UNIVERSITY OF CALIFORNIA

Los Angeles

Conceptual and Methodological Considerations Around the Measurement of Implementation in Quantitative Educational Research

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Education

by

Maria Paz Fernandez Smits

2024

ABSTRACT OF THE DISSERTATION

Conceptual and Methodological Considerations Around the
Measurement of Implementation in Quantitative Educational Research

by

Maria Paz Fernandez Smits
Doctor of Philosophy in Education
University of California, Los Angeles, 2024
Professor Jose-Felipe Martinez-Fernandez, Co-Chair
Professor Christina Christie, Co-Chair

This dissertation explores conceptual and methodological issues related to measuring implementation in quantitative educational research. Implementation, defined as the enactment of interventions in educational contexts, plays a pivotal role in understanding how programs operate and affect student outcomes. Despite its importance, there is significant variability in how implementation is conceptualized, measured, and accounted for in research, reflecting the complex relationship between interventions, schools, and outcomes.

The study examines these issues through a two-part approach. First, I conduct a systematic review to analyze the frameworks and constructs used to define implementation in a sample of grants awarded by the Institute of Education Sciences (IES). The review highlights the tensions between fidelity (adherence to program design) and adaptation (modifications to the original design). It also assesses the data collection instruments, methods of data reduction, and analytical strategies used to incorporate implementation into outcome evaluations.

Second, the dissertation uses data from the implementation of the Success for All (SFA) program to illustrate the practical consequences of methodological choices. Using correlations and hierarchical linear modeling, I estimate the association between different operationalizations of implementation and student outcomes. I also explore how implementation affects estimates of the program's impact, with a particular focus on English Language Learners (ELL) and Special Education (SPED) students. The findings indicate that while higher implementation fidelity correlates with improved outcomes for some groups, it may not benefit others equally, raising concerns about potential inequities.

This work underscores the need for nuanced approaches to measuring and accounting for implementation that reconcile fidelity and adaptation frameworks to better understand the schools and classrooms where interventions are delivered. In the context of school improvement, it is especially relevant to consider that adaptations to program design reflect teacher agency and should not be conceptualized only as deviations from the intended intervention.

The dissertation of Maria Paz Fernandez Smits is approved.

Jorge Manzi

Lucrecia Santibanez

Jonathan Schweig

Jose-Felipe Martinez-Fernandez, Committee Co-Chair

Christina Christie, Committee Co-Chair

University of California, Los Angeles

2024

Dedication

To my family, who have supported me throughout this process. ¡Los quiero mucho!

Table of Contents

List of Figures

List of Tables

Acknowledgments

I want to thank everyone who has directly or indirectly helped me in this long journey. First, my advisors, Felipe Martínez and Tina Christie, have supported and encouraged me throughout my PhD. Especially Felipe, who has given me invaluable advice, mentorship, and feedback throughout these years. Thank you so much for your patience and for always believing in me!

My friends, without whom this would have been not only a lot less fun but impossible. Jevan, Nadia, Andrés, Shujin, Lindy, Christine, Liz, Magda, Alice, Naomi, Flora, Ludovica, and Fernando. A special thank you to Marlene and María Renée for being the best support system and to Soledad, Diego, Seba, Flori, and Benja for being my family in LA.

My family in Chile, who has supported and helped me all this time. Papá, Catalina, Gonzalo, Paola, Emilia, Abuela, and Panchi, I would not have been able to do this without your love and encouragement. Mamá, gracias por tu apoyo incondicional, por siempre contestarme el teléfono, y por los 15 meses que me aguantaste en la pandemia.

Finally, I want to acknowledge the support of the Becas Chile program from the National Research and Development Agency (ANID) of the Ministry of Science, Technology, Knowledge and Innovation in Chile, which helped fund the first four and a half years of my PhD studies.

Vita

EDUCATION

2014	**Master of Public Policy (MPP).** University of California, Los Angeles.
2012	**M.A. in Political Science, Major in Public Policy.** Pontifical Catholic University of Chile.
2008	**B.A. in History.** Pontifical Catholic University of Chile.

PROFESSIONAL EXPERIENCE (RECENT)

University of California, Los Angeles

Graduate Student Researcher September 2018 – December 2024

- Worked on the development of an instrument to assess coaching for educational leaders. Project: 21st Century School Leadership Academy (21CSLA); California Department of Education.
- Developed and validated a classroom observation rubric to measure teachers' use of Computational Thinking. Project: UCLA STEM+C3; US Department of Education Research Grant (2019-2023).
- Adapted a teacher portfolio app and developed surveys to generate measures of teaching practice. Project: Teacher Cognition and Learning about Incorporating Science Representations in Elementary Classrooms; McDonnell Foundation Grant (2017-2021).

Teaching Assistant September 2022 – March 2024
September – December 2020

- Introduction to Research Design and Statistics and Linear Statistical Models in Social Science Research: Multiple Regression Analysis. Department of Education, PhD, and Master's programs.
- Introduction to Survey Research Methods, Department of Education, Educational Leadership Program (EdD).

RAND Corporation

Adjunct Researcher September – December 2021
Summer Associate June – August 2021

Researcher for the Learning Together Surveys and American Instructional Research Surveys. Lead author and co-author in two published data notes.

Research Center, Ministry of Education, Chile.

Senior Researcher July 2015 – September 2018

- Designed and implemented evaluations of national education programs, large-scale data collections, and large-scale RCTs.
- Country Manager for the OECD Reviews of National Policies for Education: Education in Chile (2017) and Chile's section of Education Policy Outlook (EPO) report.
- Coordinated the Ministry of Education's participation in the National Program Evaluation System, which is led by the Budgeting Department (DIPRES) of the Ministry of Finance.

PUBLICATIONS

Fernandez, M.P., & Martinez, J.F. (2022). Evaluating Teacher Performance and Teaching Effectiveness: Conceptual and Methodological Considerations. In Manzi, J., Sun, Y., & García, M. R.

(Eds.). *Teacher Evaluation Around the World: Experiences, Dilemmas and Future Challenges*. Springer.

Fernandez, M.P., Doan, S., & Steiner, E.D. (2021). *Use, Capture, and Value of Student Voice in Schools. Findings from the 2021 Learn Together Surveys*. RAND Corporation. DOI: https://doi.org/10.7249/RRA827-4

Kaufman, J.H., Doan, S., & Fernandez, M.P. (2021). *The Rise of Standards-Aligned Instructional Materials for U.S. K–12 Mathematics and English Language Arts Instruction. Findings from the 2021 American Instructional Resources Survey*. RAND Corporation. DOI: https://doi.org/10.7249/RRA134-11

Doan, S., Fernandez, M.P, Grant, D., Kaufman, J.H., Setodji, C.M, Snoke, J., Strawn, M., & Young, C.J. (2021). *American Instructional Resources Surveys: 2021 Technical Documentation and Survey Results*. RAND Corporation. DOI: https://doi.org/10.7249/RRA134-10

Young, C.J., Doan, S., Grant, D., Greer, L., Fernandez, M.P., Steiner, E.D., & Strawn, M. (2021). *Learn Together Surveys: 2021 Technical Documentation and Survey Results*. RAND Corporation. DOI: https://doi.org/10.7249/RRA827-2

Martinez, J.F., & Fernandez, M.P (2021). Teacher Evaluation with Multiple Indicators: Conceptual and Methodological Considerations Regarding Validity. In Manzi, J., García, M. R., & Taut, S. (Eds.) *Validity of Educational Assessments in Chile and Latin America*. Springer and Ediciones UC.

Martinez, J.F., & Fernandez, M.P. (2019). Evaluación docente con indicadores múltiples: consideraciones conceptuales y metodológicas en torno a la validez. In Manzi, J., García, M. R., & Taut, S. (Eds.) *Validez de Evaluaciones Educacionales en Chile y Latinoamérica*. Ediciones UC.

Fernandez, M.P. (2018). Mapa del estudiantado extranjero en el sistema escolar chileno, 2015-2017. *Documento de Trabajo 12*. Link to document.

Fernandez, M.P. (2017). ¿Hacia dónde avanza el sistema educativo en Chile? Análisis de las recomendaciones OCDE contenidas en *Evaluaciones de Políticas Nacionales de Educación: Educación en Chile* en el contexto de la Reforma en marcha. *Serie Evidencias 37*. Link to document.

Fernandez, M.P. (Ed.) (2017). *Reporte nacional de Chile: Revisión de las políticas educativas en Chile desde el 2004 al 2016*. Ministry of Education, Chile. Link to document.

Fernández Smits, M.P. (2011). *"Amor a palos": La violencia en la pareja en Santiago (1900-1920)*. LOM Ediciones.

Papers Presented at Conferences

Fernandez, M.P. (2025, April). *Formative Uses of a Rubric to Assess Computational Thinking*. Paper accepted at the American Educational Research Association (AERA) Annual Meeting.

Fernandez, M.P., Nava, I., Martinez, J.F., La Torre, D., Perez, L., & Betzelberger, J. (2024, April). *Measuring Computational Thinking Instruction: Reliability of a New Classroom Observation Instrument*. Paper presented at the National Council of Measurement in Education (NCME) Annual Meeting.

Fernandez, M.P. & Luo, J. (2022, September). *The Effect of School Switch in Chile: A Cross-Classified Modeling Approach*. Paper presented at the Society for Research on Educational Effectiveness (SREE) Conference.

Fernandez, M.P. (2022, April). *Is the Mean Enough?: Learning More About Teacher and Student Interactions in Math From Consensus in Students' Reports of Teaching Practice*. Paper presented at the American Educational Research Association (AERA) Annual Meeting.

Introduction

For decades, researchers and evaluators investigating educational interventions and policies have been concerned with the issue of program implementation. In the 1970s, researchers began to question the assumption that schools would passively and faithfully deliver interventions as designed, recognizing that local variability was the rule and uniformity the exception (Berman & McLaughlin, 1974; McLaughlin, 1990). This line of thinking recognizes that transferring and maintaining educational interventions from research settings to the classroom "is a complicated, long-term process that requires dealing effectively with the successive, complex phases of program diffusion" (Durlak & DuPre, 2008, p. 327).

In a context of variability in program delivery, implementation research is concerned with determining "*what* is actually enacted, *how* an innovation is enacted, and *why* the contexts, conditions, characteristics, and other influences shape innovation enactment as they do" (Century & Cassata, 2016, p. 172). Research on implementation is interested in analyzing "how well a proposed program or intervention is put into practice" (Durlak & DuPre, 2008, p. 5), examining what a program can and should be, what happens during and after it is put into practice, and what all this information tells researchers and practitioners about improvements in education (Century & Cassata, 2016).

Studying implementation can be key for understanding the delivery of the intervention, providing evidence on its impact, and validating the theory of change, which was part of the program's design. Data on implementation can help researchers determine which aspects of the intervention are most (or least) effective by shedding light on how or to what extent the intervention was actually delivered. It can also help determine whether the effects observed are due to features in the original design or model of the intervention or to adaptations that could have taken place in the field. Conversely, in cases where a program shows zero (or even negative) effects, studying implementation can help determine whether these results were due to faulty program design—indicating that the intervention should be revised—or to poor implementation —suggesting that the program could still work if it were implemented more accurately or differently (Durlak & DuPre, 2008).

The systematic study of implementation can be especially relevant in the case of complex interventions, where several components fit together to form the whole program. Educational programs are frequently designed to be implemented in different sites (e.g., school districts, schools, classrooms, etc.), each with specific characteristics that can affect the implementation of the intervention. People and organizational structures mediate the delivery of these program components across sites, so assuming that all are carried out identically or that all intended beneficiaries receive the same program is unrealistic. The study of implementation can provide insights into the conditions in which the intervention takes place, helping researchers and practitioners understand not only what works but also under which circumstances and specifically for whom (Lendrum & Humphrey, 2012; M. J. Weiss et al., 2014).

The study of implementation can also be critical in investigating the program's external validity. By providing additional information on the conditions under which an intervention has taken place, it can be replicated in other settings, ensuring its "transportability" into the real world. This can be especially relevant when replicating programs in contexts different from those where previous implementations were carried out, as the effects found in the original intervention may not be reproduced unless the implementation of the replicated program is closely aligned with the original (Wolery, 2011).

The issue of program replication has become especially relevant given the increasing push for evidence-based programs in education aimed at improving education using "rigorous and relevant research, evaluation and statistics" (What Works Clearinghouse, 2011, p. 1). In the United States, the two most recent reauthorizations of the Elementary and Secondary Education Act (ESEA) in 2001 (No Child Left Behind [NCLB] during the Bush administration) and 2015 (Every Student Succeeds Act [ESSA] under President Obama) recommended the use of programs and practices founded on "scientifically [or] evidence-based research" (Every Student Succeeds Act, 2015; No Child Left Behind Act of 2001, 2002). This, in turn, has led to the proliferation of studies designed with the explicit goal of measuring the effect of interventions on specific student outcomes.

The Department of Education created the What Works Clearinghouse (WWC) in 2002 with the goal of helping "educators, administrators, families, researchers, and policymakers make evidence-based decisions" (What Works Clearinghouse, 2022, p. 5). The WWC reviews existing research and then selects and synthesizes those studies that meet their criteria of "well-designed and well-implemented impact studies," identifying the scientifically based research required in NCLB and ESSA (What Works Clearinghouse, 2022, p. 5). Although the WWC's Procedures and Standards provide a detailed outline of the possible research designs that are acceptable for inclusion (typically, experimental or quasi-experimental methods aimed at evaluating a program's impact), until 2023, there was no explicit definition, expectations, or the criteria for acceptable implementation (Century & Cassata, 2016).

A lack of unified definitions, requirements, and expectations related to the conceptualization, measurement, and reporting of program implementation has important methodological implications that can affect the estimation of program outcomes. Different constructs can be used to measure implementation, such as fidelity, adaptation, quality, dosage, program reach, participant responsiveness, program differentiation, or monitoring of control and comparison conditions. This conceptualization determines how the research team views implementation and, therefore, what aspects of enactment are measured.

Additionally, implementation can be measured using different data collection instruments that are then aggregated through a specific method into indicators of implementation. Finally, there are more methodological decisions associated with how implementation will be accounted for and how it will be associated with program outcomes (e.g., correlational, mediation, or moderation analysis). All these methodological decisions related to measuring implementation can lead to biases in estimating its effect on a program outcome.

Research Questions

The discussion around program implementation is especially relevant when considering the inevitable variability that is to be expected in multi-site educational interventions. This dissertation aims to contribute to this debate by analyzing different methodological aspects related to the conceptualization and measurement of the implementation of interventions in educational contexts. Additionally, it uses data from the implementation of the Success for All program as an example to illustrate and model the concrete, quantifiable effects of these different methodological decisions. Specifically, this dissertation investigates the following research questions:

1. How is program implementation conceptualized and accounted for in the educational research literature?
    a. What are the frameworks and constructs used to conceptualize implementation?

    b. What data collection instruments and procedures are used to measure program implementation?
    c. What methods are used to reduce implementation data?
    d. What methods are used to account for implementation data when estimating intervention effects?
2. What are the practical consequences and implications of using these different methods to account for implementation when estimating the effects of an intervention?
    a. What are the consequences of the different approaches for data reduction on program implementation?
    b. What are the consequences of the different models that incorporate data on implementation on the estimation of the effects of an intervention?

Chapter 1 reviews the literature to compare the two main frameworks that define how implementation is conceptualized: *fidelity* (the extent to which the intervention corresponds to the originally intended program) and *adaptation* (how much the intervention has deviated from the original program; Century & Cassata, 2016; Durlak & DuPre, 2008; O'Donnell, 2008). For this study, the review will include interventions financed under the Institute of Education Sciences' funding goals number three (Efficacy and Replication) and four (Effectiveness). Only grants awarded between 2003 and 2019 will be considered to provide enough time for intervention to have publicly available evidence on its implementation. As defined in this dissertation, *interventions* or *programs* aim to change or improve Pre-K–12 classroom instruction by establishing a causal link between the intervention and student outcomes. The intervention has an explicit theory of change (including, for example, a logic model) in which the program components are outlined, and the mechanisms that will produce the expected changes in the beneficiaries are laid out (C. H. Weiss, 1995). Because the focus is on potential variability in program delivery, interventions should take place in multiple schools and include several classrooms and teachers.

The first research question examines how implementation is accounted for in the educational research literature, focusing specifically on methods for quantitative data collection and analysis. The chapter reviews data collection instruments used to collect evidence on the different implementation constructs in classrooms and schools, pointing out the advantages and limitations of each instrument for various types of programs, intervention components, implementation constructs, or people involved in the enactment of the intervention. Next, it explores the different methods used to reduce the data collected on implementation (e.g., indices, categories, etc.), and to associate these data with program outcomes.

The second question seeks to illustrate the implications of the questions and choices above. I use real-world data from the implementation of the Success for All program (Slavin et al., 2009; Slavin & Madden, 2001, 2012), specifically the 2011-2014 scale-up of the intervention (Quint et al., 2015). This example illustrates that the combination of data on implementation and how it is incorporated into quantitative models can affect the estimation of the intervention's impact on student outcomes. The significance of the work is that it explores the implications of the methodological decisions around the evaluation of program enactment, shedding light on the importance of measuring implementation and using these data to understand how and for whom the intervention works.

Understanding the interplay between fidelity, adaptations, and educational outcomes is critical for informing educational practices and program decisions. By acknowledging and investigating the nuances of implementation, researchers can understand what is happening inside the classroom and help design interventions that can fit the specific needs of different educational communities. This approach contributes to disentangling the mechanisms that generate changes in students' learning without assuming that the benchmark should be perfect adherence to the design. Adaptations can be productive, and they are inevitable, so they should be measured and accounted for.

Chapter 1: Literature Review

Studying the process of program implementation is very relevant, as research has demonstrated that levels of implementation can affect program outcomes (Durlak & DuPre, 2008). However, assessing implementation is not a straightforward formulaic process. Many methodological decisions must be made along the way, which can start from the researchers' definition and operationalization of *implementation* to the measures and instruments used to assess it.

Implementation research attempts to explain how and why programs work and fail (Scriven, 1994). According to Century and Cassata (2016), implementation research can then be understood as the "systematic inquiry regarding interventions enacted in controlled settings or ordinary practice, the factors that influence intervention enactment, and the relationships between interventions, influential factors, and outcomes" (p. 170).

Collecting and analyzing data on the implementation of educational interventions can provide insights into the conditions in which the intervention takes place to understand not only what works but also under which circumstances and specifically for whom (Lendrum & Humphrey, 2012). This becomes more relevant, considering that research on program enactment shows that variation in implementation within and across sites is ubiquitous. In contrast, uniformity and perfect implementation are practically unobtainable since interventions are rarely implemented as designed. Furthermore, empirical evidence proves that levels of implementation influence intervention outcomes. Although peer-reviewed journals are increasingly requiring authors to describe the steps taken to ensure adequate implementation, in most cases, they only state that implementation was effectively achieved and do not provide sufficient evidence to support this statement (Durlak & DuPre, 2008).

This chapter examines the literature on program implementation, beginning with a definition of the concept and the relevance of studying and measuring implementation for program evaluation. Next, it explores two different frameworks for conceptualizing implementation: fidelity and adaptation. The final part of the chapter presents methodological aspects related to measuring and analyzing program implementation, including the constructs and instruments used to measure intervention enactment and issues related to data reduction and modeling.

Defining Implementation

The beginnings of research into program implementation in the United States are associated with the advent of the Great Society programs in the 1960s, particularly their evaluation mandate and the intention to understand the mechanisms that underlie program enactment (McLaughlin, 1984). One of the first attempts at conceptualizing implementation was Berman and McLaughlin's (1974, 1978) Change Agent study prepared for the RAND Corporation, in which the authors corroborated that "implementation did not involve merely the direct application of a technology [but it] was an organizational process that implied interactions between the project and its setting" (Berman & McLaughlin, 1976). Although, from the 1960s to the 1980s, there was interest in examining implementation, particularly as it relates to adherence to original program design, "research suggests that the study of implementation had not yet been fully adopted and perhaps valued" (Dhillon et al., 2014, p. 11).

The field of implementation is based on change processes such as the *adoption*, *dissemination*, and *diffusion* model, which focuses on how, why, and when innovations are adopted (Cho, 1998). In this model, diffusion is defined as "the process by which an innovation is communicated through certain channels over time among the members of a social system"

(Rogers, 1983, p. 5). In the specific case of diffusion, communication is seen as the process by which participants create and share information about an intervention (innovation) to reach a mutual understanding. Information flows from an individual or group that knows the intervention to another individual or group that does not. After the individual or group decides to adopt the intervention, the process of implementation begins when the innovation is put into use. The rate of adoption (the speed with which the intervention is adopted) depends on the perceived attributes of the innovation (e.g., relative advantage, complexity, trialability), the type of innovation decision (optional, collective, or authority), the communication channels (e.g., mass media or interpersonal), the nature of the social system (e.g., its norms, degree of interconnectedness), and the extent of change agents' promotion efforts (Rogers, 1983).

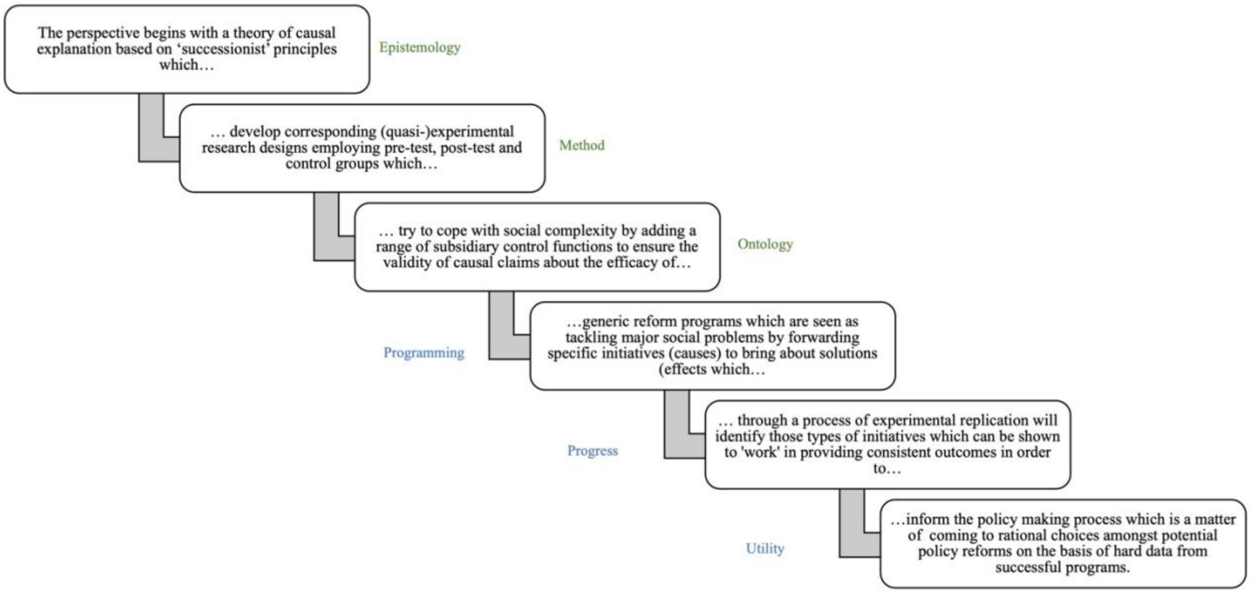
To study implementation, program developers need to specify the change model and clearly state the theory that drives the intervention. This must include a description of the mechanisms that are expected to operate within the program and during implementation to generate a change in the outcomes, especially of the most important (core) components that are expected to drive the effect of the intervention (Dhillon et al., 2014). After the theory behind the program is specified, developers may overlay a logic model that "links the outcomes (both short- and long-term) with program activities/processes and the theoretical assumptions/principles of the program" (W.K. Kellogg Foundation, 2004, p. 35). This part of program design should lay down in clear detail how and why the intervention will work (C. H. Weiss, 1995).

In research that intends to prove the existence of a causal link between the intervention and a change in a population characteristic, it is paramount to define two types of variables: the question predictor (intervention) and the outcome of interest (Murnane & Willett, 2011).

In the specific case of experimental evaluations, the causal link is proven using conditions that ensure the outcome is comparable in a group that is exposed to the intervention and a similar group that is not (Shadish et al., 2001). In experimental designs, which are considered the gold standard design for causal inference, program participants are randomly assigned to the treatment or control groups, ensuring that in these groups, the units are, on average, probabilistically similar to each other. Quasi-experiments intend to establish a causal relationship between the intervention and the outcomes, but they do so in the absence of random assignment (Murnane & Willett, 2011; Shadish et al., 2001; Slavin, 2002).

Figure 1 outlines the main methodological and policy elements behind program evaluation based on experimental designs, showing that the main goal of the model is to establish causation through experimental control.

Figure 1: An Overview of Experimental Evaluation



The perspective begins with a theory of causal explanation based on 'successionist' principles which…

… develop corresponding (quasi-)experimental research designs employing pre-test, post-test and control groups which…

… try to cope with social complexity by adding a range of subsidiary control functions to ensure the validity of causal claims about the efficacy of…

…generic reform programs which are seen as tackling major social problems by forwarding specific initiatives (causes) to bring about solutions (effects which…

… through a process of experimental replication will identify those types of initiatives which can be shown to 'work' in providing consistent outcomes in order to…

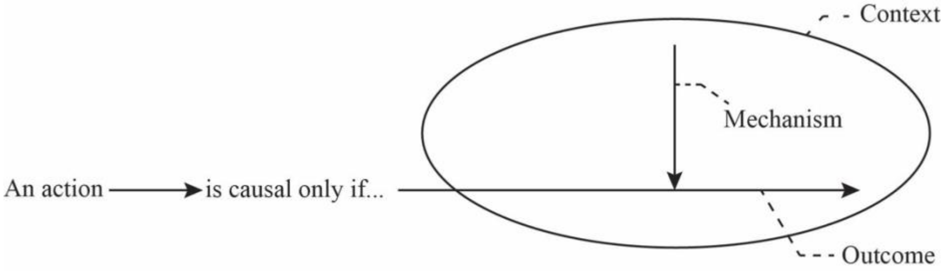
…inform the policy making process which is a matter of coming to rational choices amongst potential policy reforms on the basis of hard data from successful programs.

Source: Pawson & Tilley (1997).

In this model, implementation is meant to ensure that the causal relationship between the intervention and the outcomes is validated, linking the manipulation of the program to the outcome of interest (Shadish et al., 2001). Thus, the main goal is to determine whether the intervention works and the magnitude of its impact on a specific outcome.

Another way to conceptualize the relationship between the intervention and its outcomes is through Realistic Evaluation. This model is concerned with the mechanisms that generate the outcomes and the context in which these mechanisms operate, all within a causation model that encompasses the Context-Mechanism-Outcome (CMO) configurations (Figure 2; Pawson & Tilley, 1997).

Figure 2: Generative Causation (Realist Evaluation)



Source: Pawson & Tilley (1997), p. 58.

The figure above depicts the Context-Mechanics-Outcome (CMO) configurations in which mechanisms are frequently understood as program components. Realistic evaluation considers the relevance of the context to the causal model, assuming that the relationship between mechanisms and outcomes is context-dependent (Lemire et al., 2020; Pawson & Tilley, 1997). The goal of this type of evaluation is not only to determine whether a program works but also to determine "what is about the program that works for whom?" (Pawson & Tilley, 1997, p. 109)

These two models, which conceptualize causality in program evaluation and the corresponding role of the implementation process, help us understand the different frameworks associated with program implementation. A focus on program *fidelity* serves the experimental evaluation framework, while the generative model can be associated with an *adaptation* framework.

Intervention Components: The Building Blocks of Program Design

The literature tends to agree that one of the most important steps in program design that will consequently help inform the study of implementation is the definition of the components that make up the intervention (Blakely et al., 1987; Century & Cassata, 2016; Dhillon et al., 2014; Durlak & DuPre, 2008; Lendrum & Humphrey, 2012; Mowbray et al., 2003; Scheirer & Rezmovic, 1983). As Blakely et al. (1987) indicate, program components should be activities, materials, and facilities that are observable or verifiable through interviews with those involved with the program (implementers or beneficiaries). They should also be logically distinct from other components, or at least not depend on the implementation of other components, while also being specific to the intervention under study (i.e., not common to other programs in the developing organization). Finally, the list of components should describe the intervention exhaustively, leaving no aspects of the intervention unexplained.

There is a hierarchy of components within an intervention, as all components are not meant to be equally important for reaching the expected outcomes. *Core* components (also referred to as *essential*, *critical components,* or *active ingredients*) are the ones that program designers expect to have a more significant effect on the outcome of interest as they drive the mechanism for change. Furthermore, Fixsen et al. (2005) found evidence to indicate that a clear definition of components is associated with higher chances of successful implementation. These components are considered indispensable when implementing the intervention, “at least until empirical data prove otherwise” (Century & Cassata, 2016, p. 182). The study of implementation gives researchers data to determine which components are more strongly related to the outcomes in practice, as evidence on implementation can prove the magnitude of the relationship between the specific parts of the program and its expected results in the real world (Lendrum & Humphrey, 2012). Core components are usually defined initially from the causation model that underpins the theory of change of a program.

Although identifying and defining core components is a crucial part of program design, especially for the purposes of monitoring and analyzing implementation, it is a challenging task for intervention developers and evaluators. The difficulty in determining the hierarchy of components is independent of the complexity of the program, as it can be equally challenging for both relatively simple and multifaceted interventions. At the same time, researchers have seen that intervention designers tend to classify most components as “very important” (Mowbray et al., 2003) or to view the whole program as a whole “package,” leading to component descriptions that are vague and lack specificity. Considering these challenges, it frequently falls under the evaluators or researchers to gather data from different sources (e.g., program designers, end users, observations, artifacts, etc.) to classify the relevance of each part of the intervention (Century & Cassata, 2016).

The remaining components that are not classified as *core* are, in theory, considered to be “nonessential related components” or part of the “adaptable periphery” (Century & Cassata, 2016, p. 182). This assertion implies the notion that implementers should ensure that the *core* components are enacted as closely to the original design as possible (i.e., with higher levels of *fidelity*).

