluency intervention using a broad test of English language arts. It is unreasonable to expect the test to provide valid (i.e., accurate and useful) information about the impact of the intervention because reading fluency is not a primary determinant of success on most of the items on the test. The cognitive work that students do as they participate in an intervention must match the cognitive work required for success on the test. Sussman (2016) identified analogous issues for math and science interventions that, for example, developed students' ability to reason using mathematical or scientific information. Evaluations that use standardized tests to evaluate conceptual understanding appear to be measuring the wrong outcome relative to the goals of the intervention (Quellmalz et al., 2013; Taylor, Kowalski, Wilson, Getty, & Carlson. 2013).

Third, Sussman (2016) uncovered additional problems related to the use of standardized tests in applied research. The author identified a handful of studies where investigators examined outcome data from a standardized test and determined that it did not measure the same outcomes that were targeted by the intervention. At least one investigator admitted that (s)he could not evaluate the impact of the intervention.

The problem is a manifestation of a more general issue with the lack of attention to measurement issues in applied research. Wide agreement exists in the literature that comprehensive investigation of validity evidence is essential for the proper interpretation and use of test data (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AREA, APA, & NCME], 2014). However, in many studies that Sussman (2016) flagged for measurement issues, researchers did not attend to the validation of the test or study the inferences that could be drawn from the validation of the test scores. Although most of the studies gathered data using multiple measures, many projects positioned the outcome data from standardized tests as key evidence of educational effectiveness. Taken together, these findings suggest that the educational research community must be very cautious when researchers propose to evaluate new educational interventions using standardized test scores as the main outcome variable.

3.2.2 Exploring alignment and treatment sensitivity

In this section, I define and connect the concepts of alignment and treatment sensitivity. First I define alignment and treatment sensitivity. Second, I describe a model that connects alignment with treatment sensitivity via a theoretical mechanism of test validity. Third, I describe the results of initial efforts to empirically examine the model. Finally, I highlight the limitations of those studies and how the current research addresses those limitations

Defining and measuring alignment

The educational measurement literature typically discusses alignment as a method for judging the match, or overlap, between the contents of a standardized test and a mandated set of standards for the purpose of educational accountability (Bhola, Impara, & Buckdahl, 2003). It is important to note that the concepts and methods of alignment fully apply to the match between a test and an educational intervention (Porter, 2002). Schools and test developers typically use alignment methods to generate evidence that scores on standardized tests provide accurate and useful information for calculating the percent of students who achieve a proficiency in an academic subject. Case law has also made alignment a legal requirement for high stakes examinations in secondary education to ensure that students are tested on what they are taught, and not tested on material that they have not been taught (Hubert & Hauser, 1999).

Each alignment method provides a set of criteria that is typically used to judge the match between a test and a set of educational goals along two main dimensions: content knowledge and cognitive skills (Roach, Niebling, & Kurz, 2009). Content knowledge categorizes academic subdomains such as integers in third grade math or thermodynamics in high school chemistry. Cognitive skills include the mental activities

that one performs on content knowledge; for example, recalling facts or synthesizing information into a cogent argument. Different alignment methods have different procedures for judging alignment, with non-trivial differences between methods (Newton & Kasten, 2013; Webb, 2007).

On a theoretical level, alignment is an element of test validity (Bejar, 2010). Alignment methods are one way to generate evidence for the *construct representation* of the test. Good construct representation means that the items on the test adequately sample from the intended construct domain (Embretson, 1983). Thus, an aligned test contains items that comprehensively and representatively measure the knowledge and skills contained within the standards document or, in the case of an educational intervention, the description of educational goals for an intervention. A comprehensive discussion of the link between alignment and other validity concepts is important but beyond the scope of this empirical paper. Interested readers should see Webb (2007) and Beck (2007) for balanced perspectives.

Defining and measuring treatment sensitivity

The sensitivity of an evaluation for detecting the impact of an educational intervention is what Lipsley (1990) termed *treatment sensitivity*, defined as the likelihood that the evaluation will detect a genuine treatment effect. Many factors influence the treatment sensitivity of an evaluation, such as sample size, the magnitude of the treatment effect, the statistical model used to analyze the data and various features of the outcome measure.

Statistical power analysis provides a method for measuring treatment sensitivity in the context of an evaluation. Statistical power is the probability that a statistical test will yield statistically significant results (Cohen, 1977). Power analysis calculates statistical power in a research context defined by several variables including the reliability of measurement, the size of the research sample, and the effect size. The literature generally defines effect size as the difference between two groups relative to measurement error. Another interpretation is that the effect size is the standardized magnitude of the departure from the null hypothesis that there is no treatment effect.

Hypothesizing a relationship between alignment and treatment sensitivity

The literature that discusses the use of standardized tests in applied research contains the framework for a model that links alignment to the credibility of an evaluation. The model articulated by May et al. (2009) hypothesizes that a decrease in alignment decreases the validity of the test scores, thereby decreasing the sensitivity of the evaluation for detecting treatment effects. According to the authors, a "state assessment is most valid and reliable when it "aligns closely with the outcomes targeted by the intervention," and "the alignment of the state test can be established by determining the proportion of test items that measure skills and knowledge targeted by the intervention" (May et al., 2009, p. 7).

Thus, the model implies that low alignment leads to attenuation of the estimated impact of the intervention: all other things being equal, less alignment leads to a smaller observed effect size. The implication is that the attenuated treatment effect would be

more susceptible to Type II error (false negatives). Finally, the authors mentioned but did not develop a theoretical relationship between decreases in alignment and decreases in the reliability of the test scores. Later, I explore this connection in greater detail.

Empirical studies of alignment and treatment sensitivity

The effect of alignment on the estimated treatment effect has been investigated empirically in two studies. Somers et al. (2011) examined four IES-funded randomized experiments carried out across multiple states to determine whether interstate differences in the alignment of state achievement tests with an intervention led to variation in estimated program impacts. Additionally, Olsen et al. (2011) examined data from three small-scale evaluations of reading interventions using state achievement tests. Olsen and colleagues studied the impact of alignment on the evaluation outcomes, focusing on the (a) alignment between pre-test and post-test and (b) alignment between post-test and the intervention. Both studies uncovered evidence of misalignment but neither study detected statistically significant effects of low alignment on the impact estimates of the intervention. Consequently, these two studies downplayed the importance of alignment on evaluation outcomes.

In spite of the findings, methodological issues limited the explanatory power of these studies. First, Somers and colleagues (2011) were unable to conduct their planned analysis of alignment for three of the four studies in their sample. They based the conclusions about the practical impact of alignment on the results of a single study. Second, the alignment methods used in both studies were informal rather than systematic and hence not consistent with established methods. Somers et al examined test blueprints to investigate the percentage of test content that measured the targeted outcome. Olsen et al.'s (2011) investigation probed for evidence of misalignment with differences in the correlation between a set of tests and a benchmark outcome measure. Although the former method is often the best available option (e.g., due to the proprietary nature of test items), neither method constitutes a rigorous way to study alignment (Webb, 2007). Both should be considered *ad hoc* alignment methods and hence the findings of these studies should be interpreted with caution. The current study uses a more rigorous way to investigate alignment.

3.2.3 Modeling alignment and treatment sensitivity

In this section, I articulate an empirical model of the relationship between alignment and the sensitivity of an evaluation for detecting the impact of an educational intervention. I flesh out May et al.'s (2009) theory and conceptualize the relationship between alignment and treatment sensitivity as a specific function, eliminating the need to calculate point estimates of alignment. I begin with foundational assumptions and describe the psychometric models for data generation and analysis.

Alignment is a binary concept

A key assumption of the model is that alignment is a binary concept. An item is either aligned, by virtue of measuring skills and knowledge taught by the intervention, or it is

not aligned. A binary conceptualization limits the model to a single effect of alignment. Although this assumption may not reflect reality, as alignment is likely to be on a continuum and conditional on other variables, a binary assumption is congruent with the most common alignment methods (i.e., Roach et al., 2009).

Efficacious treatments increase the probability of success for aligned items

This research requires us to articulate a mechanism of action for the impact of alignment on test performance. I conjecture that participants who participate in an intervention will gain knowledge and skill from the intervention. What the participants learn increases the probability that they will succeed on the test items that measure the same knowledge and skill that they acquired from the intervention (i.e., aligned items). The intervention would not systematically change the probability of success on items that are not aligned with the intervention. Moreover, I assume that the impact is consistent across items and that the changes in probability are consistent with a particular model, the Rasch model, as described below.

Translating the theory into a simulation method

The research method simulates experimental (randomized) evaluations of a hypothetical educational intervention. The simulation framework permits a study of the impact of multiple conditions of alignment on dependent variables that relate to the sensitivity of an evaluation for detecting treatment effects (statistical power, treatment effect, and implied reliability). As a second independent variable, I vary the effect size of the intervention. The simulation follows the notion of Monte Carlo studies as statistical sampling experiments (Harwell, Stone, Ho, & Kirisci, 1996; Spence, 1983), exploiting the ability of simulation experiments to provide information similar to real-world experiments.

The method contains some key features that overcome traditional limitations of alignment research. A basic criticism of alignment is that that final conclusions are based, at least in part, on fallible human judgments. I exploited the power of simulation to estimate the impact of alignment for a range of alignment conditions (from one item to all items). Researchers may be able to use the function to make more informed decisions about the choice of outcome measures similarly to how they do for sample size in conventional power analysis.

Modeling changes in the probability of success related to participating in an intervention

I used item response measurement models for data generation and data analysis to estimate the link between alignment and treatment sensitivity in the context of the simulation. A key assumption is that that an educational intervention increases the latent ability of a student, for items aligned with the intervention, by the magnitude of the effect size. I used a Cohen's *d* concept of effect size consistent with others in the literature who have developed simulated evaluations analyzed through item response models (e.g., Guilleux, Blanchin, Hardouin, & Sébille, 2014; Holman, Glas, de Haan, 2003). Additionally, I assume that a "true" difference exists between the treatment group and control group on the outcome variable.

I generated the item response data in two steps using a Rasch model and random number method. The first data generation step is for misaligned items and the second is for aligned items. The misaligned step uses an initial, simulated latent ability to generate (a) all item responses for the control group and (b) item responses to misaligned items for the experimental group. Next, the aligned step simulates the experimental group's item responses to aligned items (accounting for the increase in latent ability from the intervention). After I simulated the test scores, I used the latent regression Rasch model (Adams et al., 1997; Zwindermann, 1991) to estimate the difference in mean latent ability between the experimental group and the control group.

Outcomes of alignment

I calculated the impact of alignment using three outcomes that relate to the treatment sensitivity of an evaluation: statistical power, estimated treatment effect, and implied reliability of the test. Statistical power is the percentage of simulated replications that show a statistically significant treatment effect. Recall that (a) each simulation has a deterministically set difference in mean latent ability between the treatment and the control group and (b) the independent variables include alignment and effect size. I estimated the treatment effect using the latent regression Rasch model and the significance of the parameter using a Wald test.

I also calculated the impact of alignment on the person separation reliability (Wright & Stone, 1979) and the overall reliability (Adams, 2005) for the simulated test. The person separation reliability coefficient is analogous to the KR20 measure of internal consistency in classical psychometrics and estimates the ability of an instrument to differentiate between individuals (Wright & Stone, 1999). The overall reliability estimates the reliability of the population level parameters (e.g., the latent distribution). Statistical power and reliability are related concepts (Kanyongo, Brook, Kyei-Blankson, and Gocmen, 2007). A first step towards investigating the impact of alignment on test reliability is to calculate reliability coefficients across the simulated alignment conditions.

3.2.4 Testing the model with an empirical example

I tested the accuracy of the alignment function using data from a real-world evaluation. I asked whether the alignment function could accurately predict the effect size from a misaligned assessment given two pieces of information: (a) the "true" effect size of the intervention obtained through an aligned assessment and (b) an estimate of the amount of misalignment established by panel of judges. A match between the effect size predicted by the results of the study and the observed effect size produced by a real-world evaluation with a misaligned assessment would support the accuracy of the model and the utility of the research for predicting the results of evaluations using misaligned assessments.

The empirical example is a quasi-experimental evaluation of the Learning Mathematics through Representations (LMR) curricular intervention (Saxe, Diakow, & Gearhart, 2012). LMR is a research-based unit for the teaching of integers and fractions for elementary students. Researchers evaluated the educational impact of LMR using two assessments: a research-based assessment of integers and fractions (Saxe et al., 2012) and

standardized tests in grade level mathematics administered in the prior year and the end of year (Sussman, Diakow, & Saxe 2012). The evaluation based on the research-based assessment estimated the effect size of the intervention was 0.66 ($p < 0.001$). In contrast, the evaluation based on the standardized tests estimated that the effect size of the intervention was 0.20 ($p = 0.09$).

Sussman et al. (2012) studied the alignment of the standardized tests with the LMR intervention. Two university faculty and two graduate students, including this author, conducted an *ad hoc* alignment method similar to Somers et al (2011) to estimate the percentage of items that were aligned with the intervention. The researchers reviewed test blueprints, technical manuals, and released test items to arrive at a consensus judgment that, on average, 33% of the items on the standardized test were aligned with the LMR intervention. Therefore, in this study, I asked what is the expected effect size under 33% alignment if the true effect size is assumed to be 0.66? I compared the result predicted by this study to the empirically observed treatment effect of 0.20 *SD*s.

3.3 Measurement models and statistical models

3.3.1 A modified Rasch procedure for data generation

In this section, I present the Rasch model and describe a modification that, conceptually speaking, permits unidimensional Rasch modeling with different levels of the latent trait for each item, depending on whether the item is aligned with the intervention or not. The modified Rasch model also permits the simulation of item response data that reflect the impact of an educational intervention conditional on alignment of the items with the intervention. I describe how the modifications follow a relatively straightforward mechanism of action for an educational intervention. Although this is an unconventional conceptualization of the Rasch model, the model accurately reflects measurement as it occurs in practice.

The Rasch model

The Rasch (1960/1980) model is a psychometric model that defines the link between a latent variable and item parameters. In the Rasch model for dichotomous data, the probability of answering an item correctly is given by Equation 3.1. The Rasch model estimates the probability of success on an item x_i as a function of a person parameter θ_p, which represents the magnitude of a person's latent variable (also called ability hereafter), and an item parameter δ_i, representing the difficulty of item *i.*

$$\Pr\left(x_{ip} = 1 \,\middle|\, \theta_p\right) = \frac{\exp\left(\theta_p - \delta_i\right)}{1 + \exp\left(\theta_p - \delta_i\right)}, \text{ where } \Theta \sim N(\mu, \sigma^2). \tag{3.1}$$

The probability of a correct response increases monotonically as the latent variable increases and decreases monotonically as item difficulty increases. Each θ_p is an independent realization of the random variable Θ, which is typically assumed to be normally distributed. The Rasch model relies on additional assumptions of

unidimensionality and conditional independence; item responses depend on a single, continuous latent variable that accounts for all the covariance between the items. In addition, a constraint on θ_p or δ_i must be applied for identifiability of the model. In this study, the mean of θ_p was set to zero.

A modified Rasch model

I modified the Rasch model in a way that changes change the probability of person *p*'s success for certain items. Some readers will note that the model is analogous to a model for differential item functioning (DIF; Angoff, 1993). I present the model in the following format for interpretive purposes.

I formulated the modified Rasch model as follows: for students in the control group, Equation 3.1 holds. For students in the treatment group, the item difficulties are modified by a constant value that reflects the impact of the treatment on the probability that a person *p* will succeed on item *i*, if the item is considered aligned with the treatment (Equation 3.2).

$$\Pr(x_{ip} = 1 \mid \theta_{ip}) = \frac{\exp(\theta_p - \delta_i + \Delta_{pi}\tau)}{1 + \exp(\theta_p - \delta_i + \Delta_{pi}\tau)}, \quad \text{where } \Delta_{pi} = \begin{cases} 1 \text{ if item } i \text{ is aligned} \\ 0 \text{ othterwise.} \end{cases} \tag{3.2}$$

The only change between (1) and (2) is that the item difficulties are modified by, τ, depending on whether the item is aligned or not. The magnitude of the change, τ, is a positive number that equals the theoretical effect size of the intervention. In other words, the intervention increases the probability of success on an item by τ, for items that are considered aligned with the treatment (i.e., items that measure the same knowledge and skills taught by the intervention).

The arithmetic allows us to equivalently conceptualize τ as a change in latent ability due to an educational intervention. Suppose person *p* had a latent ability, θ_p, of 0.5 logits before participating in an educational intervention. If the effect size of the intervention, τ, was 0.7 logits, the person's effective latent ability, θ_p^*, would be 1.2 logits for an aligned item and would remain at 0.5 logits for a non-aligned item. This conceptualization is more in-line with the idea that an intervention increases a person's proficiency, rather than changing an item's difficulty.

Table 3-1 shows the probability of success for this person predicted by the Rasch model. The items range in difficulty between -3 logits and 3 logits. The second column shows the probability for non-aligned items and the third column shows the probability for aligned items. The difference between the two columns shows the impact of the intervention predicted by the model. The interaction between alignment and item difficulty follows a logistic function.

3.3.2 Latent regression Rasch model for data analysis

After simulating the item response data, I used the latent regression Rasch model (Adams et al., 1997; Zwinderman, 1991) to estimate the difference between the mean ability in the experimental group and the mean ability in the control group. The group means are different by design and the goal is to determine how alignment and effect size influence the difference in the means and how often the difference is statistically significant.

The latent regression Rasch model provides benefits over a more common analytic approach that regresses the latent estimates for each person on group membership. The benefits include the elimination of bias caused by assuming the same latent distribution for both groups (Mislevy, 1987) and hypothesis testing that accounts for measurement error instead of assuming that the measurements are error-free.

A difference between the Rasch model and the latent regression Rasch model is parameterization of the latent distribution. The latent regression model in this study modifies the population mean for the experimental group. Equation 3.3 shows that the population model becomes,

$$\theta_p^* = \begin{cases} \theta_p \text{ for the control group} \\ \theta_p + \lambda \text{ for the treatment group} \end{cases}, \tag{3.3}$$

where the coefficient, λ, is the estimated difference between mean ability in the treatment group and mean ability in the control group (Adams et al., 1997).

3.4 Summary of research questions and hypotheses

This study generates empirical data to address the following research questions:

1. What is the relationship between alignment and the statistical power of an evaluation for detecting the effect of an educational intervention?
2. What is the relationship between alignment and the estimated treatment effect?
3. What is the relationship between alignment and the reliability of a test?
4. Can the alignment function predict the effect size observed from a real world evaluation with a misaligned assessment?

Research hypotheses

The first hypothesis was that that decreases in alignment will result in decreases in power and the estimated treatment effect. This is not a provocative hypothesis because this is proposed in the literature (May et al., 2009; Olsen et al., 2012; Somers et al., 2012) and mathematically determined by the model. However, the expected results will be taken as evidence that the model was successfully implemented. Second, as another verification of model implementation, I hypothesized that the mean treatment effect for each of the 100% alignment conditions should equal the deterministically fixed effect size. Third, I hypothesized that interpreting the functional form of the relationship would lead to tentative guidelines for outcome measure alignment in applied research. The guidelines are tentative pending additional research.

Fourth, I hypothesized that reliability should remain relatively constant between alignment conditions and effect size conditions. Reliability may increase for the highest effect sizes, at high alignment, due to an increase in latent variance. However, the calculation of test reliability should not be dramatically impacted by the independent variables in the simulation.

Fifth, an empirical example will help to establish the veracity of the model. I will interpret a close match between the model predicted effect size and empirically observed effect size as evidence supporting the accuracy and usefulness of the model. A model prediction within 25% of the observed estimate (0.05 *SD*s) seems like a reasonable goal.

Finally, I hypothesized that the results will uphold the practical concerns about the use of standardized tests in applied research. This hypothesis entails two main conditions. First, the results will show that close alignment is necessary for valid evaluation. Second, I will be able to point to examples in the literature where standardized tests appear to lack the level of alignment required for adequate treatment sensitivity and valid evaluation. I separated the results into two studies: the first study presents the simulation and the second study presents the empirical test of the model.

3.5 Study 1: simulation study

3.5.1 Method

The goal of the simulation study is to estimate the impact of alignment on the sensitivity of an educational evaluation for detecting the effects of an educational treatment. To this end, I have developed a Monte Carlo method that simulates experimental evaluations with known differences between treatment and control groups. I analyzed how changes in alignment impacted three outcomes: the statistical power of the evaluation for detecting treatment effects, the difference between mean ability in the experimental group and mean ability in the control group (treatment effect), and the implied reliability of the test.

The outcomes depend on a central effect: the estimated difference between mean latent ability in a treatment group and mean latent ability in a control group. In the psychometric model, decreases in alignment deterministically decrease the expected difference between groups. This decrease occurs because the hypothetical impact of the intervention (effect size) increases the probability of success for aligned items. The experimental group and control group should possess equal mean latent ability for the set of items that were considered misaligned. Mathematically, the differences between groups will increase in proportion with the increase in alignment. A central purpose of this work is to interpret this function and its implications for applied research.

Independent variables in the Monte Carlo simulation

This study adjusts two categories of variables: Monte Carlo variables and experimental specifications. The Monte Carlo variables in this study include the number of replications and the sample size. The experimental specifications encompass the outcome measure, the persons in the sample, and the impact of the evaluation.

Although conventional power analyses usually vary the sample size, I needed to fix the sample size so as to vary alignment. I selected the parameters of the evaluation

experiment with a particular type of evaluation in mind. I wanted to generalize the results to a evaluation that an investigator might design after they have conducted pilot research and wished to scale up implementation order to justify the educational efficacy of their intervention prior to a large-scale study. I reasoned that 600 participants was a practical number for an evaluation that gathers provisional evidence of efficacy needed to garner additional attention and resources. A second reason for choosing 600 participants was as an experimental control for the empirical example as that sample contained 577 participants.

Figure 3-1 contains a logic diagram for the simulation. Step 1 in the simulation set the range of the intervention's effect size from 0.1 to 2.0. The units of effect sizes are interpretable as either standard deviations (of the treatment groups) or logits when the variance of the latent distribution equals 1.0. I tested the effect size range from 0.1 to 1.0 in steps of 0.1, and also tested effect sizes of 1.5 and 2.0.

The next phase of the simulation was data generation. Step 2 controlled the degree of alignment. Here, the simulation conducted 50 replications to vary the alignment from one item (2% alignment) to all 50 items (100% alignment), inclusive. In Step 3, I simulated a latent ability for each person by random draw from a standard normal distribution. Step 4 randomly assigned half of the sample to an experimental group and half to a control group.

The following phase simulated an outcome measure with 50 items (similar to a standardized test). Step 5 simulated item difficulties for 50 items by random draw from a standard normal distribution. Although this dissertation simulated multiple item sets, future research will simulate items only once.

Step 6 simulated the educational intervention. The simulated intervention increased the probability of success on aligned items for the persons in the experimental group. The probability of success increased by the effect size of the intervention. In contrast, the item response probabilities remained the same for non-aligned items, and persons in the control group received no changes to their latent ability. In Step 7, I simulated item response data using the mirt package (Chalmers, 2016) in R (R Development Core Team, 2014).

The final task of the simulation was data analysis. For Step 8, I used the mirt package to fit a latent regression Rasch model. I estimated the latent regression parameter using 10 plausible values. Next, I stored the results in a matrix (Step 9), including the latent parameter estimate and whether it was statistically significant from zero. This was repeated for each alignment condition from 2% to 100%

The complete simulation consisted of 1000 replications for each of 50 alignment conditions and 15 effect sizes. I selected 1000 replications for each condition for practical reasons. Pilot studies showed that the resulting alignment functions were smooth, and the computing time of approximately one month was considered acceptable.

Dependent variables

This study evaluated three main dependent variables as a function of alignment: statistical power to detect treatment effects, group differences in mean latent ability, and implied reliability.

Statistical power estimation. I defined statistical power as the percent of simulated replications with a statistically significant treatment effect. For each replication, I used a Wald test to determine the significance of the latent regression parameter (e.g., Guilleux et al., 2014) under the null hypothesis Ho: $\lambda = 0$ and alternative hypothesis Ha: $\lambda \neq 0$. I calculated power based on the rejection rate of $\alpha = 0.05$. I used a two-tailed test because it is more commonly used than a one tailed test.

Group differences in mean latent ability. I also analyzed the relationship between alignment and the differences between estimated mean latent ability in the experimental group and estimated mean latent ability in the control group. To generate the final results, I calculated the mean of the latent regression parameter over 1000 replications. This calculation provides a different representation that shows the direct impact of alignment on the estimated treatment effect.

Implied reliability. I used ACER ConQuest 4 (Adams, Wu, & Wilson, 2015) to estimate the simulated reliability of the test as a function of alignment. I simulated data in R and imported the data into ConQuest 4, which calculated the person separation reliability (Wright & Masters, 1999) using weighted-likelihood estimation (Warm, 1989) and the overall reliability (Adams, 2005) under Expected A-Posteriori (EAP) estimation (Bock & Aitkin, 1981). Reliability was calculated at 20%, 40%, 60%, 80%, and 100% alignment for each of 12 effect sizes, for 60 total simulations.

3.5.2 Results

Power for detecting treatment effects

First, I analyzed how alignment influences statistical power for detecting the effects of an educational treatment. In this study, statistical power is the percent of simulated replications that produced a statistically significant treatment effect. Figure 3-2 contains a graph of the relationship between alignment and simulated statistical power. The 12 solid lines in greyscale represent the range of effect sizes. The lines were slightly smoothed using the "smooth.spline" command in R. The value of 0.8 on the *y*-axis represents 80% statistical power, which I interpret as minimally adequate. This value translates to a statistical test with 80% chance of detecting a true group effect and a 20% chance of missing a true group effect. Table 3-A1 in the Appendix contains a table of the data for the graph in Figure 3-2.

Figure 3-2 shows a positive relationship between alignment and statistical power. However, the relationship is not linear. The "S" shape of the curves reflects the mathematics of the Rasch model. For example, the deceleration in the alignment-power function at higher alignment reflects a ceiling-like effect related to the upper limit of the logistic function, where a constant change in ability produces a decreasing change in the probability of success on items. The smaller changes in probability translate to smaller real changes in the group means that drive statistical power. The diminishing return of increasing alignment on the resulting statistical power becomes noteworthy at about 0.8 power.

Figure 3-2 also shows that effect size is a critical determinant of how much alignment is necessary to achieve adequate statistical power. The slope of the alignment- power function sharpens as effect size increases. At the larger effect sizes, power rapidly reaches 0.8 with only a small fraction (~20%) of aligned items needed for adequate power. However, it is important to note that effect sizes over 0.5 *SD* are very rare for instructional interventions (Scriven, 2009). Thus, the analysis focuses on effect sizes between 0.1 to 0.5 *SD*s that are most relevant to evaluating instructional interventions.

For the smallest effect sizes (0.1 – 0.3), strong alignment appeared to be essential for adequate statistical power. Evaluations of interventions with small impacts are underpowered even at 100% alignment. High alignment is necessary for adequate power to detect small effects distributed across many items. Thus, close alignment is vital when researchers wish to conduct evaluations that can detect small effect sizes. Researchers may be able to improve sensitivity by using larger sample sizes.

For the larger effect sizes (0.4 – 0.7), statistical power remained adequate down to approximately 50% alignment. Researchers have greater flexibility to accept misalignment in an outcome measure when the impact of the intervention is expected to be moderate. Although decreases in alignment shrink the estimated treatment effect, as I will show in the next section, low power is not a central concern in this situation.

In the very rare cases where instructional interventions reach effect sizes of 0.8 – 1.0, alignment may drop to 20% while maintaining adequate power to detect differences between experimental and control groups when they exist. Interventions with effect sizes greater than 1.0, required very little alignment (< 20%) for adequate statistical power. Such large impacts are unlikely to occur for instructional interventions but may apply to certain scenarios where assessments are closely tied to instruction (e.g., Cohen, 1987).

Estimation of the treatment effect

Research Question 2 asked how the estimated treatment effect changes as a function of alignment. A model in the literature posits that group effects will decrease with decreases in alignment (May et al., 2009), but empirical research has not shown this decrease. The graph in Figure 3-3 represents the simulated group effects. The simulated group effect is the mean of the estimated group difference over 1000 replications. Each line represents a different effect size. Table 3-A2 in the appendix contains the statistics represented in Figure 3-3.

The simulated group effect is related to the effect size of the intervention, hence the differences in the slopes of the lines represented in Figure 3-3. Simple regression of the mean group effect on alignment, independently for each effect size, showed perfect fit for the linear models ($R^2 = 1.0$). Each aligned item made an equal contribution to the change in treatment effect. For example, a hypothetical treatment effect of 1.0 and 50% alignment yielded a predicted treatment effect of 0.5, with each of 25 aligned items contributing an estimated effect size of 0.02 towards the group effect. Although the results are partially an artifact of the simulation design that homogenized the variance contributed by each item (by repeated redraws of the item set instead of a single item set with differences between items as a function of position on the distribution of item difficulty), the results present an empirical model for the change in estimated treatment effects related to alignment. Decreases in alignment can lead to attenuation of the

treatment effects that would otherwise be observed using a more aligned assessment. Notably, the 100% alignment condition recovered the deterministically set effect sizes, which is evidence that the simulation functioned as expected.

Implied reliability

Research Question 3 investigated the relationship between alignment and the reliability of the test implied by the simulation. The person separation reliability and overall reliability changed very little as a function of alignment. Across effect sizes and alignment conditions, the person separation reliability ranged between 0.88 and 0.91.
The overall reliability ranged between 0.84 and 0.95, with the higher reliability typically observed in the higher alignment conditions of the large effect sizes. The lack of change was as predicted, and the increase in reliability was concomitant with increases in estimated variance.

3.6 Study 2: Empirical example

The second study tested the veracity of the alignment function using a real-world example. The goal was to determine whether the model could accurately predict the effect size from a misaligned test, given (a) an estimate of the degree of misalignment and (b) the "true" effect size obtained from an aligned test. The example comes from a quasi-experimental evaluation of the Learning Mathematics through Representations (LMR) curricular intervention. Saxe et al. (2012) evaluated LMR using a researcher-developed test that was aligned with the intervention, and the estimated effect was 0.66 *SD*s ($p < 0.001$). Sussman and colleagues (2012) evaluated LMR using data from a misaligned standardized test and estimated a treatment effect of 0.20 *SD*s ($p = 0.09$). Sussman et al. (2012) hypothesized that differential alignment might be the cause of the difference in the estimated treatment effects and their analysis determined that the standardized test was only 33% aligned with LMR. As a test to determine whether the model in this paper is realistic, I estimated the treatment effect from a misaligned standardized test, given the assumptions of (a) 33% misalignment and (b) a "true" effect size of 0.66. I compared the model prediction with the empirically observed treatment effect of 0.20 *SD*s.

3.6.1 Method

I used simple linear regression to estimate the relationship between alignment and the treatment effect. The regression had the general form shown in Equation 3.4:

$$y_i^* = \alpha + \beta_1 Alignemnt + \epsilon_i, \tag{3.4}$$

where y_i^* is the mean simulated effect size for alignment condition i, normalized for a simulated effect size of 1.0, and β_1 represents the linear impact of alignment. Thus, the magnitude of β_1 is conditional on the simulated effect size (see Figure 3-3).

I normalized the data to an effect size of 1.0, as shown in Equation 3.5, to use all available data for estimating the impact of alignment.

$$y_i^* = \bar{\lambda} * \frac{1}{ES}, \tag{3.5}$$

where $\bar{\lambda}$ is the mean of the simulated latent estimates (from Equation 3.3) and *ES* is the simulated effect size (0.1-2.0). The coefficient for alignment, β_1, is interpretable as an estimate of contribution of one percent alignment to an effect size of 1.0.

Equation 3.6 shows how the primary research result was calculated. The calculation produced an estimate for the expected effect size of LMR when evaluated using the misaligned standardized test. I used the "margins" command in Stata13 (StataCorp. 2013) to obtain the marginal estimate when the alignment was 33%. This marginal estimate is interpretable as a scale factor for the impact of decreased alignment on the "true" effect size.

$$Expected\ Effect\ Size = Marginal\ Estimate * True\ Effect\ Size. \tag{3.6}$$

3.6.2 Results

A simple linear regression of standardized mean simulated treatment effect on alignment produced an $R^2 = 0.99$, indicating excellent linear fit. The coefficient for alignment was estimated at 0.01 (SE < 0.0001). The interpretation is that each percent alignment increased the expected treatment effect by 0.01 SD, or one percent of the total treatment effect. Thus, the marginal estimate at 33% alignment was 0.33. Following Equation 6 and assuming a true effect size of 0.66, the model predicts that a test with 33% alignment would result in an estimated treatment effect of 0.22. I interpret this prediction as very close to the empirically observed value of 0.20 (i.e., Sussman et al., 2012), and within the criteria for adequacy (±0.05) specified in the research hypotheses. The empirical example supported the accuracy of the alignment function.

3.7 Discussion

The results of this study suggest that strong alignment is required for adequate statistical power to detect treatment effects of the magnitude commonly observed in education. This experiment simulated experimental evaluations of educational programs and derived simulated functions for the relationship between alignment and outcome variables associated with the sensitivity of an evaluation for detecting the effect of an educational treatment (statistical power and estimated treatment effect). The results suggest that alignment must be high for detecting smaller effect sizes of 0.1 to 0.3 *SD*s. When the treatment effects are small, any misalignment sacrifices sensitivity and threatens the validity of an evaluation. For larger effect sizes (0.4 – 0.7), the treatment sensitivity of an evaluation remains adequate under misalignment up to approximately 50%.

The results also show that decreases in alignment attenuate estimated group effects. Interventions will appear less effective when evaluated with a misaligned assessment than they will when evaluated with a fully aligned assessment.

In addition, the results showed that changes in alignment had little impact on the person separation reliability or the reliability of the population level parameters. I regard this

investigation as a first step in considering how alignment impacts measurement reliability. A next step might generate a reliability coefficient that differentiates between aligned items and misaligned items, partitioning reliability into aligned and unaligned variance. Accounting for alignment might provide a more realistic representation of reliability in evaluation scenarios.

3.7.1 Implications of the results

How do the current results help us evaluate the validity of standardized tests as outcome measures for evaluating new instructional interventions. Readers should interpret the results of this study with caution because of the exploratory nature of the method. However, cross-referencing the results of this study with the information in the literature about the use of standardized tests in applied educational research, the conclusion is that serious alignment problems exist and many evaluations have low sensitivity for detecting treatment effects.

The literature contains numerous examples where investigators use standardized tests to evaluate interventions such as brief curricular units or interventions. It is apparent that the standardized tests commonly measure more academic content and different academic skills compared to most contemporary instructional interventions. Curiously, many of the studies neglect to validate the test or discuss the overlap between what the test measures and the goals of the intervention (Sussman, 2016). Thus, to support the validity of an evaluation, it is important to ask questions about alignment: how aligned is the test, and what is the impact of alignment on the conclusions derived from the research?

To gain insight into the degree of misalignment between standardized tests and short-term curricular units, consider a fairly common situation where an investigator uses a standardized test to evaluate the learning caused by a four-week instructional intervention. Let us consider the breadth, or grain size, of each. The typical standardized test measures learning across a school year of 175 to 180 days. A month long intervention cannot tap the same breadth of content knowledge, nor is it typically designed to. It is reasonable to expect misalignment, and one might speculate that the amount of misalignment would be inversely proportional to the length of the intervention.

As further evidence of misalignment, the empirical example in this paper reported that the standardized test had 33% alignment with the educational intervention. Sussman (2016) identified a handful of IES-funded studies that could not complete an impact evaluation because the investigators learned, only after data collection, that the standardized test did not measure the appropriate outcomes relative to the purpose of the evaluation. In light of the current results, the evidence suggests that evaluation of new interventions with standardized tests leads to alignment problems in applied educational research that require closer study.

Implications for measurement in educational evaluation

The results of this study suggest that item content analysis is essential for valid measurement in applied educational research. First, stronger alignment generally leads to greater sensitivity for detecting treatment effects and more accurate estimation of the treatment effect. Second, hidden from most educational intervention research is whether

the impact of an "effective" intervention was realized across items or linked to dramatic improvements on a subset of items. I recommend item analyses that (a) investigate the alignment of the test prior to data collection and (b) review the outcome data to determine if the intervention influenced performance on many items or just a few items. An alignment study should include a mechanism of action for how the intervention helps students succeed on the test items (e.g., Sussman, 2016b). The alignment study parallels other forms of validation (AERA, APA, & NCME, 2014), which subsumes alignment as part of construct validity. However, the emphasis is on establishing the sensitivity of the items to the instruction provided by the intervention.

Researchers can reference the current results to predict the power of their study for detecting treatment effects. Researchers can generate a percentage estimate of misalignment, or estimate a range of potential alignment, and use the alignment results to predict the statistical power of their study for detecting treatment effects of various sizes. The results apply to a randomized evaluation with 600 participants, so investigators may need to use this method to generate functions for different sample sizes.

Implications for the politics of standardized testing

The current analysis points to the need to reconsider the use of standardized tests in applied educational research. Many educational leaders think of standardized tests as a "gold standard" for measuring educational achievement (Office of Technology Assessment, 1992). Some authors claim that the main alternative to standardized tests, researcher-developed tests, are prone to bias and prone to artificially inflated impact estimates because the test will be "over-aligned" with the intervention (Slavin & Madden, 2013). Indeed, federal policies reflect this perspective. For example, policy from the What Works Clearinghouse (WWC) stipulates that standardized tests are automatically valid whereas researcher-developed tests must be validated before the data are acceptable as research evidence (Song & Herman, 2010). Perhaps it is overgeneralization of this type of broad support for standardized testing that has contributed to the use of misaligned standardized tests as outcome measures.

One way to guard against the bias issue is incorporation of a construct modeling perspective that includes item analysis and careful design of measures to the learning intended by an intervention (Wilson, 2004). The key is careful measurement and knowing what you are measuring. Ultimately, I hope that many stakeholders will reexamine their beliefs about standardized tests and dedicate more effort towards appropriate measurement for evaluation of educational interventions. Many scholars in fields including educational measurement, mathematics education, and science education, have discussed concerns about the validity of standardized tests for summative evaluation of educational programs (Pelligrino, Chudowsky, & Glaser, 2001; Raudenbush, 2005; Schoenfeld, 2006). I hope that this study adds empirical clarity to the theoretical perspectives in the literature.

Implications for the concept of alignment

This study may help to advance the discussion of alignment in the literature. A basic issue with alignment is that there is no agreed-upon definition of how much alignment is enough for a given purpose. In this paper, I advanced the idea that alignment has a functional relationship with defined outcomes. The model permits a researcher to estimate how much alignment is enough for adequate statistical power to detect the effect of an educational treatment under certain conditions. It remains to be seen if this research applies to more conventional alignment research. For example, to better understand the real world consequences associated with misalignment, I plan to simulate the impact of statistical differences in alignment (Polikoff & Fulmer, 2013) on outcomes such as the mean proficiency in a school.

Finally, another basic problem in alignment research is that estimates rely on fallible human judgment. The method in this research relaxes the need for judgment, and for point estimates of alignment, because it produces a function for alignment. Although humans will always need to make final decisions about whether items do or do not align with a particular criterion, this research provides a way for individuals to conceptualize alignment within an interval and to weigh the benefits of increased attention to alignment with the cost of being at various points within the interval.

3.7.2 Limitations

This study is merely a first step in validating a new way to study alignment. Additional studies that support the current results should be considered before making broad statements about the utility or impact of the approach. Below, I highlight a handful of noteworthy limitations focused on the research method and suggest how they might be addressed by future research.

The binary conceptualization of alignment used in the current research is convenient from a modeling perspective but it is unlikely to be realistic. Alignment is more likely to be on a continuum, and partial alignment may lead to partial impact of an intervention. In addition, the efficacy of an intervention is likely to be conditional on additional variables such as prior knowledge. Future research should estimate variation in the impact of an intervention, and investigate whether the variation can be modeled as the effect of additional variables. Variables that lead to systematic (fixed) effects might prove fruitful points for interventions that boost student achievement and research that leads to basic knowledge about conditions that support the efficacy of educational interventions.

The simulation method produced data in a controlled environment relative measurement in practice. For example, I used standard normal distributions to generate both latent variables and item difficulty. I generated data that fits the Rasch model—a non-trivial requirement for real world data. The current results are conditional on a sample size of 600, and researchers who wish to generalize to other sample sizes will likely need to replicate the method. Although the modified unidimensional Rasch (DIF) model used in this study may accurately portray measurement in practice, future research should explore the multidimensional models that generate additional information through modeling changes in latent ability along two or more dimensions of alignment.

3.7.3 Conclusion

This research developed a new method for modeling the impact of alignment on the results of an evaluation. I used a Rasch item response modeling technique to simulate the mechanism of action of an intervention by increasing the probability of success on items considered aligned with the intervention. Then, I used a latent regression Rasch model to estimate the impact of the intervention. The models were fit in the context of a Monte Carlo simulation that tested multiple conditions of alignment. The results generated functions for alignment and two outcomes: statistical power and group effects.

The results provide tentative guidelines on how much alignment is required for valid evaluation. The results suggest that very strong alignment is required for adequate statistical power to detect small treatment effects (0.1 – 0.3). For the interventions with at least moderate effect sizes (i.e., > 0.4 SD), alignment can drop to about 50%. However, decreases in alignment attenuate estimated treatment effects. I tested the accuracy of the results using an empirical example and found that the function for group effects accurately predicted the effect size from a misaligned standardized test in a real-world example.

Evidence suggests that researchers should reexamine the use of standardized tests in applied research on instructional interventions. Cross referencing the current results with details from projects funded through the IES mathematics and science program suggests that, in many cases, standardized tests are not well aligned enough for use as outcome measures. To support beneficial educational reform, applied researchers should dedicate more attention to measurement, including item level alignment, to gain greater clarity on how an educational intervention or treatment impacts test scores.