Factors that Affect Implementation: Influence of the Context on Implementation

Several factors may influence the enactment of an intervention, and they can be related to the intervention itself or to the context in which it is being implemented. These factors can be grouped into *spheres of influence*, ranging from the characteristics of individual end users to organizational and environmental factors, implementation support strategies, and implementation over time (Century & Cassata, 2016). All these factors can interact to affect the way the intervention is implemented (Durlak & DuPre, 2008).

In the case of individual end users' characteristics, the analysis assumes that beneficiaries are not passive recipients since they actively interpret and make decisions that can affect implementation. When analyzing this sphere, Century & Cassata (2016) find that the literature on program implementation tends to identify two main categories that vary depending on their relationship with the intervention. In the first one, characteristics of the individual are tied to the intervention (e.g., level of understanding, expertise, prior experience, beliefs, values, attitudes, motivation, or self-efficacy). In the second, the individual's characteristics are independent of the intervention (e.g., willingness to try new things, organizational skills, classroom management style, or views about teaching and learning in general).

Organizational factors can also have an impact on the implementation of a program, both from the implementing organization or from external influences (Century & Cassata, 2016; Durlak & DuPre, 2008; M. J. Weiss et al., 2014). For educational interventions delivered at the classroom level, these types of influences are associated with schools or larger institutions (e.g., school districts), as well as individual and collective characteristics inside these organizations. Some examples of factors within school settings can include class size, resources, physical space, scheduling, or organizational structure. At the collective level, factors related to organizational culture (e.g., morale, vision, trust, collaboration, identity, commitment, supportive climate) can impact the way in which a program is delivered. At an individual level, the decision-making processes of management and administration can affect intervention adoption and use, with strong leadership associated with success in implementation (Century & Cassata, 2016; M. J. Weiss et al., 2014). These features can affect whether the organization leans more towards implementing the intervention as planned or whether it adapts it to meet local conditions and preferences (M. J. Weiss et al., 2014).

Other elements that can affect the delivery of a program are related to the characteristics of the intervention itself and people's perceptions of the program. Some researchers consider that only objective traits are attributes of a program (e.g., number of components, complexity, specification, scope of effort, empirical evidence of effectiveness, design features, and cost). In contrast, others add subjective traits to this list (e.g., level of attractiveness of the materials, ease of use, familiarity, perceived relevance, and perceived advantage over current practice). Although influences from objective and subjective characteristics are relevant, it is important to keep in mind that objective traits depend only on the program. Still, subjective characteristics will vary according to the attributes of the end user population (Century & Cassata, 2016).

Objective traits can vary significantly in their level of specificity from one program to another. Some interventions may have detailed implementation plans, while others provide only rough guidelines. The literature does not agree on how the level of detail can affect intervention enactment (M. J. Weiss et al., 2014). Some researchers argue that comprehensive plans lead to higher chances of implementation success (Fixsen et al., 2005). In contrast, others claim that there must be leeway to adapt the intervention to the requirements of each context to increase the possibility of buy-in in the implementing organization (Bardach, 1980; M. J. Weiss et al., 2014).

Factors that affect implementation can also be classified under the support strategies that program developers or intermediary organizations provide to implementers if these elements are part of the intervention design (Century & Cassata, 2016; M. J. Weiss et al., 2014). There is a wide variety of supports, and they can range from operational planning, resource provision, professional development, mentoring, strategic planning, evaluative processes, and other strategies that support ongoing implementation and improvement. The literature has several names for these factors, including implementation drivers, implementation-level activities, support systems, strategies, and implementation practices (Century & Cassata, 2016; Darrow, 2013; Domitrovich et al., 2008; Dunst et al., 2013; Fixsen et al., 2005). Even though support strategies are commonly considered very relevant in theories of change, they are often categorized as part of organizational factors and not taken as a separate sphere of influence (Century & Cassata, 2016).

Implementation phases or stages can also be a factor that affects program enactment. This sphere of influence differs from the others, as it views implementation as a long-term process, from the initial stages of adoption into the phase in which the intervention has possibly become embedded into the implementer's actions and is part of their routine. As Century and Century and Cassata (2016) note, the literature focuses on different types of implementation phases, with an emphasis on individual evolution (from routine to mechanical use and then more flexible and adaptive use as competence increases; Hall & Loucks, 1978). Other models stress the relevance of the organization, indicating that they progress through awareness, planning activities, initial implementation, and skilled implementation until the implementation of the new program becomes routine practice (Berman & McLaughlin, 1976; Fixsen et al., 2005).

Importance of Studying Implementation

Research into implementation addresses two main concerns: describing and conceptualizing the intervention itself; and identifying and organizing the influential factors that affect implementation. The focus is not placed on the decision to implement but instead on what happens next: “what is actually enacted, how an innovation is enacted, and why the contexts, conditions, characteristics, and other influences shape innovation enactment as they do” (Century & Cassata, 2016, p. 172).

Regarding the first aspect, research on implementation documents what was actually conducted in the field, contributing to supporting the project’s internal validity and providing concrete support for the causal link between the designed intervention and its outcomes (Durlak & DuPre, 2008; Lendrum & Humphrey, 2012; O’Donnell, 2008). Program theory can also be tested through the collection and analysis of implementation data, as it allows researchers to prove that the link between the designed intervention components, their effective administration, and expected outcomes was observed (Durlak & DuPre, 2008).

Information on program delivery can support the notion that a program is effective, indicating that the intervention implemented with certain specifications produces the observed outcomes. If implementation is not assessed, then the program has not been adequately tested (Durlak, 2015). In this sense, it is important to know whether the results of an intervention can be attributed to the program itself (as it was designed) or to changes that those in charge of intervention delivery incorporated in the field (Durlak & DuPre, 2008).

This is very relevant in the context of experimental research designs, where the units (e.g., students, classrooms, teachers, schools) are randomly assigned to participate in the intervention. In these designs, the causal effect is the comparison of potential outcomes, as individuals cannot be observed experimenting the treatment and control conditions simultaneously. The observed outcomes need to be measured to estimate the effect of an intervention. The stable unit treatment value assumption (SUTVA) supposes that “the potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes” (Imbens & Rubin, 2015, p. 10).

If the program is implemented without variation across sites. In that case, the SUTVA assumption is enough to estimate the *intention-to-treat* (ITT) estimate, as the assignment mechanism (randomization) accounts for the differences between the groups. However, the ITT analysis does not hold if there are differences in implementation that affect the exposure to the intervention. In this case, the alternative is an *as-treated* approach, where the assignment mechanism is determined by actually receiving the treatment (Imbens & Rubin, 2015). In this context, implementation can be seen as a way to measure the units that were treated and estimate the effect of the intervention on the individuals who were actually exposed to the intervention.

Recording and analyzing implementation can also contribute to an intervention’s external validity by helping researchers ensure that a program known or presumed to be effective is delivered closely to its original design in a different context (Rossi et al., 2019). This type of research can also inform a program’s design and development by looking into what the intervention “could and/or should be, the extent to which [it] is feasible in particular settings, and its utility from the perspective of the end users” (Century & Cassata, 2016, p. 174). These data can also inform how the different components and activities of an intervention interact with the constraints of the context that surrounds it (Lendrum & Humphrey, 2012).

If a program has passed its design and pilot phase, insights into its implementation can help answer questions related to the quality of the services provided, how well the program is organized, or how successfully it is reaching its intended beneficiaries (Rossi et al., 2019). Additionally, monitoring implementation while the program is being delivered may help practitioners make appropriate changes to ensure better outcomes (Durlak & DuPre, 2008).

Furthermore, data from monitoring implementation is frequently used in practice to explain negative or ambiguous findings when assessing program effects.

However, there is an implicit tradeoff in studying implementation. On the one hand, the necessary data collection and analyses require additional funding and time, which are not always budgeted in the evaluation from the beginning. On the other, the study of implementation can provide nuanced information on the program that may not satisfy the needs of decision-makers. For example, an impact evaluation with accompanying implementation data may indicate that the program worked for some people in particular contexts and at specific times. This may not be useful for program designers and policymakers, who require immediate data on whether the intervention works and whether it is worth scaling it up (Lendrum & Humphrey, 2012).

Conceptualizing Implementation: Fidelity and Adaptation as Frameworks

The literature on implementation research presents a historical tension between the concepts of *fidelity* and *adaptation* (Century & Cassata, 2016; Cho, 1998). These concepts underpin and define the ways in which implementation is operationalized and measured and can have a profound effect on the methods used to assess implementation (Fullan & Pomfret, 1977; O'Donnell, 2008).

Fidelity is defined as the "determination of how well an intervention is implemented in comparison with the original program design" (O'Donnell, 2008, p. 33), and it can be operationalized as adherence, compliance, integrity, or faithful replication (Durlak & DuPre, 2008). Fidelity is assessed and compared against an ideal of implementation (the originally designed program) to determine how close the delivered intervention is to the theoretical intervention. Those who advocate for implementation based on fidelity indicate that validated programs should be implemented with close correspondence to the original model, with the intention of obtaining the expected impact (Blakely et al., 1987). The assumption behind the evaluation of fidelity is that the delivered program should be as similar as possible to the design, so higher levels of fidelity are considered preferable and tend to be equated with "good" implementation (Mowbray et al., 2003).

Adaptation, on the other hand, views the implementation process from a perspective of change, considering "the need to adapt models to local conditions to maximize efficiency as well as local ownership." Implementation based on adaptation assumes that modifications to the original program are necessary due to the particular needs of the target population, differences in budget, community resources, or organizational factors (Mowbray et al., 2003, p. 327). This type of assessment may address a broad range of possible changes to the designed program, from modifications of specific components to a complete program reinvention.

Cho (1998) presents a framework to conceptualize the two paradigms of fidelity and adaptation, which helps to understand the qualitative pattern of how an intervention is transformed in each context (Table 1). Fidelity approaches the change process behind program implementation in a linear way, clearly distinguishing the researcher from the researched and facts from values. This means that researchers and evaluators can test whether a specific intervention was effective at the local level by using experimental or quasi-experimental methods, and these tested programs can then be replicated and disseminated widely to other educational contexts. Evaluating implementation from a fidelity standpoint prioritizes the developers' original intention in the relationship between an intervention and its user.

The adaptation perspective considers the context in which the intervention is being implemented, acknowledging the interactions between the program and its institutional setting. This model expects that users will and should adapt the program, appropriately interpreting it within their context (Cho, 1998).

Table 1: Implementation Paradigms

Perspectives	Fidelity	Adaptation
Paradigms	Positivism / Behaviorism	Post positivism
Ontology	Pure reflection	Negotiation/grounded
Epistemology	Objectivist	Modified dualism
Methodology	Evaluation-based	Variation-based

Adapted from Cho, (1998)

There are different ways to conceptualize fidelity and adaptation based on their relationship with the original program design and with the enactment of the intervention. On the one hand, they can be seen as two ends of a continuum, with faithful replication on one end (fidelity) and a modified version of the program (adaptation) on the other. In this sense, fidelity may be seen as the desirable end of the spec. At the same time, adaptations to the original model can be conceptualized as failures in implementation (i.e., a failure to achieve fidelity). Hall and Loucks (1978) conceptualized the fidelity-adaptation continuum in terms of a "zone of drastic mutation [beyond which] the developer will not accept what is being used as *the* innovation" (p. 18).

Conversely, fidelity and adaptation may be considered independent constructs that should be measured and related to outcomes separately (O'Donnell, 2008). This dichotomy may lead researchers to consider addressing one of the two instead of looking at how they interact when a program is implemented. Documenting and combining the adherence to the original program design and the modifications made during implementation can lead researchers to find “the appropriate combination of faithful replication and program modification that is necessary in different settings and for different innovations to achieve good outcomes” (Durlak & DuPre, 2008, p. 342).

However, there does not seem to be an adequate combination of the two for all educational programs and across all settings, with some researchers indicating that higher fidelity is better suited for well-defined programs. At the same time, adaptations are more appropriate for unstructured interventions or for those that are in the early stages of development and use (Fullan & Pomfret, 1977; Lendrum & Humphrey, 2012). There may also be degrees of acceptable modifications associated with different program components, as core components may require more fidelity than other less central aspects of the intervention (Durlak & DuPre, 2008). Consequently, researchers have not reached a consensus on the appropriate levels of fidelity and adaptation that should be allowed during implementation and on the consequences that the modifications may have for internal and external validity (M. J. Weiss et al., 2014).

Researchers who advocate for adaptation consider fidelity to be limited in its approach, as it fails to consider teachers’ knowledge utilization and their ability to modify the program to fit their students’ needs and interests (Cho, 1998). Additionally, those in favor of an adaptation framework indicate that focusing strictly on fidelity can lead researchers to overlook the potential contribution of these changes to program outcomes (Durlak & DuPre, 2008). Arguments against an exclusive focus on fidelity also point to the fact that delivering an intervention entails a “complex mix of both fidelity and adaptation” that includes interactions among “multiple factors at program, implementer and organizational level” (Lendrum & Humphrey, 2012, p. 2).

The distinction between fidelity and adaptation has several implications for conceptualizing and measuring implementation and program outcomes. This can profoundly affect the evaluation of the intervention and can change whether a program is considered effective.

Fidelity Framework

The concept of *fidelity* is based on the notion that program delivery should be compared to the original program design, with a top-down approach that favors the designers’ intentions. The study of program implementation aims to determine the extent to which the components of an intervention were enacted as stipulated in the original program model (Cho, 1998).

This line of thinking can be traced back to models that structure behavior in terms of rationality, particularly the rational-comprehensive model (Lindblom, 1959), which was, in turn, based on Dewey’s account of the problem-solving process. Dewey (1933) emphasized three aspects in his rational approach: the careful analysis of the problem, a comprehensive search for alternatives, and a rational consideration of the consequences of each alternative. As it applies to program evaluation, this theory indicates that it is possible to find a single intervention that will solve social issues and that rational adopters can and should implement it faithfully (Emshoff et al., 1987).

In education, models based on rational decision-making are associated with research on knowledge utilization and educational change developed towards the middle of the twentieth century and with the Research, Development, and Diffusion (RD&D) model, which was partly inspired by the success of the R&D efforts in space exploration. In these models, specialized research groups would develop interventions (“social technologies”) that would then go through a rigorous validation process (evaluation) conducted using experimental or quasi-experimental methods. In this model, the initiative to solve the identified social problem is with the developer, who then proceeds to create the activities that will generate a solution; finally, these activities will be disseminated to a larger population (Cho, 1998; MacDonald & Walker, 1976).

In this approach, successful programs could be replicated and widely disseminated to schools. Implementers (e.g., schools, school districts, etc.) were assumed to be passive consumers of the program who would highly value the evaluation results. The assumption behind these rational models is that research findings can be generalized to any context without modifications (Blakely et al., 1987; Cho, 1998).

This approach considers the educational improvement process “as linear and rational, concerned with faithful implementation and minimizing variation and deviation from efficacious innovation models” (Century & Cassata, 2016, p. 199). The concept of faithful replication of an intervention fits well with this view of educational research and improvement, as it allows researchers to generate seemingly objective proof of the program’s effectiveness (Cho, 1998; Fullan & Pomfret, 1977). The rationale behind fidelity of implementation indicates that an intervention that has proven to be efficacious should be replicated as closely as possible to the original design (Century & Cassata, 2016). Within this structured model, a set of research instruments is used to determine whether a well-specified intervention was implemented according to its original design, and the level of fidelity is determined using data from a set of observable variables (O’Donnell, 2008).

Assessment of implementation from the perspective of fidelity has become the dominating framework for structuring how end users should implement evidence-based interventions. Studying fidelity can help establish the efficacy and effectiveness of specific educational programs and determine whether the intervention generated any effect on an outcome measure (Century & Cassata, 2016). Efficacy studies evaluate interventions "implemented under ideal or routine conditions that are either in wide use and have not been rigorously evaluated or have evidence of promise for improving student outcomes but have not been previously evaluated for impact". On the other hand, effectiveness evaluations are intended to conduct assessments of programs that have "prior evidence of efficacy to determine its impact when implemented under routine conditions in education settings, or a follow-up study of a prior effectiveness evaluation" (Chhin et al., 2018, p. 596).

The emphasis of efficacy studies lies on the validation of the intervention's internal validity, as it is being delivered in the most favorable conditions possible. Studying implementation can contribute to confirming that the mechanisms of change defined in the program's design are operating as intended. This validates program theory, as implementation as planned can help determine that there is a link between the outcome observed and the planned intervention (Century & Cassata, 2016; Lendrum & Humphrey, 2012; O'Donnell, 2008). Monitoring fidelity ensures that the integrity of the intervention is maintained and that the role of implementers (e.g., teachers or school administrators) is contained within acceptable parameters (Snyder et al., 1992).

In the context of experimental and quasi-experimental designs, fidelity allows researchers to differentiate the treatment and control groups, using this information as a manipulation check in effectiveness research. A manipulation check is meant to corroborate whether participants received the program intended by the researcher, and when successful, it can support claims of causality supporting the hypothesis that the independent variable (intervention) causes the intended effect on the dependent variable (e.g., student outcomes). Thus, researchers recommend measuring fidelity of implementation not only for those receiving the intervention (treatment group) but also for those individuals who are not (control group) since this allows for a comprehensive assessment of the intervention and valid comparisons between groups (Hoewe, 2017; O'Donnell, 2008).

Fidelity can also be used as a manipulation check in this context. It provides information to determine whether the intended recipients experienced the stimulus (program), therefore ruling out that a lack of exposure to the intervention caused the absence of a detectable effect. In this case, the independent variable was successfully manipulated (i.e., the treatment group was exposed to the program, while the control group was not). Nonetheless, this manipulation did not change the dependent variable (e.g., student outcomes) Hoewe, 2017; Mowbray et al., 2003).

Establishing a causal connection between program components and outcomes can help reduce the possibility of type III error, which happens when researchers try to measure something that does not exist (Dobson & Cook, 1980). Claims about the effects of an intervention are made based on what the end users received. Still, these inferences are not valid if the program being measured was not delivered or if it was not measured adequately. A type III error when evaluating an intervention can happen in the following three scenarios: when researchers are trying to assess the effectiveness of a program that is not measured as implemented; if the program as designed has not been implemented; or in the case there is no testable relationship between the intervention components and the enacted activities being measured (Scanlon et al., 1977). Reducing the chances of type III error also helps differentiate failures in implementation from failures in program design, reducing the chances that a study reports that an intervention has no effects when the issues stem from problems in implementation (Harachi et al., 1999).

Information on fidelity is useful to determine the intervention's external validity, providing adequate documentation and guidelines for program replication in other contexts. Particularly in the case of research on intervention effectiveness, a detailed record of the program's implementation that evaluates adherence to the original design can provide information for future replications, informing the areas that should be prioritized to ensure that program delivery does not deviate from an innovation that has evidence of success (Mowbray et al., 2003). For programs that have already proven to be effective, the underlying rationale is that once a program is found to be efficacious, future implementations should not deviate from the established "proven" or "evidence-based" model (Blakely et al., 1987; Century & Cassata, 2016). Furthermore, the expectation is that implementers will "readily and completely" accept an educational intervention that is demonstrably effective (Marsh & Willis, 2007).

Evaluating and documenting the enactment of an intervention serves as a complement to impact evaluations that use experimental or quasi-experimental methods, as adequate delivery should not be taken for granted. Information on field execution is considered an integral part of the assessment of a program's effects to determine whether different aspects of the intervention meet an acceptable delivery standard (e.g., minimum quantity or quality) (O'Donnell, 2008; Rossi et al., 2019). A study of the impact of the intervention itself coupled with an analysis of the factors that contribute to generating the observed outcomes can be more informative than an impact evaluation alone (Durlak & DuPre, 2008; Lendrum & Humphrey, 2012; Rossi et al., 2019).

Measures of fidelity can contribute to increasing statistical power in treatment outcome studies, as they can be used to define the criteria of inclusion of sites leading researchers to exclude from the analysis those sites in which the program deviated too much from the treatment model if these settings do not reflect an example of implementation comparable to the rest of the sites (Century & Cassata, 2016; Mowbray et al., 2003). This can help reduce variation across sites, ensuring that differences in outcomes are due to assignment to treatment and not to changes to the intervention model. Additionally, removing these sites can also lead to larger effect sizes and an improvement of the statistical power for detecting genuine treatment effects by ensuring that the intervention is delivered with the intended strength (Cohen, 1988; Collins et al., 2005; Domitrovich et al., 2008; Shadish et al., 2001).

Fidelity can also be used as a mediating variable to help explain variance in outcomes by testing whether the effect of the intervention varies with levels of adherence and if the program generates a larger change in outcome for those sites where they found higher levels of adherence to the original intervention design. This provides more evidence to support the original intervention's internal validity by strengthening the link between the intervention and the observed effects, therefore substantiating the causal claims (Hansen, 2014; Mowbray et al., 2003; Teague et al., 1995). In the case of research on effectiveness, the focus may not be on monitoring and controlling levels of fidelity but on studying the differences in fidelity in a natural setting and then relating these variations to student outcomes (O'Donnell, 2008).

Research that focuses on adherence to the original research design tends to consider a lack of fidelity as negative. Evidence from studies that have found a more significant impact of previously effective programs on student outcomes when they are implemented again with high levels of fidelity helps support this claim. Furthermore, research that links fidelity of implementation to program outcomes finds a positive relationship between adherence to the original program and improvement in outcomes. Conversely, deviations from the original design were positively associated with outcomes only when measures of fidelity were held constant, indicating that fidelity moderated the relationship between intervention and outcomes (O'Donnell, 2008). Other studies suggest that problems in implementation associated with a lack of fidelity are frequently the reason for the failure of programs. In these cases, the absence of intervention effects should not be attributed to a faulty intervention design but to issues with intervention delivery (Rossi et al., 2019).

Research on program efficacy and effectiveness is increasingly reporting more information on program implementation, with a specific focus on fidelity. Federal funding agencies like the Institute for Education Sciences require researchers to collect information on program fidelity to inform the study's results (Institute of Education Sciences, 2022a, 2022b). Sanetti et al. (2011) found that between 1995 and 2008, the number of studies in the school psychology literature that reported quantitative fidelity measures increased threefold, specifically for interventions with children.

Adaptation Framework

The *adaptation* framework can be seen on the opposite side of the spectrum from *fidelity*, as it focuses on describing implementation as it was conducted without using the original model as the standard for comparison (Century & Cassata, 2016). Other terms used to exemplify program modifications made at the ground level include *mutual adaptation*, *co-construction*, *reinvention,* or *adaptive integration*. Although these concepts can be used interchangeably, they may also denote different grades of changes to the original model, ranging from additions to reinventions that do not resemble the initial design.

In contrast with the fidelity framework, which employs a top-down approach, adaptation is framed around middle-up strategies. The difference is apparent when it is contrasted with a view of implementation focused on fidelity: the latter is an "outsider perspective" in which the best person to dictate what the program should look like is the developer. While with fidelity, implementers were expected to follow the original program that the designers intended, in adaptation, the implementers (located in the middle between program designers and participants) have agency to effect changes to the intervention. Researchers and program designers can learn from implementers how to improve interventions if they collect the appropriate information on how the program was delivered (Durlak & DuPre, 2008). Although changes are expected, in this model, implementers still follow and respect what is stipulated in the written program (Buxton et al., 2015; Cho, 1998).

Research on implementation that focuses on adaptation is more concerned with the improvement of the intervention and with exploring the relationships between programs, contextual factors, and outcomes. From this perspective, it is more important to ask questions related to how the intervention's components are being used and how and why they are being adapted from their original design (Century & Cassata, 2016). This view considers that in education, program modifications are inevitable, as the concept of adaptation is associated with the idea that a teacher "cannot help but fit an innovation to his or her context" (Cho, 1998, p. 15). From this perspective, a focus exclusively on fidelity is also unrealistic, as evidence proves that some degree of adaptation is inevitable; conceptually or theoretically, opposing these changes is a futile exercise (Berman & McLaughlin, 1978; Durlak, 2015). Thus, researchers should assume that "variation in implementation is not a problem to be avoided but part and parcel of the basic operation of complex systems" (Honig, 2006, p. 21).

This model is based on a theory of bounded rationality, grounded on the assumption that factors external to the individual's intellectual capacities (e.g., social, political, economic, and organizational factors) affect their decisions (Lindblom, 1959; Simon, 1955). It is consistent with a postpositivist view of research that emphasizes the context in which the change should take place and conceptually allows for modifications to the program according to the interaction of the intervention with the institutional setting (Cho, 1998). Within this model, implementers are not concerned with delivering programs that have been empirically validated through efficacy and effectiveness with strict adherence to the original design. On the contrary, implementers "choose program components that meet the immediate needs of the organization, and they modify the program to fit the organizational constraints" (O'Donnell, 2008, p.49). Within this framework of program adaptations, the initial program design may be considered a "cognitive blueprint" for action and coordinated activity and not a manual that has to be followed literally (Marsh & Willis, 2007; Price et al., 1998).

Research that focuses on fidelity of implementation may conceptualize *adaptation* as a subconstruct of fidelity or as the absence of fidelity, therefore considering it unnecessary to define or measure it separately (Blakely et al., 1987; Hansen, 2014). Adaptation has even been described in terms of fidelity as "positive infidelity" when changes to the original design have led to improvements in outcomes (Hulleman & Cordray, 2009). Although the concept of adaptation can be defined negatively in terms of program adherence, the literature on program implementation has recently tended to agree that fidelity and adaptation are separate constructs that should be defined and measured independently (Century & Cassata, 2016; Durlak & DuPre, 2008; O'Donnell, 2008).

Berman and McLaughlin’s Change Agent study prepared for the RAND Corporation was one of the earliest works to conceptualize and openly recognize the relevance of focusing on implementation for the success of educational interventions (Berman & McLaughlin, 1974, 1978; McLaughlin, 1990). In it, the authors found three dominant patterns of implementation that combine implementation, modifications to the original design, and changes in organizational behavior. In the first pattern, *mutual adaptation*, there is an adaptation of the project design and the setting. At the same time, in *non-implementation,* there is no adaptation on the part of either the project or the setting. Finally, in *cooptation,* a one-way process takes place in which the program is adapted to the participants’ indifference and resistance to change while participants show no change (Berman & McLaughlin, 1976; Blakely et al., 1987; Lendrum & Humphrey, 2012). The type of implementation process depends on “the motivations and circumstances involved in its initiation, its substance and scope of proposed change, and its implementation strategy” (Berman & McLaughlin, 1976, p. 353).

In the particular case of mutual adaptation, Berman and McLaughlin (1976, 1978) found that the intervention was adapted to the context. At the same time, teachers and school administrators modified their practices. Some effective strategies that could promote mutual adaptation included providing teachers with necessary and timely feedback, allowing implementers to make choices, correcting errors, and encouraging commitment to the project. The authors found that the adjustments made to the intervention often caused difficulties and did not aid in the achievement of the program’s objectives. Their study concluded that the most successful implementation strategy was one that promoted mutual adaptation.

A study of adaptations can help implementers classify the types of modifications they make to the program. They can categorize them according to their alignment with program goals and theory. Changes that closely align may be considered acceptable, while changes that deviate from the goals and theory would be regarded as unacceptable (Century & Cassata, 2016; Durlak & DuPre, 2008; Hall & Loucks, 1978; Mowbray et al., 2003; O’Donnell, 2008).

As with studies focused on fidelity, those centered on adaptation examine the relationships between implementation and outcomes but do not necessarily compare the program to a theoretical ideal. The framing is more descriptive and explanatory than evaluative, with the aim of describing and understanding the extent and nature of the intervention's use in practice. Changes and omissions of core components are documented, while the factors that inhibit or promote the program's use are explored (Century & Cassata, 2016).

Critics of the view of a requirement of fidelity based on the prior effectiveness of the original model indicate that this assumption may be based on “the absence, rather than the presence, of empirical evidence about what types of adaptations are beneficial or harmful” (Century & Cassata, 2016, p. 199). This is compounded by the fact that researchers who consider adaptations as implementation failure may not be aware of the ways in which local changes to the original program design can positively affect the outcomes (Durlak & DuPre, 2008).

Although there appears to be consensus on the relevance of studying implementation to ensure that the intervention impacts educational outcomes, there is no agreement on how adaptations affect program outcomes (Durlak & DuPre, 2008). While research associated with the fidelity perspective maintains that adaptations negatively impact outcomes (O’Donnell, 2008), those who advocate for modifications have found evidence that contradicts this claim (Century & Cassata, 2016). Some studies have concluded that more adaptation is associated with positive results (Blakely et al., 1987). Research that has examined the effects across studies has found mixed results: the effects could be positive but also negative, or there could be no effects at all (Durlak, 2015; Durlak & DuPre, 2008; Hill & Erickson, 2019).

Research has indicated that one of the positive effects of adaptation is an increase in local ownership of the intervention. Organizational ownership can grow when professionals involved in the implementation of the program feel they are "exercising their judgment and expertise" (Mowbray et al., 2003, p. 335). Under this view, end users and implementers are considered active agents who are influenced by their context and whose decisions are guided by the way they make meaning of the program (Century & Cassata, 2016; Honig, 2006). The adaptation perspective considers that if those delivering the program are knowledgeable about their communities, then they should be able to make changes to the original design to make it more effective for the context (Durlak & DuPre, 2008). Additionally, modifications can make the intervention more relevant to the context by adding effective strategies or establishing that one or more components that were theorized to be indispensable are not necessary (Century & Cassata, 2016).

Translated to an educational setting, teachers and school administrators can be more engaged, motivated, and effective in implementing the program if they feel like they have agency over the delivery of the components. For example, Berman & McLaughlin (1976) found that having teachers participate in the development of intervention materials "provide[s] an important opportunity for staff to work through and understand project precepts and to develop a sense of 'ownership' in project methods and goals" (p. 361). The value of having teachers involved in the program in this way was not in the product itself but in participating in the activity. This can lead educators to see the process of implementation as an active educational effort in which they are protagonists in the efforts to improve their students' educational experience (Cho, 1998).

Additionally, modifications can increase the chances of program continuity (Blakely et al., 1987).

Researchers advocating for adaptations argue that implementation should be modified and revised according to the evolving interaction between the program and the context in which it is enacted. Adaptations may be required to meet the particular needs of the target population, as well as differences in budget or resources, or to adjust to organizational factors in a series of trade-offs within a local context (Cho, 1998; Mowbray et al., 2003). This may be particularly relevant in the case of specific racial and ethnic populations, as it can help develop more culturally sensitive and inclusive learning environments for teachers and students (American Psychological Association, 2002; Durlak & DuPre, 2008). Furthermore, a delivery perfectly aligned with the intervention's design may not be necessary for the program's success, as studies have reached positive effects with levels of fidelity that do not meet the researchers' fidelity standard (Blakely et al., 1987; Durlak & DuPre, 2008).