3.8 Table and figures

Table 3-1

Demonstrating the Relationship Between Alignment and the Probability of A Correct Answer to Non-Aligned and Aligned Items

	Probability of Success	
Item difficulty (Logits)	Non-Aligned	Aligned
-3.0	0.97	0.99
-2.5	0.95	0.98
-2.0	0.92	0.96
-1.5	0.88	0.94
-1.0	0.82	0.91
-0.5	0.73	0.86
0.0	0.62	0.79
0.5	0.50	0.69
1.0	0.38	0.57
1.5	0.27	0.45
2.0	0.18	0.33
2.5	0.12	0.23
3.0	0.08	0.15

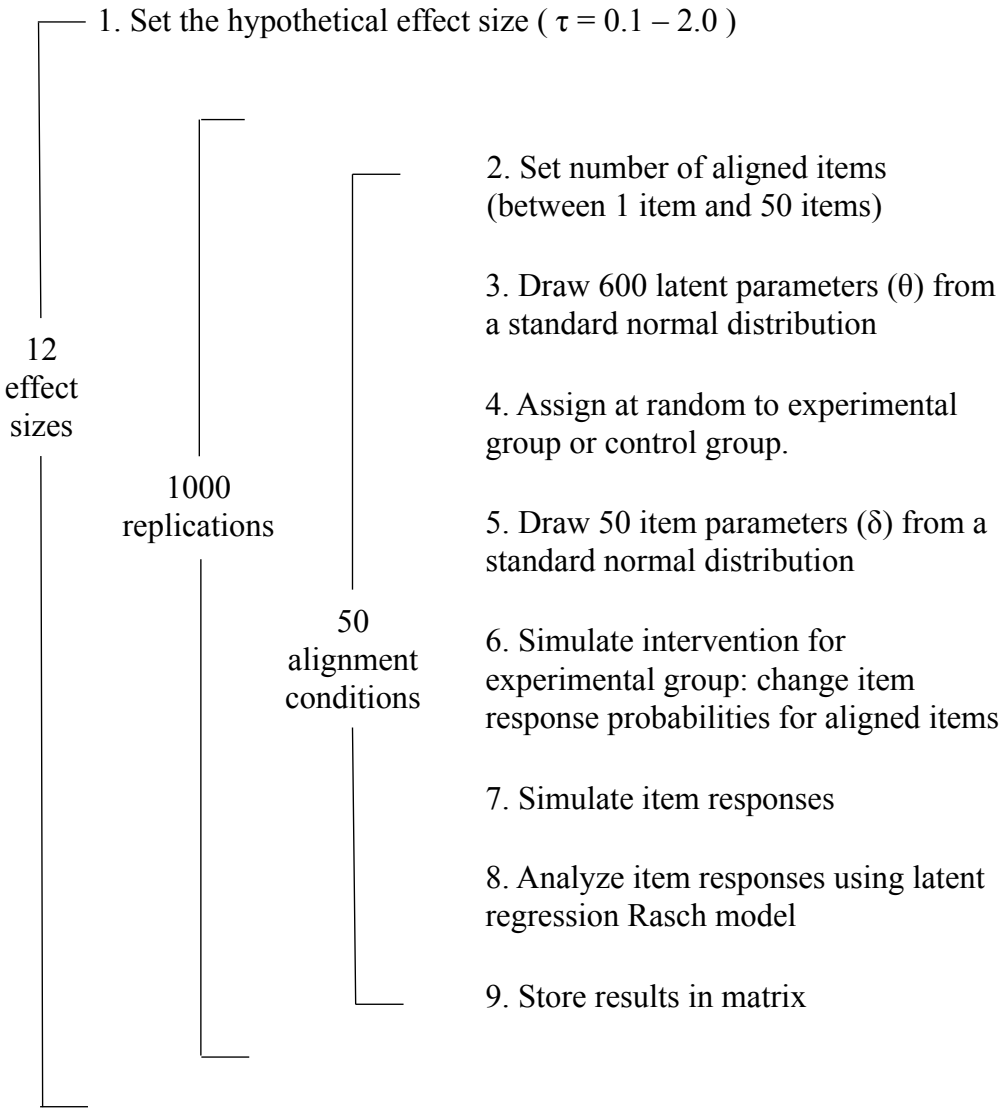

Figure 3-1. Simulation algorithm. Future research will conduct step 5 once rather than for each replication.

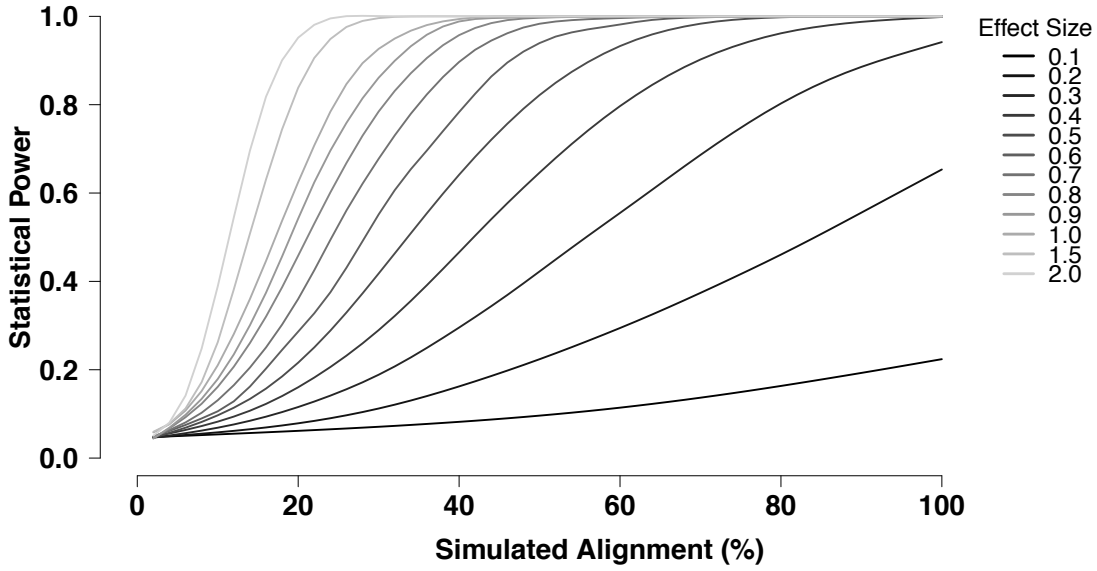

Figure 3-2. The functional relationship between alignment and statistical power for interventions of varying effect size.

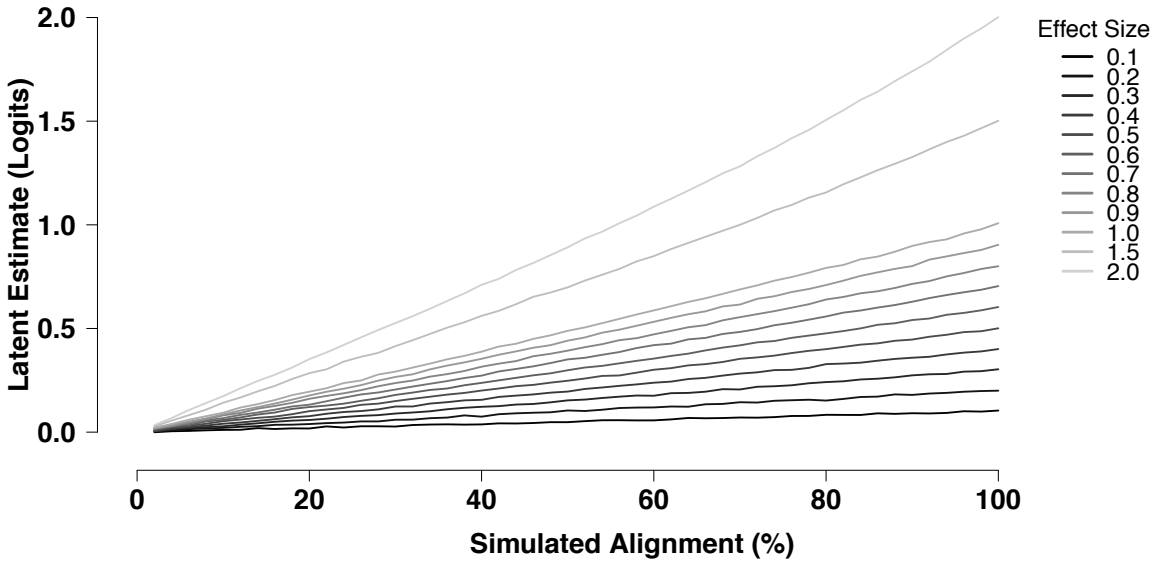

Figure 3-3. The relationship between alignment and mean estimated treatment effect for different effect sizes.

3.A Appendix

Table 3-A1 contains data for the graphs in Figure 3-2. The data are point estimates and the variance of the estimates across 1000 replications. To facilitate interpretation of the table, I shaded the threshold of 100% power in darker grey and 80% power in lighter grey. The darker shading shows the alignment conditions that have maximum statistical power to detect treatment effects. The light grey shading shows the threshold where power drops below 0.8, becoming inadequate. The variance calculations show, as expected, smaller variance in the tails of the functions and larger variance in the middle of the functions. Table 3-A2 contains the statistics represented in Figure 3-3. The table presents the mean of estimated treatment effects for 1000 simulation replications. The variance was < 0.01 for all condition

Table 3-A1

The Impact of Alignment On Statistical Power for Interventions with Various Effect Sizes

	Effect Size																					
Alignment	0.1		0.2		0.3		0.4		0.5		0.6		0.7		0.8		0.9		1		1.5		2.0	
%	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.	%	Var.
2	5.4	0.05	6.7	0.06	5.9	0.06	5.9	0.06	5.6	0.05	5.0	0.05	5.5	0.05	5.2	0.05	5.2	0.05	7.1	0.07	7.0	0.06	7.0	0.07
4	5.8	0.05	5.1	0.05	5.3	0.05	6.2	0.06	6.8	0.06	6.2	0.06	6.3	0.06	6.9	0.06	6.5	0.06	6.1	0.06	6.9	0.11	12.4	0.11
6	4.4	0.04	5.4	0.05	4.8	0.05	5.8	0.05	5.6	0.05	7.1	0.07	8.2	0.08	9.4	0.09	9.6	0.09	11.3	0.10	12.8	0.17	24.5	0.19
8	4.8	0.05	5.2	0.05	6.1	0.06	7.6	0.07	7.5	0.07	10.1	0.09	9.1	0.08	10.7	0.10	13.8	0.12	14.7	0.13	21.7	0.22	38.9	0.24
10	5.6	0.05	4.6	0.04	6.7	0.06	7.6	0.07	10.6	0.09	10.6	0.09	13.5	0.12	17.2	0.14	17.1	0.14	20.8	0.16	34.1	0.25	54.0	0.25
12	5.0	0.05	6.5	0.06	8.1	0.07	10.0	0.09	12.3	0.11	11.2	0.10	16.5	0.14	19.2	0.16	23.1	0.18	28.4	0.20	47.8	0.24	70.4	0.21
14	5.6	0.05	6.4	0.06	7.5	0.07	10.0	0.09	12.6	0.11	16.1	0.14	20.7	0.16	27.3	0.20	31.2	0.21	36.1	0.23	61.4	0.20	83.3	0.14
16	6.4	0.06	8.3	0.08	9.6	0.09	10.6	0.09	13.6	0.12	20.7	0.16	24.6	0.19	31.4	0.22	35.9	0.23	43.5	0.25	72.8	0.11	89.8	0.09
18	5.9	0.06	6.8	0.06	9.9	0.09	14.6	0.12	18.6	0.15	25.0	0.19	31.1	0.21	37.1	0.23	45.6	0.25	55.1	0.25	86.8	0.06	95.1	0.05
20	6.5	0.06	7.0	0.07	11.6	0.10	15.8	0.13	22.9	0.18	29.6	0.21	34.3	0.23	47.0	0.25	53.8	0.25	62.0	0.24	93.5	0.03	97.9	0.02
22	6.6	0.06	9.2	0.08	13.6	0.12	18.8	0.15	24.2	0.18	32.1	0.22	41.7	0.24	52.9	0.25	64.5	0.23	70.2	0.21	96.8	0.01	99.6	0.00
24	6.1	0.06	8.4	0.08	15.0	0.13	21.4	0.17	27.4	0.20	36.3	0.23	49.8	0.25	61.0	0.24	69.3	0.21	79.2	0.16	99.0	0.01	99.9	0.00
26	7.4	0.07	9.1	0.08	15.3	0.13	23.0	0.18	34.0	0.22	43.6	0.25	58.6	0.24	65.8	0.23	75.0	0.19	84.6	0.13	99.4	0.00	100	0.00
28	6.1	0.06	10.2	0.09	16.7	0.14	24.7	0.19	35.7	0.23	48.3	0.25	61.3	0.24	73.7	0.19	82.2	0.15	90.2	0.09	99.9	0.00	100	0.00

30	7.7	0.07	11.4	0.10	18.4	0.15	29.8	0.21	44.0	0.25	56.1	0.25	67.2	0.22	79.3	0.16	86.6	0.12	92.2	0.07	99.9	0.00	100	0.00
32	7.1	0.07	11.2	0.10	19.3	0.16	30.3	0.21	45.5	0.25	61.0	0.24	73.0	0.20	83.1	0.14	88.8	0.10	94.7	0.05	100	0.00	100	0.00
34	6.3	0.06	11.7	0.10	22.0	0.17	36.1	0.23	50.4	0.25	66.7	0.22	78.6	0.17	86.4	0.12	93.6	0.06	96.8	0.03	100	0.00	100	0.00
36	8.9	0.08	11.8	0.10	25.8	0.19	38.0	0.24	56.3	0.25	67.3	0.22	81.8	0.15	90.6	0.09	96.0	0.04	98.1	0.02	100	0.00	100	0.00
38	8.2	0.08	17.8	0.15	27.6	0.20	44.2	0.25	59.2	0.24	74.9	0.19	86.2	0.12	93.3	0.06	97.8	0.02	98.9	0.01	100	0.00	100	0.00
40	8.2	0.08	15.1	0.13	31.5	0.22	44.6	0.25	65.7	0.23	78.4	0.17	89.5	0.09	96.7	0.03	99.1	0.01	99.2	0.01	100	0.00	100	0.00
42	8.7	0.08	20.4	0.16	31.4	0.22	51.1	0.25	67.6	0.22	81.3	0.15	94.2	0.05	97.0	0.03	99.3	0.01	99.8	0.00	100	0.00	100	0.00
44	7.9	0.07	18.8	0.15	35.0	0.23	56.4	0.25	72.5	0.20	87.8	0.11	93.3	0.06	98.5	0.01	99.6	0.00	100	0.00	100	0.00	100	0.00
46	10.0	0.09	19.6	0.16	34.1	0.22	58.0	0.24	74.3	0.19	89.3	0.10	95.6	0.04	99.4	0.01	99.2	0.01	100	0.00	100	0.00	100	0.00
48	9.1	0.08	20.4	0.16	37.8	0.24	61.4	0.24	80.1	0.16	92.2	0.07	98.4	0.02	99.3	0.01	99.9	0.00	99.9	0.00	100	0.00	100	0.00
50	8.0	0.07	22.3	0.17	45.1	0.25	62.5	0.23	82.9	0.14	93.3	0.06	98.4	0.02	99.4	0.01	100	0.00	100	0.00	100	0.00	100	0.00
52	11.8	0.10	21.6	0.17	44.7	0.25	68.4	0.22	84.8	0.13	96.4	0.03	98.6	0.01	99.8	0.00	100	0.00	100	0.00	100	0.00	100	0.00
54	10.4	0.09	26.4	0.19	48.3	0.25	72.2	0.2	88.2	0.10	95.8	0.04	98.9	0.01	99.8	0.00	100	0.00	100	0.00	100	0.00	100	0.00
56	9.7	0.09	29.5	0.21	51.6	0.25	73.8	0.19	88.4	0.10	97.3	0.03	99.5	0.00	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00
58	10.3	0.09	28.1	0.20	53.9	0.25	76.3	0.18	91.6	0.08	97.3	0.03	99.4	0.01	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
60	10.0	0.09	28.1	0.20	52.9	0.25	81.6	0.15	94.0	0.06	97.9	0.02	99.4	0.01	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00
62	10.3	0.09	29.9	0.21	58.1	0.24	80.3	0.16	94.5	0.05	98.9	0.01	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
64	15.0	0.13	30.2	0.21	60.3	0.24	85.9	0.12	96.3	0.04	99.4	0.01	99.8	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
66	12.2	0.11	36.3	0.23	62.0	0.24	85.4	0.12	96.7	0.03	99.4	0.01	99.8	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00

68	13.7	0.12	34.7	0.23	68.4	0.22	90.0	0.09	97.4	0.03	99.7	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
70	13.9	0.12	38.7	0.24	66.9	0.22	89.8	0.09	98.5	0.01	99.7	0.00	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
72	12.9	0.11	39.2	0.24	72.9	0.20	91.9	0.07	98.9	0.01	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
74	13.4	0.12	42.4	0.24	73.3	0.20	93.8	0.06	99.3	0.01	99.8	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
76	16.1	0.14	43.4	0.25	74.3	0.19	92.8	0.07	99.4	0.01	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
78	15.9	0.13	44.2	0.25	79.4	0.16	95.4	0.04	99.4	0.01	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
80	15.7	0.13	42.1	0.24	81.2	0.15	97.1	0.03	99.7	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
82	17.1	0.14	48.3	0.25	81.6	0.15	97.1	0.03	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
84	16.0	0.13	48.7	0.25	84.2	0.13	97.3	0.03	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
86	22.2	0.17	49.7	0.25	86.4	0.12	97.7	0.02	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
88	18.7	0.15	56.7	0.25	88.0	0.11	98.2	0.02	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
90	18.7	0.15	55.3	0.25	88.6	0.10	99	0.01	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
92	16.8	0.14	57.9	0.24	89.8	0.09	99	0.01	99.9	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
94	19.5	0.16	60.4	0.24	91.3	0.08	99.2	0.01	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
96	22.8	0.18	62.0	0.24	91.7	0.08	99.4	0.01	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
98	22.2	0.17	63.3	0.23	91.6	0.08	99.7	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
100	24.1	0.18	65.0	0.23	94.3	0.05	99.5	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00

Table 3-A2

The Relationship between Alignment and Mean Estimated Treatment Effect for Interventions With Different Effect Sizes.

Alignment	Effect Size											
%	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.5	2.0
2	0.00	0.00	0.01	0.00	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03
4	0.00	0.01	0.02	0.01	0.02	0.03	0.03	0.03	0.04	0.04	0.07	0.07
6	0.01	0.01	0.02	0.02	0.03	0.03	0.04	0.05	0.05	0.06	0.10	0.11
8	0.01	0.02	0.02	0.03	0.04	0.05	0.06	0.06	0.07	0.08	0.13	0.14
10	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.09	0.16	0.17
12	0.01	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.12	0.19	0.21
14	0.02	0.03	0.04	0.05	0.07	0.08	0.09	0.11	0.12	0.13	0.22	0.25
16	0.02	0.04	0.05	0.06	0.07	0.09	0.11	0.13	0.14	0.15	0.26	0.28
18	0.02	0.04	0.06	0.07	0.09	0.11	0.12	0.14	0.16	0.18	0.29	0.31
20	0.02	0.04	0.06	0.08	0.10	0.12	0.13	0.16	0.18	0.20	0.32	0.35
22	0.03	0.04	0.06	0.09	0.11	0.13	0.15	0.17	0.19	0.21	0.36	0.38
24	0.02	0.05	0.07	0.10	0.12	0.14	0.17	0.19	0.21	0.24	0.39	0.42
26	0.03	0.05	0.08	0.10	0.13	0.15	0.18	0.20	0.23	0.25	0.42	0.46
28	0.03	0.05	0.08	0.11	0.14	0.16	0.19	0.22	0.25	0.28	0.46	0.49
30	0.03	0.06	0.09	0.12	0.15	0.18	0.21	0.24	0.27	0.29	0.48	0.53
32	0.03	0.06	0.09	0.12	0.16	0.19	0.22	0.25	0.28	0.31	0.52	0.56
34	0.04	0.07	0.10	0.14	0.17	0.20	0.23	0.27	0.30	0.33	0.55	0.59
36	0.04	0.07	0.11	0.14	0.18	0.21	0.25	0.28	0.32	0.35	0.58	0.63
38	0.04	0.08	0.12	0.15	0.19	0.23	0.26	0.29	0.34	0.37	0.62	0.67
40	0.04	0.07	0.12	0.16	0.20	0.24	0.27	0.32	0.35	0.39	0.66	0.71
42	0.04	0.09	0.13	0.17	0.21	0.25	0.29	0.33	0.37	0.41	0.69	0.74
44	0.04	0.09	0.13	0.18	0.22	0.26	0.30	0.34	0.39	0.43	0.72	0.78
46	0.04	0.09	0.14	0.18	0.23	0.27	0.32	0.37	0.41	0.45	0.75	0.82
48	0.05	0.10	0.14	0.19	0.24	0.28	0.33	0.38	0.42	0.47	0.79	0.85
50	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.39	0.44	0.49	0.82	0.89
52	0.05	0.10	0.16	0.21	0.26	0.31	0.36	0.41	0.46	0.51	0.85	0.93
54	0.06	0.11	0.16	0.22	0.27	0.32	0.37	0.42	0.48	0.53	0.89	0.97

56	0.06	0.12	0.17	0.22	0.27	0.33	0.39	0.44	0.49	0.55	0.92	1.01
58	0.06	0.12	0.18	0.23	0.29	0.34	0.41	0.45	0.51	0.57	0.96	1.04
60	0.06	0.12	0.18	0.24	0.30	0.35	0.42	0.47	0.53	0.59	0.99	1.09
62	0.06	0.12	0.19	0.24	0.31	0.37	0.43	0.49	0.55	0.61	1.04	1.13
64	0.07	0.12	0.19	0.26	0.32	0.38	0.45	0.51	0.57	0.63	1.07	1.17
66	0.07	0.13	0.20	0.26	0.33	0.39	0.45	0.52	0.58	0.65	1.10	1.21
68	0.07	0.14	0.21	0.27	0.34	0.41	0.47	0.54	0.61	0.67	1.14	1.25
70	0.07	0.14	0.21	0.28	0.35	0.42	0.48	0.55	0.62	0.69	1.18	1.28
72	0.07	0.14	0.22	0.29	0.36	0.43	0.50	0.57	0.64	0.71	1.21	1.33
74	0.07	0.15	0.22	0.30	0.37	0.44	0.52	0.58	0.66	0.73	1.26	1.37
76	0.08	0.15	0.23	0.30	0.38	0.45	0.53	0.60	0.68	0.75	1.30	1.42
78	0.08	0.16	0.24	0.31	0.39	0.47	0.54	0.62	0.69	0.77	1.33	1.46
80	0.08	0.15	0.24	0.33	0.40	0.48	0.56	0.64	0.71	0.79	1.38	1.51
82	0.08	0.16	0.25	0.33	0.41	0.49	0.58	0.65	0.73	0.81	1.41	1.55
84	0.08	0.17	0.25	0.34	0.42	0.50	0.59	0.67	0.75	0.83	1.45	1.60
86	0.09	0.17	0.26	0.34	0.43	0.52	0.60	0.68	0.77	0.85	1.49	1.64
88	0.09	0.18	0.26	0.35	0.44	0.53	0.61	0.70	0.79	0.87	1.54	1.69
90	0.09	0.18	0.27	0.36	0.45	0.54	0.63	0.72	0.80	0.90	1.58	1.74
92	0.09	0.19	0.28	0.36	0.46	0.55	0.65	0.73	0.83	0.92	1.62	1.79
94	0.09	0.19	0.28	0.37	0.47	0.57	0.66	0.75	0.85	0.93	1.67	1.84
96	0.10	0.19	0.29	0.38	0.48	0.58	0.67	0.77	0.86	0.96	1.71	1.90
98	0.10	0.20	0.29	0.39	0.49	0.59	0.69	0.79	0.88	0.98	1.76	1.95
100	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.01	1.81	2.00

Note. Variance of all estimated treatment effects was <0.01

Chapter 4

Evaluating the effectiveness of the Learning Mathematics through Representations (LMR) curricular unit for students classified as English Learners (ELs)

4.1 Abstract

Students classified as English learners (ELs) as contrasted with their English Proficient (EPs) peers show depressed achievement in mathematics. However, the literature contains few empirically supported interventions for ELs at the lesson or curricular level. The experimental study reported here documents the effectiveness of the Learning Mathematics through Representations (LMR) lesson sequence in reducing the achievement gap between ELs and EPs. Participants included 44 ELs enrolled in 11 LMR classrooms and 51 ELs enrolled in 10 comparison classrooms at the fourth and fifth grade levels. Multilevel analysis of achievement data revealed that the ELs in LMR classrooms gained more than a matched comparison group as measured by an assessment of integers and fractions ($p = 0.011$; $ES = 0.68$) and by a standardized test in mathematics ($p = 0.010$, $ES = 0.49$). I advance a theoretical mechanism that aligns with the empirical results and argues for LMR's potential as an equitable mathematics intervention for students classified as ELs.

4.2 Introduction

Students classified as English Learners (ELs) demonstrate large achievement gaps in mathematics as compared to their native English-speaking peers (Hemphill & Vanneman, 2011). National Assessment for Educational Progress data from 1996 – 2009 show that mathematics achievement gaps between Latino ELs and English proficient Latinos

averaged 0.58 standard deviations in 4th grade and 0.85 standard deviations in 8th grade. Evidence suggests that achievement gaps in math eventually lead to gaps in occupational attainment. For example, Latinos are underrepresented in scientific and technical occupations in close proportion to the amount of mathematics required for the position (Secada, 1996). The significance of the problem is increasing as Latinos are the fastest growing segment of the United States population and constitute a sizable majority of ELs (Garcia & Cuéllar, 2006). Extrapolation to non-Latino populations of ELs (e.g., Lee & Jung, 2004; Warren & Miller, 2014) suggests that ELs' difficulties in mathematics should be urgently addressed in order to prevent further exacerbation of the gaps in achievement and attainment.

In spite of the long-standing achievement gaps between ELs and native English speaking students, mathematics education for ELs is an underdeveloped area of the research literature (Janzen, 2008). Empirical data that bear on empirically supported practice are sparse (Goldenberg, 2014) and research on effective math interventions for ELs at the lesson or curricular level is similarly underdeveloped (Khisty & Viego, 1999; Secada, 1996). A handful of instructional approaches for enhancing mathematics education for ELs exist (Echevarria, Vogt, Short, Avila, & Castillo, 2010; Chamot & O'Malley, 1994), but the empirical support is limited, especially for the relationship between instructional approach and gains in math achievement (e.g., Chamot, 2007; Short, Echevarria, and Richards-Tutor, 2011). The lack of empirical data is cause for concern that recent educational reforms, such as the Common Core State Standards in mathematics, will fail to meet the needs of ELs. Given the persistent achievement gaps and the projections for increasing number of ELs in classrooms in the United States, developing curricula to support high quality mathematics education in all students, regardless of language classification, should be a priority.

This study documents the effectiveness and the equity of a research-based lesson sequence on integers and fractions, Learning Mathematics through Representations (LMR), for promoting learning gains in students classified as ELs. LMR's developers (Saxe et al., 2013; see http://lmr.berkeley.edu) intended for the curriculum to meet the needs of all students. They documented the effectiveness of LMR by using a measure of integers and fractions achievement, contrasting students who participated in LMR with those who participated in matched classrooms using *Everyday Mathematics* (University of Chicago School Mathematics Project, 2007), a highly regarded curriculum unit. However, in their analysis of achievement, Saxe et al. (2013) did not consider EL status. In the present study, I produce analyses of two measures of mathematics achievement with a focus on ELs who participated in LMR and comparison ELs who participated in Everyday Mathematics classrooms. The first measure was that used by Saxe et al. (2013); however, unlike the Saxe et al. analysis, I disaggregated the original data to evaluate the impact of LMR specifically for students classified as ELs. Second, I add new data to the evaluation, which was previously collected but not published, on LMR participants: student scores on the state assessment in mathematics.

4.2.1 Learning Mathematics through Representations (LMR)

LMR is a research-based curriculum unit designed to support upper elementary students' understandings of integers and fractions (Saxe et al., 2013). The designers' intent is to foster deep conceptual understanding of the mathematics though rich instructional tasks

and pedagogical techniques that engage students in productive whole class and small group mathematical discussions. Students who participate in LMR complete number line problems in the context of a supportive, resource rich instructional environment for mathematical problem solving. The curricular design principles encourage constructive discussions and interplay between instructional resources for mathematical communication and sensemaking. As the lessons progress, the curricular resources shift to more complex and more elaborated problems supporting the development of all students. In the following sections I elaborate upon LMR's curricular and pedagogical features and describe how they support ELs in particular.

LMR's design principles

Table 4-1 depicts LMR's design principles, stated as simple maxims, that support mathematical teaching and learning for all students. The table highlights both specific tools and classroom norms that support integration of all the resources in service of mathematical problem solving.

An organized lesson structure supports student participation. One way that LMR accomplishes active participation of all students in rich discussions centered on mathematical problem solving is through a five-phase lesson structure that blends whole-class, group, and individual instruction (Gearhart & Saxe, 2014). Lessons begin with an opening problem that focuses on core mathematical ideas and elicits variations in student thinking that teachers subsequently use to structure the lessons. After opening problems, lessons follow a structured sequence of discussions that each support listening, deliberation, and building mathematical consensus. The teacher facilitates student discussion about problematic ideas, steers the lesson toward important mathematical concepts, and supports resolution of mathematical disagreements through student discussion. In the final phase, students individually solve closing problems, and teachers can use their solutions as information for structuring future discussions. This structure is used over each of the 19 LMR lessons.

The number line as a central representational context for the LMR lesson series. LMR's designers intended to create inclusive environments that would support learning for all students. The curriculum incorporates research-based instructional strategies that support students' mathematical development. LMR makes extensive use of visual representations. For example, students who participate in LMR complete number line problems such as splitting a single unit interval on a number line into subunit intervals to provide a context for making sense of unit-subunit relations and the connections between integers and fractions. Number lines are the focal point for constructing a system of meaning for teaching math and for coordinating the multiple resources in the classroom.

Mathematical definitions as resources to support mathematical communication. LMR provides support for language development with publically displayed, expanding set of mathematical definitions. During LMR lessons, teachers introduce and explain mathematical definitions to the class. For example, teachers provide formal definitions for *unit interval* and *multiunit interval*. The definitions serve multiple roles, including being a starting point for mathematical discussions whose purpose is problem solving and

construction of new knowledge (Saxe et al., 2010). LMR teachers post written definitions in plain view and mark them as important reference information. The list of definitions grows as students progress through the curriculum.

In this way, LMR's definitions provide an initial basis for mathematical reasoning and, over the course of the curriculum, become powerful resources for ongoing inquiry and sensemaking. At the outset, the definitions are a set of basic mathematical concepts and a form of initial *common ground* (Saxe et al., 2015a) that supports the mutual understanding and equitable participation of the entire class. However, LMR treats the definitions as dynamic concepts and thus over time, students participate in the use and refinement of the definitions in the context of an ongoing discussion. The class negotiates the meaning of the definitions, and empirically tests the logical confines of the definitions while engaging in mathematical inquiry with the number line representations.

Use of manipulatives as support for mathematical teaching, learning, and communication. Students use Cuisenaire rods (c-rods) as tools for mathematical sensemaking. In LMR lessons, C-rods allow students to manipulate linear magnitudes. For example, students may physically iterate, or displace, a c-rod to achieve an understanding of continuous linear distance. Similarly, students may test their understanding of equivalent fractions by splitting the same unit interval (i.e., same point on a line) into multiunit intervals of different lengths and counting the number of subunits. The c-rods provide opportunities for students to use physical tools to display their mathematical reasoning though actions.

Productive classroom norms that support effective instructional practices and student learning. The intent of the designers was for the LMR sequence to foster *sociomathematical norms* (Yackel & Cobb, 1996) that support the coordinated actions of classrooms in collective activity. For example, LMR fosters a norm around high expectations for mathematical communication. Teachers who follow LMR expect students to use the formal mathematical definitions to support their reasoning, especially when students are explaining their thinking or justifying their statements. In addition, teachers commonly expect students to present solutions to the entire class, not just the teacher. For the teacher's part, the norm is that the teacher will play a central role in orchestrating productive mathematical discussions. Teachers should continually encourage students to make conjectures and test their hypotheses using the classroom resources. Teachers will also provoke cognitive conflict by encouraging students to consider alternative (and often problematic) solutions to number line problems. When LMR lessons are skillfully orchestrated, the shared expectation is that mathematical discussions will take place and that student talk is a central part of those lessons.

Coordination between resources that support mathematical learning. The productive interplay between the elements over time results in powerful learning experiences for students (Saxe et al., 2015a). LMR's mathematical definitions provide an initial network of resources for students' communication and solving of the number line problems. The use of c-rods supports students' ability to use physical (sensorimotor) actions to test the meanings and limits of the verbal definitions. For example, students' may use c-rods to perform actions such as counting, splitting, and displacing on the

number line representation to test the entailments of the mathematical definitions while solving number line problems (Saxe, de Kirby, Kang, Le, & Schneider, 2015b). The teacher guides students through this process of negotiating the meaning and utility of the definitions for solving mathematical problems. The crosstalk between definitions, sensorimotor actions, and the number line representation constitutes a nexus of mathematical sensemaking.

A system of definition-action relations emerges as the class engages in mathematical inquiry. The interplay between various resources for mathematical problem solving supports the progressive coordination of increasingly sophisticated mathematical ideas. Form-function relations shift by design, for example when students transition from integers to fractions. In sum, the individual and group activities that occur as part of the LMR lesson are quite different than what has been reported as typical for ELs experiences with mathematics education.

4.2.2 Theoretical perspectives on teaching and learning mathematics

The design of LMR is consistent with current perspectives on effective mathematics education and on designing supportive environments for ELs. First, LMR's approach to mathematical teaching and learning parallels the current consensus model that emphasizes mathematical communication and problem solving (National Council of Teachers of Mathematics [NCTM], 2000, Schoenfeld, 2002). In mathematics classrooms driven by the current consensus model, students participate in mathematical discourse practices such as presenting arguments, describing mathematical hypotheses, explaining solutions, and proving conclusions (Moschkovich, 2002). Communication is emphasized because it is viewed as a primary mechanism for mathematical development (i.e., Cobb, Boufi, McClain, & Whitenack, 1997).

High-quality mathematics education should center on communication but also draw from a variety of resources for mathematical sensemaking. Utilizing an array of resources can mean a language-rich mathematics lesson that refers to visual representations and incorporates students' actions as tools for mathematical sensemaking (Schleppegrell, 2007). Visual representations may take the form of graphs and figures that help students attend to key information. Actions such as gestures and manipulatives can be powerful resources for communicating mathematically, especially for students who struggle with language (Bustamante & Travis, 1999).

Finally, high quality mathematics instruction provides classroom norms, routines, and expectations that support students' experiences with mathematical communication and regulate their mathematical development (Yackel & Cobb, 1996). The general idea is that norms convey the shared expectation among teachers and students about the process and goals of communicating mathematically. For example, effective norms might cultivate a classroom environment where lessons lead to discussions that support reflective inquiry and the progressive construction of knowledge that is useful for solving mathematical problems (Hiebert et al., 1996).

4.2.3 Mathematics instruction for English Learners (ELs)

The consensus in the literature is that mathematics education for ELs should support their full participation in activities that focus on mathematical communication and

sensemaking (Moskchovich, 2012). However, language minority students have, by and large, been excluded from participating in high-quality programs of mathematics education (Anderson & Tate, 2008). Historically, many math classrooms that served ELs, especially low-income ELs of Latino descent, neglected to facilitate their participation in mathematical discussions and instead emphasized lower-level skills such as computation, rote memorization of math facts, and practice of word problems (Darling-Hammond, 2007; Khisty & Viego, 1999; Moskchovich, 2002; Ortriz-Franco, Hernandez, & De La Cruz, 1999).

Unfortunately, earlier theories that shaped educational programs for ELs overemphasized the importance of rote aspects of learning vocabulary words and underemphasized their use in rich mathematical discussions (Moskchovich, 1999). As a result, ELs often experience narrow math lessons that are lecture-based, textbook-centered, and make little use of external resources (Ramirez and Bernard, 1999). For example, Brenner (1998) studied a math classroom that served a high proportion of ELs and found that students had very little opportunity for mathematical communication. The talk that did occur was low-level and focused on simple answers rather than extended, elaborated discussion. Opportunities for building from prior knowledge were slim: lessons had little sense of purpose, little connection to key ideas, and little concept development over time.

Increasing ELs' access to the mathematics curriculum

Mathematics curriculum and instruction can improve access for ELs by drawing on the full range of resources available for the purposes of mathematical communication and sensemaking. The question of how one provides access for ELs in K-12 mathematics education is an under-studied area of the literature (Ortiz-Franco, Hernandez, & De La Cruz, 1999). However, scholars have provided some general recommendations. For example, Moschkovich (2012) recommended access to curricula, instruction, and teachers who are demonstrably effective in supporting academic success for this population. Moschkovich defined access as the provision of "abundant and diverse opportunities for speaking, listening, reading, and writing" (p. 18). In addition, instruction that provides access should help students actively engage with, and think deeply about, the mathematics. It should "encourage students to take risks, construct meaning, and seek reinterpretations of knowledge within compatible social contexts" (Garcia & Gonzalez, 1995, p. 424). In sum, the prescription is to fully engage ELs in mathematical communication and to form productive norms that encourage ELs in sensemaking activities.

LMR offers ELs access to the curriculum through visual, verbal, and physical resources for mathematical problem solving. First, in LMR, the number line is a visual resource that serves as the principal representational context for supporting all students' problem solving. Central use of the number line contrasts with some traditional educational experiences that make little reference to visual representations. Second, the mathematical definitions provide a foothold for ELs to enter into discussions and a way to bootstrap mathematical knowledge. Teachers articulate definitions early on and they remain on public display as the class references them in an ongoing process of negotiating the definitions. Third, the C-rods provide opportunities for students to use physical tools to display their mathematical reasoning though actions. The use of c-rods

provides another opportunity for ELs to access the mathematics in a way that does not disadvantage them due to differences in language proficiency. Finally, LMR's designers intended to promote the interplay between resources and instructional strategies, leading to productive educational experiences for all students, including ELs. I posit that LMR supports a resource rich environment and productive sociomathematical norms that level the playing field between ELs and native English speakers.

Systems for teaching ELs mathematics

Two commercially available systems of instruction exist for teaching ELs across all content areas. The Cognitive Academic Language Learning Approach (CALLA) is a recognized method for teaching ELs in the content areas, including mathematics (Kersaint, Thompson, & Petkova, 2013). However, scant empirical data supports the association between CALLA and increases in ELs' math achievement (i.e., Chamot, 2007; Chapiro, & Ball, 1999). An alternative is the SIOP model (Sheltered Instruction Observation Protocol; Echevarria et al., 2010), which was originally developed for instruction in English language arts and later applied to mathematics instruction. As with CALLA, an empirical evaluation of SIOP did not find support for ELs' achievement gains in mathematics (Short et al., 2011). A third commercially available curriculum focusing on mathematics only, *Math Pathways and Pitfalls*, has limited research support for equity with ELs (i.e., Curtis, Heller, Clarke, Rabe-Hesketh, & Ramirez, 2009).

Curricular interventions for teaching ELs mathematics

The literature contains a small number of curricular intervention studies that reported noteworthy gains in mathematics achievement for ELs (Doty, Mercer, & Henningsen, 1999; Fuson, Smith, and Lo Cicero, 1997; Lane et al., 1995; Warren and Miller, 2014). These studies evaluated the impact that educational interventions lasting at least one year had on the mathematics achievement of Latino ELs. Fuson et al. (1997) conducted a year-long teaching experiment with first graders from a predominantly Latino neighborhood and assessed their progress along a hypothetical developmental continuum for place value, addition, and subtraction knowledge. The results from experimenter-administered tasks suggested that the first graders' performance was "substantially above that reported in other studies for U.S. first graders of higher SES and for older U.S. children" (p. 738).

Doty et al. (1999) reported on project QUASAR, a teacher professional development intervention carried out for multiple years in schools that served predominantly Latino or predominantly African American populations. The authors evaluated the performance of one QUASAR school, where 75% of students were classified as ELs, against other schools in the district. The results from a standardized test in mathematics showed that the QUASAR school outperformed the district average and a matched comparison school across all three years of the investigation. Additionally, the results from a study of two QUASAR schools using a researcher-developed assessment of mathematical thinking and communication showed that the gains' in bilingual students' achievement exceeded the gains made by English speaking students (Lane et al., 1995).

A more recent example is Warren and Miller (2014), who presented evidence on the efficacy of *Representations, Oral Language and Engagement in Mathematics* (RoleM) a three-year mathematics curriculum for Australian ELs in kindergarten (foundation)

through second grade. The authors compared the average performance of 328 ELs with 133 mainstream students, all from the same 15 low-SES schools, using a set of researcher-developed mathematics tests modeled on previous Australian national and state tests. Analysis of pre-tests and post-tests showed that EL students displayed greater gains than the mainstream students, and ELs who participated in RoleM were meeting grade-level expectations. Taken together, I interpret these studies as proof in concept that research-based curricula can boost the achievement of ELs.

LMR shares similarities with the empirically supported interventions for ELs in the literature but it also differs in important ways. Rich theoretical perspectives on teaching and learning guide all of the studies. LMR and RoleM both share an emphasis on progressive coordination of representations to support mathematical development. However, the key difference is that LMR is a relatively brief lesson sequence of 19 lessons whereas the empirically supported math interventions for ELs in the literature operated for a year or more. Additionally, LMR has empirical support from a quasi-experimental study, which is, in theory, more robust to threats to validity than the sampling designs in the literature I cited. Finally, the use of two outcome measures, both of which have been previously validated, makes this a methodologically robust contribution to the research.

4.2.4 The current study

Measuring the efficacy and equity of LMR

The purpose of the current analysis is to investigate whether LMR is effective and equitable for students classified as ELs. I evaluated LMR using data from both the LMR assessment and a standardized test in mathematics. The conclusions will involve a synthesis of the results from both measures. I will consider LMR effective if the results show that, on average, ELs who participated in LMR gained more in achievement than ELs who participated in the comparison group.

I also evaluated the equity of LMR. Equity can be understood as the narrowing or elimination of achievement gaps between groups of students defined along ethnic/racial or language proficiency boundaries (Gutiérrez, 2008). Many studies adopt a weaker criterion of equity, considering a curriculum that promotes equal attainment between groups (i.e., maintaining but not exacerbating preexisting achievement gaps) as equitable (e.g., Curtis et al., 2009). To accommodate both of these perspectives, I consider a statistically significant narrowing of achievement gaps as evidence that LMR is a strongly equitable curriculum for ELs. I further interpret evidence that LMR supported equal growth for ELs and their native English-speaking peers as an educationally important, but weaker form of equity. This general framework is in-line with other evaluation research with ELs (e.g., Thomas & Collier, 2001), although the parsing into weak and strong equity is, to the best of my knowledge, a new addition to the literature.

Research Questions

This study addresses the following research questions:

1. What is the impact of LMR on the achievement of ELs compared to the achievement of ELs who participate in the regular math curriculum?