Critics of the fidelity perspective argue against decontextualizing interventions, considering that it is not a realistic assumption. Implementation is always different from the theoretical ideal, as it is a complex process; reducing this complexity and decontextualizing it can limit the applicability of implementation research results. Therefore, evidence emanating from efficacy and effectiveness studies based on a requirement of fidelity should be considered only informative and not definitive, as they are implemented in a specific context with a particular population (Century & Cassata, 2016).

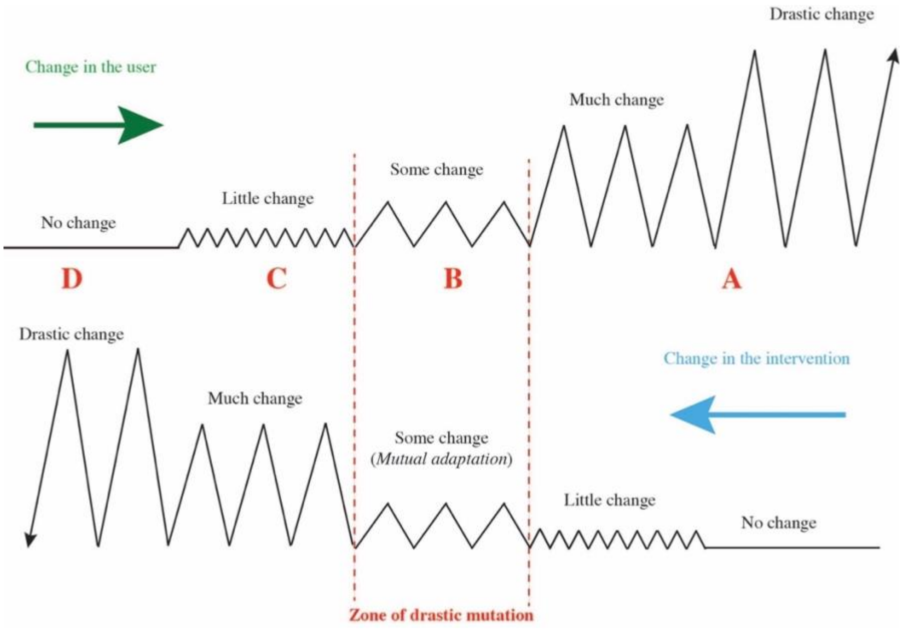
Adaptation can play an important role during program implementation, and researchers must account for this when designing and implementing educational interventions. The discussion of fidelity versus adaptation is based on fundamental differences about the role of implementation in program evaluation, particularly as to whether changes to program design should be considered flaws in implementation or an inevitable consequence of enactment that must be accounted for. Nonetheless, the divide does not have to be so clear, as there can be space for a view that reconciles both frameworks.

Reconciling Fidelity and Adaptation

The tension between fidelity and adaptation is based on different ways to conceptualize and account for program implementation. On the side of fidelity, interventions implemented with adherence to the original design help prove causal claims, generating models of interventions that, when replicated with close alignment, should produce the expected impact on outcomes. However, a strict focus on fidelity can be seen as "ethically indefensible" if it does not allow implementers to adapt the program to the recipients' specific needs. Adaptation, on the other hand, has a more flexible approach towards on-site modifications, assuming they are not only inevitable but often beneficial for end users. At the same time, modifications "cannot be merely a way of avoiding fidelity to a curriculum that scrupulously and rigorously reveals to students the actuality of the larger world in which they live in" (Marsh & Willis, 2007, p. 228). Since there are advantages and disadvantages to each model, a middle ground that combines both frameworks may be most appropriate (Century & Cassata, 2016; Marsh & Willis, 2007).

Hall and Loucks (1978) bring together the fidelity and adaptation frameworks, depicting the levels of modifications that may be acceptable for program designers. They identify the "zone of drastic mutation" as, the area beyond which the designer does not consider what is being implemented as the intended program. For these authors, adaptations are acceptable up to the point where the intervention's effectiveness and integrity are compromised. Figure 3 presents the different combinations of people-change and innovation-change, illustrating the process of adaptation and change as a continuum, with the area between the dotted lines as the zone of drastic mutation. The cases shown in the figure indicate a negative association between program adaptation and changes in the user. In case A, the intervention is implemented with high levels of fidelity, producing significant changes in the user. On the opposite side of the spectrum, in cases C and D, the intervention has gone beyond the zone of drastic mutation, which is linked to decreasing levels of change in the user.

Figure 3: Change in the user and change in the intervention



Source: Hall & Loucks (1978).

In this model, the area of change is conceptualized as a "zone" and not a point, as there may not be agreement between designers and implementers as to when the modifications are too extreme. As depicted in Figure 3, changes in the intervention affect the users. Nonetheless, this model operates under the assumption that higher levels of change to the program are associated with fewer changes in the user. This may be considered positive, as more change in the program can reflect ownership of the intervention (i.e., the activities are aligned with the implementer's method of working and may not require profound changes). On the other hand, it may also be evidence of resistance to implementing the new intervention. The concept has other important shortcomings, as the zone of drastic mutation is not readily identifiable since it depends on the specificity of the program's design (especially the critical components) and on the implementers reaching consensus on what constitutes a breach of this zone into an intervention that can no longer be associated with the original (Mowbray et al., 2003).

In reality, fidelity and adaptation co-occur, and both can have a relevant effect on program outcomes, as implementers "often replicate some parts of programs but modify others" (Durlak & DuPre, 2008, p. 341). In educational contexts, teachers' and administrators' knowledge of the conditions that surround the implementation allows them to have the appropriate pedagogical judgment that external researchers may lack (Cho, 1998). Allowing teachers to exercise their professional autonomy may have positive effects on the program and can also benefit the students.

The literature indicates that it is necessary to find a balance between requiring strict adherence and allowing any change. However, with the current information available, it is not possible to know what the exact combination is that will lead to increased outcomes. Considering the many variables that interact in an educational intervention, it may be pointless to attempt a definition a priori, as the combination will depend on the intervention itself and the context in which it is delivered. However, the literature concedes that one way to come closer to this balance is through a clearly articulated theory of change that specifies the core and the non-essential related components and how they should be enacted (Lendrum & Humphrey, 2012). Other approaches recommend that core components be implemented with more fidelity than non-essential ones (Durlak & DuPre, 2008). Programs may also be differentiated according to their level of specificity, with those that are outlined in detail requiring higher levels of implementation fidelity (Lendrum & Humphrey, 2012).

Collecting data on fidelity and adaptation while the intervention is being implemented provides information on how the program components were delivered. In turn, these measures can be associated with program outcomes to understand how fidelity and adaptation affect them. This can help researchers have a better understanding of the mechanisms of change behind the program by seeing what the delivered intervention really looked like on the field. By recording how the implementers have adapted their intervention to their context, program designers can modify the intervention to fit educational communities' needs better, potentially achieving higher impacts for students. All while allowing for agency and input from educational professionals who own the intervention.

Methodological Aspects Related to the Measurement of Implementation

The study of implementation is complicated, as it must consider many aspects of program delivery. Consequently, researchers who wish to assess implementation must make methodological decisions in the different stages of program design and delivery (Durlak, 2015).

By the 1980s, there was agreement on the importance of measuring program implementation in educational research. However, Scheirer and Rezmovic (1983) acknowledged the absence of "standard methodological paradigms for constructing implementation measures," instead noticing that researchers "tended to create ad hoc implementation indicators consistent with their intuitions or their research budgets" (p. 600). More recently, other authors that examine the evolution of the assessment of program implementation have found inconsistencies in the operationalization of relevant terms for implementation (e.g., fidelity) in the data sources used and the methods used to analyze the data. Furthermore, they have identified no consensus on the way in which fidelity of implementation is described and reported in the educational research literature (Century & Cassata, 2016; Hansen, 2014).

However, it is unlikely that evaluators will agree on a unique way of conceptualizing and measuring implementation. Educational programs are multidimensional and complex and consist of several components that are implemented in different contexts, all of which negate the possibility and convenience of developing a standardized way to measure implementation across settings and interventions (Durlak & DuPre, 2008; O'Donnell, 2008). Given these complex and variable conditions, the recommended practice is to employ multiple data collection methods and sources to evaluate the implementation of a program (Mowbray et al., 2003).

Although there is no agreed-upon methodology, researchers have created guidelines that recommend the most appropriate steps to follow when measuring program implementation. Century and Cassata (2016) summarize these guidelines into seven steps that outline the methodological decisions researchers must make to measure program implementation. In their view, step one is to determine the core components of the program. In contrast, in step two, program designers should determine fidelity benchmarks "or expectations for component enactment" (p. 193). The third step involves the development of the theoretical causal model that links the program components with their expected outcomes, as well as the variables that are expected to mediate the relationship between components and outcomes. Next, researchers should specify the methods and data sources that they will use to measure each core component and then select the appropriate time frame for data collection. In the sixth step, Century and Cassata recommend ensuring that the data collected is reliable and valid. Finally, step seven focuses on determining how the data will be summarized or reduced for its analysis.

Constructs to Operationalize Implementation Fidelity

Program implementation needs to be operationalized into measurable constructs that can be estimated and evaluated. Durlak and DuPre (2008), relying partly on the work of Dane and Schneider (1998), identify seven aspects of program implementation: adherence, dosage, quality, participant responsiveness, program differentiation, monitoring of control and comparison conditions, and program reach.

Table 2: Constructs to Measure Implementation Fidelity

Construct	Description
1. Adherence	Extent to which the intervention corresponds to the originally intended program
2. Quality	How well the different program components have been conducted
3. Dosage	How much of the original program was delivered
4. Program reach	Rate of involvement and representativeness of program participants
5. Participant responsiveness	Degree to which the program stimulates the interest or holds the attention of participants
6. Program differentiation	Extent to which a program's theory and practices can be distinguished from other programs
7. Monitoring of control/comparison conditions	Nature and amount of services received by members of each group

Sources: Adapted from Dane & Schneider (1998); Durlak & DuPre (2008)

Adherence. In the specific case of fidelity, program theory should be clearly specified, as a detailed theory and well-defined model contribute to the development and identification of fidelity criteria and scales. Measuring fidelity can be increasingly difficult in the case of complex programs that depend, for example, on practitioner decision-making, on adapting components to individuals' needs, or on the coordination of multiple services. Furthermore, to measure fidelity, there must first be a precise specification of an intervention that serves as a benchmark to compare what is enacted in the field to an "ideal" implementation. Interventions that have already been tested and proven successful (e.g., through efficacy studies) can use this implementation under controlled conditions as a model to compare replications or extended implementations (e.g., through effectiveness studies; Mowbray et al., 2003).

Adaptation. Adaptation contrasts with the prescriptive approach of fidelity, as it explores the ways in which the intervention was enacted in practice without having an ideal of implementation to be measured against. A line of researchers considers that modifications are inevitable in complex social interventions, so they must be addressed and measured. Nonetheless, not all types and levels of adaptation are deemed desirable or productive. Researchers discuss the limits of these adaptations and the effects they have on internal validity, as deviations may limit the ability to determine the effectiveness of a specific program.

Quality. Implementation research can also be studied from the perspective of quality, looking at how well the components were delivered. This relevant construct can be operationalized in different ways, which include collecting data to determine whether intervention delivery was conducted with skill and understanding (e.g., looking at facilitator enthusiasm, clarity, and teaching techniques) or if it approximates an external theoretical ideal of delivery (Hansen, 2014; Moore et al., 2013). Within the fidelity framework, quality may be equated with fidelity, as the original intervention design can be seen as the standard of quality that the implementation should achieve (Carroll et al., 2007).

Dosage. Another relevant construct is dosage, which is concerned with the quantity and strength of the intervention. Dosage can be measured in terms of the prevalence, frequency, intensity, and duration of services received (Durlak & DuPre, 2008; M. J. Weiss et al., 2014). This construct can be conceptualized from the perspective of what the beneficiaries receive or what is delivered to them. For the former, in the context of education, dosage can be measured as the number of lessons students are exposed to using a curriculum intervention or the number of professional development sessions teachers attend. In the case of the delivery, dosage can be operationalized as the number of lessons teachers delivered using the new curriculum or the number of professional development sessions that the research team delivered to teachers (Century & Cassata, 2016; Hansen, 2014; Hill & Erickson, 2019). It can also be defined as "teacher opportunity to learn or to use program materials" (Hill & Erickson, 2019, p. 591).

Program reach. A closely related construct is program reach, which assesses participation rates or program scope. This can provide insights, for example, into how many of the eligible people received the intervention, considering their willingness to participate or the program's capacity to reach them. While dosage is concerned with the program's intensity in the services offered, reach is concerned with collecting data on whether the intervention was able to reach its potential adopters and beneficiaries (Century & Cassata, 2016). Reach can be operationalized as the proportion of the eligible population who participated in the intervention and their characteristics (Durlak & DuPre, 2008).

Participant responsiveness. This construct intends to determine the level of interest the program generates. The degree of responsiveness depends on beneficiaries' acceptance of the program and the program's acceptability. Responsiveness is then contingent on the intervention's relevance for beneficiaries and their perception of the intervention's pertinence for their lives (Carroll et al., 2007). This is closely related to Rogers' (1983) diffusion of innovations model, particularly to the attribute of compatibility and how it affects the program's rate of adoption. Compatibility is defined as "the degree to which an innovation is perceived as consistent with the existing values, past experiences, and needs of potential adopters" (p. 223). Rogers asserts that higher perceived compatibility is associated with higher levels of adoption. In the case of school-based interventions, an example of the operationalization of responsiveness can look at teachers' engagement in the program as they deliver the lessons or students' interest as the final beneficiaries (Century & Cassata, 2016).

Program differentiation. The seventh construct is program differentiation, which intends to identify the program's core components to determine which elements are essential for its success. This enables researchers to find which components are associated with larger changes in outcomes, which in turn can help focus the efforts of implementation support or monitoring implementation in future replications or intervention scale-ups. From a fidelity perspective, researchers may want to ensure that these components are implemented with higher levels of fidelity (Carroll et al., 2007; Dusenbury, 2003).

The study of treatment and control conditions can be applied to studies based on experimental or quasi-experimental designs, where researchers are interested in comparing those who have been exposed to the intervention with those who have not. Measuring this construct can help researchers ensure that the beneficiaries in an experimental condition received the planned intervention (Dane & Schneider, 1998). Conversely, researchers should evaluate what is happening in the control condition to determine that they are not exposed to an alternative intervention. Some of the relevant constructs to investigate include treatment contamination, usual care, and alternative services (Durlak & DuPre, 2008).

From the perspective of fidelity, these constructs cover aspects of structure (framework for service delivery) and process (the way in which services are delivered; Mowbray et al., 2003). In consequence, adherence and dosage can be grouped under fidelity to structure; quality, program differentiation, and treatment and control conditions belong to fidelity to process; and participant responsiveness and reach take on characteristics of both (O'Donnell, 2008). Other researchers have separated dosage into a different category from structure and process, indicating that it is meant to record whether a program was accessible to those meant to implement it (Hill & Erickson, 2019).

Since implementation is a multifaceted concept, the recommendation is generally to assess more than one aspect of the same program (Durlak & DuPre, 2008). Research focused on how implementation is conceptualized and measured in educational contexts suggests that program evaluations use a combination of these constructs to assess implementation (Dane & Schneider, 1998; Hill & Erickson, 2019; Mowbray et al., 2003; Ruiz-Primo, 2006).

Instruments to Collect Data on Implementation

Types of Instruments. A vast range of data sources can be employed to collect the data to analyze the constructs presented in the previous section (Century & Cassata, 2016; Hansen, 2014; Hill & Erickson, 2019; Ruiz-Primo, 2006; Scheirer & Rezmovic, 1983). The following section will present some of the instruments and methods that can be used to obtain information on program enactment. Additionally, it will consider issues such as validity, reliability, and timing of data collection, which can affect the quality of the data gathered.

Broadly, data collection methods for program implementation can be classified as direct observations conducted by expert raters, user interviews, self-reported surveys, and collection of institutional records or document analysis (Century & Cassata, 2016; Ruiz-Primo, 2006). Although many studies combine qualitative and quantitative data (Mowbray et al., 2003), this dissertation will focus on quantitative methods or methods coded quantitatively since the analysis looks at how implementation can be incorporated into models that quantitatively estimate the impact of educational programs.

An argument in favor of document analysis and observations conducted by independent raters is their perceived “objectivity”, particularly when compared to other methods. Ruiz-Primo (2006), relying partly on Scheirer and Rezmovic (1983), orders the different methods used to measure the degree of fidelity of implementation based on four categories: extent of judgment (based on the level of objectivity of the method); directedness (degree to which the method directly captures implementation); sensitivity (extent to which the method can detect the behavior of providers and participants); and alignment to the program (degree to which the instruments are able to detect program characteristics). In descending order, document analysis, rating scales, and direct observations (rating scales and checklists) are considered the most objective, direct, and sensitive instruments.

Direct observations are the most used way to collect information on program implementation, and they have frequently been considered the preferred method (Fullan & Pomfret, 1977; Ruiz-Primo, 2006). These types of instruments are considered “the most direct measures of practice, the most rigorous, and the most objective with respect to implementation quality, compared to self-reported data” (Century & Cassata, 2016, p. 193). Hill and Erickson (2019) corroborated the preference for direct observations when evaluating program implementation. These authors found that out of 65 studies that reported fidelity in IES grants awarded from 2002 to 2011, classroom observation was the most frequently used method (46% of projects), followed by teacher self-reports typically in the form of logs and surveys (29%). A smaller proportion used both methods (18%).

Observations conducted by external raters may be used to assess different constructs of program implementation, such as fidelity, quality, dosage, or participant responsiveness. For example, they can be used to evaluate the general quality of teaching practices, to rate teachers’ effectiveness and enthusiasm, or to measure student engagement. Furthermore, observations are used to measure adaptations, as observers can document the amounts and types of additions and modifications made by teachers during their lessons. Many studies rely on ad hoc checklists that trained observers complete when they visit the implementation sites; these checklists are intended to assess the delivery of specific program components or activities (Dusenbury et al., 2005; Hansen et al., 2013; Hansen, 2014).

Even though observations may be preferred to other methods for their perceived objectivity, they have important practical drawbacks. Observations are considered expensive and time-consuming, and their feasibility is compromised with large samples. Additionally, observations may not be the most suitable data collection method to measure some program components or constructs. For example, in the case of dosage, observers may not be present for all the lessons students participate in, so observations would not provide the most accurate information on how much time was devoted in each lesson to the intervention. Finally, observations may not be precise enough, or people may act differently if they are being observed (Century & Cassata, 2016; Hansen, 2014; Ruiz-Primo, 2006).

Surveys can provide information on participants' perceptions or experiences with the program. Teacher and student self-reports can be used to assess participant engagement, quality, or dosage. The advantages of this type of instrument lie in their relatively low cost compared to other data collection methods (e.g., observations) and in the proximity of the informant with the program (e.g., students are experiencing the implementation of the intervention the most). Additionally, since surveys may reach more participants and occasions, they may be the only instrument that provides a sample large enough to achieve appropriate statistical power (Century & Cassata, 2016; Hansen, 2014).

In contrast with observations, self-reports can be regarded as more problematic as they rely on an individual’s account of their own behavior. Researchers have manifested their concern with the potential inaccuracy of these methods, as they can be more prone to social desirability bias if participants or those delivering the intervention inflate their scores. Additionally, a lack of variability in scores is not informative and cannot be used to analyze the effects of implementation on outcomes (Century & Cassata, 2016; Fullan & Pomfret, 1977). The same can be said for a large variability when assessing a single practice or behavior (e.g., students reporting on their teacher’s instructional practices).

Document analysis and surveys can also help researchers collect information related to the individuals exposed to the control condition. Follow-up surveys can be conducted on a random subsample of people, providing additional data on the situation for those sites that were not closely monitored as part of the treatment group. Administrative records are also a less expensive way to gain insights into those sites (M. J. Weiss et al., 2014).

Each data collection instrument has advantages and disadvantages. To overcome these shortcomings, the general recommendation is to use a multimethod approach to collecting data on program implementation. Combining methods can enhance measurement quality, provide a broader view of the implementation process, and help reduce bias (Century & Cassata, 2016; Durlak & DuPre, 2008; Dusenbury et al., 2005; Hansen et al., 2013; Mowbray et al., 2003; Ruiz-Primo, 2006, 2006; Scheirer & Rezmovic, 1983).

Methodological considerations. Another important concern that researchers should address when studying implementation is that the instruments to collect data and measure the relevant implementation constructs are reliable and valid for their intended use (Durlak & DuPre, 2008; Mowbray et al., 2003; O’Donnell, 2008). However, research on implementation shows that the psychometric properties of fidelity measures are rarely reported (Century & Cassata, 2016; Scheirer & Rezmovic, 1983), and when they are, the information may be questionable (Lendrum & Humphrey, 2012). Those that do report these properties mainly provide information on the following measures of reliability: interrater or interobserver agreement, assessment of intraclass correlations among raters, and test-retest reliability. Validity is often observed by examining the internal structure of the data through measures of internal consistency or confirmatory factor analysis (Century & Cassata, 2016; Mowbray et al., 2003; Scheirer & Rezmovic, 1983).

The timing of data collection is also a relevant aspect that must be taken into consideration when evaluating program implementation. Firstly, the intervention should be ready to be assessed, and enough time must pass to allow for its effective implementation. Other relevant considerations include the frequency with which these instruments should be administered, which must consider the structure of the program components and their recurrence. Considering the complexity of educational interventions that have multiple components, experts recommend that implementation should be measured at more than a single time point, but the exact frequency and number of measurements will depend on the specific characteristics of the intervention and its components (Durlak & DuPre, 2008; Lendrum & Humphrey, 2012). Program evaluators should think this through carefully, as measuring too frequently may result in unmanageable and uninterpretable data and unnecessary use of resources (Mowbray et al., 2003).

Measurement Issues in Data Reduction

Once data on implementation is collected, researchers must condense the information so it can be analyzed, entailing a series of methodological decisions (Mowbray et al., 2003). The framework used to assess implementation affects the methodological decisions related to data reduction (Durlak & DuPre, 2008). From a fidelity perspective, researchers are required to define an ideal of implementation that will be used to compare the different implementation constructs and components of the enacted intervention (Durlak, 2015; Mowbray et al., 2003; Ruiz-Primo, 2006). This is not necessary from an adaptation framework because the focus is on implementation as it happens, not on a theoretical ideal. Literature on program implementation presents two primary approaches to data reduction: creating indices or categories (Century & Cassata, 2016).

The level of aggregation may vary depending on the characteristics of the program that is being evaluated. In the specific case of curriculum interventions, there are several potential levels at which the data can be aggregated: classroom, teacher, school, district, etc. O'Donnell (2008) observed that most of the literature aggregated data at the classroom level.

Indices. One alternative to condensing the implementation data collected during program enactment is to generate indices or scales that quantify the level of implementation reached in practice. Researchers may choose to compute indices by intervention component or implementation construct or by a combination of these two dimensions. Alternatively, they may opt for an overall implementation index that gathers data from all the intervention components or implementation constructs into a single indicator.

Generating a single score for all components or constructs brings together all the data into a single number that can be easily interpreted. Nonetheless, this simplicity dilutes the nuances in implementation across components and constructs. This may lead to different intervention components, implementation constructs, or whole interventions that were implemented differently in practice to receive the same score, thus obscuring their differences (Mowbray et al., 2003). Consequently, it may be preferable to construct different indices for each major component of the program or each implementation construct being measured. This may pose other issues related to the complexity of the measurement model and the interpretation of the results of the intervention for each implementation index (Century & Cassata, 2016).

Hulleman and Cordray (2009) explore two different ways to construct an implementation index: as a percentage of the ideal implementation (absolute fidelity index) or as an independent score (average implementation index). In the first approach, the index is conceived as an indicator of the level of the theoretical program that was achieved in practice. To construct the index, the authors operationalized the maximum value as the maximum possible response scores in each instrument used to measure the implementation of the treatment and control groups separately. The index was then constructed by dividing the average observed score across all instruments by the theoretical maximum, generating a fidelity index for the treatment group and another for the control group. The values were rescaled from 0 to 100 to create a percentage of adherence.

The second approach (average implementation index) used the same method of averaging across treatment and control conditions but did not use an ideal implementation benchmark as a denominator. The scale of this index is directly tied to the scale of the instruments used to measure implementation, which in this case ranged from 0 to 4 points (Hulleman & Cordray, 2009). This approach can be used to determine the level of implementation for each component, to understand which aspects of the program are linked to the outcomes, or what level of the intervention is enough to produce changes in the beneficiaries. It is important to note that researchers do not have to choose between indices with or without a benchmark (fidelity), as implementation data can be used to compute both types of statistics. However, in practice, literature on program implementation has observed that studies tend to use the former approach more frequently (Century & Cassata, 2016).

Categories. In the second approach, researchers create and assign implementation categories to the constructs, components, or interventions. This differs from the index approach, as the variables designed to reduce the implementation data are not continuous but categorical or dichotomous.

As with fidelity indices, an alternative is to create categories that designate the level of implementation achieved with respect to an ideal of enactment. A common way to operationalize this is to generate implementation levels (e.g., low, medium, high). These categories can be arbitrary, and they may not transfer from one study to another (Durlak & DuPre, 2008).

One example is the concept of *innovation configurations*, which is defined as “the operational patterns of the innovation that result from selection and use of different innovation component variations” (Hall & Loucks, 1978, p. 9). Innovation configurations take into account program adaptations and generate categories of implementation based on the differences between types of intervention delivery. The categories emerge by identifying common variations across the relevant units (e.g., schools, teachers). To construct the typologies, the researcher must use interviews and checklists to record which critical and related components are being used.

The creation of categories of implementation (two or more) can be problematic when measuring program effects. By design, the control group did not implement the intervention, but it would have received an implementation score if it had implemented the program. For example, when using a dichotomous indicator, schools in the control group will likely have “potential higher implementers” and “potential lower implementers” if they had been assigned to the treatment group. In this case, it would be more appropriate to compare high and low implementers in the treatment group with an expected comparison in the control group and not with the complete control group (Unlu et al., 2016).

Hulleman and Cordray (2009) employed a different approach that constructed an implementation indicator that modeled program implementation elements in both the treatment group and control group. For the treatment group, data on implementation was derived from the enactment of the intervention, but for the control group, the aspects of implementation were theorized to occur naturally. This process generates a treatment-control contrast that resembles the hypothesized one created through random assignment. The authors specified a minimum cut-off point that designated a level of fidelity of implementation that was enough to conclude that participants were exposed to the intervention. The cut-off point was determined empirically from the distribution of scores observed in a previous implementation of the intervention,

Modeling Implementation Data

Since the implementation of educational intervention is likely not homogeneous across all sites, this variation should be considered in models that estimate program effects. Although the evidence presented above indicates, this is an important practice (or that, at least, it should be considered in the analysis), “conventional analytic methods that have been used to address these types of questions have been limited and flawed” (Unlu et al., 2016, p. 100).

Variables related to implementation may be correlated with different outcomes to determine if any implementation constructs are associated with larger effects (Century & Cassata, 2016). If the implementation variable is continuous (indices), the level of the variable is associated with the outcome. For categorical variables, ANOVA can be used to compare the results of each implementation group (Hall & Loucks, 1978; O’Donnell, 2008). Although both approaches can provide information on the relationship between implementation constructs and program outcomes, the first approach has more statistical power (Durlak & DuPre, 2008). If implementation data is available only for the treatment group, then these associations will only apply to the group exposed to the intervention. Some authors have manifested concern over these methods, as they are likely reflecting pre-existing differences between treatment group members that were precisely what led to variations in implementation (Unlu et al., 2016).

Correlational analysis can be conducted to assess the relationship between the information on the implementation of specific components. For example, a variable containing an index of fidelity in implementing a component can be correlated with a measure of dosage to determine whether higher levels of fidelity are associated with increased dosage. This can be very relevant for potential replications of the intervention, as it can guide implementers on the essential program components (Durlak, 2015).

Researchers commonly use multiple regression to estimate the effects of an intervention on the program outcomes. In these models, individual factors that affect implementation may be conceptualized as moderating variables, as they can affect the direction or strength of the relationship between the predictor variables (i.e., program components) and the expected outcomes. In this sense, the impact of the program depends on the levels of the moderator variables (Century & Cassata, 2016; M. J. Weiss et al., 2014). Variables can be used as moderators under the condition that they are observed prior to the intervention and that they are not affected by its implementation (Century & Cassata, 2016; Raudenbush & Bloom, 2015).

The most used moderating variables are end-user characteristics (e.g., gender, age, educational level, race/ethnicity, social background, risk of failure, etc.). Other variables that may be observed less often include those related to the context in which the program is implemented (M. J. Weiss et al., 2014). Nonetheless, no matter what type of moderating variable is incorporated into the model, it is infrequent to “find an evaluation that is founded on an explicit moderation theory" with an a priori hypothesis of how the program may affect individuals differently (Raudenbush & Bloom, 2015, p. 477).

Within a fidelity framework, adherence is commonly used as a moderator to understand how fidelity to the original program design affects outcomes (Dusenbury, 2003; Hansen, 2014; Hansen et al., 1991; Raudenbush & Bloom, 2015). Furthermore, constructs like quality of delivery or participant responsiveness can also be used as interaction terms as they influence the level of adherence program enactment can achieve (Carroll et al., 2007). Implementation variables can be held constant to determine if the outcome varies at the same level of a specific level of an enactment construct (Blakely et al., 1987). In the context of experimental designs, the dichotomous variable that indicates assignment to treatment ("intent to treat") can be replaced with an implementation index score to obtain a more precise estimate of the "treatment on the treated" (Cook, 2005).

Data on fidelity can also be incorporated into models that measure intervention effects as mediators or variables that are affected by the intervention and that then influence the outcomes. These variables can be defined as short-term outcomes that the intervention modifies and that are then theorized to work in a causal chain to affect outcomes (Hansen, 2014; Hansen et al., 1991; O'Donnell, 2008; Stein et al., 2008). For example, using an experimental research design, LoCasale-Crouch et al. (2018) examines whether fidelity of implementation (measured in the treatment and control groups) mediates the impact of the Banking Time program, which is intended to improve the quality of teacher-child interactions. However, it is hard to draw valid causal inferences from this relationship, as there may be many confounding factors that affect implementation quality that may not necessarily be related to the intervention (e.g., teachers' previous experience and skills; Raudenbush & Bloom, 2015).