2. How equitable is the impact of LMR between groups of students classified as ELs and English proficient students?
3. Does participation in LMR narrow achievement gaps between ELs and EP students?

4.3 Method

This study is an extension of the study by Saxe et al. (2013). I utilized the same participants and procedures described in the original article. Saxe and colleagues' evaluated the impact of LMR using data from the LMR assessment as the outcome. The authors also disaggregated data for prior achievement levels, and produced different subscales for number line and non-number line problems as well as integers and fractions achievement. In this paper, I conduct a related analysis by disaggregating Saxe et al.'s impact evaluation for students classified as ELs. I replicate their measurement model on the data and develop a statistical model to analyze data from the LMR assessment that results in a modified version of Model 3 in their paper[1]. In addition, this article contributes new data from a state-administered standardized achievement test in mathematics and a new statistical model for pre/post analysis. The combined data permit us to conduct a rigorous impact evaluation of LMR supported by evidence from two measures of mathematics achievement.

4.3.1 Participants and experimental design

The participants were 571 students from three urban and suburban school districts in the San Francisco Bay Area. Fourth and fifth grade teachers using the same math curriculum (Everyday Mathematics) were matched on background indicators and then assigned to either the LMR experimental group (n = 11) or the Comparison group (n = 8 with 10 classrooms). The sample contained 315 fourth and 256 fifth grade students of 19 teachers with 21 classrooms. Eleven teachers (11 classrooms and 292 students) were assigned to the LMR group to teach the LMR curriculum. Eight teachers (10 classrooms and 279 students) were assigned to the Comparison group who received *Everyday Mathematics*. LMR class size ranged from 15 to 32 students, with an average class size of 26.5 (SD = 4.7). Comparison class size ranged from 20 to 32 students, with an average class size of 27.6 (SD = 4.0). Readers can find additional details about sampling, experimental design, and fidelity of implementation in Saxe et al. (2013).

Student background

The students in this study reflect the ethnic/racial and socioeconomic diversity of the school districts in which this study was conducted. School officials provided demographics information for 86.2% of the participants, from which 38.2% of the students identified as White, 21.7% identified as African American, 18.7% identified as

[1] Model 3 is a longitudinal growth model that estimates the difference between students who participated in LMR and students in a matched Comparison group. The model accounts for different levels of prior ability in mathematics.

Asian or Pacific Islander, 11.8% identified as Latina/o, and 9.6% of the sample identified as another ethnicity. Information release forms could not be located for one classroom assigned to the comparison group (5.8% of the total sample) and therefore school officials were unwilling to release demographics information for those students.
The remaining missing data (8% of the sample) was missing from school district databases.

English learners (ELs). The sample contained 95 students classified as ELs (16.6% of the sample). All school districts classified participants into four categories of English language proficiency: English learner, initially fluent English proficient, reclassified fluent English proficient, and English only. I collapsed the latter three categories into a single category of English proficient (EP) students. Thus, the sample of EP students included individuals formerly classified as ELs who were reclassified when they met California standards for English language proficiency (cf., Saunders & Marcelletti, 2012 for a discussion of known measurement issues). The results section presents descriptive statistics on the sample of ELs.

4.3.2 Procedure

Detailed information about implementation of LMR is available in Saxe et al. (2013). Briefly, LMR teachers implemented nine lessons on integers followed by ten lessons on fractions. LMR teachers received professional development consisting of a three-day institute prior to implementation followed by four evening meetings during implementation. Two of the evening meetings focused on integers and two on fractions. Teachers in both LMR and comparison groups were required to cover the content of Everyday Mathematics according to their district pacing guide. Saxe and colleagues documented that teachers implemented LMR (via review of videos, analysis of self-report teacher surveys, and review of student work).

Assessments and data collection

Students in all classrooms completed a set of researcher-developed mathematics assessments named the LMR Assessment and a state-administered standardized test in mathematics. In this section, I briefly describe the LMR assessment and the standardized test.

LMR Assessment. Saxe and colleagues (2012) developed a paper-pencil student assessment to investigate the effectiveness of LMR in supporting growth in student achievement. They adapted items from multiple sources including the LMR curriculum, the Everyday Mathematics curriculum, and other sources such as released items from the NAEP and California's statewide testing programs. The items focused on two mathematics content domains: integers (positive integers, negative integers) and fractions (fractions of various values, equivalent fractions, ordering fractions, benchmark fractions). In each domain, item formats balanced number lines with formats that did not include number lines (e.g., numbers only or area models). Saxe and colleagues selected items from different sources and balanced item formats to guard against the threat to the validity of treatment group comparisons posed by items that too closely resembled the

material in the LMR curriculum.

Saxe et al. (2012) constructed a set of four assessments to gauge student progress at key points in the study. The authors used a set of 18 common items in all assessments to link student scores from the different time points using item response models. Then, for each assessment, the developers included an additional 11–14 items to optimize the match between the test and the students' learning opportunities. The additional items were easier at pretest and harder at final test, and the content emphasis shifted over time from integers to fractions in step with the curricular sequence. The composition of items in each assessment addressed potential validity concerns with floor, ceiling, and practice effects. The result was a set of three linked written assessments of 29–32 questions each. The post-unit assessment and end-of-year assessment was identical.

Standardized test. I measured student performance in grade-level mathematics achievement with scores from the mathematics portion of the California Standards Test in mathematics (CST; Educational Testing Service, 2011). District officials provided end-of-year standardized mathematics test scores from the students' current grade as well as the standardized mathematics test scores from prior year. The statistical models included scores from the prior year to control for prior mathematics achievement.

Analyzing invariance between grades. The sample of participants contained 4th and 5th grade students who completed different pairs of standardized tests. Students in 4th grade during the study received the 3rd grade standardized test in the prior year and the 4th grade standardized test at the end of the year. Students in 5th grade during the study received the 4th grade test in the prior year and the 5th grade test at the end of the year.

An analysis that treats two sets of pre/post tests as a single set is potentially problematic because the measurement model used by the state generated scaled scores with an item response theory model that does not allow direct comparisons of scores between grades. Although the statistical model used in this paper does not require the pre and posttests to be on the same scale, it may not be robust to differences in the pre to post differences between the two grade bands. In other words, the average pre to post difference should be similar for the two grade bands. It is possible that an interaction between grade level and mean pre to post difference could lead to artifacts in the data and threaten the validity of the main analysis.

I conducted testing to determine whether it would be reasonable to treat the grade levels (i.e., measurement scales represented by the respective pre- and posttests) as equivalent for the statistical analysis. I investigated whether the mean difference from 3rd to 4th grade was the same as the mean difference from 4th to 5th grade. Table 4-2 contains the means and standard deviations for 4th grade and 5th grade in the prior year and end-of-year tests.

The means of all four tests are within five points, which is a relatively insignificant difference given the range (200-600) and standard deviations of the tests (78 to 92) points. I determined that the mean grade-level difference in the difference between prior year and end-of-year was negligible. However, the SDs showed some (~10-15%) differences. To investigate the impact of different SDs between tests, I compared model results using standardized scores with model results using unstandardized scores. I standardized scores by subtracting the state mean and dividing by the state SD for each

grade and year. Then, I fit the final statistical model to the standardized data. I compared the standardized and unstandardized results and found no statistical or practical differences between the models. Therefore, I determined that I could treat the scales as equivalent and retained the unstandardized values for the analysis to maintain interpretability of the coefficients as points on the test.

Missing data

Missing data arose because of practical difficulties obtaining the data and the longitudinal nature of the study. I could not locate release forms for one classroom assigned to the comparison group (5.8% of the sample) and therefore I did not obtain demographics or standardized test data for those students. In addition, some students entered or left the classroom during the year. For some new arrivals to the district, prior-year standardized test data were unavailable. Finally, a small number of students were absent during administration of LMR assessment and on corresponding make up dates.

Cases with complete data. For the LMR assessment, I collected complete assessment data and demographic information including EL status for 77.6% of LMR and 73.7% of comparison participants. Complete LMR assessment data included scores at four time points for the LMR group and three time points for the comparison group. For the CST, I collected complete assessment data from the prior year and end-of-year tests, and demographic information including EL status, for 73.5% of LMR and 71.0 % of comparison students. I obtained a *full set* of data (complete LMR data, complete standardized test data, and demographic information including EL status) from 65.0% of the LMR students and 66.4% of the comparison students.

Within the sample of ELs, I collected complete LMR assessment data for 88.6% of LMR and 92.2% of comparison participants. I obtained complete standardized test data for 82.3% of LMR ELs and 72.7 % of comparison ELs. Finally, 68.2% of the ELs in LMR and 76.5% of the ELs in the comparison group supplied a full set of data.

Investigating the effect of missing data. I investigated the potential effect of missing data on the results. I fit the final statistical models on both the entire sample and the subsample of participants who had a full set of data. I also fit the final models to the ELs only, for both the entire sample of ELs and the subsample of ELs who had a full set of data. The results were similar between models, and, where they differed, the differences were small and did not change statistical significance. Based on this situation, I can state that the participants with partial data did not influence the outcome of the evaluation. Further, the unbalanced design with more numerous EP students than ELs did not impact the statistical significance of estimated parameters for the ELs in the sample. I therefore report the results from the entire sample because these results utilize all available information.

4.3.3 Multilevel analysis

The student-level mathematics achievement data are considered hierarchical, or nested, because students existed within classrooms and I collected data on both on student-level variables (whether or not the student is an English language learner) and at the

classroom-level (e.g., whether the classroom is an LMR or a comparison classroom). It is not appropriate to use ordinary linear regression because I cannot assume that students within the same classroom are independent of one another. The regression would yield incorrect standard errors and *p* values that were too low (Rabe-Hesketh & Skrondal, 2012). Thus, I used multilevel models, also known as hierarchical linear models (HLM), to analyze these nested data (Raudenbush & Bryk, 2002). I did not include school-level or district-level variables because there were few teachers per school and only three districts in total. I lacked the statistical power to include grade level effects in the model. I applied a 5% level of significance for all analyses.

Measurement model for the LMR assessment

The measurement model used the item response data to estimate the mathematics ability of each student at each time point. I replicated the measurement model from Saxe et al (2012). The LMR assessment data led to a nested (assessments within students within classes), unbalanced (classes of different sizes; students with different numbers of assessments), and linked (assessments with different items at different times) data structure, necessitating the use of a hierarchical item response model for estimating student achievement. I fit the measurement model using ACER ConQuest 4 (Wu, Adams, & Wilson, 2015). I estimated student achievement at each time point using a longitudinal Rasch model (Andersen, 1985; Pastor & Beretvas, 2006; Wilson, Zheng, & McGuire, 2012) parameterized as a special case of the multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997). I used concurrent estimation for all time points and linking items were restricted to the same difficulty between time points. I used weighted likelihood estimates (WLEs; Warm, 1989) to estimate each student's achievement at each time point.

Statistical model for the LMR Assessment (Model 1)

I used a longitudinal growth model to estimate the changes in student achievement during the study. The longitudinal growth model in this paper is similar to Model 3 in the Appendix of Saxe et al. (2013). I modified their model by (a) adding an indicator variable for EL status, (b) removing two indicator variables for achievement level and (c) adding three-way cross-level interactions between LMR/comparison group membership, EL status, and time. To ensure that the sample size of ELs was adequate for HLM, I compared the results of models fit to the entire sample of participants (n = 572) and on the sample of ELs only (n = 95). I observed no changes in statistical significance and the parameters were nearly identical.

I estimated changes in student achievement over time using a longitudinal growth model specified as a three-level HLM where time was nested within students, nested within classrooms. Each model included a classroom-level random intercept (Level-3) and a student-level random intercept (Level-2). The response variable was the student WLEs from the measurement models. The model compared changes in achievement over time between students classified as ELs with native English speaking students. The model included covariates for test occasion, group assignment, EL status, and their 2- and 3-way interactions. Linear combinations of the interaction coefficients provided the estimated differences in change over time between ELs and native English speakers.

I used *Stata13* (StataCorp. 2013) to estimate the longitudinal growth model (Rabe-Hesketh & Skrondal, 2012). I estimated parameters by maximum likelihood estimation using the "mixed" command. I used the *Stata* command "lincom" to obtain simple effect comparisons that resulted in estimates of student achievement at each time point.

I specified Model 1 in three levels as follows:

Level 1:

$$wle_{tsc} = \pi_{0sc} + \pi_{1sc} inter_{tsc} + \pi_{2sc} post_{tsc} + \pi_{3sc} final_{tsc} + \epsilon_{tsc}$$

Level 2:

$$\pi_{0sc} = \beta_{p0c} + \beta_{p1c} proficient_{sc} + \zeta_{psc} \quad \text{for} \quad p = 0, 2, 3$$

$$\pi_{1sc} = \beta_{10c} + \beta_{11c} proficient_{sc}$$

Level 3:

$$\beta_{00c} = \gamma_{000} + \gamma_{001} comp_c + \zeta_c$$

$$\beta_{01c} = \gamma_{010} + \gamma_{011} comp_c$$

$$\beta_{1qc} = \gamma_{1q0} \quad \text{for} \quad q = 0, 1$$

$$\beta_{pqc} = \gamma_{pq0} + \gamma_{pq1} comp_c,$$

for $p = 2, 3 \ and \ q = 0, 1,$ and where wle_{tsc} is the WLE estimate of student achievement for student *s* in classroom *c* at time *t*. The indicator variables for test occasion are $inter_{tsc}$, $post_{tsc}$, and $final_{tsc}$. $proficient_{sc}$ is an indicator variable for a student being classified as English proficient. $comp_c$ is an indicator for classroom *c* belonging to the comparison group. ζ_c is a random classroom-level effect assumed to follow a normal distribution. ζ_{psc} is a student-level random effect assumed to follow a normal distribution. ε_{tsc} is a time-specific error also assumed to follow a normal distribution.

Statistical model for the standardized achievement test (Model 2).

I analyzed participants' scores on the California Standards Test (CST) in mathematics administered in the year 2010, which I refer to as the *prior year* test, and 2011, which I refer to as the *end-of-year* test. The correlation between prior year scores and end of year scores was $r = 0.78$.

I used HLM to predict students' the end-of-year scores on the standardized achievement test. I used HLM because the data are nested and I cannot assume that students within the same classroom are independent observations. I used *Stata13* and the "mixed" command with mle. I used the "lincom" command for contrasts, and the "margins" command to obtain adjusted means and standard errors for the groups of

students that differed by English proficiency and assignment to the LMR or the comparison group.

Mean centering of prior year scores. For interpretive purposes, I centered prior-year standardized test scores to have a mean of zero. I subtracted the prior-year grand mean from each student's prior-year test score. By centering the pretest data, I obtain adjusted posttest means from the model, which I interpret as the estimated posttest means for students with mean pretest scores.

The design results in a two level hierarchical structure with students nested in classrooms. I specified Model 2 as follows:

Level 1:

$$end_{sc} = \pi_{0c} + \pi_{1c}(prior_{sc} - prior_{..}) + \pi_{2c}(prior_{sc} - prior_{..})^2 + \pi_{3c} proficient + \epsilon_{sc}$$

Level 2:

$$\pi_{0c} = \gamma_{00} + \gamma_{01} LMR + \zeta_c$$

$$\pi_{pc} = \gamma_{p0} \quad \text{for} \quad p = 1, 2$$

$$\pi_{3c} = \gamma_{p0} + \gamma_{p1} LMR \quad \text{for} \quad p = 3$$

I modeled the expected end-of-year standardized test score for student (*s*) with teacher (*t*). I parameterize the model to set the reference group (i.e., the model intercept) to ELs in the comparison group. The student variables at Level 1 are as follows:

1. end_{sc}. End-of-year test score for each student.
2. $prior_{sc}$. The prior year test score for each student.
3. $prior_{sc}^2$. The square of each student's prior year test score, which relaxes the linear assumption of the model.
4. $proficient_{sc}$. An indicator variable for whether or not the student is classified as English proficient.
5. $prior_{..}$ is not a model parameter, it is the grand mean of the prior year test score.

The classroom-level indicator variable at Level 2 is as follows:

6. LMR_c. Indicator variable for whether the classroom participated in LMR or not

The model included one cross-level indicator variable:

7. $proficient_{sc}$*LMR_c. Is an indicator variable for the EL-by-treatment group interaction. I constructed this variable by multiplying the $proficient_{sc}$ variable by the treatment group variable LMR_c.

Marginal effects for estimating achievement gaps on the standardized test. I estimated achievement gaps as the marginal fixed effects of the HLM using the post-estimation "margins" command in *Stata13*. Marginal effects provide an intuitive way to interpret the practical significance of LMR's impact on achievement gaps. Analysis using

marginal effects is in line with other reports in the literature (e.g., Aud, Fox & Kewal-Ramani, 2010; Hemphill & Vanneman, 2011).

I predicted marginal effects for prior year scores and end-of-year scores using two separate models, Model 3 and Model 4 respectively. Both are two level HLMs with a classroom-level random intercept. Model 3 predicted students' prior year scores and included two indicator variable covariates: assignment to treatment group and English learner classification. The model for end-of-year data, Model 4 predicted the end-of-year scores with the same two covariates. The Model 4 is identical to Model 2 without the two covariates that controlled for prior achievement: prior year score and the square of prior year score.

Comparing HLM with OLS (results not reported in this paper). Concerns about low numbers of ELs within each classroom (Level 2 cluster) prompted us to compare the HLM results with results from an OLS regression that modeled variance only at the individual level instead of the classroom level as well. I fit the OLS regression in *Stata13* with cluster-specific robust standard errors that relaxed the assumption of independence of the observations (observations were assumed independent between clusters). I fit the OLS models using the entire sample and the sample of ELs only.

The estimates for ELs were very similar across all model comparisons: OLS on the full sample vs. HLM on the full sample; OLS on the sample of ELs vs. HLM on the full sample; and OLS on the full sample vs. OLS on the ELs. Statistical significance was not impacted.

The model, Model 5, is specified below.

$$end_i = \alpha + \beta_1(prior_i - \overline{prior}) + \beta_2(prior_i - \overline{prior})^2 + \beta_3 LMR_i + \beta_4 proficient_i +$$

$$\beta_5 LMR_i * proficient_i + \epsilon_i.$$

The OLS model consisted of five main predictor variables, the mean-centered prior year test scores (β_1) the square of the mean-centered prior year test scores (β_2), plus three additional predictors:

1. β_3. Treatment group indicator variable for being in the LMR group.
2. β_4. English proficiency indicator variable for being classified as English proficient.
3. β_5. Treatment-by-proficiency interaction. I constructed this indicator variable by multiplying the treatment group indicator by the English proficiency indicator.

For this model, α is the expected score for comparison students who are English learners. β_1 is the linear change in post-intervention CST score when grand mean centered pre-intervention CST score increases by one point. β_2 is the quadratic change in mean of y_i when $(x_i - \overline{x})$ increases by a unit. β_3 is the expected difference between LMR and comparison students who are English learners. β_4 is the expected difference between ELs and English proficient students in the comparison group. β_5 is the LMR by native English interaction interpreted as the difference in "effect" of LMR for English proficient students in the LMR group relative to the reference (ELs in the comparison) group.

Calculating effect size statistics

If an indicator variable predictor was statistically significant, then I calculated an effect size statistic (*ES*). I calculated the *ES* for the LMR assessment model by dividing the estimated regression coefficient by 1.33, the *SD* for the sample at pretest. For the standardized test, I divided the estimated regression coefficient by 75[2]

4.4 Results

I divided the results into three sections. The first section contains a descriptive analysis of the sample and assessment data. I analyzed key variables such as ethnicity, gender, grade level, and EL status, including the distribution of ELs across classrooms. I also presented a descriptive analysis of the data from the LMR assessment and from the standardized test. The remaining sections contain the multilevel analyses that address the research questions about the efficacy (Research Question 1) and equity (Research Questions 2 and 3) of LMR. I interpreted the results from LMR assessment and standardized test separately. Section two contains the analysis of the LMR assessment data using an HLM for longitudinal growth. In section three I analyzed the standardized test data using a random intercept HLM. The discussion contains a synthesis of the research evidence.

Descriptive Analysis

This section contains the following descriptive analyses:

1. Demographic information on participants' ethnicity, gender, grade level, and EL status
2. Demographics for only the participants classified as ELs
3. Distribution of ELs across classrooms
4. Participants' scores on the LMR Assessment and the standardized test

Student demographics. Table 4-3 reports participants' grade level, sex, ethnicity, and language proficiency separated by group assignment. The participants were ethnically diverse and the sample contained 16.6% ELs (n = 95). I tested the two-way associations between each demographic variable and group assignment using chi-squared tests that ignored missing data. The LMR and comparison groups were balanced for grade level, gender, and language proficiency. There was statistically significant difference in the ethnicity of the LMR and comparison groups, $\chi^2(4) = 12.48$, $p = 0.014$. The largest difference between expected and observed frequencies was in the number of African American and Asian/Pacific Islander (API) students. Students who identified as African American were more numerous in the LMR group, and students who identified as API were more numerous in the Comparison group.

---

[2] This value is a proxy for the within sample pooled *SD* across the four CST administrations that would, in theory, provide a more accurate effect size. The proxy value comes from a review of CST technical manuals from 2011-2013. I have the *SD*s from the technical manual and the *n*s are buried in, but extractable from, a publically available dataset.