For educational interventions delivered at the classroom level, the most appropriate model will consider the nested structure of the data. These models allow researchers to analyze data at different levels, such as classroom-level data with student-level outcomes. Importantly, these methods also provide information on between-site variation in implementation constructs, showing the heterogeneous nature of intervention enactment in educational settings (Mowbray et al., 2003).

Data on implementation can be analyzed to provide important insights into the mechanisms of change that generate changes in outcomes. People in charge of implementing the intervention can empirically determine whether a combination of implementation constructs and program components is expected to produce a larger effect on the outcomes of interest.

Chapter 2: Conceptualizing, Measuring and Accounting for Implementation in IES-Funded Studies

In this chapter, I examine how program implementation is accounted for in educational research and evaluation. To do this, I examine the literature, focusing specifically on conceptualizations of implementation and related constructs, methods, and instruments used to collect evidence of implementation in classrooms and schools. I also investigate the analytic models used to synthesize implementation data and associate it with program outcomes.

Introduction

Educational programs are inherently multidimensional and typically comprise a variety of components implemented across diverse classrooms and schools. This complexity poses significant challenges to establishing a standardized methodology for conceptualizing and measuring implementation fidelity (Durlak & DuPre, 2008; O'Donnell, 2008). The current state of knowledge reveals considerable inconsistency and uncertainty without consensus on the appropriate ways to conceptualize and measure implementation. This uncertainty underscores the importance of reviewing existing literature to identify prevailing practices, challenges, and gaps in measuring program implementation.

The objective of this chapter is to examine how program implementation has been conceptualized and accounted for in the literature that quantitatively evaluates the impacts of educational programs. Specifically, the chapter seeks to address the following research question:

3. How is program implementation conceptualized and accounted for in the educational research literature?
    a. What are the frameworks and constructs used to conceptualize implementation?

b. What data collection instruments and procedures are used to measure program implementation?

c. What methods are used to reduce implementation data?

d. What methods are used to account for implementation data when estimating intervention effects?

I explore these questions in the context of grants funded under the Institute of Education Sciences, delving into the publications generated from the research associated with these studies. Addressing these objectives contributes to a deeper understanding of the complexities involved in implementation measurement and helps generate better evaluation practices for educational interventions.

Methods

To investigate these research questions, I conducted a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page, McKenzie, et al., 2021; Page, Moher, et al., 2021). The PRISMA guidelines provide "a transparent, complete, and accurate account of why the review was done, what [the authors] did (such as how studies were identified and selected) and what they found (such as characteristics of contributing studies and results of meta-analyses)" (Page, McKenzie, et al., 2021, p. 1).

I opted to use a systematic analysis and not a meta-analytic approach to synthesize the results across studies and to explore the relationship between fidelity and student outcomes. Previous efforts, such as those by Hill and Erickson (2019), have sought to standardize and synthesize implementation measures across educational interventions through a meta-analysis. Although this could have provided more statistical evidence regarding the relationships between implementation and educational outcomes, the literature pointed towards a wide variation in the instruments and constructs used to measure implementation. This introduces significant restrictions on the analyses that can be conducted, as it is complicated to aggregate data from different studies. Additionally, having the grants as the unit of analysis provides a small sample size, limiting statistical power and, therefore, the possible analyses. Furthermore, access to complete datasets that contain data on program implementation (including fidelity and outcome measures) is very limited.

Sample

The review includes interventions financed under the Institute of Education Sciences' (IES) Education Research Grants Program (CFDA 84.305A), which is part of IES' National Center for Education Research (NCER). NCER finances studies that aim to improve education outcomes and access to education opportunities. Its goal is "to identify what works for whom, in what context, and why in order to provide reliable information about how to improve education outcomes and narrow achievement gaps for U.S. students" (Institute of Education Sciences, 2020, p. 1).

The sample of eligible studies encompasses Education Research Grants' funding Goals Two (Development and Innovation), Three (Efficacy and Replication), and Four (Effectiveness),[1] as these are the funding streams that support the evaluation of interventions that intend to positively affect student outcomes when implemented in "authentic education settings" (Institute of Education Sciences, 2019, p. 55, 2019, p. 64, 2019, p. 80). Goal Two aims explicitly to finance the "development of new interventions and the further development or modification of

[1] The names of goals 3 and 4 have changed slightly over time. For example, Goal 3 was called *Efficacy and Replication* until 2019, when it was renamed to *Efficacy and Follow-up*. Goal 4 began as *Effectiveness Evaluations* and then was renamed *Scale-Up Evaluation* (2008-2012), *Effectiveness* (2013-2018), and *Replication: Efficacy and Effectiveness* (2019).

existing interventions" (Institute of Education Sciences, 2019, p. 55). Goal Three, on the other hand, "supports the evaluation of fully developed education interventions that have not been previously evaluated using a rigorous design" (Institute of Education Sciences, 2019, p. 64). Finally, Goal Four provides "evidence on education interventions that have been shown by prior rigorous research to produce positive impacts on student outcomes" (Institute of Education Sciences, 2019, p. 80). Goals One (Exploration) and Five (Measurement) were excluded from the analysis as they do not focus on evaluating interventions or consider an implementation component (Institute of Education Sciences, 2019).[2]

The research funded under Goals Two, Three, and Four is relevant for the analysis in the present study as it is focused on evaluating the effect of a specific intervention, requiring researchers to establish a causal link between the program and student outcomes that meets the What Works Clearinghouse (WWC) standards (Institute of Education Sciences & National Science Foundation, 2013). Eligible studies for the WWC, in turn, include randomized controlled trials (RCTs), quasi-experimental designs (QEDs), regression discontinuity designs (RDDs), and single-case designs (SCDs; What Works Clearinghouse, 2022). The requirements for implementation in authentic education settings and for an adequate sample size to detect meaningful changes in outcomes mean that the interventions are implemented in multiple schools and include several classrooms and teachers, leading to potential variability in program delivery.

The Requests for Proposals (RFPs) for the Education Research Grants Program indicate within their methodological requirements that researchers must examine fidelity of implementation. In the RFPs, *fidelity of implementation* is defined as "the extent to which the

[2] Since 2020, the numbered *Goal* structure has become lettered *Project Types*, with C (*Development and Innovation*) and D (*Initial Efficacy and Follow-Up*).

intervention is being delivered as it was designed to be by end users in an authentic education setting" (Institute of Education Sciences, 2015, p. Glossary ii, 2016, p. Glossary ii, 2017, p. Glossary ii, 2018, p. Glossary iii, 2019, p. Glossary iii). Measuring fidelity of implementation has been a requirement for IES grants from 2007 to 2020 for RFPs financed under Goal Three. RFPs for Goals Two and Four do not mention fidelity until 2011 and 2013, respectively; from then on, it is part of the requirements.

I consider grants awarded between 2007 and 2020, as IES established the Education Research Grants competition in fiscal year 2007 (U.S. Department of Education, personal communication, February 2, 2024). The closing year (2020) provides enough time for the grant to have publicly available evidence of its implementation as of 2024. 807 IES grants meet the eligibility criteria (awarded under Goals Two, Three, or Four between 2007 and 2020), constituting the universe from which I selected the studies for this review.

Study Eligibility Criteria

Establishing eligibility criteria to ensure the inclusion of relevant, high-quality studies is essential for a systematic review. Table 3 summarizes the specific eligibility criteria.

Table 3: Summary of Study Inclusion and Exclusion Criteria

Dimension	Inclusion criteria	Exclusion criteria
Topic	- Intervention based on curriculum or instructional strategies. - Any subject.	- Does not include curriculum or instructional component.
Program Design	- Activities or components are explicitly laid out.	- Loose set of instructional goals only.
Target Grades	- PK–12.	- Pre-school only (before pre-kindergarten). - College students. - Adult learners.
Program Delivery	- Delivered by students' teacher(s).	- Delivered only by people external to the school or school administrators, etc.

Dimension	Inclusion criteria	Exclusion criteria
Level of Implementation	- Classroom intervention with teacher-delivered component(s).	- Implemented and measured only at school level.
Mode of Implementation	- Delivered in person or online with teacher as mediator.	- Based exclusively on technology.
Target Student Population	- Intended to benefit the general student population.	- Intended to benefit students with physical or cognitive disabilities or other special populations. - Intended to benefit incarcerated students.
Outcomes	- Effectiveness measured through change in cognitive or academic student outcomes.	- Focused only on non-cognitive or non-academic outcomes. - Not measured at student level.

Eligible studies for this review include those with a curriculum-based component or set of instructional strategies. However, not all the components in the intervention must be focused on curriculum or instruction for the study to be eligible. The selection process excludes interventions based on models for school change or those that focus only on student assessment, career training, or changes in behavior (e.g., providing information to change students' behavior). There are no restrictions regarding the subject that the curriculum intervention can address.

Since the analysis focuses on implementation, eligible interventions must have an explicit design, with components, activities, or lessons laid out for the implementer to follow. These activities may be developed exclusively by a research team or in conjunction with teachers or schools. I do not consider interventions that do not outline their activities (i.e., those that contain only a loose set of instructional goals) as part of the analysis, as it is not possible to incorporate them into the fidelity/adaptation framework.

Selected interventions include those implemented in school settings in one or more grades from pre-kindergarten to 12th grade. This excludes programs delivered exclusively in preschools before pre-kindergarten (early childhood education or infant and toddler care) and interventions delivered in higher education institutions (Community Colleges, Colleges, Universities, and Vocational or Technical Schools) and adult education settings.

The interventions in the analysis must be delivered by the student's teacher of record or the teacher who regularly teaches the subject in the school. This criterion aims to include educators who interact with students regularly and are not directly linked to the research team. The review excludes interventions delivered only by classroom aides, as they are not the primary educators in charge of the class and may not be credentialed education professionals. As the analysis intends to identify variations in implementation, the selection does not consider interventions delivered exclusively by research team members, external tutors, external expert teachers, or others hired solely to deliver the program. This group may be more incentivized to abide by the original program design, as their main purpose in the school is to deliver the intervention.

The IES grants eligible for the analysis must evaluate an intervention implemented in classrooms. The program must contain at least one teacher-delivered component, although not all components must be delivered by teachers inside a classroom. I excluded interventions delivered by other school personnel (e.g., school administrators) or implemented only at the school level (without any classroom-level components). The selection criterion for programs based solely on delivering professional development to teachers depended on the study's design. I included grants that measured classroom practices or teacher-student interactions but excluded those without data on teacher implementation.

Eligible interventions can be delivered in person or online, with the teacher acting as a mediator or facilitator. Studies that rely solely on technology, such as computer or online tutors or games, web-based applications, or any computer-based methods without teacher mediation, are excluded from the analysis.

I incorporate studies intended to serve the general student population, including English-language learners (ELLs) or students at risk of academic disabilities. However, I exclude interventions meant exclusively for students with physical or cognitive disabilities or other special populations (e.g., students diagnosed with autism spectrum disorder (ASD), attention-deficit/hyperactivity disorder (ADHD), or visually or hearing-impaired students, etc.).[3] Additionally, studies implemented in juvenile correction settings intended for incarcerated students are omitted from the analysis. Although these are very relevant topics, I decided to exclude them because their implementation may entail challenges different from those faced by interventions delivered to the general student population, and these differences are not the focus of this dissertation.

The effect of the eligible interventions must be measured through a change in a cognitive student outcome, assessed through a quantitative instrument. These instruments may have been previously developed and validated (e.g., state academic assessments or other pre-existing instruments), or they may have been designed exclusively for the intervention. I exclude studies focusing solely on improving non-cognitive or non-academic student outcomes, such as school retention, student behavior, or health. Other ineligible grants only measure outcomes at the classroom level, for example, focusing on improvements in classroom quality (e.g., measured using the CLASS).

[3] The studies intended to benefit these student populations are financed under the National Center for Special Education Research's (NCSER) Special Education Research Grants Program (CFDA 84.324A).

Search Strategy

The first step of the data collection process is to download the list of all grants awarded under the NCES program between 2007 and 2020 from the IES website (https://ies.ed.gov/funding/grantsearch/), with 847 grants screened. Next, I select the grants from Goals Two, Three, and Four, resulting in 533 grants. Using the grant abstracts, I could exclude 328 grants that do not meet the eligibility criteria, leaving 205 grants for document retrieval and review.

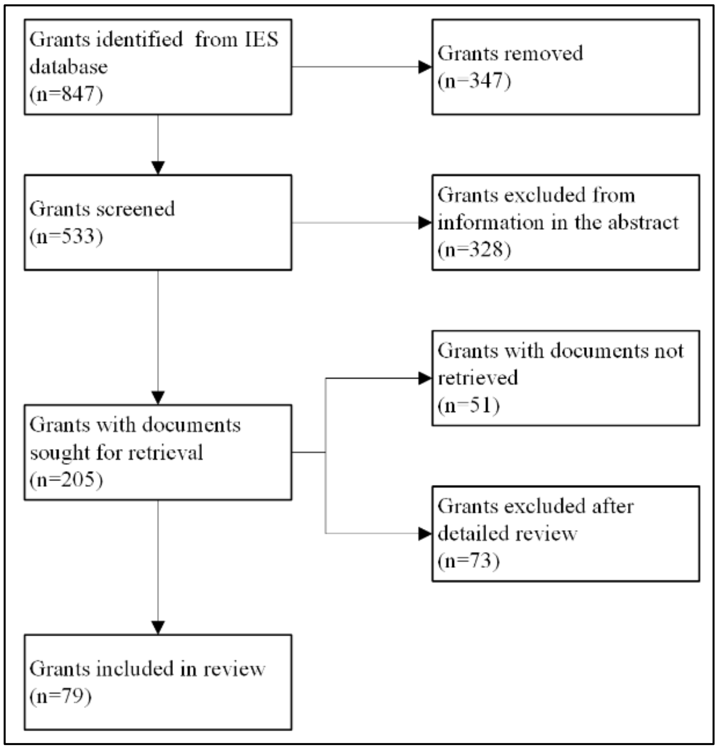
The document review includes academic publications and gray literature from eligible grants. Gray literature is defined as "produced on all levels of government, academics, business, and industry in print and electronic formats, but not controlled by commercial publishers" (Schöpfel & Farace, 2017, p. 1746). The term traditionally covers three categories of documents: conference proceedings, reports, and doctoral dissertations, which I include in the review.

I conducted the first part of the document search on the page for the grant on the IES website, where all grants awarded report a summary of the study. Some of these pages report the publications derived from the grant, which I use as a first approach to the document search.[4] I searched for documents online for all the grants, initially using Google and Google Scholar. The search was done sequentially, using keywords from each grant. First, I used the grant's award number (R305A#####), which returned documents in which the authors explicitly reported the funding source. Next, I searched the names of the Principal and Co-Principal Investigators in the grant, looking for matches between the author's name and the name of the program being evaluated. Finally, if the name of the intervention is reported in the grant summary, I use it to search for related publications. For 51 grants, I could not find any publications related to the

[4] Reporting publications is not a requirement, as many grants do not include them on their IES page.

evaluation of the intervention, so these grants are not part of the analysis. In the case of 73 grants, a more detailed review of the documents revealed that the intervention does not meet the eligibility criteria for this dissertation, so I removed them from the analysis. This leaves 79 eligible grants with documents that report the results of an evaluation.

Figure 4: Flow Diagram of Grant Selection


Grants identified from IES database (n=847)
Grants removed (n=347)
Grants screened (n=533)
Grants excluded from information in the abstract (n=328)
Grants with documents not retrieved (n=51)
Grants with documents sought for retrieval (n=205)
Grants excluded after detailed review (n=73)
Grants included in review (n=79)

The final review sample considers a total of 143 documents, including papers published in peer-reviewed journals (107), conference papers and posters (21), research reports (12), and others, including book chapters, documentation on data collection instruments, and websites (3).

Coding

The following step in a research synthesis consists of extracting the relevant information from the studies included in the review sample. This involves coding the data and using categories derived from the literature presented in Chapter 1 to classify the methodological decisions the researchers made when measuring program implementation in these studies. A summary of the characteristics and findings from the sample of studies that measure implementation is available in Table A 1 in the Appendix, and a list of the categories is included below (Figure 5).[5]

Figure 5: Coding Categories for IES Studies

1. Title
2. Year awarded
3. Study design
4. Implementation framework and constructs
5. Data collection instrument(s) to measure implementation
6. Method(s) to aggregate implementation data
7. Associates implementation data with outcomes
8. References

The first two categories (Title and Year awarded) provide general information to identify each study. The next category identifies the study's design, indicating whether it uses experimental (RCT) or quasi-experimental methods to estimate program impacts.

The fourth category describes the main framework used to assess implementation in each study (*fidelity* or *adaptation*) and the specific constructs it measures. The list of constructs comes from the aspects of enactment that are considered relevant in the literature and that are described in detail in Chapter 2. Durlak and DuPre (2008), relying partly on the work of Dane and Schneider (1998), identify eight aspects that comprise program implementation: adherence, adaptation, dosage, quality, participant responsiveness, program differentiation, monitoring of control and comparison conditions, and program reach. When the literature associated with the

[5] Table A 1 summarizes the data collected. The details of the information gathered are provided below.

IES grants does not explicitly identify these constructs, I derive them from the descriptions of the methods used to assess implementation.

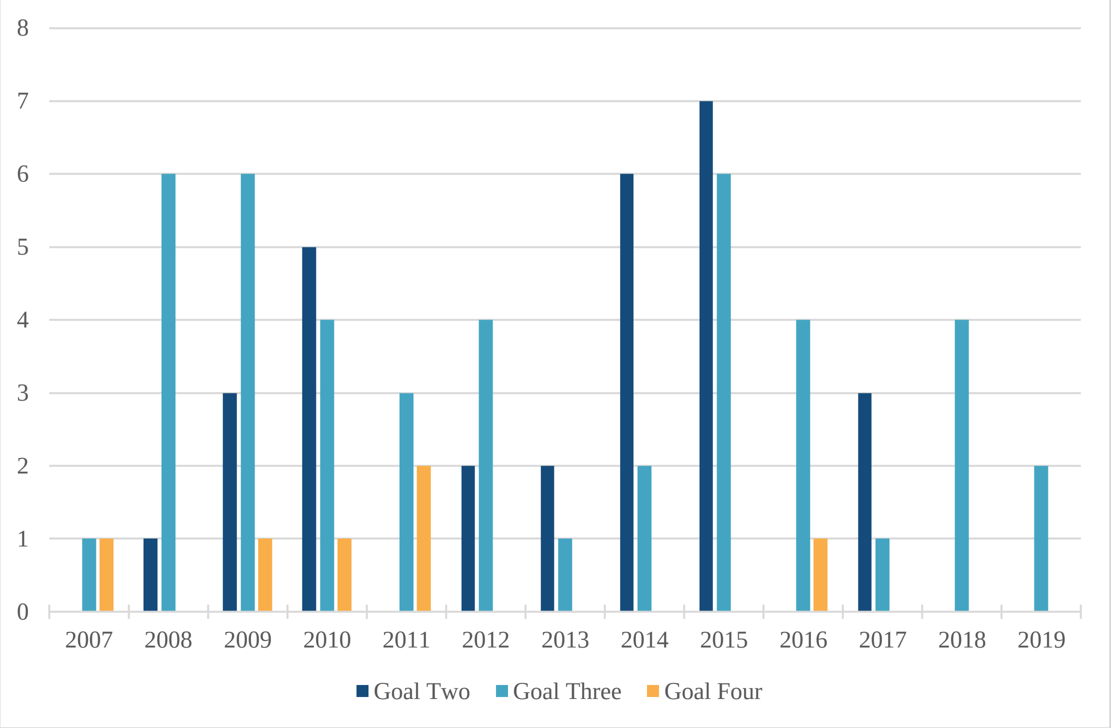
The next category lists the different instruments used to measure implementation (observation rubrics, surveys, logs, artifacts, or interviews). I gathered additional data for each instrument, including the implementation framework it is related to (fidelity or adaptation) and the type of information it collects (quantitative or qualitative). I also recorded each instrument’s target group (treatment or control groups, or both) and its origin (developed ad hoc for the study, adapted from a previous study, or used as-is from an earlier study).

The sixth category looks at the methods used to reduce implementation data, identifying if it is a simple aggregation (sum or average) or if it is based on latent variables. I collected additional data on the type of model used to generate the latent variables (e.g., confirmatory factor analysis, latent class analysis, etc.). The following category (7) indicates whether the study associates the data on implementation with student outcomes. I also classified the type of analysis, detailing the methods used (e.g., correlational analysis, regression analysis). Finally, I included a list of the relevant publications for each study.

Results

Of the 79 eligible grants, most belong to Goals Three (Efficacy and Replication), with 44 grants (56%), and Two (Development and Innovation), with 29 grants (37%). Goal Four (Effectiveness) had comparably fewer eligible grants, with only 6 (8% of the total). The distribution of eligible grants across years varies from 13 in 2015 to 2 in 2007 and 2019.

Figure 6: Number of Eligible IES Grants by Year and Goal



A high proportion of the eligible studies used experimental designs through RCTs to estimate the impact of the intervention on students (62 grants). Other studies used non-experimental methods, such as QEDs, RDDs, and SCDs (17 grants). Two grants used a combination of experimental and non-experimental methods. In one of these studies, the researchers used an experimental design in the first year of implementation and a non-experimental one (regression analysis controlling for baseline characteristics) in year two due to high levels of attrition experienced after year one. In the other study, a single-case design was used to compare two intervention versions, while researchers used experimental methods to compare one version to the business-as-usual condition.

Most of the research derived from the eligible studies measure implementation. Out of the 79 studies, 66 report measures of implementation in publications (85%), while 13 do not address implementation (15%). Grants awarded under Goal Two tend to focus less on measuring implementation, as there is no evidence of this in 7 of the 29 awarded grants (24%). The number of studies that do not measure implementation is smaller for Goals Three (five grants, 11% of all the eligible grants in this Goal) and Four (one grant, 17% of all the eligible grants in this Goal).

Of the 13 grants that do not report measuring implementation, nine are based on RCTs, while the remaining four use other pre-post comparisons to estimate intervention effects. Although this number is larger for RCTs, proportionally, more studies that do not report implementation use methods that are not experimental to estimate program impacts. 23.5% of the studies that use non-experimental designs do not measure implementation (four of 17 grants), compared to 14.5% for studies that use an experimental design (nine of 62 grants).

Fidelity

Table 4 presents the frequency with which each fidelity construct appears in the research related to the eligible IES studies that measure fidelity of implementation. Adherence is the most frequently measured construct, with 79% of the 66 grants assessing this construct. This is aligned with the definition of fidelity of implementation in the RFPs, which focuses on the similarities between design and enactment. Therefore, the research emphasizes this construct as the primary implementation measure.

Table 4: Fidelity Constructs Reported in Grants that Measure Fidelity of Implementation

Construct	Number of grants	As proportion of grants measuring fidelity (%) (n=66)
Adherence	52	79
Quality	23	35
Dosage	22	33
Monitoring of control/ comparison conditions	20	30
Participant responsiveness	10	15
Program reach	0	0

Construct	Number of grants	As proportion of grants measuring fidelity (%) (n=66)
Program differentiation	0	0

Quality, Dosage, and Monitoring of control/ comparison conditions are used with similar frequency across studies. In the specific case of quality, the analysis reveals several issues in how studies conceptualize and measure this construct within the context of implementation fidelity. *Quality* can be defined as “how well the different program components have been conducted” (Dane & Schneider, 1998; Durlak & DuPre, 2008). However, many studies measure and report “quality” as an implementation construct for treatment and control classrooms, even though only the control group experienced the intervention. This happens particularly in studies that assess quality in the treatment and control groups, as researchers cannot measure implementation features specific to the intervention in classrooms that, by design, should not implement it (i.e., the control group).

Some studies measure “quality” as a component of fidelity, incorporating factors such as pacing, use of feedback, frequency of practice opportunities, teacher preparedness, clarity of questions, explicit instruction, and teacher enthusiasm into a single factor score representing implementation fidelity (K. R. Harris et al., 2023; Swanson et al., 2024). This conflation may be a consequence of the need to provide information on the instructional practices in control classrooms using instruments that do not exclusively measure intervention components. However, aggregating general quality measures into estimates of implementation fidelity may not align conceptually or theoretically, as *quality of implementation* typically refers to a specific intervention, whereas *instructional quality* encompasses broader pedagogical practices. This approach raises questions about whether these measures accurately capture general instructional quality or reflect adherence to the particular intervention, potentially blurring the conceptual boundaries between adherence, dosage, and quality. Therefore, it is relevant to distinguish quality in the context of *implementation* from *instructional* quality, as instruments must appropriately capture relevant practices without conflating different constructs that are associated only with the intervention under evaluation.

The construct Monitoring of control/comparison conditions is intended to ensure that a difference exists in the instruction of students in the treatment and control groups, so it is also frequently used to support the inferences about the impact of the intervention. In the case of Dosage, there is a connection with adherence, as the implementation should follow a specified exposure to the intervention. This may obscure the prevalence of Dosage as a construct if studies conceptualize measures related to the quantity of the intervention under the general umbrella of adherence.

Participant Responsiveness is a less relevant construct in the context of IES grants. This indicates that studies do not tend to collect data on the experiences of those taking part in the intervention, choosing to focus more on measures that relate to program delivery and not its reception. Program reach and Program differentiation are not measured in any of the grants in the sample. This may be due to the focus of the IES grants, which evaluate the effect of an intervention within an experimental or quasi-experimental design.

Instruments Used to Measure Fidelity of Implementation. Out of the 66 grants that indicate they measure fidelity of implementation, 61 report the instruments researchers used to assess the intervention’s enactment.

Classroom observation rubrics (including those administered live in a classroom or through a video or audio recording of the lesson) are the most frequently used data collection tool, with 82% of the studies that report instruments indicating they use them (50 grants). These instruments are aligned with the intervention's design, aiming to provide information on how teachers implement the program. This matches what Hill and Erickson (2019) observed, as they also found that classroom observations were the most prevalent instrument when measuring fidelity of implementation, although the proportion is lower (46%). Similarly, others have shown that these instruments are the most frequently used when assessing teachers and teaching (Bell et al., 2019).

Teacher logs and surveys are tied in second place in terms of frequency, each used in 15 grants. These instruments involve self-reports of implementation by program participants. This contrasts with observation rubrics, which are usually administered by external observers, typically members of the research team or trained raters. Logs and surveys are less resource-intensive than classroom observations, so they can be used to collect information from schools with more frequency. However, they are susceptible to social desirability bias, potentially compromising the validity of the data (Fernández & Martínez, 2022; Muijs, 2006). Although classroom observations help mitigate the issue of social desirability bias often associated with self-reported data, there is a trade-off in funding as these instruments tend to be more costly to administer. Despite higher costs, the preference for classroom observations in many of the studies analyzed may be attributed to the funding from IES, enabling researchers to adopt more rigorous data collection methods.

Other instruments are used less frequently to collect data on fidelity of implementation. These instruments include teacher and student artifacts, teacher interviews, qualitative classroom observations, and surveys responded to by participants other than teachers. Nine grants (15% of all the grants that report instruments) indicate they use any of these methods.

Most data collection instruments are developed ad hoc for the study and are meant to measure fidelity of implementation for the particular intervention. Overall, 84 out of the 99 reported instruments are developed by researchers specifically to measure fidelity (85%). Only seven instruments are adapted or used directly from a previous study (7%), and there is no information for eight.

Just over half of the studies that measure fidelity use one instrument to do so (35 studies or 57%). The most frequently used single instrument is observation rubrics (through video or live classroom observations), employed in 27 studies. Teacher logs and surveys are used as a single data collection instrument in only three and two grants, respectively. Conversely, 26 of the 61 studies that report the instruments used to collect data on implementation use a combination of two or more instruments (43%). Classroom observation rubrics and teacher logs are used together in eight studies, while observations and teacher surveys are used jointly in seven.

Most of the instruments collect data from schools where the intervention is implemented (treatment group). Out of the 61 studies that report data on the instruments they use to measure fidelity, 36 have instruments that gather information only on treatment classrooms (59%). Conversely, 22 studies declare that at least one of their instruments is used in treatment and control schools (36%). Out of the 93 instruments reported across all studies, 59 are used only in treatment classrooms (63%), and 33 are administered to participants in both groups (35%). Only one instrument is used exclusively in control schools.

Issues with the concept of *Quality* in the context of implementation are also reflected in the instruments used to measure this construct. This is particularly the case with observation tools designed for fidelity assessment that are used to measure instructional quality across different conditions. For example, the Classroom Observation of Early Mathematics –

Environment and Teaching (COEMET) is utilized across three different IES grants. It "measures the quality of the mathematics environment and activities (…) and is not connected to any specific curriculum" (Sarama et al., 2016, p. 39). Some researchers use this instrument as a measure of instructional quality that is not conceptually associated with fidelity of implementation (Clements et al., 2011, 2013; Sarama et al., 2012; Whittaker et al., 2020). Conversely, others determined that the COEMET can be used to measure implementation as a way to monitor control conditions (Sarama et al., 2016) and assess adherence and dosage (Clements et al., 2020).

Methods to Aggregate Fidelity Data. In general, studies report implementation scores calculated by averaging implementation data across constructs and teachers or using a simple sum of scores (44 studies). One study reports implementation scores computed using a weighted average based on the difficulty the intervention designers assigned to implementing each program activity (Meador et al., 2015), which is also the method used in the evaluation of Success for All analyzed in the following chapter of this dissertation (Balu & Quint, 2015; Quint et al., 2015).

Several studies employ latent variable techniques to assess implementation fidelity. Some use principal component analysis (PCA) to reduce dimensionality and identify underlying factors within fidelity measures (De La Paz et al., 2014; Stylianou et al., 2019). However, the most commonly used method is confirmatory factor analysis (CFA), which generates factor scores representing fidelity (Rimm-Kaufman et al., 2014; Sullivan et al., 2016; Swanson et al., 2024; Vaden-Kiernan et al., 2018; Vaughn et al., 2013). Other latent variable methods include latent profile or latent class analyses, which categorize teachers based on their fidelity scores (Gómez et al., 2023; Sullivan et al., 2016).