Demographics of ELs. I inspected the demographic information for the ELs in the sample for imbalances that could threaten the internal validity of the evaluation (Table 4-4). Forty-four ELs participated in LMR classrooms and 51 ELs participated in comparison classrooms. This difference was not statistically significant. Chi-squared tests showed no statistical difference between LMR and comparison ELs for sex or for ethnicity. However, it is worth noting that the LMR group contained half the number (13) of Latino ELs as the comparison group (26). A chi-squared test showed there was a statistical association between grade level and assignment to treatment, $\chi^2(1) = 9.89$, $p = 0.002$. Additionally, Table 4-4 shows that LMR group contained more EL students in 4th grade, and the comparison group contained more EL students in 5th grade.

Distribution of ELs across classrooms. All classrooms contained participants classified as ELs (Table 4-5). With the exception of LMR classroom number 10 for which I was not able to obtain data, the percentage of ELs within each class ranged from 6.2% to 45.5%, translating into a range of 2 – 10 participants.

Student achievement. Table 4-6 contains descriptive statistics for LMR assessment scores and standardized test scores. I included cases with missing data to use all available information as I did for the final statistical models. Mean scores on the LMR assessment increased at each time point. For the calculation of change on the LMR assessment, I subtracted each participant's pretest score from their final test score. Students who participated in LMR, whether classified as EL or EP, gained more in achievement than students who participated in the comparison group. EL students who participated in LMR gained, on average, 0.05 logits less during the school year than EP students who participated in LMR.

The bottom section of Table 4-6 contains descriptive statistics for the standardized test in mathematics. On average, students who participated in LMR scored higher on the end-of-year test than they did on the prior year test. The ELs who participated in LMR gained more in achievement than the EP students who participated in LMR. In contrast, students who participated in the comparison group scored lower on the end-of-year test than they did on the prior-year test.

4.4.1 Efficacy of LMR

The first research question asked whether LMR was effective for participants classified as ELs. I compared the achievement of ELs who participated in LMR with the achievement of ELs in in the comparison group. I used a longitudinal growth model to estimate student achievement on the LMR assessment for the two groups at each of four time points. I tested the statistical significance of the differences in achievement between the groups at each time point using linear combinations of regression coefficients. I was especially interested in the outcome at the final test because this contrast measured the durability of LMR's efficacy months after the intervention concluded.

Figure 4-1 contains a graph of the mean logit scores on the LMR assessment for LMR students classified as ELs and comparison students classified as ELs. The figure shows two patterns of growth for the groups of students. The ELs who participated in LMR, represented by the black line, showed steady growth from the pretest to the posttest phase in which LMR was implemented. Although the slope of the line was free to vary at the interim test, it remained consistent. The black line shows that the ELs in LMR grew less

during the 5-month gap between posttest and final test in which they received the regular mathematics curriculum (*Everyday Mathematics*).

In contrast, the ELs who participated in the comparison group represented by the grey line showed less growth from pretest to posttest and sharper growth from posttest to final test. The comparison group did not receive the interim test; thus the slope was not free to vary at this point. These growth patterns are identical to those shown in Saxe et al. (2012), who explained the patterns as differences in the opportunities to learn the assessment topics of integers and fractions. The LMR sequence covered the assessment topics in the fall, and the *Everyday Mathematics* curriculum covered these topics in the spring.

Table 4-7 contains the results of the multilevel analysis for the LMR assessment data. There was no statistical difference between the groups at pretest ($p = 0.736$). However, on the posttest, the predicted achievement of ELs who participated in LMR was 1.52 logits higher than the predicted achievement of ELs assigned to the comparison group ($ES = 1.14$, $p < 0.001$). Results from the final test, administered five months after LMR lessons concluded, estimated that the average achievement of LMR ELs was 0.90 logits higher than the average achievement of comparison ELs' ($ES = 0.68$, $p = 0.011$). In sum, ELs in LMR gained significantly more in achievement than ELs in the comparison group on the LMR assessment. The evidence indicates that LMR provided effective instruction in integers and fractions for students classified as ELs.

Efficacy of LMR measured by the standardized test in mathematics

Next, I investigated whether LMR supported ELs' growth on a standardized test as it did on the LMR assessment. I used scores from a state-administered standardized test that measured grade-level mathematics achievement. I predicted students' end-of-year scores, controlling for prior mathematics achievement using students' prior year scores. I interpreted the model results as the value added from participating in LMR, controlling for prior achievement.

Table 4-8 contains results from the multilevel model that address the efficacy of LMR for participants classified as ELs (Research Question 1). The expected end-of-year score for ELs in LMR was 37.01 points higher than the expected end-of-year score for ELs in the Comparison group, after controlling for prior achievement ($p = 0.010$, $ES = 0.49$). The results suggest that LMR added statistically significant value to ELs achievement in grade level mathematics. Furthermore, the predicted gains were educationally significant. The expected score for EL students in LMR fell in the *advanced* score range set by the state of California, whereas the expected score for the EL students in the comparison group fell one category below in the *proficient* score range. The results suggest that LMR could serve as a tool to address equity concerns of direct interest to policymakers.

4.4.2 Equity of LMR

Equity of LMR as measured by the LMR assessment

Research Question 2 asked whether LMR supported similar rates of learning for both ELs and EP students. I compared the average growth of ELs who participated in LMR with the average growth of EP students who participated in LMR. Figure 4-2 graphs the growth on the LMR assessment for ELs represented by the black line and EP students

represented by the grey line. The growth trajectories were similar between groups. The final model fit to subsamples of the data (ELs only, participants with full data only) produced identical trajectories.

Table 4-9 contains the estimated achievement for ELs who participated in LMR and EP students who participated in LMR. A statistically significant achievement gap existed at pretest ($ES = 0.47$, $p = 0.002$) and persisted across time, ranging from 0.45 to 0.58 *SD*s. The bottom row of the table shows that the difference between ELs and EP students' gains was not statistically significant ($p = 0.937$). Therefore the results show that ELs' achievement in integers and fractions grew at a rate similar to EP students'. I interpreted the results as evidence that LMR was equally effective for teaching integers and fractions to both groups of students.

Equity of LMR as measured by the standardized test

I asked whether LMR also supported equal gains for ELs and EP students in grade level mathematics. I compared the estimate of LMR's value added for ELs with the estimate of LMR's value added for EP students. Table 4-10 contains these estimates, which are derived from linear combinations of regression coefficients in Table 4-8. The 11.61-point estimated value added of LMR for EP students was not statistically significant ($p = 0.216$). Further, the estimated mean difference between ELs and EP students (i.e., LMR's value added for ELs relative to EP students) of 25.4 points was not statistically significant ($p = 0.060$, $ES = 0.34$).

The results lead to a complicated interpretation. Some evidence supports the ability of LMR to narrow achievement gaps. Participation in LMR was associated with a statistically significant 37.01-point improvement in standardized test scores for ELs, whereas the 11.61-point improvement for EP students was not statistically significant. This piece of evidence suggests that LMR is equitable for ELs in the sense that it provided more value for ELs than for EP students. However, other results are more ambiguous. The estimated difference between the value added for ELs and value added for EP students (25.4 points) was not statistically significant. This piece of information is cause for cautious interpretation of the results.

In addition, I cannot rule out the influence of ceiling effects for EP students as a rival hypothesis. Initial student achievement, which was higher for EP students, was negatively correlated with the change from prior year to end of year ($r = -0.32$, $p < 0.001$). It may be that the relatively low bar for mathematics achievement set by many state-administered standardized tests (Fuller, Gesicki, Kang, & Wright, 2006) suppressed the average gain of the relatively higher achieving EP student groups. Taken together, the results tentatively suggest that LMR was more effective for increasing the general achievement of ELs than for EP students. However, given that a nontrivial effect size of 0.34 *SD*s was not statistically significant, I expect that replicating the study with a larger sample of ELs would provide useful increases in statistical power.

4.4.3 The impact of LMR on achievement gaps

Assessment of integers and fractions knowledge.

Research Question 3 asked whether participation in LMR narrowed the preexisting achievement gaps between ELs and EP students. The relevant comparison was between ELs who participated in LMR and EP students in the comparison group. Following these groups longitudinally showed the achievement gap at pretest and tracked the change in the achievement gap over the course of the year.

Figure 4-3 contains a graph of the mean logit scores on the LMR assessment for LMR ELs represented in black and comparison EP students represented in grey. I interpret the graph as the expected change in achievement for the groups of students over the school year. The results show a pretest achievement gap, with EP students demonstrating greater achievement then ELs. Recall that the pretest occurred before the LMR intervention, so group assignment is irrelevant at pretest.

During the first part of the academic year, the achievement of ELs in LMR grew sharply compared to EP students in the comparison group. The expected achievement gap narrowed and equaled zero around the interim test. The gap reversed when ELs had greater average expected achievement than EP students. The reverse gap widened until the post-test. Then, it narrowed between the posttest and final test. The specific change trajectories are partially artifacts of the method (i.e., the trajectory can only change at a point of measurement). However, the patterns of change were consistent with the classroom experiences of students (e.g., the opportunities to learn integers and fractions earlier on for LMR students and later in the year for comparison students).

Table 4-11 contains the statistics represented in the graph that show the change in achievement gaps. At pretest, there was a statistically significant achievement gap between ELs who participated in LMR and EP students who participated in the comparison group of 0.79 logits ($SE = 0.29$) in favor of the EP students ($p = 0.005$). The effect size of 0.60 *SD*[3] is similar to the 13-year average achievement gap on the 4th grade NAEP mathematics test of 0.58 *SD*. The change between the pretest and the final test is the important comparison because *Everyday Mathematics* curriculum introduced relevant concepts late in the year.

The results show that the achievement gaps narrowed during the school year. Over the year, ELs in LMR gained more in integers and fractions achievement than comparison EP students. ELs in LMR gained 2.10 logits ($SE = 0.14$) whereas EP students in the comparison group gained 1.15 logits ($SE = 0.07$). ELs in LMR gained 0.95 logits more than EP students in the comparison group ($SE = 0.16$). This difference was highly statistically significant ($p < 0.001$) and corresponded to an effect size of 0.72. As a result of the larger gain for LMR students, the final test administered at the end of the school year showed no statistical difference between the achievement of ELs in LMR and the achievement of EP students in the comparison group ($p = 0.608$).

The evidence suggests that LMR eliminated the achievement gap in integers and fractions achievement between ELs who received LMR and the general population of EP students. As mathematics educators consider integers and fractions foundational for

[3] The *ES* of 0.6 is based on an assumed SD of 75. A more precise estimate may be calculated through analysis of primary data available for download from the state of California.

learning higher mathematics, the implication is that LMR could be highly effective as a targeted intervention for ELs.

Standardized test in math.

Next I asked whether LMR narrowed achievement gaps in grade-level mathematics achievement. Again, I compared ELs who participated in LMR with EP students who participated in the comparison group. I estimated the achievement gaps using the marginal fixed effects from two separate HLMs that predicted prior year test scores and end-of-year test scores, respectively. HLM estimated the baseline achievement gap between all ELs and all EP students in the sample (before assignment to treatment) was 47.3 points ($p < 0.001$), which corresponds to a prior year achievement gap of $ES = 0.63$. This gap is congruent with the mathematics achievement gaps for ELs reported in Hemphill & Vanneman (2011).

Figure 4-4 contains a graph that represents the change in achievement gaps. ELs are represented in black and EP students are represented in grey. The results from the prior year show that, within language status, the LMR and comparison groups were roughly equivalent. The ELs who participated in LMR showed the most dramatic change from prior year to end-of-year. The positive slope suggests that participating in LMR narrowed the achievement gap between ELs and EP students. I also saw the achievement gap between ELs in the comparison group and all EP students widen over the year. The results suggest that ELs in the comparison group fell further behind their EP peers during the year.

Table 4-12 contains the statistics represented in Figure 4-4. First, I assessed a threat to the internal validity of the study posed by groups of students (EL and EP) with different levels of average achievement at pretest. For example, evidence that the group of ELs assigned to LMR had different levels of initial achievement than the group of ELs assigned to the comparison condition would lead to concerns about the homogeneity of the sample and the reliability of the estimates. I subtracted the estimated prior year score for ELs who participated in LMR from the estimated prior year for ELs who participated in the comparison group. The difference was 3.28 points, which was not statistically significant ($p = 0.892$). Likewise, the prior year difference between EP students who participated in LMR and EP students who participated in the comparison group was 1.71 points and not statistically significant ($p = 0.918$). These results suggest that the treatment and control groups had equivalent math achievement at the outset of the study.

Second, I estimated the baseline achievement gap between ELs and EP students without the influence of LMR. The estimated achievement gap between ELs who participated in the comparison group and EP students who participated in the comparison group was 49.27 points. This was statistically significant ($p < 0.001$) and corresponds to an achievement gap of 0.66 *SD*s.

Third, I assessed the educational impact of LMR on the achievement gaps between ELs and EP students. I compared the estimated score for ELs who participated in LMR to the average estimated score for EP students who participated in the comparison group (i.e., treatment as usual). The estimated gap was 18.00 points, which is not statistically significant ($p = 0.349$).

Finally, I interpreted the difference between the baseline and the post intervention achievement gaps. The results suggest that participation in LMR shrunk the mathematics

achievement gap between ELs and EP students from 49.27 to 18.00 points, a 63.5 percent reduction. Whereas the achievement gap was highly statistically significant at pretest, it was not after the LMR intervention. The implication is that LMR may help ELs catch up to their native English-speaking peers in grade-level math achievement. Notably, A handful of studies have reported similarly dramatic gains on ELs math achievement following research-based mathematics intervention (Doty et al., 1999; Fuson, et al., 1997; Lane et al., 1995; Miller & Warren, 2014; Silver & Stein, 1996). I explore plausible mechanisms in the discussion.

4.5 Discussion

In this article, I reported results from an experimental study that supported the efficacy and equity of the Learning Mathematics though Representations (LMR) curriculum for students classified as ELs. LMR is a lesson sequence designed to support upper elementary students understandings of integers and fractions. Saxe and colleagues (2012) showed that LMR was effective for teaching integers and fractions for students who enter with different levels of academic achievement. This study modified Saxe et al.'s analysis, disaggregating the results for students classified as ELs. I supplemented the analysis with new data from a standardized test in mathematics that allowed us to investigate whether LMR also supported growth in grade level mathematics achievement.

The theoretical analysis suggests that LMR could be an effective and equitable curricular intervention for students classified as ELs. In the introduction to this paper I posited that the affordances of LMR match the needs of ELs in the mathematics classroom. LMR contains curricular and pedagogical supports that are widely recommended for all students. LMR supports rich mathematical discussions in the context of problem solving activities. Lesson structures include varied formats and promote useful sociomathematical norms that, for example, encourage all students to participate in mathematical practices such as describing conjectures, explaining solutions, and proving conclusions.

Second, LMR provides a resource-rich environment for communication and problem solving that is particularly well suited to the needs of ELs. LMR provides visual, verbal, and sensorimotor supports for mathematical sensemaking that offer students varied opportunities to access the curriculum. LMR's uses the number line as a central representational context, publically displayed written definitions of mathematical concepts that represent shared, negotiated understandings of key mathematical ideas that serve multiple roles. For example, mathematical definitions are starting points for mathematical discussions as well as references for testing the validity of specific mathematical conjectures. LMR also incorporates the use of actions in mathematical activities on the number line representation. The visual, verbal, and physical actions co-regulate each other. For example, a student may iterate c-rods on the number line representation to empirically test the entailments of a mathematical definition cited as evidence by another student. The interaction of resources over time forms an especially rich environment for mathematical development that gives EL students ways to access the mathematics that are less dependent on mastery of the English language. A purpose of this study was to investigate whether evidence from student assessments implied that the theoretical benefits translated to real world outcomes for ELs.

Summary of Key Findings

This study analyzed data from Saxe et al.'s (2012) study of the efficacy of LMR. The research design sampled volunteer teachers in 4th and 5th grade from districts that used the *Everyday Mathematics* curriculum. Teachers used LMR as a supplemental curricular intervention during their regular instructional time reserved for mathematics. Saxe and colleagues developed a set of paper and pencil assessments of integers and fractions (the LMR assessment) that gathered data on student achievement at four time points throughout the school year: prior to LMR, in the middle of LMR, at the end of LMR, and at the end of the school year approximately five months after conclusion of LMR. For the current study, I gathered additional data from school districts: scores on participants' state-administered standardized test in mathematics from the prior year and the end-of-year.

Effectiveness of LMR for students classified as ELs. The results provided evidence that ELs who participated in LMR gained more in achievement than ELs who participated in the comparison group. The relatively brief intervention of 19 lessons led to gains in the average achievement of ELs on both the LMR assessment that measured achievement in integers and fractions and on the standardized test in mathematics that measured grade-level mathematics achievement. The end-of-year gains were large: 0.68 *SD*s on the LMR assessment and 0.49 *SD*s on the standardized test in mathematics. The combined evidence suggests that LMR was effective for ELs in multiple domains of mathematics.

Equity of LMR for students classified as ELs. The findings also documented that LMR met criteria for an equitable mathematics curriculum. I defined two types of equity: weak equity and strong equity. A weakly equitable curriculum would support equal gains in achievement for ELs and EPs. Indeed, the evidence from the LMR assessment suggests that LMR was effective for teaching integers and fractions to both groups of students who learned at the same rate (i.e., weakly equitable).

If equity is defined as the elimination of achievement gaps, a strongly equitable curriculum would help ELs would learn at a higher rate than EP students. The results from the standardized test in math provide evidence that LMR was strongly equitable for the general achievement outcome. The value added by LMR was greater for ELs than for EP students for grade-level mathematics achievement. Although the statistical significance was only marginal (p = 0.06) I interpret the difference of 0.34 SD as potentially educationally relevant. Additional research with larger samples of ELs is warranted.

Explaining the mechanism of action for LMR

It is easy to explain how participation in LMR was associated with gains in integers and fractions achievement. The LMR assessment measured precisely what LMR aims to teach. It is more difficult to explain how LMR may influence general mathematics achievement. LMR's impact on grade-level achievement might be transmitted through changes in the teacher or the student. Teachers may generalize the rich curricular and

pedagogical principles to other content areas in mathematics. Another possibility is that LMR invokes changes in the student that, for example, lead to differences in the ways that students approach math. This study raises additional research questions about LMR's mechanism of action for ELs that could be addressed through additional quantitative and qualitative and research.

Limitations

The sampling of ELs is a primary limitation of this study. Although LMR was designed to be inclusive, Saxe et al.'s (2012) research design was not optimized for studying the effects of LMR on ELs. This sampling limitation leads to suboptimal statistical power and inability to model grade level effects. Other measurement issues exist, most notably dichotomizing a heterogeneous population into EL and non-EL groups (Saunders & Marcelletti, 2012). Although the number of ELs was adequate for a robust evaluation, and the numbers were greater than similar evaluations in the literature (e.g., Lane, Silver, & Wang, 1995), the sample size was smaller than desirable as evidenced by relatively large standard errors for estimated effects. The measurement error theoretically leads to low statistical power to detect average differences between groups (i.e. Type-2 error). Additional research with a larger sample of ELs (e.g., Curtis, et al., 2009) would increase confidence in the results.

Contributions and Implications

This paper contributes to the empirical research that documents the effectiveness and equity of the Learning Mathematics through Representations (LMR) curricular intervention. I documented the potential of LMR to provide effective and equitable instruction for students classified as ELs. I evaluated LMR with data from two assessments and an outcome-based definition of educational equity. The results suggest that LMR supported ELs achievement in the focal areas of integers and fractions and in grade-level mathematics achievement. LMR significantly narrowed achievement gaps in mathematics between ELs and their native English-speaking peers. The empirical results are notable because LMR's core strategies, which research suggests are effective for all students (Saxe et al., 2013), align with the ELs' needs in the classroom. Thus, the results and theory align to support the ability of LMR to provide ELs access to the mathematics curriculum.

This study responds to a key need in the literature for empirical data on empirically supported strategies for ELs. The results are notable given the historically intractable achievement gaps between ELs and their native English-speaking peers. LMR may provide ELs the access to mathematics education from which they have traditionally been marginalized (Secada, 1996). This paper joins the small number of mathematics curricula with empirical support for ELs. Although a handful of studies suggest that long-term (six months or more), research-based mathematics interventions can have statistically significant and educationally relevant impact on students achievement (Fuson, et al., 1997; Lane et al., 1995; Miller & Warren, 2014; Silver & Stein, 1996), this study is, to my knowledge, the first of its kind to warrant the efficacy of a brief curricular unit for supporting similar gains in ELs.

Future research should document how the forms of participation in LMR influence the access that ELs have to the mathematics curriculum. Most of the research on Latino ELs has examined students engaged with traditional math problems, not in mathematical discussions and argumentation (Moschkovich, 1999). Thus, research on how the various forms of participation in the mathematics support or do not support ELs' development is needed.

Conclusion

Schools in the U.S. have struggled to support ELs development in mathematics, and it is widely believed that the long-standing achievement gaps will not narrow without intensive support (Ivory, Chaparro, & Ball, 1999; Khisty, 1995; Secada, 1996). This study documents the potential efficacy of LMR for populations of students classified as ELs. I presented a rigorous study of the efficacy and equity of LMR for students classified as ELs. The analysis revealed that LMR had a noteworthy impact on ELs' achievement. Participation in LMR narrowed achievement gaps between students classified as ELs and their English proficient peers. I introduced a theoretical argument that connects LMR's design features with increased access to the curriculum for ELs. This research should spread interest in the use of LMR in classrooms that contain ELs and should catalyze additional research into the features of LMR that influence ELs' access to the mathematics.