Studies report specific values for implementation and then proceed to indicate whether these levels meet an appropriate standard for fidelity. The method used to determine adequate levels of implementation varies across studies. Some researchers set pre-specified theoretical thresholds to define acceptable implementation. For example, studies can use benchmarks such as 75% or 80% compliance to determine acceptable fidelity levels (Crosson et al., 2024; S. B. Piasta et al., 2023; Vitale & Romance, 2013b; Zucker, Cabell, et al., 2021). A similar approach represents the highest implementation score based on the completion of all activities that teachers are expected to implement (Babendure et al., 2011). An alternative is to classify implementation levels by combining thresholds across different fidelity constructs, categorizing teachers according to the levels reached in each construct, such as "good" or "high" (Bae et al., 2022; Solomon et al., 2019). Other studies reported achieving "high," "good," "moderate," or "acceptable" levels of implementation without specifying expected levels or thresholds (Babinski et al., 2018; Cabell, 2020; Doabler et al., 2014; S. I. Gray et al., 2024; Murphy et al., 2022; S. Piasta et al., 2015; Proctor et al., 2021; Rohrer et al., 2020; Starkey et al., 2022; Zucker et al., 2019).

Some researchers generate classifications ex post (after the intervention is implemented) using distribution-based categories. This can include implementation categories using fidelity score distributions, such as means or quartiles, or the percentage of completed activities relative to the total agreed upon by program participants and designers (Barber et al., 2015; De La Paz et al., 2014; A. M. Gray et al., 2022; Marti, Melvin, et al., 2018).

A notable pattern across studies is the reporting of "adequate" levels of fidelity without referencing predefined standards or thresholds necessary to consider the implementation successful. Others classify implementation levels based on distribution metrics like quartiles without explaining what these categories signify in terms of the intervention's essential components or desired instructional practices. Although some studies suggest specific benchmarks, they often do not justify why these levels are deemed sufficient or how they relate to the different implementation measures. This obscures the relationship between program theory (the hypothesized effect of the intervention on student outcomes) and the enactment of the intervention (the instructional practices the students were exposed to).

The analysis of multiple studies reveals significant variation in the way programs are enacted in practice, suggesting differences in the teaching practices related to the program under evaluation. For example, variations in individual teachers' fidelity (adherence or dosage) can range from 56 percentage points (40% for the teacher with the lowest fidelity to 96% for the highest fidelity) to 27 percentage points (60% to 87%; (De La Paz et al., 2017; S. B. Piasta et al., 2021; Proctor et al., 2021; Stevens et al., 2022). Some studies publish the standard deviation of their implementation measures, providing information on the dispersion of the data. The coefficient of variation (the standard deviation expressed as a percentage of the mean) can go from 27% to 42%, indicating that there is a wide dispersion around the mean (Clements et al., 2011; Jitendra et al., 2015).

Variations can also be expressed in a different metric, with the number of activities related to the intervention implemented by each teacher ranging from an average of 0.5 to 25.9 strategies per week (Deussen et al., 2015). Similarly, one study observed variation in the number of intervention activities each teacher implemented across the three years of the intervention (Wakabayashi et al., 2020). In year 1, the mean was 84.4 activities (range 35–176); in year 2, the mean was 87.4 (range 72–108); and in year 3, the mean was 79.8 (range 62–89). They suggested that this variation might be due to teacher attrition and fluctuations in the number of teachers per classroom. Nonetheless, they concluded that the frequency of activities and coaching contacts was consistent with the intended dosage specified by the developers.

Variations may also be a result of the instruments used to measure enactment. Fidelity of implementation for an academic vocabulary intervention was measured using classroom observations completed by external observers and through teacher logs (Lesaux et al., 2014). On the one hand, teachers reported completing an average of 94% of lessons (standard deviation = 6%; range = 22%), while observers rated 86.7% of lesson components as completed (SD = 12.5%; range = 45.8%). The authors did not address the inconsistencies between teacher self-reports and observer ratings and only indicated that "there was high fidelity of implementation in treatment classrooms" (p. 1178).

Collectively, these examples highlight a common pattern present in many studies. While they report mean implementation scores and acknowledge variability in fidelity, few delve into the implications of this variability for program effectiveness and student outcomes. Although this issue is not observed in all studies, as there are examples of implementations where variation may not be a concern, there is an important lack of in-depth analysis in several instances where the authors report data that reflects these differences.

Furthermore, the issue of variability or low levels of fidelity is prevalent across many studies. In some instances, studies may report wide variations in implementation across sites but conclude that fidelity was sufficient without addressing how this variability might affect student experiences and outcomes. For other grants, researchers provide detailed implementation data in specialized papers but omit this information from publications estimating program effects, thereby limiting the insights into how fidelity levels influence outcomes. From the information observed, the incentive is to report adequate levels of fidelity, supporting this with an estimate that is not necessarily tied to a pre-defined benchmark to justify the internal validity of the study. Problematizing the causal relationship between the intervention and the measured outcomes introduces additional nuance that is not always the focus of the research that intends to provide evidence on the effectiveness of educational programs.

Associating Fidelity Data with Student Outcomes. Of the 66 grants that report measuring implementation, 23 use different methods to associate these data with student outcomes. Overall, the evidence proves that the relationship between implementation and student test scores is complex. The results are mixed, as studies report positive, negative, or no associations.

Some authors determine positive associations between measures of fidelity and student outcomes. For example, higher adherence to program design can be positively correlated with larger increases in test scores in a pre-post study (Duncan Seraphin et al., 2017). Similarly, other researchers report positive associations between measures of fidelity and student outcomes when fidelity is incorporated as a predictor (Gropen et al., 2011; Wasik & Hindman, 2020).

Several studies reported minimal or no significant relationships between fidelity and student outcomes. Some examples do not find any significant association between fidelity and outcomes (Apthorp et al., 2012; Pollard-Durodola et al., 2018; Weiland et al., 2011), while others report limited correlations between fidelity and student outcomes (Chaparro et al., 2022; S. I. Gray et al., 2024; Karam et al., 2017; S. Piasta et al., 2015; Rimm-Kaufman et al., 2014; Zucker et al., 2019). Moreover, others see a negative association between fidelity and outcomes, indicating that higher levels of fidelity could be detrimental to student outcomes (Rimm-Kaufman et al., 2014).

However, the direction of the relationship is not always stable within the same intervention. A group of studies finds that some fidelity subconstructs are positively associated with outcomes, while others show no detectable relationship (De La Paz et al., 2014, 2017; Stylianou et al., 2019; Sullivan et al., 2016). For example, within a single study, adherence and dosage can be positively associated with specific tests and the constructs they measure (expressive vocabulary, basic math, and math ability) while showing a negative association with some outcomes (phonological awareness) and no association with others (pre-literacy and math ability [short form]; (Marti, Melvin, et al., 2018). Other studies find positive relationships between higher fidelity (adherence and dosage) and social-emotional learning outcomes but not academic outcomes (Gómez et al., 2023; Rimm-Kaufman et al., 2021).

Differences in fidelity across implementations of the same intervention further complicate this relationship, as two enactments of the same intervention can show different relationships between fidelity and outcomes. Roberts et al. (2023) observe lower fidelity levels in an effectiveness study (scale-up of the intervention) compared to an initial efficacy trial (pilot of the intervention with a smaller sample size). Despite this, student outcomes were similar or better in the effectiveness study. The authors attribute these differences to the complex interaction between "implementation fidelity and teacher quality" Roberts et al. (p. 679) but do not explore this further.

All this proves that the association between fidelity and outcomes depends not only on the specific intervention but on the conceptualization of fidelity and its measures. The outcome measures add another layer of complexity, as implementation may have differing effects on the tests and constructs. Additionally, other considerations such as context (e.g., sites, participants, etc.) and scale (e.g., number of sites, sample size, etc.) can also play an essential part in this relationship.

Adaptation

An examination of educational research studies reveals varied approaches to conceptualizing and measuring teacher adaptations during the implementation of interventions. Among the 67 grants that measure implementation, nine also address the concept of adaptation alongside fidelity, and six of them detail the data collection instruments they use. Researchers predominantly adopt an exploratory approach to measuring modifications, often employing qualitative data collection methods such as interviews and open-ended questionnaires. Specifically, out of the six grants, four report using qualitative measurement instruments: two utilize interviews, and two employ open-ended questionnaires. Two studies use quantitative data collected through a survey (close-ended questions) and a classroom observation rubric. This reliance on qualitative methods reflects a focus on gaining in-depth insights into teachers' instructional practices and the nuanced ways in which they modify and adapt lessons.

Studies that thoroughly analyze adaptations concentrate on small samples of teachers, delving deeply into their instructional practices through interviews and observations. Thus, they can provide detailed insights into how teachers modify and adapt lessons. These analyses find that teachers tend to add or extend existing activities rather than omit them, suggesting a proactive engagement with the curriculum to enhance student learning (Burkhauser & Lesaux, 2017; Firetto et al., 2019; Monte-Sano et al., 2014).

Researchers find that teachers make these productive adaptations to suit their students' specific needs better. Two studies identified that teachers modified the intervention lessons to make activities more accessible to struggling students (Burkhauser & Lesaux, 2017; Monte-Sano et al., 2014). Similarly, McKeown et al. (2023) find that teachers do not implement the lesson as intended because they adapt it to prioritize higher-level thinking skills. Despite deviating from the prescribed lesson plan, the authors conclude that student learning was not negatively impacted on that day. Monte-Sano et al. (2014) found that teachers' understanding of their students, pedagogical content knowledge, and content knowledge were key in allowing them to make "successful modifications"(p. 560). This finding suggests that teacher adaptations can be beneficial, emphasizing the need to consider the quality and intent of modifications rather than solely focusing on adherence to the original design.

As mentioned earlier, there are two examples of the use of quantitative instruments to collect data on teachers' adaptations. In the survey, teachers are asked to rate the extent to which they implemented the intervention as written and designed (Spencer et al., 2020). In this case, teachers in the treatment group reported an average rating of 4.67 on a scale from 1 ("not at all") to 5 ("very much") regarding adherence to the prescribed curriculum. They also indicate whether they shortened lessons or incorporated new materials, with an average rating of 2.00 on the same scale. However, the reliance on self-reported measures raises concerns about social desirability bias, as teachers might feel compelled to report higher adherence to meet perceived expectations (Fernández & Martínez, 2022).

In a separate study, a classroom observation instrument included an item that asked external raters to indicate whether the teacher "conducted the activity as written in the curriculum or made positive adaptations to it (not changes that violated the spirit of the core mathematical activity)" (Clements et al., 2011, p. 137). The authors do not report the information for this item separately, but they add it to a fidelity construct that incorporates all 52 items in the instrument. This approach acknowledges adaptations but may conflate them with adherence, potentially obscuring the distinct effects of teacher modifications on implementation fidelity and student outcomes. This type of item can introduce measurement error, as there may be a lack of agreement between the two observers who are rating the lesson or between the observer and the researchers as to what changes are positive or do not violate the core of the activity.

Overall, the varied approaches to conceptualizing and measuring adaptations further reflect the complexity of capturing teacher practices in educational research. The predominant use of qualitative methods underscores the value placed on rich, contextualized data to understand the intricacies of adaptation. However, quantitative methods can provide information on a larger number of teachers, allowing for more generalizable inferences about modifications to the intervention.

Although some studies measure adaptations (whether using a single item or an entire instrument), most IES studies conceptualize adaptations as a lack of fidelity. In these studies, any deviation from the prescribed intervention is often viewed negatively, as it signifies a departure from the intended implementation. This perspective frames adaptations primarily as reductions in fidelity rather than considering them as purposeful modifications that teachers make to meet the specific needs of their students.

Conclusion

The instruments used to measure implementation are diverse, ranging from structured observation protocols and standardized assessment tools to teacher self-reports and qualitative interviews. Each method has inherent strengths and limitations, and practical considerations like resource availability and feasibility often influence choices. The methods for aggregating implementation data also vary, with some studies employing simple averages, others using latent variable models, and still others integrating multiple constructs into composite scores. These methodological differences impact the reliability and validity of the implementation measures and affect how fidelity and adaptation are accounted for when evaluating the effectiveness of interventions.

The sample of IES-related studies analyzed tends to conceptualize implementation as fidelity, with a particular focus on adherence. This aligns with the requirements of IES RFPs, which expect investigators to assess how similar the enacted intervention is to the original design. Following the most frequently reported information in this chapter, an average IES grant measures fidelity of implementation (conceptualizing it as adherence), collecting data through a classroom observation rubric. The information on implementation is aggregated into a single implementation construct calculated as a simple sum across all items of the rubric and reported as a percentage of the maximum possible score. This percentage reflects whether the intervention reached an adequate level of fidelity, although the cutoff is not defined explicitly. The data on implementation is not incorporated into the model that estimates program impacts.

The analysis indicates the importance of conceptualizing and accurately describing the implementation constructs in a study. Imprecisely defined or overlapping constructs (such as *quality* in the context of *implementation* and *instruction*) may lead to fusing concepts that do not measure the same features of the intervention. Similarly, aggregations of data on implementation should consider whether the combination makes sense conceptually and in terms of the intervention. Therefore, it is important to have a clear definition of implementation, indicating the different constructs and items that comprise it and determining the relative importance of the intervention components based on the program’s theory of change.

To illuminate the intervention's implementation and its relationship with student outcomes, researchers should determine the expected levels of fidelity for each component and provide an example of what the enacted intervention should look like to generate changes in the students' outcomes. This would strengthen the intervention's internal validity while providing a framework for the expected relationship between program activities (and the specific classroom and instructional practices associated with them) and the beneficial effects they should generate on students.

Overall, while variability in fidelity is acknowledged, there is often a lack of comprehensive analysis regarding how differing levels of fidelity affect program outcomes. Discussions frequently focus on overall implementation rates or mean scores without examining the nuances of variability and its influence on effectiveness. This gap suggests a need for more in-depth research to understand the relationship between implementation fidelity (or a lack thereof) and educational outcomes. A deeper exploration of how variations in fidelity impact results could inform strategies to enhance program implementation and ultimately improve student learning experiences.

Another way to understand implementation in more depth would be to triangulate data from diverse sources to explore potential differences among them. This may be useful, as classroom observations capture only what happens in a single lesson, while responses from logs and surveys may capture a larger sample of instructional practices. This is especially relevant, considering there is evidence of disagreement in one of the studies (Lesaux et al., 2014). By combining information, researchers may obtain a more accurate measure of implementation that reflects actual instructional practices over time.

Implementation measurement is further complicated when the concept of adaptation is introduced, which challenges the definition and measurement of fidelity. Adaptation in educational interventions is a complex construct that is often intertwined with fidelity measures.

Although some IES-funded research explicitly measures adaptations—whether using single items or comprehensive instruments—most studies conceptualize adaptations as deviations from fidelity. In this perspective, any modification made by teachers is viewed as a reduction in fidelity, potentially overlooking the intentional and beneficial adjustments educators make to meet their students' unique needs. Although adaptation can be disregarded as the absence of fidelity, in-depth research into program implementation proves that these modifications can be productive and lead to enhancements to the intervention.

Research indicates that teachers frequently adapt interventions based on their perceptions of students' needs and interests. However, by equating adaptations with diminished fidelity, studies may fail to recognize the nuanced ways in which teacher-initiated changes can positively influence educational outcomes. At the same time, this may prevent researchers from understanding the classroom practices that are generating improvements in students' outcomes. Recognizing that teachers often adapt curricula to suit their students' needs better—and that such adaptations can enhance learning—is crucial for developing interventions that are both effective and adaptable to diverse classroom contexts. At the same time, measuring and accounting for these changes can help researchers understand what is happening inside the classroom.

Nonetheless, measuring and accounting for adaptations is a challenging task in the context of IES-funded grants. On the one hand, the goal of this research is to provide a quantifiable estimate of the intervention's effect on students, which requires investigators to prove the internal validity of their study (i.e., that their program generated the changes observed in students). Examining adaptations may open the possibility of questions about the theory of change and the causal link between program and outcomes, which can invalidate the study's findings. At the same time, the focus of the RFAs on fidelity (mainly conceptualized as adherence) does not encourage researchers to explore the differences across teachers and the possible modifications to their designed program.

Chapter 3: Measuring Implementation and its Implications for Estimating the Effects of the Success For All Program

This chapter addresses the second research question of this dissertation, exploring how different methodologies and analytical approaches to measuring program enactment can influence our understanding of the program’s effectiveness and its relationship with implementation. Specifically, it examines how varying measures of implementation and different approaches to data reduction impact understanding of a program’s effects on students’ reading test scores. Additionally, it explores the consequences of various modeling approaches for incorporating implementation when estimating an intervention's effects.

To illustrate the implications of these methodological choices, I use real-world data from the implementation of the Success for All program (SFA, Slavin et al., 2009; Slavin & Madden, 2001, 2012), specifically the 2011-2014 scale-up of the intervention (Quint et al., 2015). In the analyses, I explore the associations between implementation data and student outcomes, focusing on the consequences of employing distinct approaches to reduce information on program enactment.

The first section of the chapter describes the SFA program, including the instruments used to measure implementation and outcomes. Next, the analysis is structured around the constructs to assess the implementation of SFA and the data collection instruments associated with them. First, I explore the data collected using an instrument intended to measure adherence to SFA in treatment schools, employing correlational approaches and hierarchical linear models (HLM) to determine if there is an association between implementation and students’ reading outcomes. Second, I use data from instruments used to measure implementation in treatment and control schools to explore if enactment moderates the effects of SFA.

Description of Success for All

The Success For All (SFA) program aims to ensure that every child learns to read well in the elementary grades. It was originally designed and first implemented in 1987 by researchers at Johns Hopkins University. Since 1996, it has been developed and disseminated by the Success for All Foundation (SFAF). Its basic design and operation have remained constant throughout the years (Cheung et al., 2021). SFA combines “a challenging reading program, whole-school reform elements, and an emphasis on continuous improvement” (Quint et al., 2015, p. iii). The intervention focuses on “providing a curriculum with a strong emphasis on phonemic awareness and phonics and using proven instructional methods such as cooperative learning and effective classroom management methods” for students in grades kindergarten through 6 (Cheung et al., 2021, p. 91). The Success for All Foundation (SFAF) has defined specific eligibility criteria for schools that intend to implement the program. Firstly, schools must serve students from kindergarten through grade 5, and at least 40% of the student population must be eligible for the free or reduced-price lunch program. Additionally, schools are required to submit their participation to an internal vote, in which at least 75% of the teachers must choose to adopt the program.

In 2010, the SFAF received a scale-up grant under the U.S. Department of Education’s Investing in Innovation (i3) program to implement SFA from 2011 to 2014. Five school districts met the eligibility criteria to participate in the intervention; each district had between 4 and 17 schools for a total sample of 37 schools. In the 2011-2012 school year, schools were randomly assigned to treatment (19 schools) or control (18 schools) conditions. Most schools were in large or mid-size cities in the South of the United States.

Table 5: Selected Characteristics of Schools in the Study Sample (2010-2011)

Characteristics	Study sample
Geographic region (% of schools)	
Northeast	16.2
South	67.6
Midwest	0.0
West	16.2
Urbanicity (% of schools)	
Large or midsize city	62.2
Urban fringe or large town	21.6
Small town or rural area	16.2

Source: Quint et al. (2015)

All the schools participating in this implementation of SFA were classified as Title I, and approximately half of all the students were eligible for free or reduced-price lunch. About 85% of the students in both groups were Black or Hispanic, with a majority of Hispanic students (62%). None of the differences between the treatment and control groups in the variables listed in Table 6 are statistically significant.

Table 6: Selected Characteristics of Schools in the Treatment and Control Groups (2010-2011)

Characteristics	Treatment group	Control group	Estimated Difference	P-value for Estimated Difference
Title I status (% of schools)	100	100	0.0	
Free or reduced-price lunch (school average % of students)	56.1	56.3	-0.2	0.928
Race/ethnicity (school average % of students)				
White	13.1	13.9	-0.7	0.496
Black	23.0	21.3	1.8	0.671
Hispanic	62.1	63.1	-1.0	0.823
Asian	0.6	0.8	-0.2	0.542
Other	1.2	1.0	0.2	0.436
Male (school average % of students)	51.6	51.0	0.6	0.407
Total school enrollment	558.4	533.8	24.6	0.548
Number of full-time teachers	32.8	31.7	1.1	0.598
Number of schools	19	18	1.4	0.595

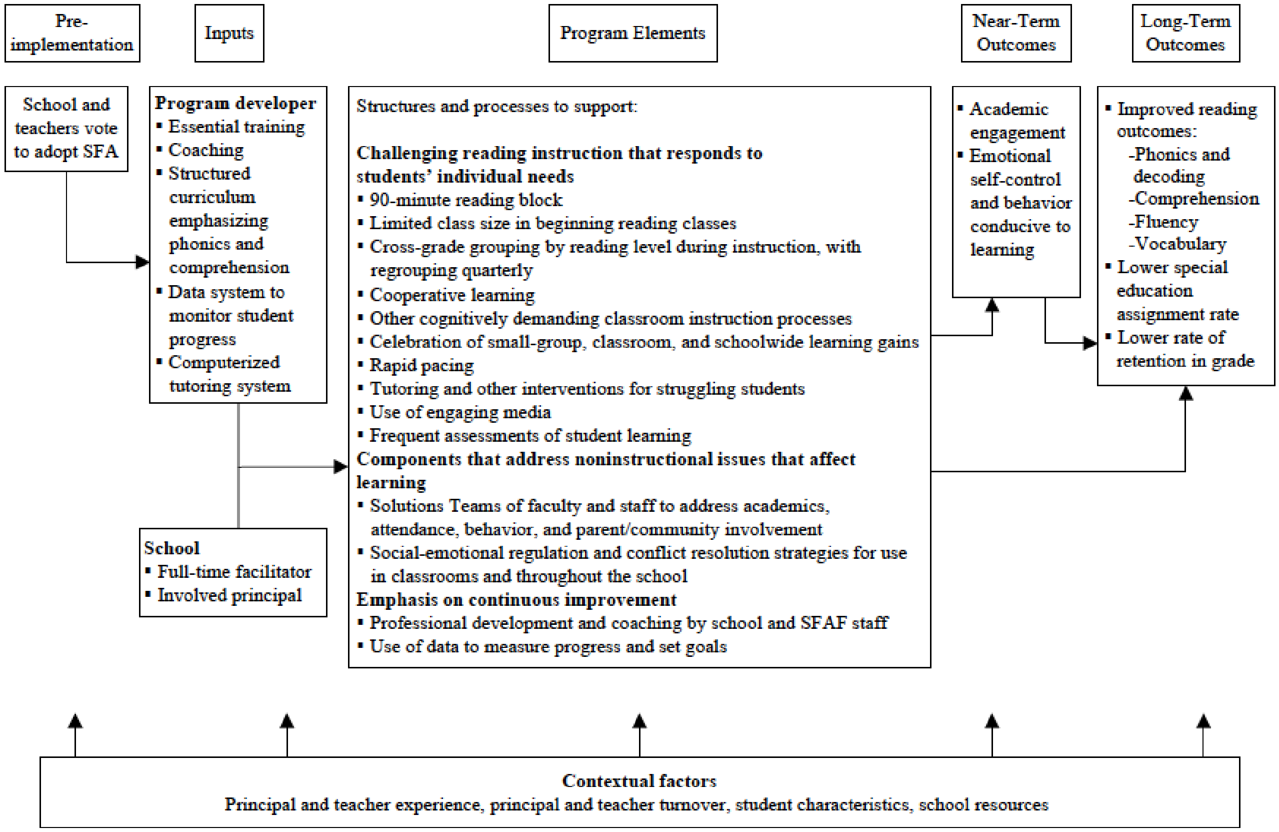
Source: Quint et al. (2015)

Program components

SFA has several components, including activities aimed at students, teachers, and the whole school. These components can be grouped into three categories: challenging reading instruction that responds to students' individual needs, noninstructional issues that affect learning, and continuous improvement.

The reading program, meant for students from kindergarten through grade 6 (K-6), emphasizes phonics for beginning readers and comprehension for all students. SFA has different programs, according to the student's grade levels: *KinderCorner* (or *Descubre Conmigo*) program in kindergarten, *Reading Roots* (or *Lee Conmigo*) in grades 1 and 2, and *Reading Wings* in grades 2 and above (for students who have tested out of *Reading Roots*). Instruction in SFA "is characterized by 'scripted,' briskly paced lesson plans that make extensive use of cooperative learning in pairs and small groups" (Quint et al., 2015, p. 1). Scripted lessons are meant to allow teachers to "achieve a rapid pace of instruction and interaction (…) [while reducing] the time that teachers need to spend in preparation" (Slavin et al., 2009, p. 96). Figure 7 presents the intervention's logic model, including all components and the inputs the program developer and the schools provide.

Figure 7: Success for All Logic Model


Pre-implementation
Inputs
Program Elements
Near-Term Outcomes
Long-Term Outcomes
School and teachers vote to adopt SFA
Program developer
▪ Essential training
▪ Coaching
▪ Structured curriculum emphasizing phonics and comprehension
▪ Data system to monitor student progress
▪ Computerized tutoring system
School
▪ Full-time facilitator
▪ Involved principal
Structures and processes to support:
Challenging reading instruction that responds to students' individual needs
▪ 90-minute reading block
▪ Limited class size in beginning reading classes
▪ Cross-grade grouping by reading level during instruction, with regrouping quarterly
▪ Cooperative learning
▪ Other cognitively demanding classroom instruction processes
▪ Celebration of small-group, classroom, and schoolwide learning gains
▪ Rapid pacing
▪ Tutoring and other interventions for struggling students
▪ Use of engaging media
▪ Frequent assessments of student learning
Components that address noninstructional issues that affect learning
▪ Solutions Teams of faculty and staff to address academics, attendance, behavior, and parent/community involvement
▪ Social-emotional regulation and conflict resolution strategies for use in classrooms and throughout the school
Emphasis on continuous improvement
▪ Professional development and coaching by school and SFAF staff
▪ Use of data to measure progress and set goals
▪ Academic engagement
▪ Emotional self-control and behavior conducive to learning
▪ Improved reading outcomes:
-Phonics and decoding
-Comprehension
-Fluency
-Vocabulary
▪ Lower special education assignment rate
▪ Lower rate of retention in grade
Contextual factors
Principal and teacher experience, principal and teacher turnover, student characteristics, school resources

Source: Quint et al. (2015)

SFA requires schools to group students across grades according to their reading abilities during reading instruction. These groups are different from the class in which students spend most of the day (where they are grouped according to their age), and they should contain fewer students than regular classrooms. Students should be informally assessed in their groups daily and weekly, while formal summative assessments will be conducted quarterly. Schools are expected to use this information to regroup students according to their reading level and to determine which students require additional tutoring. Grouping should be evaluated quarterly, using information from student assessments to regroup them when needed.

The program also considers additional individual and small-group tutoring for students who need further assistance. This support is meant to complement and support in-classroom learning, so tutors should work in close alignment with teachers. Tutors must identify students' learning problems and use different strategies to teach the same content as classroom teachers. For the i3 scale-up implementation, SFA tutors should use a computerized tutoring system explicitly developed for SFA reading programs (Team Alphie). The SFA guidelines recommend that schools in high-poverty areas make sure that they provide enough tutoring for 30% of first graders, 20% of second graders, and 10% of third graders (Quint et al., 2015; Slavin et al., 2009).

The SFA guidelines consider appropriate adaptations to the program in the context of grouping and individual tutoring. Schools are encouraged to regroup students across classes and grades according to the results of frequent reading assessments. The 2009 program guidelines also state that additional adaptations for students below grade level can be made in one-on-one tutoring sessions (Slavin et al., 2009).

To ensure the proper implementation of SFA, teachers are offered professional development on pacing, assessment, classroom management, and cooperative learning, among others. The initial training (led by the Success For All Foundation) usually lasts two days and is conducted in the summer before the first year of implementation. Schools also have an on-sit— facilitator in charge of ongoing training (Slavin et al., 2009).

In the logic model, the program elements are expected to affect the near and long-term outcomes. In contrast, the near-term outcomes mediate the relationship between the program elements and the long-term outcomes. Contextual factors such as "staff turnover, student characteristics, and schools' access to resources" affect program implementation and outcomes (Quint et al., 2015, p. 6).

Outcomes

In the near term, the program emphasizes influencing non-cognitive student outcomes, including academic engagement, emotional self-control, and behavior conducive to learning. Academic engagement is measured using surveys in which teachers indicate their perceptions through items that ask teachers to report whether their students, in general, are engaged during their reading class, if the reading program gets students excited about reading or learning how to read, and if their students are well-behaved during their reading class.

Students in treatment and control schools took two assessments (Peabody Picture Vocabulary Test and Woodcock-Johnson Letter-Word Identification [WJLWI] test) in the fall of 2011 before the intervention began its implementation. The Test of Word Reading Efficiency (TOWRE) and Woodcock-Johnson Passage Comprehension (WJPC) test were administered at two different time points (Spring of 2012 and 2014), while students took the Woodcock-Johnson Word Attack (WJWA) test three times (Spring of 2012, 2013 and 2014). The only test administered in the four time points (Fall of 2011 and Spring 2012, 2013, and 2014) was the Woodcock-Johnson Letter-Word Identification (WJLWI) Test.

Table 7: Student Reading Outcome Measures

Outcome	Reading skills measured	Description	Time administered
Peabody Picture Vocabulary Test	Receptive vocabulary	The student is given a word spoken by the examiner and four pictures on a single card. The student selects the picture that best represents the examiner's spoken word.	Fall 2011 (baseline)
Woodcock-Johnson Letter-Word Identification (WJLWI)	Reading decoding and sight word recognition	The student is asked to identify letters that appear in large type and is then asked to pronounce words correctly. Items become increasingly difficult as the selected words appear less and less frequently in written English.	Fall 2011 (baseline) Spring 2012 Spring 2013 Spring 2014

Outcome	Reading skills measured	Description	Time administered
Woodcock-Johnson Word Attack (WJWA)	Reading decoding and phonetic coding	The student is asked to produce the sounds for individual letters, then read aloud letter combinations that are regular patterns in English but are nonwords or low-frequency words.	Spring 2012 Spring 2013 Spring 2014
Test of Word Reading Efficiency (TOWRE)	Efficiency of sight word recognition and phonemic decoding	Assessment is based on the number of real words the student can identify within 45 seconds, as well as the number of pronounceable nonwords the student can accurately decode within 45 seconds.	Spring 2013 Spring 2014
Woodcock-Johnson Passage Comprehension (WJPC)	Reading comprehension and verbal language comprehension	The student is asked to match pictographic representations of words with actual pictures of the object, choose pictures represented by a phrase, and read several short passages and identify missing key words.	Spring 2012 Spring 2014

Sources: Campbell & Dommestrup (2010); Quint et al. (2015); Quint (2016f); Wills & Wolf (2021).

All the tests used to measure student outcomes assess constructs related to reading. In this sense, it is expected that students' scores on the different tests will correlate with each other as they measure skills around a single ability. All the correlations among tests prove to be large ($r >$ 0.7), indicating a high alignment in the skills they measure. The WJLWI presents the highest associations with other tests ($r > 0.8$). This could be explained by the fact that the WJLWI and WJWA assess the same key area of reading (Alphabetic Principle; (Wendling et al., 2007) and are intended to measure reading decoding. Similarly, both the WJLWI and TOWRE evaluate students' sight word recognition.

Table 8: Correlations Among Test Scores (2014)

	WJLWI	WJWA	TOWRE	WJPC
WJLWI	1			
WJWA	0.845	1		
TOWRE	0.867	0.765	1	
WJPC	0.817	0.704	0.794	1

Source: Quint (2016a).

The analyses in this chapter include only students with valid scores in each test's initial and final administration. The sample is described below.

Table 9: Students with Valid Test Scores in the First and Last Administration, by group

Test	Treatment	Control	Total
Woodcock-Johnson Letter-Word Identification	891	819	1,710
Woodcock-Johnson Word Attack	721	663	1,384
Test of Word Reading Efficiency	829	762	1,591
Woodcock-Johnson Passage Comprehension	715	664	1,379

Source: Quint (2016a).

Implementation measures.

Data on program implementation is collected for schools in the treatment and control groups. The administration of each instrument varies according to its purpose and the target informant. In the case of treatment schools, the research team gathered information to monitor the implementation of SFA using a researcher-developed checklist. Data from principal and teacher surveys and teacher logs were collected in treatment and control schools to determine whether instructional and administrative practices differed between the schools implementing SFA and those that did not.

Table 10 summarizes the data collection instruments used to gather information on reading programs in schools that implemented SFA (treatment schools) and those that did not (control schools).

Table 10: Data Collection Instruments Related to the Enactment of Reading Programs

Instrument	What does it measure?	Who reports information?	Sample	Sample size (2014)
School Achievement Snapshot	- Monitors implementation of SFA	SFA staff	Treatment schools	19 schools
Principal Surveys	- Perceptions of and attitudes toward reading programs.	Principals	Treatment and Control schools	27 Principals (Treatment = 14; Control = 13)

Instrument	What does it measure?	Who reports information?	Sample	Sample size (2014)
Teacher Surveys	- Perceptions of and attitudes toward reading programs. - Perceptions of student engagement.	Teachers	Treatment and Control schools	373 Teachers (Treatment = 199; Control = 174)

Created using data from Quint (2016c, 2016e, 2016b, 2016d) and Quint et al. (2015).

School Achievement Snapshot (SAS). This instrument is designed to monitor the enactment of SFA in treatment schools and to provide a quick and immediately interpretable indicator of where each school is and where it is going in its implementation of SFA (Slavin et al., 2009). Since this instrument measures implementation of SFA, information from the Snapshot is only available for schools participating in the SFA program. Success for All The SAS administered in the 2013-2014 school year consists of 47 items grouped into 13 subconstructs. The subconstructs are organized into three constructs (Challenging Individualized Instruction, Non- Instructional Issues that Affect Reading Instruction, and Continuous Improvement of Students and Staff; a detailed description of the instruments is available in the Appendix, Table A 2).

The variables in the School Achievement Snapshot are aggregated using theoretical weights determined by the research team (MDRC) and the SFA staff. Items that are “more central to the SFA reading program” (*Importance* weights) or that apply “to reading levels that cover a larger percentage of the student population” (*Reading-Level* weights) receive a double weight (2x). Items considered central in *Importance* and *Reading Level* are multiplied by four (2x for *Importance* * 2x for *Reading Level* = 4x; J. Quint, 2016e, p. 11). The items in the Challenging Individualized Instruction construct account for the most significant proportion of the score (58.8% of the 97 points), followed by Non-Instructional Issues that Affect Reading Instruction (23.7%) and Continuous Improvement of Students and Staff (17.5%).

Teacher survey. Teacher surveys were administered to individuals in the treatment and control groups. This instrument measures teachers' perceptions, attitudes, and teaching practices related to the reading programs being implemented in their schools. In treatment schools, teachers report on SFA, while in control schools, they report on the reading programs the school is using. The questionnaire has 124 items and is divided into four sections.

Table 11: Number of Items in the Teacher Questionnaire by Section and Subsection

Sections and Subsections	Number of items	
Your Background and Current Responsibilities		10
The Reading Program at Your School		74
The Current Program	62	
Use of Data	12	
General School Functioning		39
Climate	29	
The Success for All Program	10	
General open ended		1
Total		124

Source: (Quint, 2016e)

Eleven items appear only in surveys for teachers in schools participating in SFA (all the items are in the subsection on the SFA program, and one question is in the Your Background and Current Responsibilities section). For a list of all the items and scales, refer to Table A 3 in the Appendix.

Principal survey. This survey aims to learn how principals lead their schools and support reading instruction. As with the teacher survey, questions about the reading program are not specific to SFA, so they can be applied to any reading program the school is implementing. The questions are then valid for treatment and control schools, allowing for comparisons between the two groups. The principal questionnaire has 180 items and is divided into the same sections as the teacher survey (see Table 12). For a list of all the items and scales, refer to Table A 4 in the Appendix.[6]

Table 12: Number of Items in the Principal Questionnaire, by Section and Subsection

Section	Number of items	
Background and Broad Responsibilities		17
The Reading Program at Your School		82
Instruction	15	
Tutoring/ Intervention	23	
Use of Data	26	
Grouping	2	
General School Functioning		80
Teams/ Functions	7	
Staffing	14	
Student Health Policies	12	
Attendance, Parental Involvement	7	
Funding Support	13	
The Success For All Program	14	
General Open Ended		1
Total		180

Source: (Quint, 2016b)

Implementation-Related Findings from the Evaluation of Success for All

The evaluation of the i3 scale-up of SFA (Quint et al., 2015) found that the intervention positively impacted students who had the maximum possible exposure to the program in the context of this study (with an increase of 0.15 standard deviations in their test scores), as they were enrolled in participating schools from kindergarten through second grade. However, the study did not find statistically significant effects of SFA on the other tests (WJLWI, TOWRE, and WJPC), indicating that the program, on average, helped improve students' phonics and decoding skills but not other reading-related skills.

[6] The research design also included collecting data related to teacher practices through teacher logs (Quint, 2016d). These items are not explicitly mapped to SFA components, so there is no way of knowing which teacher practices measured in the logs are expected to be affected by SFA by design. Thus, the information from the logs will not be incorporated into the implementation model.

In terms of program implementation, by the end of the final year of the grant (2013-2014 school year), the report indicates that 17 out of the 19 schools were judged to have met SFA's standards for adequate implementation fidelity, reaching at least 50% of the maximum score in the implementation index.

Information collected for the final year of implementation (2014) shows that teachers are changing the prescribed curriculum. In SFA schools, 55% of teachers agreed that they should change the parts of their school's reading program (strongly agree and agree) that do not work for their students. This adaptation trend is notably lower than in the control schools (88%), which indicates there could be a higher emphasis on following the prescribed curriculum in SFA schools (see Table 13).

Table 13: Distribution of Responses to Selected Items in the Teacher Questionnaire by Treatment Status (2014)

Item	Treatment Status	Percentage (%)				
		Strongly Disagree	Disagree	Agree	Strongly Agree	Total
You change the parts of your school's reading program that do not work for your students.	Treatment	8.1	37.1	48.7	6.1	100
	Control	1.7	9.8	72.4	16.1	100
The reading program at your school is too rigid or scripted.	Treatment	3.0	49.7	35.7	11.6	100
	Control	19.0	66.1	14.4	0.6	100

Source: Quint (2016e)

Although there is information to determine that an important part of teachers is making adaptations, the data collection instruments do not delve deeper into what these changes may look like in each classroom. The main differences in the teaching of reading between SFA and control schools were related to instructional strategies, particularly in three critical components. SFA classrooms were more likely to focus on cooperative learning and to use educational media and cross-grade grouping. However, no differences were observed between SFA and control schools in other critical instructional practices emphasized in the design of SFA, such as extended reading periods, data usage, and tutoring for struggling students. The lack of clear differentiation in some of the practices between treatment and control groups may have affected the intervention's impact on student outcomes, blurring the potential effects of SFA.

The analyses in this chapter aim to provide further insight into the relationship between the enactment of SFA and reading scores by summarizing information on implementation from treatment and control schools into different types of indices and then modeling the relation between these indices and student outcomes.

Methods

Measuring Implementation Adherence in Treatment Schools

I use data from the 2014 School Achievement Snapshot (SAS) to generate an implementation score that measures program enactment at the school level in SFA schools.[7] Although the dataset for the SAS contains item-level variables, the program's documentation reports the weights only at the subconstruct level. Thus, I generate implementation scores for each school using the aggregated values for each subconstruct.

Reduction of Implementation Data from the SAS. Information on implementation in treatment schools measured with the SAS is aggregated through linear composite indices. I compute the indices using three methods: one based on program theory, an unweighted index, and an index based on empirical weights.

Theoretical index. The theoretical index I_{SAS-T} is calculated using the implementation subconstructs from the SAS and the theoretical weights as determined by the program. Table 14

[7] The research team did not expect schools to implement higher-level practices (levels 2 and 3) during the first years of enactment, so some of the items in the School Achievement Snapshot were not rated for those years (Quint et al., 2015, p. 25). 2014 presents the most complete reflection of implementation, as measured by the items in the School Achievement Snapshot.

presents the weights reflecting the relative importance of each subconstruct as determined by the research and program teams, with a maximum of 97 points across all subconstructs.

Table 14: Success for All School Achievement Snapshot constructs, subconstructs, and weights

Constructs	Subconstructs	Number of Items	Theoretical weight
1. Challenging Individualized Instruction	Cooperative Learning	5	23
	Cognitively Demanding Instruction	3	8
	Pacing	1	4
	Media Use	1	4
	Grouping	3	4
	Tutoring	4	4
	Celebration	2	10
Subtotal		19	57
2. Non- Instructional Issues that Affect Reading Instruction	Solutions Teams	4	12
	Parent/ Community Involvement	3	5
	Attendance	1	1
	Behavior	5	5
Subtotal		13	23
3. Continuous Improvement of Students and Staff	Use of Data	10	15
	Professional Development	2	2
Subtotal		12	17
Total		44	97

Source: (Quint, 2016f)
NOTE. The number of Items column indicates how many SAS items are aggregated to generate the subconstruct.

In (Equation 1), x_i represents the value for subconstruct *i* (n = 3) in the SAS, and w_i is the weight assigned to this subconstruct *i*.

$$I_{SAS-T} = \sum_{i=1}^{n}(x_i * w_i)$$

(Equation 1)

Unweighted index. A second linear composite index (I_{SAS-U}) is calculated as the unweighted summation of the schools' scores in the School Achievement Snapshot items. Each item in the SAS has a value between 0 and 1, with a maximum score of 44 across all items.

Additionally, I calculate a sub-index for each implementation construct, with a maximum score of 19 points for construct 1, 13 for construct 2, and 12 for construct 3.[8]

Relationship Between Implementation in Treatment Schools and Student Outcomes.

Correlational Analysis. I calculated a naïve estimate of the program's effect by computing the difference between students' initial and final scores for each test. These differences were then averaged at the school level to generate a comparable estimate to the school-level implementation score in the SAS. Finally, I estimate correlation coefficients to assess the association between aggregate student outcomes in the four tests and adherence to the implementation model in treatment schools.

Multilevel Models. I further explored the association between implementation and student outcomes using multilevel models, treating schools as random effects to account for variability in intercepts across schools and student and school-level covariates that help explain the variation. The models in Equations 2 and 3 mirror the one used in the final evaluation of i3-SFA but incorporate information on implementation that is not considered in the original report (Quint et al., 2015, pp. 171–172).

The first model in *(Equation 2* explores the relationship between implementation indices (I_{SAS-T} and I_{SAS-U}) and student outcomes through a two-level model where students are nested in schools.[9]

[8] I attempted to use Factor Analysis to estimate empirical weights for the constructs. A unidimensional model and a three-factor model reflecting the theoretical constructs in the program design presented a poor fit. Similarly, two- and three-factor exploratory models showed poor fit (RMSEA > 0.18 for all CFA and EFA models). Empirical weights cannot be calculated from individual items, as the model is not identified (47 variables with n=18).

[9] The intraclass correlation coefficients (ICC) for the different tests indicate that between-school variation in scores fluctuates between 13.1% (Woodcock-Johnson Passage Comprehension test) and 8.6% (Test of Word Reading Efficiency). Although the ICC is relatively low, indicating that most of the variation in test scores is due to within-school differences, the fixed effects (implementation constructs) can be estimated

$$Y_{ik} = \gamma_{00} + \gamma_{01}I_{1k} + \gamma_{02}I_{2k} + \gamma_{03}I_{3k} + \gamma_{10}X_{ik} + \sum_{l} \gamma_{c0}\, W_{ick} + u_{0k} + r_{ik}$$

(Equation 2)

Y_{ik} represents the score of student *i* from school *k* in the 2014 test administration. The effect of the implementation constructs for SFA schools is captured by γ_{01} (Challenging Instruction, I_{1k}), γ_{02} (Non- Instructional Issues, I_{2k}) and γ_{03} (Continuous Improvement, I_{3k}). The implementation variables are rescaled, so 1 unit represents 10 percentage points in the implementation index. The model adjusts for students' prior achievement (X_{ik}, with effect γ_{10}) and a set of covariates W_{cik} for student characteristics (English language learner [ELL] status, special education [SPED] status, age, and gender). School- and student-level random error terms ($\mu_{0k}$and r_{ik}, respectively) are assumed to be independently and identically distributed.

Next, I investigate the effect of program implementation replacing the three implementation constructs with the overall implementation level for school *k*, using the theoretical index (I_{SAS-T}), represented by $I_{k,}$ with effect γ_{01}.

$$Y_{ik} = \gamma_{00} + \gamma_{01}I_{k} + \gamma_{10}X_{ik} + \sum_{l} \gamma_{l0}\, W_{ilk} + u_{0k} + r_{ik}$$

(Equation 3)

Equation 4 incorporates the interaction of the implementation variable with two student-level covariates to determine whether the effect of implementation varies as a function of students' ELL (γ_{02}) and SPED (γ_{03}) status.

$$Y_{ik} = \gamma_{00} + \gamma_{01}I_{k} + \gamma_{02}(I_k \times ELL_{ik}) + \gamma_{03}(I_k \times SPED_{ik}) + \gamma_{10}X_{ik} + \sum_{l} \gamma_{l0}\, W_{ilk} + u_{0k} + r_{ik}$$

(Equation 4)

more precisely by incorporating random effects at the school level that can accounts for unobserved heterogeneity across schools.

Measuring Implementation in Treatment and Control Schools

Reduction of Implementation Data from Principal and Teacher Surveys. The assessment of SFA uses principal and teacher survey data to evaluate implementation constructs that can be observed in schools in the control group (Quint et al., 2015). The analyzed sample contains 27 schools (14 in the treatment group and 13 in the control group). I removed ten schools from the original sample (37 schools), as six schools did not return surveys for teachers and principals (four SFA and two control group schools), and four principals did not return surveys (two treatment and two control schools).[10] Based on program theory, the 17 implementation variables from the teacher and principal surveys are grouped into six theoretical constructs: Length of the reading block, Small class size, Grouping, Cooperative learning, Tutoring, and Use of educational media/technology (for a full description, see Table A 5 in the Appendix).[11]

Unweighted Implementation Index. The implementation items from the teacher and principal surveys (y_k) are aggregated at the school level (*k)* into an unweighted index (I_{Impact}) *(Equation 5)*.[12]

$$I_{Impact} = \frac{\sum_{k=1}^{n}(y_k * w)}{\sum_{i=1}^{n} \max(y_k)}$$

(Equation 5)

[10] The items "Does your school group students who are in the same reading class into smaller groups according to their ability level?" and "Does your school group students who are in the same grade into separate reading classes according to their ability level?" were missing responses from one principal each. I imputed these values using predictive mean matching (Enders, 2022), and calculated the final values by pooling across five imputations.

[11] The items from the teacher surveys are aggregated at the school level so as not to give more weight to schools with more teachers.

[12] I attempted to estimate empirical weights of the items from the hypothesized factor structure from the implementation constructs of the Teacher and Principal surveys using Confirmatory Factor Analysis. However, the model did not converge, and the solutions were still unstable after several attempts which included increasing the number of iterations and trying different optimization algorithms.

Relationship Between Implementation and Program Effects. The following models explore the relationship between the implementation index (*I_{Impact}*) and students' reading scores. Two-level Hierarchical linear models (HLM) are fit to the data, with students (*i*) nested in schools (*k*; Raudenbush & Bryk, 2002). In the baseline model I consider implementation sites (schools) as random effects and do not incorporate implementation data.

$$Y_{ik} = \gamma_{00} + \gamma_{01}T_k + \sum_{m=1}^{5} \gamma_{0m} D_{mk} + \gamma_{10}X_{ik} + \sum_l \gamma_{l0}\, W_{cik} + u_{0k} + r_{ik}$$

(Equation 6)

In the model in *(Equation 6,* Y_{ik} represents the score for student *i* from school *k* in the four tests intended to measure program outcomes (WJLWI, WJWA, TOWRE, and WJPC). Treatment effects are represented as γ_{01}. The coefficients for the school district indicators (district *D*) are γ_{0m}, representing the effects of districts 1 through 5 (m).[13] The model controls for student prior achievement (X_{ik}) and a set of student-level covariates (W_{cik}; students' English language learner [ELL] status, special education [SPED] status, age, and gender). School-level and student-level random error ($\mu_{0k}$and r_{ik}, respectively) are assumed to be independently and identically distributed. Model 7 incorporates implementation data as a control variable (I_{Impact}, derived from the aggregated implementation index described previously), modeled as γ_{08}, to test whether implementation affects the program's impact.

The final model (Model 8) incorporates the interaction of the implementation variable (I_{Impact}, modeled as γ_{08}) with two student-level covariates (ELL and SPED). This is intended to determine whether the effect of implementation varies as a function of students' English

[13] Randomization of schools into the treatment and control groups was conducted within each school district. The original model to estimate the impacts of SFA (Quint, 2015) treats districts as fixed effects, not as a third level. This is intended to represent the impact of the intervention in the average SFA school within the five study districts.

Language Learner ($\gamma_{04}(I_{Impact} \times ELL_{ik})$) and Special Education ($\gamma_{05}(I_{Impact} \times SPED_{ik})$) status.

Results

Implementation Adherence in Treatment Schools

Comparison of the Theoretical and Empirical Implementation Indices. In Table 15, I compare the two implementation adherence indices ($I_{SAS\text{-}T}$ and $I_{SAS\text{-}U}$) that can be computed from the data. Schools are ranked from highest to lowest score in the theoretical implementation index ($I_{SAS\text{-}T}$) and unweighted index ($I_{SAS\text{-}U}$).

Table 15: Level of Implementation (as Percentage of Maximum Score) and Ranking of Schools According to Their School Achievement Snapshot Implementation Scores, Theoretical and Unweighted Indices

School id	Theoretical Index ($I_{SAS\text{-}T}$)		Unweighted Index ($I_{SAS\text{-}U}$)	
	Level of implementation (%)	Ranking	Level of Implementation (%)	Ranking
7	95.3	1	95.8	1
32	94.2	2	95.8	1
37	90.1	3	93.2	3
10	88.7	4	92.1	4
22	87.2	5	91.1	5
26	84.1	6	88.9	6
35	82.3	7	80.8	8
36	79.6	8	74.1	9
12	75.7	9	70.8	12
31	73.4	10	81.1	7
30	73.2	11	69.4	15
24	71.1	12	74.1	9
5	69.1	13	70.8	13
6	67.2	14	71.1	11
18	67.2	14	69.5	14
14	62.9	16	66.8	16
13	54.8	17	48.5	17
17	48.5	18	45.3	18
3	36.1	19	32.0	19

Source: Computed from Quint (2016c).

The information above shows that the levels of implementation are similar across the two indices. The linear correlation between the two indices is 0.977 ($p < 0.001$), implying an almost perfect association. Additionally, Kendall's τ —which measures the degree of correspondence between two ranked variables (Turner, 2014)— is 0. 873 ($p < 0.001$). This shows a high correlation between the two methods to measure the levels of program adherence in treatment schools and their respective rankings (Cohen, 1988).

Although there is a clear association, there are differences in scores and rankings for specific schools. School positions tend to be similar at the top and bottom of the rankings in both indices, but there are differences towards the middle. For example, school 30 is placed 11th with the $I_{SAS\text{-}T}$ and 15th with the $I_{SAS\text{-}U}$. However, in terms of the overall implementation score, this school's level of implementation (score achieved as a proportion of the maximum possible score) presents a difference of 3.8 percentage points between the $I_{SAS\text{-}T}$ and $I_{SAS\text{-}U}$, very close to the average difference of 3.6 percentage points across all schools. The most significant differences in levels of implementation are in schools 31 with 7.6 percentage points (73% in the $I_{SAS\text{-}T}$ and 82% the $I_{SAS\text{-}U}$) and 13 with 6.4 percentage points (55% $I_{SAS\text{-}T}$ and 48% in the $I_{SAS\text{-}U}$). In the case of school 31, the gap is mainly due to the difference in scores in construct 2 (14 percentage points higher for the $I_{SAS\text{-}U}$), as the scores are similar for constructs 1 and 3 (difference of 1 percentage point for construct 1 and 3 percentage points for construct 3).

The larger gap in school 13 is due to differences in the implementation levels for construct 2 (15 percentage points higher in the $I_{SAS\text{-}U}$) and construct 1 (12 percentage points higher in the $I_{SAS\text{-}T}$). For school 31, the difference can be attributed to the large gap in construct 2; the slight difference in construct 1 does not offset that. Therefore, this school presents a higher level of implementation in the $I_{SAS\text{-}U}$. Conversely, in school 13, the higher score in the $I_{SAS\text{-}U}$ for construct 2 does not offset the higher level in construct 1 in the $I_{SAS\text{-}T}$, so the overall implementation score is higher in the $I_{SAS\text{-}T}$ as a function of the weight of construct 2. Other schools present relatively large differences in the scores for each construct but similar overall scores. For example, school 35 reported a level of implementation of 82% for construct 1 in the *I_{SAS-T}* and 66% in the *I_{SAS-U}*, with a difference of 16 percentage points (higher *I_{SAS-T}*). Conversely, for construct 2 it obtained a score of 70% in the *I_{SAS-T}* and 85% in the *I_{SAS-U}* (15 percentage points higher in the *I_{SAS-U}*). However, the school has an overall implementation level of 82% in the theoretical index and 81% in the unweighted index.

The SFAF and the research team set the threshold for the adequate level of implementation at 50% of the total score. In the case of the *I_{SAS-T},* this is a score of 48.5 and 22 for the *I_{SAS-U}*. Using the unweighted index, three schools do not meet this threshold (13, 17, and 3). However, using the theoretical index, only two of these schools are below (17 and 3). In this case, using different methodological approaches to compute the total implementation adherence scores would affect the determination of which schools meet the appropriate threshold, as school 13 is compliant with *I_{SAS-T}* but not with *I_{SAS-U}*.

Relationship Between Implementation in Treatment Schools and Student Outcomes.

Correlational Analysis. Table 16 presents the correlations between student test score change and the implementation index constructed using the theoretical program weights (*I_{SAS-T}*). The correlation coefficients indicate a statistically significant linear association between the Challenging Individualized Instruction implementation subconstruct and WJWA (r=0.395, *p*=0.094).

Table 16: Correlation Coefficients Between Test Scores and the Theoretical Implementation Index, I_{SAS-T} *(p-values in parenthesis)*

Construct	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
1. Challenging Individualized Instruction	0.076 (*p*=0.758)	0.395* (*p*=0.094)	0.180 (*p*=0.460)	0.313 (*p*=0.192)

Construct	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
2. Non- Instructional Issues that Affect Reading Instruction	-0.120 (*p*=0.624)	0.206 (*p*=0.397)	0.375 (*p*=0.114)	0.356 (*p*=0.134)
3. Continuous Improvement of Students and Staff	-0.090 (*p*=0.713)	0.228 (*p*=0.347)	0.266 (*p*=0.272)	0.157 (*p*=0.521)
All implementation	-0.017 (*p*=0.944)	0.332 (*p*=0.165)	0.273 (*p*=0.258)	0.309 (*p*=0.198)

Note. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 17 displays the correlation coefficients for the test scores and the unweighted index (I_{SAS-U}) for each construct and across all implementation items. The results indicate that there is no linear association between implementation scores in the unweighted index and test scores (as measured by a difference between the initial and final administration of each test), except for Non- Instructional Issues that Affect Reading Instruction, with a moderate correlation of 0.403 (statistically significant at the 0.1 level).

Table 17: Correlation Coefficients Between Test Scores and the Unweighted Implementation Index, $\mathrm{I}_{\mathrm{SAS-U}}$ *(p-values in parenthesis)*

Construct	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
1. Challenging Individualized Instruction	0.129 (*p*=0.600)	0.283 (*p*=0.241)	0.177 (*p*=0.470)	0.214 (*p*=0.378)
2. Non-Instructional Issues that Affect Reading Instruction	-0.098 (*p*=0.689)	0.302 (*p*=0.208)	0.370 (*p*=0.119)	0.403* (*p*=0.087)
3. Continuous Improvement of Students and Staff	-0.114 (*p*=0.642)	0.207 (*p*=0.396)	0.283 (*p*=0.240)	0.133 (*p*=0.585)

Construct	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
All implementation	-0.034 (*p*=0.891)	0.283 (*p*=0.241)	0.299 (*p*=0.213)	0.266 (*p*=0.271)

Note. * p < 0.1, ** p < 0.05, *** p < 0.01.

The information presented in Tables 12 and 13 indicates that there tends to be no linear association between the theoretical and empirical implementation indices for most measures of implementation (across all constructs and items) and all the tests. $I_{SAS\text{-}T}$ and $I_{SAS\text{-}U}$ present moderate correlations between one construct and a specific test, but each index has different associations. In the case of the theoretical implementation index, the correlation is significant for the relationship between the Challenging Individualized Instruction construct and the difference in scores in the WJWA test. For the unweighted index, there is a significant correlation between the Non-Instructional Issues that Affect Reading Instruction and the WJPC test.

Multilevel Models. The results of the model testing the association between implementation (by construct) and student outcomes are presented in Table 18. Only the coefficient associated with construct 1 (Challenging Individualized Instruction) in the $I_{SAS\text{-}U}$ and the Woodcock-Johnson Letter Word Identification (WJLWI) test is statistically significant (0.166, *p*=0.0991). None of the other coefficients are statistically significant for the theoretical or unweighted indices ($I_{SAS\text{-}T}$ and $I_{SAS\text{-}U}$). Thus, controlling for student characteristics, following the SFA curriculum more closely (in terms of challenging instruction) by 10 percentage points is associated with an increase of 0.166 standard deviations in test scores for the WJLWI. However, this positive association does not hold for the $I_{SAS\text{-}T}$, as the coefficient is not statistically significant (0.077, *p*=0.394).

Table 18: HLM Regression Coefficients (and Standard Errors) for Student Test Scores with Implementation Scores –SAS Theoretical Implementation Index ($I_{SAS\text{-}T}$*) and Unweighted Implementation Index, by Construct*

	Woodcock-Johnson Letter-Word Identification		Woodcock-Johnson Word Attack		Test of Word Reading Efficiency		Woodcock-Johnson Passage Comprehension	
	Theoretical weights ($I_{SAS\text{-}T}$*)*	*Unweighted (*$I_{SAS\text{-}U}$*)*	*Theoretical weights (*$I_{SAS\text{-}T}$*)*	*Unweighted (*$I_{SAS\text{-}U}$*)*	*Theoretical weights (*$I_{SAS\text{-}T}$*)*	*Unweighted (*$I_{SAS\text{-}U}$*)*	*Theoretical weights (*$I_{SAS\text{-}T}$*)*	*Unweighted (*$I_{SAS\text{-}U}$*)*
Intercept	0.845 (0.664)	0.689 (0.633)	-0.021 (0.656)	0.036 (0.653)	0.717* (0.382)	0.763* (0.372)	-0.454 (0.646)	-0.434 (0.637)
Implementation								
1. Challenging Individualized Instruction	0.077 (0.088)	0.166* (0.095)	0.062 (0.054)	0.057 (0.063)	-0.003 (0.042)	-0.031 (0.046)	0.071 (0.065)	0.079 (0.074)
2. Non-Instructional Issues that Affect Reading Instruction	-0.077 (0.076)	-0.088 (0.066)	-0.034 (0.048)	-0.008 (0.045)	0.041 (0.036)	0.048 (0.032)	-0.013 (0.058)	0.001 (0.052)
3. Continuous Improvement of Students and Staff	0.016 (0.057)	-0.0144 (0.051)	0.014 (0.035)	-0.003 (0.034)	0.001 (0.027)	0.007 (0.024)	-0.012 (0.043)	-0.027 (0.040)
Student's Earliest Available Test Score	0.282*** (0.025)	0.284*** (0.025)	0.488*** (0.028)	0.489*** (0.028)	0.788*** (0.017)	0.787*** (0.017)	0.403*** (0.026)	0.404*** (0.026)
Student's ELL status	-0.556*** (0.065)	-0.564*** (0.065)	-0.349*** (0.068)	-0.355*** (0.069)	-0.065 (0.043)	-0.059 (0.043)	-0.595*** (0.064)	-0.600*** (0.065)
Student's SPED status	-0.452*** (0.098)	-0.454*** (0.098)	-0.189* (0.113)	-0.190* (0.113)	-0.096 (0.066)	-0.096 (0.066)	-0.380*** (0.104)	-0.383*** (0.104)
Age	-0.120* (0.067)	-0.123* (0.067)	-0.025 (0.081)	-0.025 (0.081)	-0.110** (0.043)	-0.109** (0.043)	0.040 (0.075)	0.040 (0.075)
Gender (male)	-0.014 (0.049)	-0.013 (0.049)	0.082 (0.052)	0.083 (0.052)	0.041 (0.032)	0.041 (0.032)	0.056 (0.049)	0.057 (0.048)

Notes. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Source: Quint (2016a, 2016c).