4.6 Tables and figures

Table 4-1

Learning Mathematics through Representations (LMR) Design Principles That Support Mathematical Teaching and Learning

1. A five-phase lesson structure blends whole-class, group, and individual instruction to support the active participation of all students.
2. The number line serves as a principal representational context for mathematical teaching and learning.
3. Mathematical definitions are classroom resources that support student argumentation, generalization, and problem solving.
4. Cuisenaire rods provide manipulative support for mathematical teaching, learning, and communication.
5. A set of classroom norms and routines supports students' productive use of multiple resources for mathematical communication and problem solving.

Table 4-2

Comparing the Means and Standard Deviations for the Different Forms of the Standardized Test in Mathematics

Grade Band	Prior Year		End-of-year	
	Mean	*SD*	Mean	*SD*
3rd to 4th grade	395	92	392	78
4th to 5th grade	390	79	393	92

Table 4-3

Student Demographics and Missing Data

	LMR (n = 294)		Comparison (n = 277)		Total (N = 571)	
	Pct	N	Pct	N	Pct	N
Grade						
4th	57.5	169	52.7	146	55.2	315
5th	42.5	125	47.3	131	44.8	256
Sex						
Female	47.3	139	44.0	122	45.7	261
Male	46.9	138	38.6	107	42.9	245
Missing	5.8	17	17.3	48	11.4	65
Ethnicity*						
African American	22.8	67	14.4	40	18.7	107
Asian/Pacific Islander	13.6	40	18.8	52	16.1	92
Latino	8.5	25	11.9	33	10.2	58
White	36.4	107	29.2	81	32.9	188
Other	10.2	30	6.1	17	8.2	47
Missing	8.5	25	19.5	54	13.8	79
Language Proficiency						
English Learner	15.0	44	18.4	51	16.6	95
English Proficient	76.5	225	62.8	174	69.9	399
Missing	8.5	25	18.8	52	13.5	77

Note. * Difference between LMR and Comparison groups is statistically significant.

Table 4-4

English Learners' Demographics and Missing Data

	LMR (n = 44)		Comparison (n = 51)		Total (N = 95)	
	Pct	N	Pct	N	Pct	N
Grade*						
4th	63.6	28	31.4	16	46.3	44
5th	36.4	16	68.6	35	53.7	51
Sex						
Female	40.9	18	49.0	25	45.3	43
Male	59.1	26	51.0	26	54.7	52
Ethnicity						
African American	9.1	4	3.9	2	6.3	6
Asian/Pacific Islander	40.9	18	31.4	16	35.8	34
Latino	29.5	13	51.0	26	41.1	39
White	9.1	4	9.8	5	9.5	9
Other	4.5	2	0.0	0	2.1	2
Missing	6.8	3	3.9	2	5.3	5

Note. * Difference between LMR and Comparison groups is statistically significant.

Table 4-5

Distribution of English Learners Across Classrooms

Class	English Learner		English Proficient		Missing Data	
Comparison	%	*n*	%	*n*	%	*n*
1	26.9	7	69.2	18	3.8	1
2	17.2	5	72.4	21	10.3	3
3	45.5	10	45.5	10	9.1	2
4	22.7	5	72.7	16	4.5	1
5	17.9	5	75.0	21	7.1	2
6	7.1	2	75.0	21	17.9	5
7	6.2	2	93.8	30	0.0	0
8	25.8	8	61.3	19	12.9	4
9	25.9	7	66.7	18	7.4	2
10	0.0	0	0.0	0	100.0	32
LMR						
1	17.9	5	60.7	17	21.4	6
2	16.1	5	74.2	23	9.7	3
3	15.6	5	68.8	22	15.6	5
4	7.4	2	85.2	23	7.4	2
5	8.0	2	84.0	21	8.0	2
6	20.8	5	75.0	18	4.2	1
7	6.5	2	90.3	28	3.2	1
8	20.7	6	72.4	21	6.9	2
9	20.8	5	75.0	18	4.2	1
10	7.1	2	85.7	24	7.1	2
11	33.3	5	66.7	10	0.0	0
Total	16.6	95	69.9	399	13.5	77

Table 4-6

Scores on the LMR Assessment and the Standardized Test

	English Learner						English Proficient					
	LMR			Comparison			LMR			Comparison		
	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*	Mean	*SD*	*n*
LMR Assessment												
Pre	-0.49	1.13	42	-0.57	0.97	49	0.28	1.35	208	0.44	1.30	166
Interim	0.53	1.21	42	.	.	0	1.23	1.35	207	.	.	0
Post	1.33	1.50	43	-0.19	1.11	47	2.23	1.64	207	0.75	1.35	167
Final	1.66	1.29	42	0.76	1.26	51	2.36	1.63	214	1.58	1.29	165
Change†	2.07	0.95	40	1.31	0.90	49	2.12	1.02	199	1.17	0.86	159
Standardized Test												
Prior Year	371.1	88.88	34	372.6	71.7	43	417.5	93.1	191	420.0	90.6	159
End-of-Year	394.9	87.43	40	363.0	63.9	49	424.6	96.3	215	421.1	84.3	169
Change	12.9	64.54	32	-20.0	42.8	42	7.0	64.1	184	-3.6	53.3	155

Note. LMR assessment score is represented in logits; Standardized test score has a range from 200 to 600. † Change = Final – Pre.

Table 4-7

Comparing the Estimated Average Achievement of English Learners who Participated in LMR With the Estimated Average Achievement of English Learners who Participated in the Comparison Group

	Estimated Achievement (*SE*)				Difference (*SE*)		Std. Effect Size. [a]	*p*-value
	LMR		Comparison					
Pretest	-0.43	(0.23)	-0.54	(0.23)	0.11	(0.32)	0.08	0.736
Interim test	0.52	(0.23)						
Posttest	1.35	(0.26)	-0.17	(0.25)	1.52	(0.36)	1.14	< 0.001
Final test	1.67	(0.25)	0.77	(0.25)	0.90	(0.35)	0.68	0.011

Note. [a] Standardized effect sizes calculated using the pretest score standard deviation for the LMR assessment

Table 4-8

Multilevel Model Results Using the Standardized Test as an Outcome Variable

Fixed effects	Estimate	*SE*	*p*-value
Intercept (ELs in Comp)	396.28	9.97	< 0.001
Prior year test	0.81	0.03	< 0.001
Nonlinear change	<-0.01	<-0.01	< 0.001
ELs in LMR	37.01	14.28	0.010
EP students in comparison group	22.44	9.24	0.015
EP students in LMR	34.06	11.73	0.004

Random effects	Estimate
Random intercept variance	278.3071
Level-1 residual variance	2542.235

Table 4-9

Comparing the Impact of LMR for English Learners Versus English Proficient Students on Average Estimated Achievement in Integers and Fractions

	Estimated Achievement (*SE*)				Difference (*SE*)		Std. Effect Size. [a]	*p*-value
	English Learner		Eng. Proficient					
Pretest	-0.43	(0.23)	0.19	(0.15)	0.62	(0.21)	0.47	0.002
Interim test	0.52	(0.23)	1.11	(0.15)	0.59	(0.21)	0.45	0.004
Posttest	1.35	(0.26)	2.12	(0.16)	0.77	(0.24)	0.58	0.002
Final test	1.67	(0.25)	2.28	(0.16)	0.61	(0.24)	0.46	0.010
Estimated Change	2.10	(0.14)	2.09	(0.06)	-0.01	(0.16)	-0.01	0.937

Note. [a] Standardized effect sizes calculated using the pretest score standard deviation for the LMR assessment

Table 4-10

Comparing LMR's Value Added for English Learners and English Proficient Students on Grade Level Mathematics Achievement.

Value Added from LMR (*SE*) [a]				Difference (*SE*)		Std. Effect Size. [b]	*p*-value
English Learner		Eng. Proficient					
37.01	(14.28)	11.61	(9.38)	25.40	(13.50)	0.34	0.060

Note. [a] Model controls for prior achievement. [b] Std. effect sizes calculated using the within-sample pooled *SD* for the 4th and 5th grade CST in mathematics administered in 2011.

Table 4-11

Comparing the Estimated Average Achievement of English Learners who Participated in LMR With the Estimated Average Achievement of English Proficient Students in the Comparison Group

	Estimated Achievement (*SE*)				Achievement Gap (*SE*)		Std. Effect Size. [a]	*p*-value
	English Learner in LMR		Eng. Proficient in comp.					
Pretest	-0.43	(0.23)	0.36	(0.17)	0.79	(0.29)	0.60	0.005
Final test	1.67	(0.25)	1.51	(0.18)	-0.16	(0.31)	-0.12	0.608
Estimated Change	2.10	(0.14)	1.15	(0.07)	-0.95	(0.16)	-0.72	< 0.001

Note. [a] Standardized effect sizes calculated using the pretest score *SD* for the LMR assessment

Table 4-12

Comparing the Estimated Average Achievement of English Learners who Participated in LMR With the Estimated Average Achievement of English Proficient Students in the Comparison Group

	English Learner				English Proficient			
	LMR		Comparison		LMR		Comparison	
Standardized Test	Mean	*SE*	Mean	*SE*	Mean	*SE*	Mean	*SE*
Prior Year	371.40	17.41	374.68	16.79	418.99	11.21	417.28	12.34
End-of-Year	398.17	15.66	366.90	15.18	424.42	9.97	416.17	11.12
Change	26.77		7.78		-5.43		1.11	

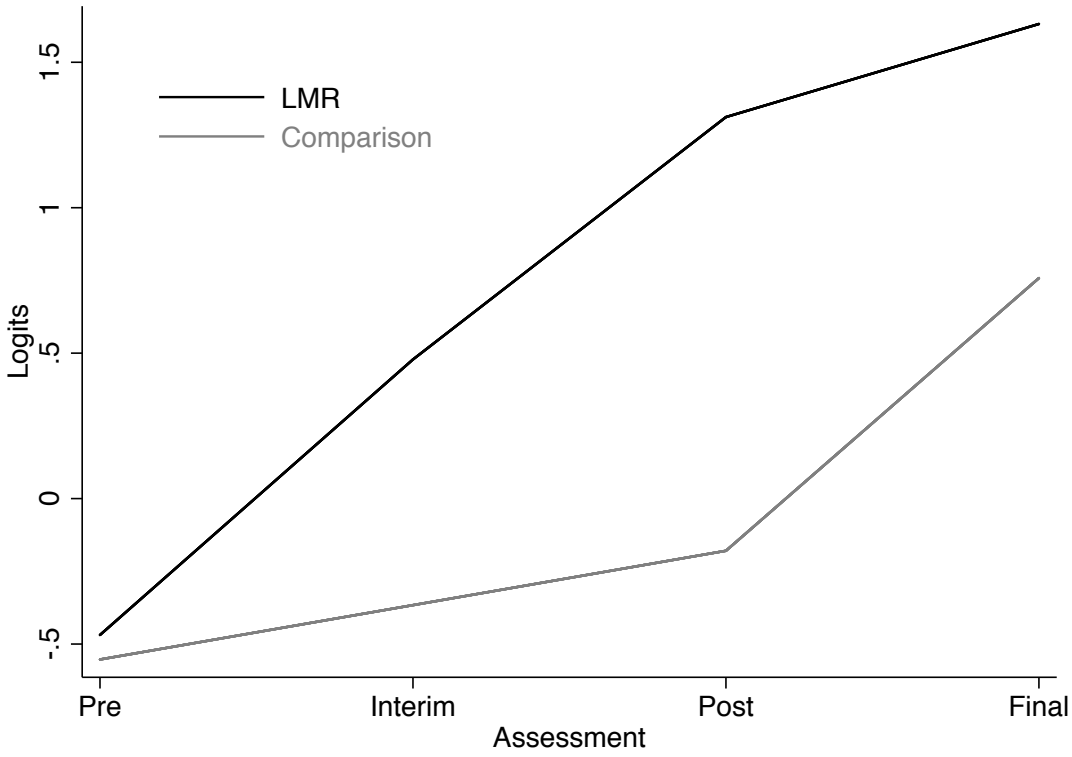

Figure 4-1. Average expected growth in student achievement in integers and fractions, comparing English learners in the LMR group with English learners in the comparison group.

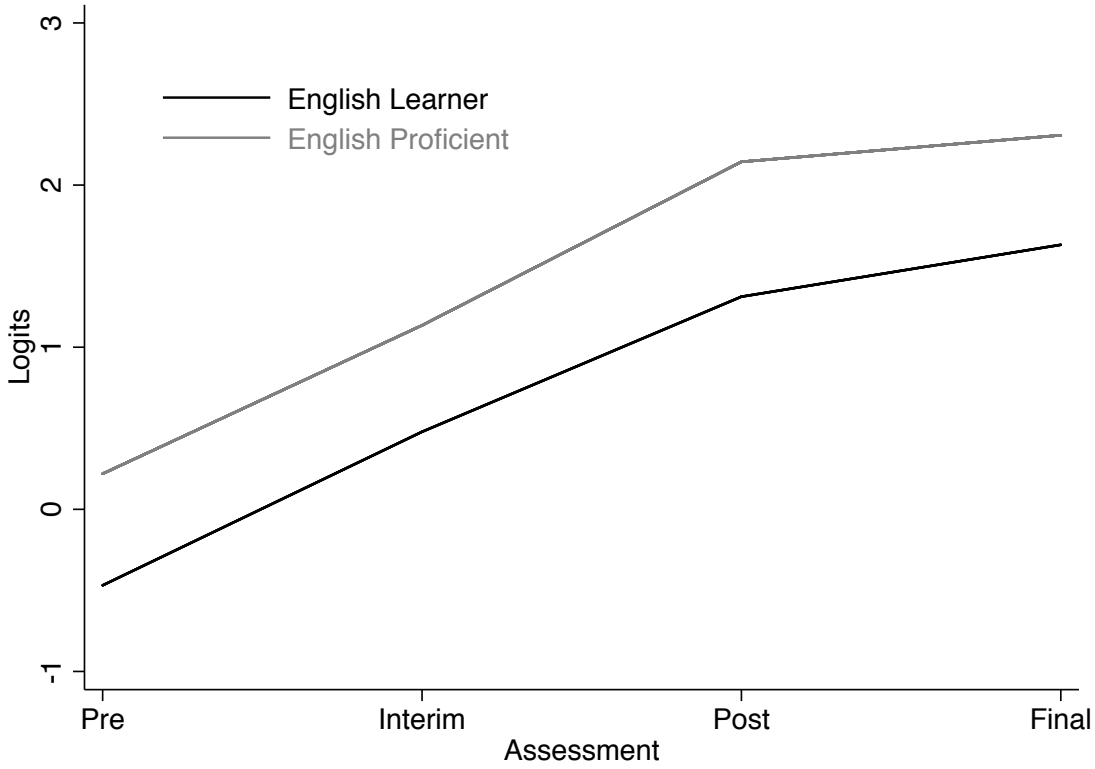

Figure 4-2. Average expected achievement in integers and fractions comparing the growth of English learners who participated in LMR with English proficient students who participated in LMR.

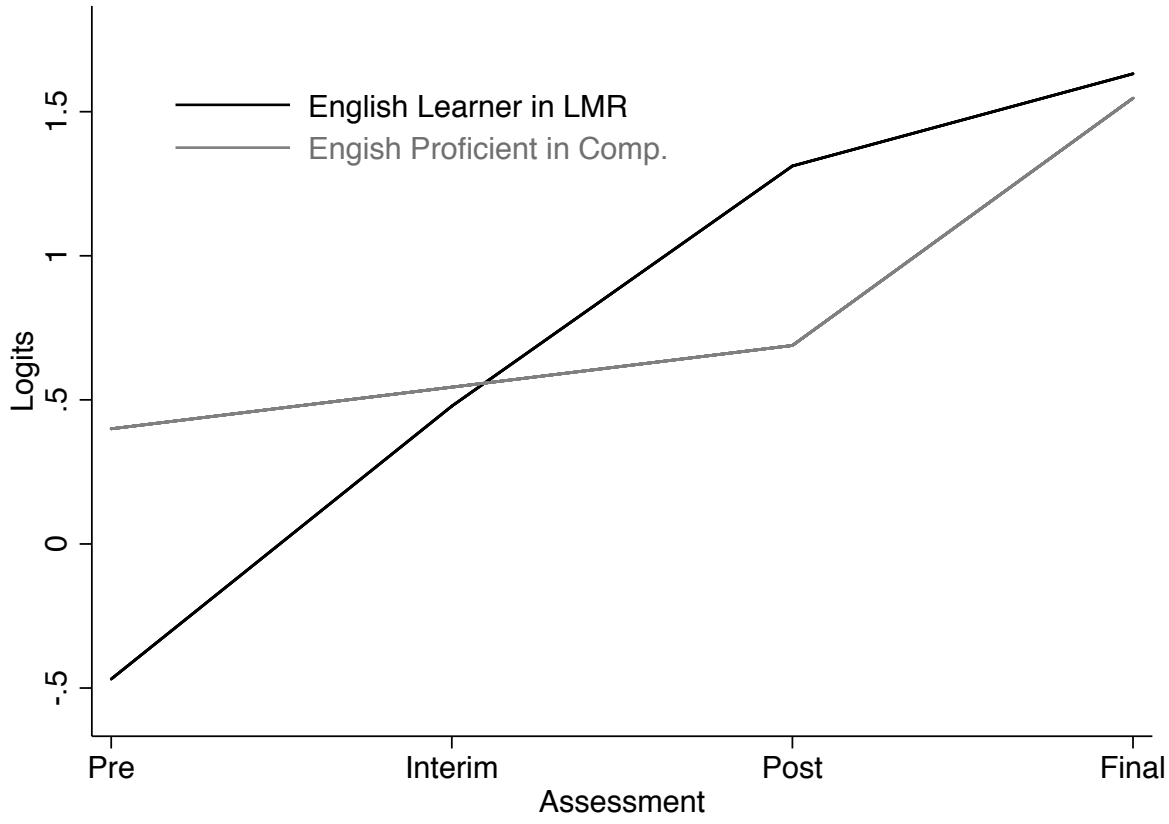

Figure 4-3. Average expected achievement in integers and fractions, comparing the growth of English learners who participated in LMR with English proficient students who participated in the Comparison groups.

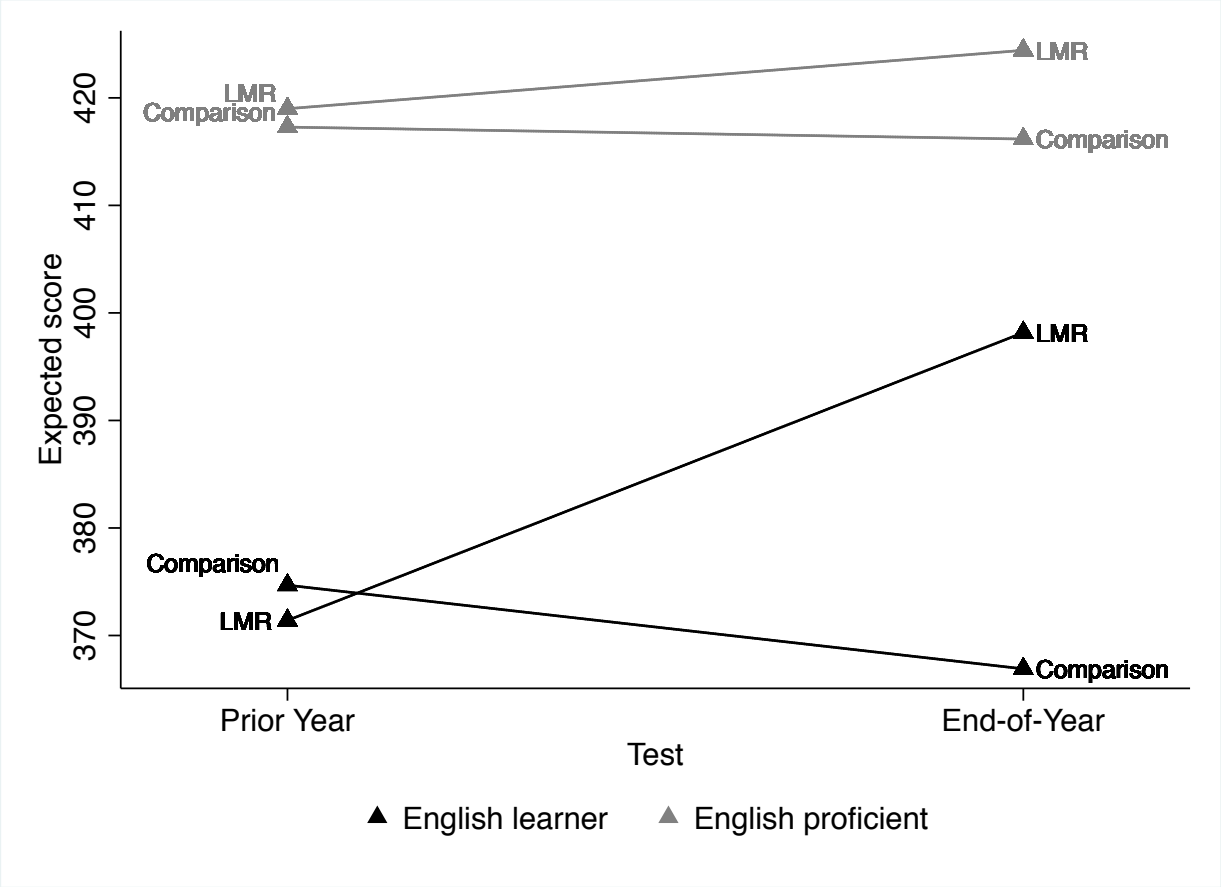

Figure 4-4. Expected achievement gaps in grade-level mathematics comparing average student achievement on the prior year and the end-of-year standardized tests in mathematics.