The following models (Table 19) incorporate the overall implementation index as a moderator (model 3) and an interaction term to determine whether the effect of implementation varies as a function of students' ELL or SPED status (model 4). Implementation is significantly associated with students' outcomes only for the TOWRE (0.042, *p*=0.077), but no significant association exists with any other tests. While there is no main effect of implementation across all tests, the interactions are significant across most outcomes. This reveals that the level of implementation is differently associated with outcomes for different groups of students. The main effects show no significant differences for ELL students, except in the WJLWI test, where ELL students score higher on average (0.665, *p*=0.028). However, a higher level of implementation adherence is associated with a smaller difference in scores, as the regression coefficients are negative and statistically significant for the interaction for most tests. Overall, higher levels of implementation adherence are associated with a reduction in the test scores for ELL students, except for the TOWRE. The gap is most prominent for the WJLWI test, where an increase of 10 percentage points in the overall $I_{SAS\text{-}T}$ is associated with a decrease of 0.119 standard deviations in the advantage that ELL students see in their scores.

This is followed by the WJPC test, where an increase of 10 percentage points in implementation is associated with a decrease of 0.074 standard deviations in average test scores between ELL and non-ELL students (0.372, *p*=0.215).

Table 19: HLM Regression Coefficients (and Standard Errors) for HLM Models 3 and 4–SAS Theoretical Index (I_{SAS-T}*)*

	Woodcock-Johnson Letter-Word Identification		Woodcock-Johnson Word Attack		Test of Word Reading Efficiency		Woodcock-Johnson Passage Comprehension	
	Model 3	*Model 4*	*Model 3*	*Model 4*	*Model 3*	*Model 4*	*Model 3*	*Model 4*
Intercept	1.063*	0.769	0.098	0.031	0.633*	0.744**	-0.308	-0.444
	(0.618)	(0.609)	(0.640)	(0.647)	(0.363)	(0.366)	(0.619)	(0.622)
Implementation (I_{SAS-T})	0.008	0.047	0.037	0.053	0.042*	0.783	0.032	0.401
	(0.049)	(0.047)	(0.030)	(0.032)	(0.023)	(0.017)	(0.035)	(0.026)
Student’s Earliest Available	0.282***	0.282***	0.491***	0.483***	0.786***	0.028***	0.405***	0.057***
Test Score	(0.025)	(0.025)	(0.028)	(0.028)	(0.017)	(0.024)	(0.026)	(0.037)
Student’s ELL Status	-0.556***	0.665**	-0.348***	0.394	-0.065	-0.066	-0.597***	0.372
	(0.065)	(0.302)	(0.068)	(0.314)	(0.043)	(0.196)	(0.064)	(0.299)
Student’s SPED Status	-0.453***	-1.455***	-0.187*	-1.735***	-0.096	-1.501***	-0.380***	-2.024***
	(0.098)	(0.414)	(0.113)	(0.476)	(0.066)	(0.279)	(0.104)	(0.436)
Age	-0.121*	-0.121*	-0.025	-0.033	-0.109**	-0.109**	0.040	0.033
	(0.067)	(0.067)	(0.081)	(0.080)	(0.043)	(0.043)	(0.075)	(0.074)
Gender (male)	-0.013	-0.012	0.083	0.077	0.041	0.039	0.057	0.050
	(0.049)	(0.049)	(0.052)	(0.052)	(0.032)	(0.031)	(0.048)	(0.048)
Implementation (I_{SAS-T}) *		-0.167***		-0.102**		-0.001		-0.132***
Student’s ELL Status		(0.040)		(0.042)		(0.026)		(0.039)
Implementation (I_{SAS-T}) *		0.143**		0.220***		0.197***		0.234***
Student’s SPED Status		(0.056)		(0.065)		(0.038)		(0.060)
-2 log Likelihood	3065.2	3040	2330.8	2313	1761.12	1734.28	2160.6	2133.2

Notes. * p < 0.1, ** p < 0.05, *** p < 0.01.
Model 4 shows a better fit than Model 3 for all tests.

Contrary to what can be observed for ELL status, there is a positive association between implementation adherence and test scores for students in SPED across all the tests. The effect is more significant in the WJPC test, where an increase of 10 percentage points in the overall $I_{SAS\text{-}T}$ is associated with an increase of 0.292 standard deviations in the difference that students in SPED have in their scores compared to their non-special education peers (-2.024 standard deviations). The positive association holds for the other tests: 0.190 standard deviations for the WJLWI, 0.273 for the WJWA, and 0.225 for the TOWRE.

Implementation in Treatment and Control Schools.

Comparison of Implementation Scores in Treatment and Control Schools. Implementation scores show that schools in the treatment group obtain higher levels of implementation than their counterparts in the control group. This indicates that treatment schools are completing activities aligned with the SFA model at a higher percentage than control schools, as measured by the variables that the program designers determined were relevant. Although this is not meant to be an explicit measure of implementation adherence, the questionnaire items are aligned with practices that should be observed in SFA schools, so treatment schools are expected to show higher scores. Table 20 presents the mean scores averaged at the school level, with the unweighted index (I_{Impact}).

Table 20: Implementation Scores from the Unweighted Implementation Index ($\mathrm{I}_{\mathrm{Impact}}$*) by Treatment Status*

Group	Mean		Minimum		Maximum		Difference in means (T-C)
	Percentage (%)	Raw Score	Percentage (%)	Raw Score	Percentage (%)	Raw Score	
Treatment	82.7	11.579	57.5	8.054	94.6	13.244	3.907*** (*p*<0.001)
Control	54.8	7.672	37.5	5.256	67.1	9.389	

Note. * p < 0.1, ** p < 0.05, *** p < 0.01.
Source: Quint (2016b, 2016e)

The information above reflects a considerable difference in the mean, minimum, and maximum implementation scores between treatment and control schools, indicating that practices in SFA schools are more closely aligned with the intervention.

Relationship Between Implementation and Student Outcomes. In Table 21, models 6 and 7 estimate the impact of SFA on the different reading tests. Model 7 incorporates implementation as a covariate through the I_{Impact} index.

Table 21: HLM Regression Coefficients (and Standard Errors) for HLM Models 6 and 7 –Impact Implementation Index

	Woodcock-Johnson Letter-Word Identification		Woodcock-Johnson Word Attack		Test of Word Reading Efficiency		Woodcock-Johnson Passage Comprehension	
	Model 6	*Model 7*	*Model 6*	*Model 7*	*Model 6*	*Model 7*	*Model 6*	*Model 7*
Intercept	1.191*** (0.391)	1.423*** (0.434)	0.218 (0.509)	0.313 (0.556)	0.984*** (0.282)	1.028*** (0.315)	-0.253 (0.482)	-0.162 (0.517)
Treatment Status (Treatment schools)	0.102 (0.076)	0.247* (0.139)	0.115 (0.085)	0.171 (0.162)	0.056 (0.053)	0.083 (0.101)	0.108 (0.070)	0.165 (0.134)
Implementation (I_{Impact})		-0.055 (0.045)		-0.021 (0.052)		-0.010 (0.033)		-0.022 (0.043)
District 1	0.089 (0.154)	0.125 (0.152)	0.073 (0.171)	0.085 (0.177)	0.256** (0.106)	0.263** (0.110)	0.355** (0.142)	0.368** (0.146)
District 2	0.644*** (0.144)	0.729*** (0.156)	0.295* (0.160)	0.325* (0.180)	0.117 (0.099)	0.132 (0.112)	0.611*** (0.132)	0.643*** (0.149)
District 3	-0.172 (0.121)	-0.125 (0.123)	-0.064 (0.133)	-0.048 (0.142)	0.263*** (0.083)	0.271*** (0.089)	0.072 (0.112)	0.090 (0.119)
District 4	0.265* (0.145)	0.327** (0.149)	0.198 (0.160)	0.220 (0.172)	0.218** (0.100)	0.229** (0.108)	0.376*** (0.134)	0.400** (0.144)
Student's Earliest Test Score	0.322*** (0.021)	0.321*** (0.021)	0.529*** (0.023)	0.528*** (0.023)	0.791*** (0.014)	0.791*** (0.014)	0.425*** (0.022)	0.425*** (0.022)
Student's ELL status	-0.453*** (0.055)	-0.449*** (0.055)	-0.186*** (0.057)	-0.185*** (0.057)	-0.048 (0.036)	-0.048 (0.036)	-0.496*** (0.054)	-0.494*** (0.054)
Student's SPED status	-0.376*** (0.081)	-0.377*** (0.081)	-0.160* (0.086)	-0.161* (0.086)	-0.085 (0.053)	-0.085 (0.053)	-0.335*** (0.081)	-0.335*** (0.081)

	Woodcock-Johnson Letter-Word Identification		Woodcock-Johnson Word Attack		Test of Word Reading Efficiency		Woodcock-Johnson Passage Comprehension	
	Model 6	*Model 7*	*Model 6*	*Model 7*	*Model 6*	*Model 7*	*Model 6*	*Model 7*
Age	-0.152*** (0.050)	-0.152*** (0.050)	-0.029 (0.066)	-0.029 (0.066)	-0.144*** (0.036)	-0.144*** (0.036)	0.021 (0.063)	0.021 (0.063)
Gender (male)	0.007 (0.040)	0.007 (0.040)	0.118*** (0.042)	0.118*** (0.042)	0.027 (0.026)	0.027 (0.026)	0.057 (0.040)	0.057 (0.040)
-2 log Likelihood	4220.4	4218.2	3216.4	3216	2399.4	2399.4	3056	3055.6

Notes 1. * p < 0.1, ** p < 0.05, *** p < 0.01.
2. There are no differences in model fit between Model 6 and Model 7 for any of the tests.

None of the coefficients for the treatment school indicator are statistically significant for any of the four tests (model 6). This matches what was observed in the evaluation report (Quint et al., 2015), where students exposed to SFA did not, on average, achieve higher scores in any of the reading tests compared to their peers in schools that implemented other reading programs. Conversely, student-level covariates such as previous test scores, ELL and SPED status, and age are significant predictors of final test scores.

Nonetheless, in model 7, the effect of SFA is positive for the Woodcock-Johnson Letter-Word Identification test, as students in treatment schools score on average 0.247 standard deviations higher than their peers exposed to other reading programs. This indicates that controlling for the level of program implementation across schools and comparing schools with the same level of implementation, SFA has a positive impact on reading decoding and sight word recognition. Adding implementation as a covariate and controlling for this variation in the model may help to reveal the effect of SFA on this specific student outcome.

A comparison of models 6 and 7 shows that incorporating implementation as a covariate does not improve model fit (differences in -2 Log likelihood model fit are not statistically significantly different from zero). Although implementation is not associated with program impacts, the SFA has a positive impact on reading scores in the WJLWI (that measures reading decoding and sight word recognition) when controlling for implementation (0.247 standard deviations compared to control schools, p=0.091).

Model 8 (Table 22) adds two interaction terms to model 7 to determine whether the effect of implementation as measured by the impact index (I_{Impact}) in treatment and control schools varies as a function of students' ELL and SPED status.

Table 22: HLM Regression Coefficients (and Standard Errors) for HLM Model 8 –Impact Implementation Index

	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
Intercept	1.305*** (0.438)	0.147 (0.561)	1.024*** (0.318)	-0.310 (0.518)
Treatment Status (Treatment schools)	0.251* (0.140)	0.175 (0.165)	0.085 (0.102)	0.169 (0.131)
Implementation (I_{Impact})	-0.027 (0.031)	-0.002 (0.036)	-0.008 (0.022)	-0.002 (0.029)
District 1	0.111 (0.153)	0.056 (0.180)	0.258** (0.111)	0.352** (0.144)
District 2	0.710*** (0.157)	0.294 (0.184)	0.130 (0.113)	0.620*** (0.147)
District 3	-0.130 (0.124)	-0.056 (0.144)	0.271*** (0.089)	0.086 (0.116)
District 4	0.322** (0.150)	0.210 (0.175)	0.228** (0.108)	0.398*** (0.141)
Student’s Earliest Test Score	0.322*** (0.021)	0.530*** (0.023)	0.792*** (0.014)	0.426*** (0.022)
Student’s ELL status	-0.006 (0.235)	0.505** (0.235)	0.075 (0.156)	-0.013 (0.223)
SPED status	-0.131 (0.327)	-0.471 (0.337)	-0.487** (0.211)	-0.060 (0.321)
Age	-0.148*** (0.050)	-0.021 (0.066)	-0.142*** (0.036)	0.026 (0.063)

	Woodcock-Johnson Letter-Word Identification	Woodcock-Johnson Word Attack	Test of Word Reading Efficiency	Woodcock-Johnson Passage Comprehension
Gender (male)	0.009 (0.040)	0.123*** (0.042)	0.027 (0.026)	0.062 (0.040)
Implementation (I_{Impact}) * Student's ELL Status	-0.044* (0.023)	-0.069*** (0.023)	-0.012 (0.015)	-0.048** (0.022)
Implementation (I_{Impact}) * Student's SPED Status	-0.026 (0.033)	0.033 (0.034)	0.042* (0.022)	-0.029 (0.033)
-2 log Likelihood	4213.8	3206.2	2394.8	3049.6

Notes 1. * p < 0.1, ** p < 0.05, *** p < 0.01.

2. Model 8 shows a better fit than Model 7 for the WJWA (χ^2=9.831, *p*=0.007) and WJPC (χ^2=6.069, *p*=0.048).

The model shows that higher levels of fidelity of implementation negatively affect ELL students. This is evident in the WJLWI, WJWA, and WJPC tests, where these students score lower than their peers when schools show higher levels of fidelity. The case of the WJWA is different from the other tests, as, on average, ELL students score higher than their peers in this assessment (the coefficient of the main effect is 0.505 standard deviations, *p*=0.03). However, this advantage is reduced at higher levels of implementation fidelity, with a decrease of 0.069 standard deviations in this gap for each 10 percentage-point increase in the impact implementation index. Additionally, incorporating the two interaction terms (model 8) improves model fit when compared to model 7, indicating that the interaction between fidelity of implementation and ELL status explains a significant part of the variance in WJWA and WJPC test scores.

Conversely, more fidelity leads to higher TOWRE scores for SPED students. The main effect of SPED on TOWRE scores is negative, with SPED students scoring 0.487 standard deviations lower than their peers (*p*=0.02). However, an increase of 10 percentage points in implementation fidelity is associated with a reduction of 0.042 standard deviations in the gap between SPED and non-SPED students. Model 8 does not show a better fit than model 7, indicating that the interaction terms do not significantly contribute to explain more of the variance in TOWRE test scores.

Discussion

The analysis revealed several key insights into the relationship between program implementation and student outcomes.

Implementation Adherence in Treatment Schools

Firstly, the comparison of the indices used to measure implementation adherence in treatment schools as measured by the SAS highlights the relevance of the selected weighting scheme. Scores from the theoretical index proved very similar to the unweighted scores, particularly for schools obtaining the highest and lowest scores. However, some schools show different positions in the rankings for each scale (especially towards the middle of the distribution). Similarly, some schools with varying scores in each implementation construct have similar overall implementation levels, while others with similar levels in some of the constructs have different final scores.

Although, in this case, both methods yield similar results (as shown by the high correlation between the indices), the differences in rankings towards the middle of the scale could have practical implications for schools. For example, researchers may use this information to make decisions about providing additional implementation support to schools, selecting them based on their rank in the index. Even though for this enactment of SFA, the weights of the implementation variables do not make a relevant difference in terms of the score, the decisions related to the weighting scheme could have a more significant impact for other programs.

The theoretical index places greater importance on the components that program designers determine are relevant and should affect implementation and outcomes more. This type of schema can follow program theory more closely, as it may assume that the components that are hypothesized to have a stronger effect on outcomes should account for a larger proportion of the final implementation score.

The correlational analyses suggested a potential link between higher levels of adherence to the program and student reading outcomes. Specifically, the Challenging Individualized

Instruction construct in the theoretical implementation index ($I_{SAS\text{-}T}$) showed a moderate association with improved scores on the WJWA test, indicating an association between adherence to the school-level instruction model in SFA and students' word identification skills (r=0.395, p=0.094). Conversely, the Non-Instructional Issues that Affect Reading Instruction subconstruct in the unweighted implementation index ($I_{SAS\text{-}U}$) showed a moderate positive association with the WJPC test. These were the only significant correlations for each index, proving that the associations depend on the method used to construct the index. However, these moderate and significant associations do not hold in the multilevel models, implying that the covariates added (students' first available test score and ELL and SPED status) explain a more significant proportion of the variation in test scores.

In the models that incorporate the overall implementation scores ($I_{SAS\text{-}T}$), higher levels of implementation adherence are associated with a reduction in the test scores for ELL students, except for the TOWRE. The effect of implementation and ELL on test scores is most prominent for the WJLWI test, where an increase of 10 percentage points in the overall $I_{SAS\text{-}T}$ is associated with a decrease of 0.119 standard deviations in the advantage that ELL students see in their scores. This effect may be due to a potential misalignment between the SFA curriculum and ELL students' needs, especially those measured in the WJWLI, WJWA, and WJPC.

Although the SFA program contains Spanish-language interventions (e.g., *Descubre Conmigo* and *Lee Conmigo*), there is no specific information available to determine whether ELL students were exposed to these versions. Qualitative data collected as part of the evaluation of SFA indicates that teachers have concerns about its suitability for ELL students, particularly regarding the quality and availability of materials in Spanish. Some teachers even said they opted not to use SFA materials, instead using the district text or creating their own (Quint et al., 2015).

This aligns with the findings presented, as the negative relationship between implementation fidelity and test scores for ELL students supports teachers' concerns about the benefits of SFA for students who are not proficient in English.

Contrary to ELL, higher levels of adherence are associated with a decrease in the difference in test scores between students with and without SPED status. On average, students in SPED tend to score lower than the group that is not under this classification on all the tests. At very high levels of implementation, this difference in scores may be significantly reduced, although not eliminated, especially in the TOWRE. This is especially interesting when contrasted with the general program findings, as 45% of the teachers in treatment schools indicated that they did not agree that their reading program adequately serves SPED students (Quint et al., 2015, p. 62). Similarly, the program expected to lower special education assignment rates as part of their long-term outcomes, but the data showed SFA had no effect. Although, on average, teacher reports point towards a misalignment between SFA and the needs of SPED students, the evidence indicates that higher levels of fidelity can lead to a reduction in the gap in test scores between SPED and non-SPED students. Nonetheless, additional information would need to be collected to determine whether specific components or instructional practices associated with SFA benefit these students.

The models show that the relationship between adherence to the SFA design is not the same for the TOWRE as the other tests. This is evident in that implementation adherence was not associated with ELL students only for this test, and higher adherence was linked to a more significant reduction in the gap in TOWRE test scores for SPED students. These differences may be attributed to an alignment among the specific reading skills measured in the TOWRE (efficiency of sight word recognition), the SFA curriculum, and these students' learning needs.

However, test scores across the four tests are highly correlated ($r > 0.7$), indicating that it is likely they all measure similar abilities in students, so the alignment between tests and SFA components should be analyzed in detail. The lack of a comparison group means that these associations should not be interpreted as causal.

Implementation in Treatment and Control Schools.

An underlying hypothesis of the implementation of the SFA intervention is that a high level of adherence to the instructional model—as reflected by higher implementation scores—should be associated with an increase in student test scores when compared to the group that received their usual reading instruction. Exploring implementation data from both treatment and control schools, using composites from principal and teacher surveys, offered additional perspectives on how variations in program delivery might explain differences in student outcomes.

The considerable difference in the mean, minimum, and maximum implementation scores between treatment and control schools in the I_{Impact} index points to a clear differentiation in practices between the two groups of schools. However, this only partially aligns with other data on implementation available in the evaluation report, as Quint et al. (2015) found no differences in key SFA practices (e.g., extended reading periods, data usage, and tutoring for struggling students) when using data from surveys and teacher logs. This may indicate that an aggregated implementation index does not reflect the nuances in teacher practices that may be happening within and between schools. At the same time, the lack of differentiation between the two groups may lead to questions about the internal validity of the study, opening the possibility that students in the treatment schools were not exposed to SFA-specific practices that are required to validate the causal link between the intervention and the outcomes. Similarly, the absence of a clear distinction may point to students in control schools being exposed to similar practices as those in treatment schools, indicating that SFA is not sufficiently different from other curricula for the potential effects to be attributable to its implementation.

Regardless of this, model 7 shows a positive effect of SFA on the WJLWI test when controlling for implementation, as students in treatment schools obtain better scores (0.247 standard deviations higher) than their peers who do not participate in SFA. This positive effect holds in model 8, which adds the interaction between implementation and students' SPED and ELL status into the model. SFA appears to have a positive impact on foundational reading skills (measured by the WJLWI) and not on more advanced reading competencies such as comprehension (WJPC; (Wendling et al., 2007)).

The effect of fidelity of implementation to the SFA program is complex, especially for ELL and SPED students. Overall, these students tend to score lower than their non-ELL or SPED peers on most tests, but implementation has different effects on their scores. Higher levels of fidelity of implementation negatively affect ELL students, increasing the gap in their scores with non-ELL students. However, ELL students present an advantage over their peers in the WJWA (0.5 standard deviations), which decreases by 0.07 standard deviations for each 10-percentage point increase in implementation scores, showing that higher adherence to the SFA design does not benefit them. The type of test could explain this, as the WJWA measures students' ability to translate nonwords, such as "nat" or "ib," into sounds (Wendling et al., 2007). Furthermore, the fact that the test evaluates sounds and not actual words may benefit students who are not fluent in English.

For SPED students, model 8 shows that higher levels of fidelity of implementation lead to an increase in their TOWRE scores. These students obtain scores on average 0.49 standard deviations lower than their peers. However, this disadvantage is offset by 0.04 standard deviations for every 10-percentage point increase in the fidelity on the implementation scale. This relationship may be explained by the constructs measured in this test, which focuses on students' ability to read and pronounce words and nonwords (Tarar et al., 2015). The convergence between specific SFA practices and SPED students' reading abilities could be explored further to determine if any particular components in the SFA curriculum help them improve these specific reading skills.

As to the overall impact of implementation on student outcomes, the coefficients for the implementation index I_{Impact} in models 7 and 8 are negative for all the tests, although they are not statistically significant. These non-significant coefficients do not allow for inferences about the precise relationship between adherence to the SFA curriculum and students' reading scores. Still, the negative signs for both models and across all tests indicate that the association may not be positive. In terms of the statistical analysis, the absence of a significant coefficient may mean there was not enough power in this sample to find an effect, considering there were only 27 schools (14 in the treatment and 13 in the control group).

Conclusion

In this chapter, I examined the implications of various methodologies and analytical approaches to measuring implementation to understand the effects of SFA on students' reading test scores. The analysis highlighted how different measurement strategies can influence the interpretation of the program's effectiveness and its impact on diverse student populations.

The results indicated that high levels of implementation adherence were positively associated with test scores for special education (SPED) students but negatively for English language learner (ELL) students. This differential relationship suggests that close adherence to the SFA program could benefit specific student groups while potentially placing others at a disadvantage. The positive outcomes for special education students may indicate that the structured and consistent implementation of SFA aligns well with their learning needs, providing the necessary support to improve their reading skills. Conversely, the negative association observed for ELL students raises important questions about the SFA program's suitability for learners who face language barriers. This finding warrants further investigation into potential aspects within the SFA model that could be contributing to these differences. Specifically, exploring whether there is a relationship between SFA and the skills measured in each test for these students is crucial, as it could shed light on which components of the program are effective and which may need adaptation for ELL students.

Moreover, instructional practices in treatment and control schools could be explored further to determine if and how they affect students' outcomes and the effectiveness of SFA. For example, the evaluation reported that SFA schools focused more on cooperative learning and cross-grade grouping but found no differences between SFA and control schools in some core components (extended reading periods and tutoring for struggling students) (Quint et al., 2015). These qualitative insights could reveal underlying factors that contribute to the observed differences in outcomes and inform strategies to understand the mechanisms that generate the impacts (or lack thereof) of SFA on students.

Overall, the analysis reveals a complex relationship between implementation and test scores. The findings suggest differential effects for subgroups of students facing specific academic challenges, particularly ELL and SPED students. Moreover, the consistent pattern across models indicates that higher implementation fidelity does not necessarily yield greater benefits for students' reading scores. Consequently, further research is warranted, utilizing data from additional implementations of the SFA program, to determine whether this pattern represents a persistent trend or is unique to the current dataset.

Summary of Findings and Conclusion

Measuring implementation fidelity and adaptation in educational interventions presents substantial methodological complexities. The multifaceted nature of educational programs, which often encompasses multiple components delivered in diverse contexts, contributes to significant variability in how fidelity and adaptation are conceptualized and assessed. This complexity is evident in the frequency with which the constructs to define implementation are used, the diversity of instruments employed for data collection, the various methods for aggregating implementation data, and the different approaches to incorporating fidelity measures into analyses of program effectiveness.

Research on implementation provides insights into the challenges and complexities of enacting a program across diverse educational settings and helps clarify the mechanisms that affect an intervention's success. In this dissertation, I explored the two main frameworks used to assess implementation (fidelity and adaptation), looking at how they are conceptualized and measured in practice. My goal was to gain a better understanding of how variation in implementation across sites is conceptualized and measured and whether the methodological decisions behind the measurement of implementation can affect the way we view the relationship between implementation and outcomes.

The first research question examines how implementation is conceptualized and measured in educational research. To accomplish this, I studied a large sample of grants awarded through the Institute of Education Sciences' Education Research Grants between 2007 and 2019. While not representative in a statistical sense, this sample of studies is highly informative, as it selects studies from a highly competitive grant program that awards funding to evaluate the effectiveness of educational interventions quantitively (Whitehurst, 2018). I analyzed the different constructs, data collection instruments, and aggregation methods used and how these studies associate implementation data with student outcomes. This provides a methodological foundation for understanding how researchers address the complex issue of program implementation in diverse contexts.

With the second research question, I investigated the practical implications of these methodological decisions, using data from the implementation of the Success for All (SFA) intervention. This section of the dissertation expands on the evaluation by Quint et al. (2015) by comparing treatment effect estimates from models that incorporate implementation indices to those of the original assessment, which did not include implementation data in their models. I sought to determine whether including implementation data provides additional explanatory power regarding the results of the SFA program.

Summary of Findings

Understanding Implementation Fidelity in IES-Funded Research.

The systematic review of IES-funded studies highlighted a strong emphasis on fidelity, particularly close adherence to prescribed program components, as a primary measure of implementation. This aligns with the requirement in the Requests for Applications (RFAs), where the IES asks investigators to include an assessment of fidelity in their research design. The other constructs that are often measured (dosage and quality of delivery) also suggest that researchers prioritize ensuring that program components are enacted consistently across settings.

Despite the emphasis on fidelity (frequently understood as adherence), I found significant variability in how this construct is measured in published research. Studies identify different specific constructs (e.g., adherence, quality, dosage, etc.), use various instruments to assess implementation (e.g., observation rubrics, surveys, logs, etc.), and employ diverse methods to aggregate and report the data. This makes it challenging to generate a single definition of fidelity in educational research and provides further evidence of the methodological complexities behind measuring implementation. While fidelity measures aim to assess the extent to which programs are delivered as intended, the variability in implementation and the lack of predefined standards challenge the interpretation of what constitutes adequate fidelity.

Additional difficulties arise when researchers compare implementation in sites that are using their intervention (treatment) and those in the comparison condition (control), as the data collection methods need to be adjusted to account for practices that can be observed in classrooms that are not using the intervention. This is particularly the case with constructs that aim to capture *quality* of delivery in classroom settings, as this construct may refer to two distinct concepts: *instruction* and *implementation*. While *quality of instruction* may be intervention-agnostic and observed across classrooms, it does not necessarily provide information on fidelity of implementation to the original design. This conflation of the two is conceptually unclear, as it can muddle the relationship between general instructional practices and adherence to the program design.

The lack of in-depth analysis of implementation—especially regarding its variation across sites—suggests that studying implementation is often treated more as a procedural requirement to prove internal validity, rather than as a genuine effort to understand classroom dynamics and the mechanisms through which interventions produce changes in students' learning. Similarly, the thresholds that determine the acceptable levels of implementation are frequently unclear, which points towards a general lax association between program theory and its enactment in practice. In this context, priority is typically placed on using information on implementation to show that adequate levels of fidelity were reached to support the internal validity of treatment estimates, with little attention given to gaining a better understanding of what instruction looks like inside the classroom. A thorough analysis of implementation could provide a greater insight into the instructional practices that teachers are using inside the classroom while at the same time providing information on potentially productive adaptations that can enhance the intervention.

The examination of fidelity and adaptation in educational research underscores the complexity of implementing interventions in real-world classroom settings. While fidelity measures aim to assess the extent to which programs are delivered as intended, the lack of predefined standards and the natural variability in implementation expected across classrooms challenge the interpretation of what constitutes adequate fidelity. At the same time, an in-depth analysis of variations across and within schools and their potential implications for the intervention could threaten the causal link between the intervention and the outcomes. Although, understandably, researchers may not want to expose the intervention to these types of challenges, it is important to question the usefulness of this approach to educational improvement as a whole and to teacher practice in particular.

Adaptation as Absence of Fidelity.

Adaptation adds another layer of complexity to implementation research. Teachers naturally modify interventions to align with their student's needs in the context of their classroom, and these adaptations can positively impact learning outcomes. This was evident in the IES studies and the implementation of SFA. For example, some teachers indicated they found the SFA materials inadequate for ELL students, so they had to adapt to better suit their students' needs (Quint et al., 2015). In the context of IES studies, research on adaptation indicated that teachers modified the curriculum by adding or extending program activities to enhance student learning (see e.g., Burkhauser & Lesaux, 2017; Firetto et al., 2019; Monte-Sano et al., 2014).