Chapter 5

Conclusion

The overarching goal of this three-paper dissertation was to generate knowledge that informs the use of standardized tests as outcome measures for evaluating instructional interventions in mathematics and science. In the first paper, I problematized the use of standardized tests as outcome measures. The second paper explored the consequences associated with the use of a misaligned test. The third paper presented a method for evaluating an equitable instructional intervention by (a) coordinating multiple sources of evidence and (a) accounting for the alignment between each assessment and the intervention in the interpretation of the results.

The first two studies provide new ways to show what scholars already know about problems with using standardized tests as outcome measures. Many authors have warned against, and provided examples of, the misuse of data from standardized tests (e.g., Hubert & Hauser, 1999). The results of this dissertation suggest that most investigators use data from standardized tests to evaluate new interventions in mathematics and science. Further, the investigators often neglect to establish validity evidence to support the use of the standardized tests. Some investigators learned, only after data collection, that the standardized test that they used as an outcome measure did not actually measure the same skills and knowledge taught by the intervention. The situation where an IES funded investigator cannot conduct an evaluation because of this type of measurement problem should be alarming, even if it is not surprising.

The second paper showed the consequences of misalignment between an outcome measure and an intervention. The results of the simulation model suggested that a high degree of alignment is necessary for adequate statistical power to detect the effects of an instructional intervention with a typical effect size of 0.1-0.5 *SD*s. The results support the idea that alignment should be a primary concern of applied educational researchers. When selecting outcome measures, investigators must carefully weigh the costs of test development with the costs of misalignment typically associated with off-the-shelf assessments.

The third paper documented the efficacy and equity of the Learning Mathematics through Representations (LMR) curriculum unit for students classified as English Learners (ELs). The results demonstrated that LMR was efficacious for ELs, whether the

outcome was an assessment of integers and fractions knowledge or a standardized test that measured grade-level achievement in mathematics. The results suggested that LMR narrowed achievement gaps between ELs and their English proficient peers. The paper articulated a theoretical mechanism of action for LMR that connected LMR's design features to the needs of ELs in the classroom. The combined evidence supports the promise of LMR for teaching integers and fractions with language minority students.

In addition, the third paper is an example of a evaluation that interpreted test scores with attention to the alignment between the test and the intervention. The research-based assessment of integers and fractions was the primary outcome of interest. Researchers developed the LMR assessment specifically to evaluate the impact of the intervention. The standardized test in mathematics was considered more of a distal outcome measure, also called a measure of learning transfer. Whereas both assessments showed improvements for the treatment group relative to the Comparison group, it is an open question as to how a brief, targeted intervention such as LMR might have produced a type of learning transfer from the specific learning goals of integers and fractions knowledge to the more distal construct of grade-level mathematics achievement. Speculatively, this could be evidence that supports the power of core generative ideas for supporting general, long-term learning. Regardless, the paper demonstrates that the results of evaluations are more interpretable and potentially meaningful when the reports articulate and justify test score interpretations in light of the theoretical construct measured by the test.

In closing, I provide three recommendations for research and policy that extend from the findings. One, stakeholders might narrow the gap between measurement research and measurement in practice by making changes to research proposals. Proposals for impact evaluation research should require investigators to discuss measurement in detail. The proposal should, at minimum, require discussions about (a) the construct that a test measures (b) how the items instantiate the construct, and (c) a mechanism of action that connects the program to success on the test. The additional focus on measurement is likely to improve the rigor and quality of educational research. Two, researchers who wish to measure complex forms of learning must address the difficult question of how much alignment is enough for adequate construct representation, and how much is too much for valid extrapolation inferences. The issues are complex, but pointed research questions, asked in the context of specific evaluation scenarios, may address basic problems in test validity and applied measurement. Third, the measurement community should increase its efforts to narrow the gap between research and practice by providing additional, concrete, guidance on how to synthesize and interpret data from multiple assessments in the context of impact evaluation. Increased collaboration between the evaluation and the measurement communities seems necessary encourage a subtle shift from the general idea that investigators should collect data from multiple measures to the idea that investigators must carefully select meaningful measures.

References

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory into Practice*, *42*, 18–29. doi:10.1353/tip.2003.0001

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162–172. doi:10.1016/j.stueduc.2005.05.008

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23. doi:10.1177/0146621697211001

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER ConQuest: Generalised item response modeling software* [computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

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

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16. doi:10.1007/bf02294143

Anderson, C. R., & Tate, W. F. (2008). Still separate, still unequal: democratic access to mathematics in U.S. schools. In L. English (Ed.), *Handbook of International Research in Mathematics Education* (pp. 299–318). New York, NY: Routledge. doi:0.4324/9780203930236.ch13

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Dublin, Ireland: Educational Research Center. doi:10.4324/9780203357811

Aud, S., Fox, M. A., & Kewal-Ramani, A. (2010). *Status and trends in the education of racial and ethnic minorities* (NCES 2010-015). Washington, DC: National Center for Education Statistics.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, *17*, 37–45. doi:10.1111/j.1745-3992.1998.tb00627

Beck, M. D. (2007). Review and other views: “Alignment” as a psychometric issue. *Applied Measurement in Education*, *20*, 127–135. doi:10.1080/08957340709336733

Bejar, I. I. (2010). Application of evidence-centered assessment design to the Advanced Placement redesign: A graphic restatement. *Applied Measurement in Education*, *23*, 378–439. doi:10.1080/08957347.2010.510969

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

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.x

Billingham, C. M., & Hunt, M. O. (2016). School racial composition and parental choice new evidence on the preferences of white parents in the United States. *Sociology of Education*, *89*, 99–117. doi:10.1177/0038040716635718

Bock, R. D., & Aitkin M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Brenner, M. E. (1998). Development of mathematical communication in problem solving groups by language minority students. *Bilingual Research Journal*, *22*, 149–174. doi:10.1080/15235882.1998.10162720

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

Bustamante, M. L., & Travis, B. (1999). Teachers’ and students’ attitudes towards the use of manipulatives in two predominantly Latino school districts. In L. Ortiz-Franco, N. G. Hernandez, & Y. De La Cruz (Eds.), *Changing the faces of mathematics: Perspectives on Latinos* (pp. 81–84). Reston, VA: National Council of Teachers of Mathematics.

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

Chalmers, P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–29. http://www.jstatsoft.org/v48/i06/.

Chamot, A. U. (2007). Accelerating academic achievement of English language learners: A synthesis of five evaluations of the CALLA Model. *Springer International Handbooks of Education, 15*, 317–331. doi:10.1007/978-0-387-46301-8_23

Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley Publishing Company.

Cobb, P., Boufi, A., McClain, K., & Whitenack, J. (1997). Reflective discourse and collective reflection. *Journal for Research in Mathematics Education*, *28*, 258–277. doi:10.2307/749781

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press. doi:10.1016/c2013-0-10517-x

Cohen, S. A. (1987). Instructional alignment: searching for a magic bullet. *Educational Researcher*, *16*, 16–20. doi:10.3102/0013189X016008016

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences*, *13*, 15–42. doi:10.1207/s15327809jls1301_2

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.

Curtis, D. A., Heller, J. I., Clarke, C., Rabe-Hesketh, S., & Ramirez, A. (2009, April). *The impact of Math Pathways and Pitfalls on students' mathematics achievement.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Darling-Hammond, L. (2007). Race, inequality and educational accountability: the irony of "No Child Left Behind." *Race Ethnicity and Education*, *10*, 245–260. doi:10.1080/13613320701503207

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

Doty, R. G., Mercer, S., & Henningsen, M. A. (1999). Taking on the challenge of mathematics for all. In L. Ortiz-Franco, N. G. Hernandez, & Y. De La Cruz (Eds.), *Changing the faces of mathematics: Perspectives on Latinos* (pp. 99–113). Reston, VA: National Council of Teachers of Mathematics.

Echevarria, J., Vogt, M., Short, D., Avila, A., & Castillo, M. (2010). *The SIOP model for teaching mathematics to English learners*. New York, NY: Pearson.

Educational Testing Service. (2011). *California Standards Tests technical report Spring 2011*. Retrieved from the California Department of Education website: http://www.cde.ca.gov/ta/tg/sr/technicalrpts.asp

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197. doi:10.1037/0033-2909.93.1.179

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*, 27–32. doi:10.2307/1176716

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.

Fulmer, G. W. (2011). Estimating critical values for strength of alignment among curriculum, assessments, and instruction. *Journal of Educational and Behavioral Statistics*, *36*, 381–402. doi:0.3102/1076998610381397

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement* (Working Paper 06-1). Policy Analysis for California Education. Retrieved from http://eric.ed.gov/?id=ED492024

Fuson, K. C., Smith, S. T., & Lo Cicero, A. M. (1997). Supporting Latino first graders' ten-structured thinking in urban classrooms. *Journal for Research in Mathematics Education*, 28, 738–766. doi:10.2307/749640

Garcia, E. E., & Cuéllar, D. (2006). Who are these linguistically and culturally diverse sudents? *Teachers College Record*, *108*, 2220–2246. doi:1467-9620.2006.00780

Garcia, E. & Gonzalez, R. (1995). Issues in systemic reform for culturally and linguistically diverse students. *Teachers College Record, 96*, 418–431. doi:10.1111/j.1467-9620.2006.00780

Gearhart, M., & Saxe, G. B. (2014). Differentiated instruction in shared mathematical contexts. *Teaching Children Mathematics*, *20,* 426–435. doi:10.5951/teacchilmath.20.7.0426

Goldenberg, C. (2013). Unlocking the research on English learners: what we know–and don't yet know–about effective instruction. *American Educator*, *37*, 4–11.

Guilleux, A., Blanchin, M., Hardouin, J. B., & Sébille, V. (2014). Power and sample size determination in the Rasch model: evaluation of the robustness of a numerical method to non-normality of the latent trait. *PloS One*, *9*, e83652. doi:10.1371/journal.pone.0057279

Gutiérrez, R. (2008). A "gap-gazing" fetish in mathematics education? Problematizing research on the achievement gap. *Journal for Research in Mathematics Education*, *39*, 357–364.

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*. Stanford, CA: National Academy of Education.

H.R. 2632. Elementary and Secondary Education Act (ESEA) of 1965 (89th Congress). Retrieved from https://www.gpo.gov/fdsys/pkg/STATUTE-79/pdf/STATUTE-79-Pg27.pdf

H.R. 3801. Education Sciences Reform Act of 2002 (107th Congress). Retrieved from https://www2.ed.gov/policy/rschstat/leg/PL107-279.pdf

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education, 85*, 130–146. doi:10.1080/01619561003685270

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125. doi:10.1177/014662169602000201

Hemphill, F. C., & Vanneman, A. (2011). Achievement gaps: how Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress (NCES 2011485). Washington, DC: National Center for Education Statistics. doi:10.1037/e595292011-001

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

Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press. doi:10.17226/6336

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K., Human, P., Murray, H., Oliver, D., & Wearne, D. (1996). Problem solving as a basis for reform in curriculum and instruction: the case of mathematics. *Educational Researcher*, *25*, 12–21. doi:10.3102/0013189x025004012

Holman, R., Glas, C. A., & de Haan, R. J. (2003). Power analysis in randomized clinical trials based on item response theory. *Controlled clinical trials*, *24*, 390–410. doi: 10.1016/s0197-2456(03)00061-8

Hubert, J., & Hauser, R. (Eds). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press. doi:10.17226/6336

Institute of Education Sciences. (2015). Department of Education, Institute of Education Sciences, fiscal year 2015 budget request. Washington DC: Author.

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

Ivory, G., Chaparro, D., & Ball, S. (1999). Staff development to foster Latino students' success in mathematics: insights from constructivism. In L. Ortiz-Franco, N. G. Hernandez, & Y. De La Cruz (Eds.), *Changing the faces of mathematics: Perspectives on Latinos* (pp. 113–122). Reston, VA: National Council of Teachers of Mathematics.

Janzen, J. (2008). Teaching English language learners in the content areas. *Review of Educational Research*, *78*, 1010–1038. doi:10.3102/0034654308325580

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

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: how measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, *6*, 81–90.

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

Kersaint, G., Thompson, D. R., & Petkova, M. (2014). *Teaching mathematics to English language learners*. New York: Routledge. doi:10.4324/9780203098561

Khisty, L. L., & Viego, G. (1999). Challenging conventional wisdom: A case study. In L. Ortiz-Franco, N. G. Hernandez, & Y. De La Cruz (Eds.), *Changing the faces of mathematics: Perspectives on Latinos* (pp. 71–79). Reston, VA: National Council of Teachers of Mathematics.

Khisty, L. L. (1995). Making inequality: issues of language and meanings in mathematics teaching with Hispanic students. In W. G. Secada, E. Fennema, & L. B. Adajian (Eds.), *New Directions for Equity in Mathematics Education* (pp. 279–301). New York, NY: Cambridge University Press.

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

Lane, S., Silver, E. A., & Wang, N. (1995, April). An Examination of the performance gains of culturally and linguistically diverse students on a mathematics performance assessment within the QUASAR project. Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Lee, H.-J., & Jung, W. S. (2004). Limited-English-Proficient (LEP) students and mathematical understanding. *Mathematics Teaching in the Middle School*, 9, 269–272.

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

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4–14. doi:10.2307/1177052

Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational* Assessment*, 7*, 29–38. doi:10.1207/s15326977ea0701_4

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

Lissitz, R. W. (Ed.). (2009). *The concept of validity: revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing. doi:10.1007/s11336-010-9180-6

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. (1992). The influence of testing on teaching math and science in grades 4–12: report of a study funded by the National Science Foundation (SPA8954759). Boston, MA: The Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

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 2009-13). 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

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*, 81–91. doi: 10.1177/014662168701100106

Moschkovich, J. (1999). Supporting the participation of English language learners in mathematical discussions. *For the Learning of Mathematics*, *19*, 11–19.

Moschkovich, J. (2002). A situated and sociocultural perspective on bilingual mathematics learners. *Mathematical Thinking and Learning*, *4*, 189–212. doi:10.1207/s15327833mtl04023_5

Moschkovich, J. (2012). Mathematics, the Common Core, and language. *Teachers College Record*, *96*, 418–431.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author. doi:10.17226/9870

National Mathematics Advisory Panel. (2008). *Foundations for success: the final report of the National Mathematics Advisory Panel*. Washington DC: National Academies Press.

National Research Council (NRC). (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academies Press.

National Research Council (NRC). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press. doi:10.17226/9822

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

Newton, J. A., & Kasten, S. E. (2013). Two Models for Evaluating Alignment of State Standards and Assessments: Competing or Complementary Perspectives? *Journal for Research in Mathematics Education*, *44*, 550–580. doi:10.5951/jresematheduc.44.3.0550

Noble, T., Suarez, C., Rosebery, A., O'Connor, M. C., Warren, B., & Hudicourt-Barnes, J. (2012). "I never thought of it as freezing": How students answer questions on large-scale science tests and what they know about science. *Journal of Research in Science Teaching*, *49*, 778–803. doi:10.1002/tea.21026

Office of Technology Assessment (OTA). (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.

Ortiz-Franco, L., Hernandez, N. G., & De La Cruz, Y. (Eds.). (1999). *Changing the faces of mathematics: Perspectives on Latinos*. Reston, VA: National Council of Teachers of Mathematics.

Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*, 100–120. doi:10.1177/0146621605279761

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press. doi:10.1037/e302582005-001

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, *43*, 119–134. doi:10.1080/15391523.2010.10782565

Perie, M., Park, J., & Klau, K. (2007). *Key elements for educational accountability models*. Washington, DC: Council of Chief State School Officers.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29*, 3–14. doi:10.1111/j.1745-3992.2010.00189

Polikoff, M. S., & Fulmer, G. W. (2013). Refining methods for estimating critical values for an alignment index. *Journal of Research on Educational Effectiveness*, *6*, 380–395. doi:10.1080/19345747.2012.755593

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., 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

Quinn, H., Schweingruber, H., & Keller. T. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata*. *Third edition.* College Station, TX: Stata Corp.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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

Saunders, W. M., & Marcelletti, D. J. (2012). The gap that can't go away: the catch-22 of reclassification in monitoring the progress of English learners. *Educational Evaluation and Policy Analysis*, *35*, 139-156. doi:10.3102/0162373712461849

Saxe, G. B., de Kirby, K., Kang, B., Le, M., & Schneider, A. (2015b). Studying cognition through time in a classroom community: the interplay between "everyday" and "scientific concepts." *Human Development*, *58*, 5–44. doi:10.1159/000371560

Saxe, G. B., Earnest, D., Sitabkhan, Y., Haldar, L. C., Lewis, K. E., & Zheng, Y. (2010). Supporting generative thinking about the integer number line in elementary mathematics. *Cognition and Instruction*, *28*, 433–474. doi:10.1080/07370008.2010.511569

Saxe, G. B., de Kirby, K., Le, M., Sitabkhan, Y., & Kang, B. (2015a). Understanding learning across lessons in classroom communities: A multi-leveled analytic approach. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Approaches to qualitative research in mathematics education* (pp. 253–318). New York, NY: Springer. doi:10.1007/978-94-017-9181-6

Saxe, G., Diakow, R., & Gearhart, M. (2013). Towards curricular coherence in integers and fractions: a study of the efficacy of a lesson sequence that uses the number line as the principal representational context. *ZDM - The International Journal on Mathematics Education*, 1–22. doi:10.1007/s11858-012-0466-2

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: a research review. *Reading & Writing Quarterly*, *23*, 139–159. doi:10.1080/10573560601158461

Schoenfeld, A. H. (2002). Making mathematics work for all children: issues of standards, testing, and equity. *Educational Researcher*, *31*, 13–25. doi:10.3102/0013189x031001013

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

Scriven, M. (2009). Demythologizing causation and evidence. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice* (pp. 134–152). doi:10.4135/9781412995634.d14

Secada, W. G. (1996). Urban students acquiring English and learning mathematics in the context of reform. *Urban Education*, *30*, 422–448. doi:10.1177/0042085996030004004

Shepard, L. A. (2010). What the marketplace has brought us: item-by-item teaching with little instructional insight. *Peabody Journal of Education*, *85*, 246–257. doi:10.1080/01619561003708445

Short, D. J., Echevarria, J., & Richards-Tutor, C. (2011). Research on academic literacy development in sheltered instruction classrooms. *Language Teaching Research*, *15*, 363–380. doi:10.1177/1362168811401155

Silver, E. A., & Stein, M. K. (1996). The Quasar project: The "revolution of the possible" in mathematics instructional reform in urban middle schools. *Urban Education*, *30*, 476–521. doi:10.1177/0042085996030004006

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37, 5–14. doi:10.3102/0013189x08314117

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.

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

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement*, *7*, 405–425. doi:10.1177/014662168300700403

StataCorp. (2013). *Stata Statistical Software: Release 13*. College Station, TX: Author.

Sussman, J. (2016). *The use and validity of standardized tests for evaluating new curricular interventions in mathematics and science* (Doctoral Dissertation).

Sussman, J. (2016b). *Evaluating the effectiveness of the Learning Mathematics through Representations (LMR) curricular unit for students classified as English Learners (ELs)* (Doctoral Dissertation).

Sussman, J., Diakow, R., & Saxe, G. (2012). The efficacy of the Learning Mathematics through Representations (LMR) lesson sequence using California Standards Test (CST) as an outcome measure (LMR Technical Report 1). Berkeley: Authors.

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.

Thier, H. D. (2004). Structuring successful collaborations between developers and assessment specialists. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 250-263). Chicago, IL: University of Chicago Press. doi:10.1111/j.1744-7984.2004.tb00060.x

Thomas, W. P., & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Santa Cruz, CA: University of California at Santa Cruz, Center for Research on Education, Diversity, and Excellence.

University of Chicago School Mathematics Project. (2007). *Everyday mathematics* (Third edition). New York, NY: McGraw-Hill Education.

Wang, L., Beckett, G. H., & Brown, L. (2006). Controversies of standardized assessment in school accountability reform: A critical synthesis of multidisciplinary research evidence. *Applied Measurement in Education, 19,* 305–328. doi:10.1207/s15324818ame1904_5

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. doi:10.1007/bf02294627

Warren, E., & Miller, J. (2014). Supporting English second-language learners in disadvantaged contexts: learning approaches that promote success in mathematics. *International Journal of Early Years Education, 23*, 1–17. doi:10.1080/09669760.2014.969200

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*, 7–25. doi:10.1207/s15324818ame2001_2

Wiggins, G. (1993). Assessment worthy of the liberal arts; authenticity, context, and validity. In Author (Ed.), *Assessing student performance* (pp. 34–71). San Francisco, CA: Josey-Bass.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. New York, NY: Psychology Press, Taylor & Francis Group. doi:10.4324/9781410611697

Wilson, M., Zheng, X., & McGuire, L. (2012). Formulating latent growth using an explanatory item response model approach. *Journal of Applied Measurement*, *13*, 1–22.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials*. Wilmington, DE: Wide Range.

Wu, M. L., Adams, R., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest: generalized item response modeling software* (Version 2). Camberwell, Australia: ACER Press, an Imprint of Australian Council for Educational Research Ltd.

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, *27*, 458–477. doi:10.2307/749877

Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika*, *56*, 589–600. doi:10.1007/bf02294492