Adaptation — the degree to which educators modify programs to fit their unique contexts — was measured less often and less systematically than fidelity. This finding aligns with the literature that views fidelity and adaptation as distinct constructs and not two ends of an implementation continuum. In a context where studies are required to measure fidelity (often conceptualized as adherence), adaptations are viewed as deviations from fidelity, so they are conceptualized as an absence of fidelity. This approach overlooks the potential benefits that these changes to the original design may have on students and their academic outcomes. Recognizing the role of teacher agency and the dynamic nature of classroom environments is essential for understanding how interventions function in practice. However, measuring and accounting for adaptations is complicated in the context of IES research, as the incentive is placed on estimating the impact of a pre-designed intervention, validating the causal link in its theory of change.

Nonetheless, researchers can benefit from placing a stronger emphasis on understanding the interplay between fidelity, adaptations, and educational outcomes, as this can offer critical insights informing educational practices and program decisions. By acknowledging and investigating the complexity of implementation, researchers can better understand what is happening inside the classroom and design interventions that can fit the specific needs of different educational communities. This approach will contribute to disentangling the mechanisms that generate changes in students' learning without assuming that the benchmark should be higher adherence to the design. Adaptations can be productive, and they are inevitable, so they should be measured and accounted for.

SFA Implementation: Fidelity and Adaptation.

The examination of the SFA program provided a concrete example of the complexities involved in program implementation and its measurement. The SFA model emphasizes fidelity to its structured curriculum, with scripted lessons and specific grouping strategies intended to improve reading skills in elementary school students. However, this structured approach often clashes with the practical realities of diverse classroom environments. Implementation data collected from SFA schools showed that teachers frequently adapted parts of the program to meet their students' perceived needs. For instance, around 55% of teachers in treatment schools reported making changes to the SFA reading program, albeit to a lesser extent than teachers in control schools. This finding suggests that even within a structured program like SFA, educators perceive a need to modify it in response to the specific conditions and challenges they encounter in the classroom.

The first part of this dissertation compared two methods commonly used to summarize information about implementation adherence in treatment schools. I constructed the implementation indices using theoretical weights (determined by program designers) and no weighting schemes (unweighted index). My analyses revealed few differences when considering overall implementation scores across scales (expressed as a percentage of the maximum score). However, there were some discrepancies in schools' rankings, particularly in the middle of the distribution, where schools saw differences in their relative positions when using the weighted versus unweighted indices. This could have meaningful implications in practice, as the SFAF (or a similar program) could plausibly make decisions to provide additional implementation support to schools based on an overall assessment of their implementation levels.

Moreover, while weighted and unweighted indices provided broadly consistent estimates of total implementation in the case of SFA, this may not be the case for other interventions, as different constructs, data collection instruments, and aggregation methods can lead to different results. For example, an item or construct that presents high variability across sites could yield different levels of implementation using a weighted index (that gives more preponderance to this item or construct) or an unweighted index (where all items or constructs have the same importance).

Relationship Between the Implementation of SFA and Student Outcomes.

The investigation revealed a complex association between adherence to SFA components and student reading outcomes. Using only information from treatment schools and implementation indices built with adherence to SFA, correlational analyses indicate that higher levels of adherence to the program's prescribed components are associated with improved performance in specific reading skills. On the one hand, higher implementation measured with the theoretical index (which uses the weighting scheme that program designers assigned to the implementation constructs) shows a raw positive association with phonics and decoding skills, as measured by the Woodcock-Johnson Word Attack (WJWA) test. Similarly, the unweighted index (in which all constructs contribute equally to the index) is positively correlated with an increase in reading and language comprehension assessed by the Woodcock-Johnson Passage Comprehension (WJPC) test.

However, these significant associations disappear when student and school-level covariates are incorporated into a hierarchical linear model (model 3, with students nested in schools). This indicates that variables other than implementation may explain a more significant portion of the variation in student test scores. These models show a positive association between

SFA and the Test of Word Reading Efficiency (TOWRE), which assesses phonological decoding ability and sight word reading fluency.

In multilevel models that incorporate data from treatment and control schools, which explore the causal link between SFA and student outcomes (model 7), SFA has a positive effect only on the WJLWI test scores for schools at the same level of implementation. However, fidelity of implementation does not have a significant direct effect on outcomes. This indicates that fidelity of implementation is not a direct predictor of students' reading outcomes, or that a closer adherence to the SFA model does not lead to increases in test scores. This finding challenges the assumption that fidelity to the SFA intervention directly affects the outcome, suggesting that the relationship between the program, its implementation, and students' reading outcomes is more complex.

The results from the impact model (model 7) can be compared to those obtained in the model estimated with data only from treatment schools (model 3), where there was a significant association between implementation and the TOWRE. Both the TOWRE and WJLWI assess sight word reading efficiency or students' ability to recognize and pronounce words (Tarar et al., 2015; Wendling et al., 2007). This suggests that enactment that is closer to the original SFA design may not affect all reading skills equally, with higher levels of fidelity leading to improvements in specific reading skills over others.

Impact of SFA for students classified as ELL and SPED.

The results of model 4, which estimates the relationship between SFA and outcomes in treatment schools, indicate that high levels of adherence to SFA are associated differently with reading outcomes for students classified as ELL and those in SPED. For students with ELL status, high levels of implementation adherence to the program design tend to be negatively associated with test scores. In contrast, the relationship is positive for students with SPED status. In models that estimate the impact of SFA on students' test scores using data from treatment and control schools (model 8), the association holds for ELL students. Higher levels of adherence to the SFA model lead to a decrease in test scores for ELL students, as assessed by the WJLWI, WJWA, and WJPC. Conversely, higher levels of fidelity to SFA practices have a positive effect on TOWRE scores for SPED students. This pattern holds across models that use different implementation data ($I_{SAS\text{-}T}$ and I_{Impact}), which further suggests that fidelity of implementation may indeed have a differential effect on these student groups.

Evidence that the intervention may affect different student groups differently is important information for the program and schools. On the one hand, changes in program design or its components may be warranted to benefit students with an ELL status. This prompts further exploration into how the program can be tailored or supplemented to address the unique challenges faced by ELL students without diminishing the gains experienced by SPED students. For instance, high levels of implementation might advantage SPED students but not ELL students, indicating that the program components align more closely with the needs of SPED students. This analysis may lead to modifying instructional strategies, incorporating additional language support, or adjusting program components to enhance its effectiveness across diverse learner groups.

The differential effect of SFA on ELL and SPED students underscores the need for a nuanced approach to educational interventions. This involves examining the program itself and considering the broader school contexts that affect implementation levels. By understanding these dynamics, adjustments can be made—such as modifying instructional strategies, incorporating additional language support, or refining program components—to enhance the program’s effectiveness across diverse learner groups.

The differential effects observed highlight that high implementation levels do not automatically result in better outcomes for all student groups. Instead, they may reflect underlying differences between schools or indicate that the program’s design currently favors some students over others. Recognizing this allows for a more informed and equitable approach to designing and adapting educational interventions that benefit all learners.

Relationship Across Findings: Discussion and Final Reflection

This research adds to the literature by highlighting the methodological challenges associated with measuring and modeling implementation fidelity. The diversity of measures and constructs used to assess fidelity across IES-funded studies reflects an ongoing debate within the field about the best ways to capture the complexities of program implementation. The lack of consensus on measurement practices limits the comparability of findings across studies, creating challenges for meta-analyses and syntheses that seek to generalize the impact of educational interventions.

The findings from this research underscore the critical importance of balancing fidelity and adaptation in implementing educational programs like SFA. While fidelity is essential to validate the causal link between the intervention and its outcomes, the rigid application of program guidelines can be counterproductive in complex, real-world educational settings. The mixed results of SFA implementation suggest that strict adherence may limit educators’ ability to respond effectively to their students’ needs and context-specific challenges.

This is reflected in the low prevalence of research on adaptations in IES-related publications. The focus is placed chiefly on fidelity (conceptualized as adherence), as this is what the funder requires from researchers. However, the variability in levels of implementation within and between studies shows that the demand for high levels of adherence may be not only unrealistic but also uninformative and restrictive in the context of educational improvement. Furthermore, it can limit the researchers' interest in understanding the adaptations (or deviations from fidelity), as the incentive is placed on bringing forth evidence to prove the internal validity of the study. An approach that acknowledges that teachers will make adaptations and that recognizes the relevance of teacher agency could contribute more to improving teacher practice and education in general.

Moreover, the findings suggest that high adherence may, in fact, be detrimental to some students, as was the case for ELL students in the context of SFA. Teachers indicated their dissatisfaction with the intervention's materials for ELL students (Quint et al., 2015), so there was evidence that the intervention did not match these students' needs. Expecting and demanding adherence to the design in these conditions may not allow schools to provide the best learning opportunities for specific groups of students, and this should be taken into consideration. Such an approach may be more realistic and effective in ensuring the success of educational interventions in diverse settings.

Conversely, the positive effects on SPED students can shed light on the more promising aspects of SFA. Researchers should look into the SFA data in more depth to determine if any particular elements of the intervention may be helping these students with their reading outcomes. An exploration of teaching practices may also provide insights into potential adaptations that may be behind these positive outcomes. As with ELL students, this information can help improve the intervention, using the implementation as a learning opportunity for the designers and researchers while keeping the focus on benefitting students, increasing their exposure to effective teaching practices, and, ultimately, improving their academic outcomes.

Rather than viewing fidelity and adaptation as opposing forces, my research points to the need for a more comprehensive model in which both can coexist. Fidelity and adaptation can be seen as parts of a continuum, and both constructs can provide researchers with important information to understand how the intervention works in real-world settings. Equating *good* implementation to fidelity (more specifically, adherence to the intervention's design) leaves little space for productive adaptations. Investigation into adaptation allows researchers to learn from teacher agency instead of limiting it by expecting high levels of adherence to their design. This balance ultimately depends on the goal of the research. If the objective is to prove the effectiveness of an intervention that was externally designed, thoroughly tested, and refined in a variety of settings, then there may be less space or need to include measures of adaptation. Conversely, a focus on school and teacher improvement that considers the complex contexts in schools and that acknowledges teacher agency and expertise can help researchers understand the mechanisms that lead to positive changes. In this context, an expectation of adherence to program design would only limit this two-way learning process.

Limitations and Future Opportunities

While this research provides valuable insights, several limitations must be acknowledged. The investigation I present exemplifies how we may systematically analyze methodological issues around fidelity and adaptation within the context of quantitative studies that focus on estimating program impacts. The sample of IES studies has a clear focus on quantitative measures of fidelity, which may not capture the full complexity of implementation processes. This sample is biased towards measures of fidelity and adherence, as this is what this type of causal research (and the IES program funding requirements specifically) demands of investigators. The emphasis is on providing evidence that supports the estimates of the impact of the intervention, proving that the observed differences in scores are attributable to the program under evaluation. This methodological approach limits the opportunities for researchers to explore deviations from the original design, so it follows that constructs like Adaptation would be observed with less frequency in the research.

Secondly, all the IES-related documents were coded by one rater. Having at least two raters can help ensure the consistent application of the coding criteria, reduce individual biases, and introduce an external check. I was unable to include a second rater as part of my dissertation work. I explored the possibility of using artificial intelligence to code the studies, but none of the available tools gathered the information I needed with sufficient depth. Furthermore, I could not find a tool that was flexible enough to adapt to the coding scheme. Although I was unable to do this now, the rapid advances in IA may soon allow for an additional virtual rater.

As to the second research question, methodological limitations prevented me from calculating empirically derived indices for the SFA data. I tested several models with data from the School Achievement Snapshot and the teacher and principal surveys, but the model fit was very poor. I acknowledge that it would have been helpful to compare the weighted and unweighted implementation scores with an empirically derived index to check whether scores and rankings were more sensitive to this aggregation method. Estimating these latent factors would shed light on the empirical structure of the data and help understand the relationships among items and constructs. This would also provide an empirical model that can be used to contrast the theoretical and unweighted implementation indices. In addition, employing these empirically derived indices may have led to different estimates in the correlation and hierarchical linear models.

The effect of fidelity in the implementation of the SFA model was relevant for two specific groups of students (ELL and SPED). It would have been interesting to explore this further, using qualitative data to understand teachers' and students' experiences with the intervention. This would have allowed me to gain a deeper understanding of the instructional practices that took place in this implementation of SFA to explore if any specific intervention components or teaching practices may explain the differing effects for students. At the same time, qualitative data could contain more information on potential adaptations to the SFA intervention, shedding light on specific changes that teachers made to adapt the intervention to their students' needs. Consequently, the explanations that I offer for the effects of implementation on student outcomes are speculative, as this is what the publicly available data for SFA allows for.

Finally, the findings suggest opportunities for future research to explore the role of adaptive program models that balance fidelity with flexibility. An excessive focus on fidelity can lead to validating internal program theory but may not be the most productive framework to assess implementation in schools. A more flexible approach that collects data on teaching practices and analyzes them from an exploratory perspective, aimed at learning about teachers' changes to the intervention, can help researchers understand the mechanisms that are generating the changes that they estimate in the impact models. In the context of quantitative research that aims to find effective educational interventions, research should investigate the benefits of moving away from an implementation framework that views fidelity (adherence) as the goal and adaptations as deviations from an ideal. A model that views implementation as a continuum and that is open to exploring adaptations may be more helpful for school improvement, as it incorporates dynamic school contexts and teacher agency.

In conclusion, this research contributes to the fields of evaluation and implementation science by emphasizing the importance of both fidelity and adaptation in the successful application of educational interventions. By addressing the methodological challenges associated with measuring fidelity and highlighting the need for adaptive program models, this research offers a path forward for developing and implementing educational programs that are both effective and contextually relevant. Through continued exploration of these concepts, educational researchers and practitioners can work toward interventions that achieve meaningful and lasting improvements in student outcomes across diverse educational settings.

Appendix

Appendix 1

Table A 1: Summary of Characteristics and Findings for Studies Measuring Implementation

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
1.	The Efficacy of the Responsive Classroom Approach for Improving Teacher Quality and Children's Academic Performance	2007	RCT	Fidelity (adherence, monitoring control conditions)	Observation rubric, teacher surveys	Latent variables	Yes	Abry et al. (2013), Rimm-Kaufman et al. (2014)
2.	Effectiveness of Cognitive Tutor Algebra One Implemented at Scale	2007	RCT	Fidelity (adherence, dosage/frequency, monitoring control conditions)	Teacher surveys	Simple sum or average	No	Daugherty et al. (2012), Karam et al. (2017), Pane et al. (2014a, 2014b)12/7/2024 12:39:00 AM
3.	Increasing the Efficacy of An Early Mathematics Curriculum with Scaffolding Designed to Promote Self-Regulation	2008	RCT	Fidelity (adherence, monitoring control conditions)	Observation rubric	Simple sum or average	No	(Sarama et al. (2016), Germeroth et al. (2019), Clements et al. (2020)
4.	Efficacy of Read It Again! In Rural Preschool Settings	2008	RCT	Fidelity (adherence, exposure, quality, participant responsiveness)	Observation rubric, teacher logs	Simple sum or average	Yes	Mashburn et al. (2016), S. Piasta et al. (2015)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
5.	Increasing Opportunities-to-Learn in Urban Middle Schools	2008	RCT	- Fidelity (adherence, quality, monitoring control conditions). - Adaptation	Observation rubric, teacher logs	Simple sum or average	No	Burkhauser & Lesaux (2017), Lesaux et al. (2014)
6.	An Efficacy Trial of Robust Vocabulary Instruction	2008	RCT	Fidelity (dosage, quality, adherence monitoring control conditions)	Observation rubric, teacher surveys, teacher logs	Simple sum or average	Yes	Apthorp et al. (2012)
7.	Project Collaborative Strategic Reading (CSR): Interventions for Struggling Adolescent and Adult Readers and Writers	2008	RCT	Fidelity (adherence, quality, monitoring of control conditions)	Observation rubric, teacher logs	Simple sum or average, latent variables	Yes	Vaughn et al. (2011), Vaughn et al. (2013)
8.	Education Research - BioBridge Teacher Quality	2008	Quasi-experimental	Fidelity (adherence)	Observation rubric	Simple sum or average	No	Babendure et al. (2011), Peterman et al. (2014)
9.	Early Learning in Mathematics: Efficacy in Kindergarten Classrooms	2008	RCT	Fidelity (adherence, monitoring control conditions)	Observation rubric	Simple sum or average	No	Doabler et al. (2014), Doabler et al. (2016)
10.	Systems and Cycles: Using Structure-Behavior-Function Thinking as a Conceptual Tool for Understanding Complex Natural Systems in Middle School Science	2009	Quasi-experimental	Not reported	–	–	–	Hmelo-Silver et al. (2017)
11.	Assessing the Efficacy of a Comprehensive Intervention in Physical Science on Head Start Teachers and Children	2009	RCT	- Fidelity (quality, dosage)	Observation rubric	Not reported	Yes	(Gropen et al. (2011), (Gropen et al.(2017)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
12.	National Randomized Controlled Trial Study of SRA/McGraw-Hill Open-Court Reading Program	2009	RCT	Fidelity (adherence, exposure, quality, participant responsiveness, monitoring control conditions)	Observation rubric, teacher surveys, teacher interviews	Latent variables	No	Sullivan et al. (2016), Vaden-Kiernan et al. (2018)
13.	A Multi-Part Intervention for Accelerating Vocabulary Acquisition through Inductive Transfer	2009	RCT	Fidelity (no constructs reported)	Observation rubric	Not reported	No	Vitale & Romance (2013b, 2013a)
14.	Word Generation: An Efficacy Trial	2009	RCT	Not reported	–	–	No	Lawrence et al. (2015, 2017), Lin et al. (2016)
15.	Experimental Validation of the Tools of the Mind Prekindergarten Curriculum	2009	RCT	Fidelity (adherence, dosage)	Observation rubric, teacher surveys	Simple sum or average (weighted), latent variables	Yes	Farran et al. (2015), Meador et al. (2015)
16.	Disciplinary Writing Instruction for the Social Studies Classroom: A Path to Adolescent Literacy	2009	Quasi-experimental	- Fidelity (adherence, quality, participant responsiveness) - Adaptation	Observation rubric, student artifacts, teacher interviews	Simple sum or average, latent variables	Yes	De La Paz et al. (2014), De La Paz et al. (2017), Monte-Sano et al. (2014)
17.	Preparing to Succeed: An Efficacy Trial of Two Early Childhood Curricula	2009	Quasi-experimental	Fidelity (adherence, dosage, quality)	Observation rubric	Latent variables	Yes	Weiland et al. (2011), Weiland & Yoshikawa (2013)
18.	Efficacy of the Science Writing Heuristic Approach	2009	RCT	- Fidelity (dosage, quality)	Observation rubric	Simple sum or average	No	Hand et al. (2013), Hand, Park, et al.

	Title	Year Awarded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
								(2018), Hand, Shelley, et al. (2018), Bae et al. (2022)
19.	Promoting Science among English Language Learners (P-SELL): Efficacy and Sustainability	2009	RCT	Not reported	–	–	No	Diamond et al. (2014), Maerten-Rivera et al. (2016)
20.	Fostering Reading Engagement in English-Monolingual Students and English Language Learners Through a History Curriculum	2010	Quasi-experimental	Fidelity (quality)	Observation rubric	Simple sum or average	No	Barber et al. (2015)
21.	Toward High School Biology: Helping Middle School Students Make Sense of Chemical Reactions	2010	RCT	Not reported	–	–	–	Herrmann-Abell et al. (2016), Roseman & Herrmann-Abell (2016)
22.	Improving the Teaching and Learning of English Language Learners: The Instructional Conversational Model	2010	Quasi-experimental	Fidelity (adherence, monitoring control conditions)	Observation rubric, teacher logs	Not reported	No	Gonzalez Canche et al., (2014), Hendy & Cuevas, (2020), Mellom et al. (2018), Portes et al., (2018)
23.	The Connected Chemistry Curriculum	2010	Quasi-experimental	Not reported	–	–	–	Stieff (2011, 2019)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
24.	National Randomized Control Trial of Everyday Mathematics	2010	RCT	Fidelity (dosage, adherence, quality, and participant responsiveness).	Observation rubric, teacher surveys, teacher interviews	Latent variables	Yes	Vaden-Kiernan et al. (2015)
25.	Using an Empirically-supported Teacher Consultation Model to Facilitate the Implementation of an Integrated Social-emotional Learning and Literacy Curriculum in Urban Elementary Schools	2010	Quasi-experimental	Fidelity (no constructs reported)	Not reported	Not reported	No	Doyle et al. (2023)
26.	ECHOS: Early Childhood Hands on Science	2010	RCT	Fidelity (adherence and quality).	Observation rubric	Not reported	No	(Herrmann-Abell et al., 2019)
27.	Accessible Professional Development for Teaching Aquatic Science Inquiry	2010	Quasi-experimental	Fidelity (adherence)	Teacher logs	Simple sum or average	Yes	(Duncan Seraphin et al., 2017)
28.	Efficacy of Rich Vocabulary (RVOC) Instruction for Classrooms	2010	RCT	Fidelity (adherence, dosage)	Observation rubric, teacher logs	Not reported	Yes	(Vadasy et al., 2015)
29.	An Efficacy Study of Project GLAD	2010	RCT	Fidelity (adherence, monitoring control conditions)	Observation rubric, teacher surveys	Simple sum or average	No	(Autio et al., 2014), (Deussen et al., 2014), (Deussen et al., 2015)
30.	Longitudinal Study of a Successful Scaling-Up	2011	RCT	Fidelity (adherence)	Observation rubric	Simple sum or average	No	(Clements et al., 2011),

	Title	Year Awarded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
	Project: Extending TRIAD							(Sarama et al., 2012)
31.	Learning of Ratio and Proportion Problem-Solving Using Schema-Based Instruction: Efficacy and Sustainability	2011	RCT	Fidelity (adherence, monitoring control conditions)	Observation rubric	Simple sum or average	No	(Jitendra et al., 2015)
32.	Numbers Plus Efficacy Study	2011	RCT	Fidelity (dosage)	Teacher logs	Simple sum or average	No	(Wakabayashi et al., 2020)
33.	WORLD Efficacy Study	2011	RCT	Fidelity (quality, adherence, participant responsiveness)	Observation rubric	Simple sum or average	Yes	(Gonzalez et al., 2024), (Pollard-Durodola et al., 2018)
34.	Scale-up Evaluation of Reading Intervention for First Grade English Learners	2011	RCT	Not reported	–	–	No	(Barr et al., 2019)
35.	Developing Consultation and Collaboration Skills: ESL and Classroom Teachers Working Together with Students and Families	2012	RCT	- Fidelity (adherence, monitoring control conditions) - Adaptation	Observation rubric	Simple sum or average	No	(Babinski et al., 2018)
36.	Investigation of the Efficacy of the JUMP Program of Mathematics Instruction	2012	RCT	Fidelity (adherence, monitoring control conditions)	Observation rubric	Simple sum or average	Yes	(Solomon et al., 2019)
37.	Efficacy Study of a Pre-Algebra Supplemental Program in Rural Mississippi Schools	2012	RCT	Fidelity (no constructs reported)	Teacher logs	Not reported	No	(Clark et al., 2015)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
38.	A Randomized Study of the Efficacy of a Two-Year Mathematics Intervention for At-Risk Pre-Kindergarten and Kindergarten Students	2012	RCT	Fidelity (adherence, dosage)	Observation rubric	Simple sum or average	No	(Klein et al., 2008), (Starkey et al., 2022)
39.	Getting Ready for School: An Integrated Curriculum to Help Teachers and Parents Support Preschool Children's Early Literacy, Math, and Self-Regulation Skills	2012	Quasi-experimental	Fidelity (adherence, dosage, participant responsiveness)	Teacher surveys, survey (others)	Simple sum or average	Yes	(Marti, Melvin, et al., 2018), (Marti, Merz, et al., 2018)
40.	Efficacy Trial of MyTeachingPartner-Mathematics and Science Curricula and Implementation Support System	2012	RCT and Quasi-experimental	Fidelity (adherence, quality)	Observation rubric	Simple sum or average	No	(Barton et al., 2017), (Whittaker et al., 2016), (Whittaker et al., 2020)
41.	GlobalEd 2	2013	RCT	- Fidelity (adherence) - Adaptation	Teacher logs (qualitative)		No	(Lawless et al., 2018), (Riel et al., 2016)
42.	First Grade, Second Language: Uniting Science Knowledge and Literacy Development for English Learners	2013	Quasi-experimental	Fidelity (no constructs reported)	Not reported	Not reported	No	(Billman et al., 2018)
43.	Quality Talk: Developing Students' Discourse to Promote Critical-Analytic Thinking, Epistemic Cognition, and High-Level Comprehension	2013	Quasi-experimental	- Fidelity (adherence) - Adaptation	Observation rubric	Not reported	No	(Firetto et al., 2019), (Murphy et al., 2022)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
44.	An Elementary-age Origami and Pop-up Paper Engineering Curriculum to Promote the 3-D Spatial Thinking and Reasoning Underlying STEM Education	2014	Quasi-experimental	Not reported	–	–	–	(Burte et al., 2017, 2020)
45.	Testing the Integration of an Empirically-supported Teacher Consultation Model and a Social-emotional Learning and Literacy Intervention in Urban Elementary Schools	2014	RCT	Fidelity (dosage, adherence, participant responsiveness)	Observation rubric, teacher surveys	Latent variables	Yes	(Corbin et al., 2020), (Gómez et al., 2023)
46.	Story Talk: A Cognitive Research-based Vocabulary Intervention for Preschoolers	2014	RCT	Fidelity (adherence, monitoring control conditions)	Observation rubric	Simple sum or average	Yes	(Wasik & Hindman, 2020, 2023)
47.	Development of a Dual Language Narrative Curriculum	2014	RCT	Fidelity (adherence, participant responsiveness, quality, dosage)	Observation rubric, teacher surveys, teacher logs	Simple sum or average	No	(Spencer et al., 2020)
48.	The CLAVES Intervention Project: Developing a Supplemental Intervention for Comprehension, Linguistic Awareness, and Vocabulary in English for Spanish Speakers	2014	Quasi-experimental	Fidelity (no constructs reported)	Observation rubric	Simple sum or average	No	(Proctor et al., 2020), (Proctor et al., 2021)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
49.	The Impact of a Teacher-Led Early Algebra Intervention on Children's Algebra-Readiness for Middle School	2014	RCT	Fidelity (adherence, quality)	Observation rubric, teacher surveys	Latent variables	Yes	(Blanton et al., 2019), (Stylianou et al., 2019)
50.	Mathematics and English Language Development for English Language Learners: Project MELD for ELLs	2014	RCT	Fidelity (adherence)	Observation rubric	Simple sum or average	No	(August et al., 2023)
51.	Web-mediated Literacy Coaching for High-quality Reading Comprehension Instruction	2014	RCT	Not reported	–	Simple sum or average	Yes	(Correnti et al., 2021), (Matsumura et al., 2008), (Matsumura et al., 2019)
52.	Examining the Efficacy of Differential Levels of Professional Development for Teaching Content Area Reading Strategies	2015	RCT	Fidelity (adherence, quality, monitoring control conditions)	Not reported	Simple sum or average, latent variables	Yes	(Swanson et al., 2021), (Swanson et al., 2024), (Vaughn et al., 2022)
53.	Improving Children's Understanding of Mathematical Equivalence: An Efficacy Study	2015	RCT	Fidelity (dosage, monitoring control conditions)	Teacher surveys, teacher logs, student artifacts	Simple sum or average	No	(Davenport et al., 2022)
54.	Teaching Together: A Multimedia School-Home Intervention for Young Children At-Risk for Academic Difficulties	2015	RCT	Fidelity (dosage, adherence)	Observation rubric, teacher logs	Simple sum or average	No	(Zucker, Cabell, et al., 2021)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
55.	An Efficacy Trial of the HighScope Preschool Curriculum (HSPC)	2015	RCT	Fidelity (adherence, dosage, quality)	Not reported	Not reported	No	(American Institutes for Research, n.d.), (Howard et al., 2020, 2021)
56.	For Argument's Sake: Applying Questioning the Author Techniques to Move from Comprehension to composition of Written Arguments	2015	RCT	Fidelity (adherence)	Observation rubric, student and teacher artifacts	Simple sum or average	No	(Crosson et al., 2024)
57.	Improvement of Elementary Fractions Instruction: Randomized Controlled Trial Using Lesson Study with a Fractions Resource Kit	2015	RCT	Fidelity (adherence)	Observation rubric	Simple sum or average	No	(Schoen et al., 2024)
58.	Language for Reading: Building Vocabulary Through Engaged Learning	2015	Quasi-experimental	Fidelity (adherence)	Observation rubric	Simple sum or average	Yes	(Hadley et al., 2022), (Hassinger-Das et al., 2016)
59.	Building Students' Understanding of Energy in High School Biology	2015	RCT	Not reported	–	–	–	(Herrmann-Abell et al., 2019)
60.	Word Learning Strategies: A Program for Upper-Elementary Readers	2015	RCT	Fidelity (dosage, monitoring control conditions)	Observation rubric, teacher logs, teacher interviews	Not reported	No	(Li et al., 2019)
61.	An Investigation of Direct Instruction Spoken English for At-Risk English Learners	2015	RCT	Fidelity (adherence, quality)	Observation rubric	Not reported	Yes	(Chaparro et al., 2022), (Gunn et al., 2021)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
62.	Improving Teacher Capacity to Implement High Quality Service Learning in Elementary Science Classrooms	2015	RCT	Fidelity (adherence, dosage, monitoring control conditions)	Teacher surveys, teacher logs	Simple sum or average	No	(Rimm-Kaufman et al., 2021)
63.	Seeds of STEM: The Development of an Innovative Pre-Kindergarten STEM Curriculum	2015	RCT	Fidelity (adherence)	Observation rubric	Not reported	No	(Sibuma et al., 2018)
64.	Returning to Our Roots: Development of a Morphology Intervention to Bolster Academic Vocabulary Knowledge for Adolescent English Learners	2015	RCT and Quasi-experimental	Fidelity (adherence)	Observation rubric	Simple sum or average	No	(Crosson & Moore, 2017), (Crosson et al., 2019), (Crosson et al., 2021)
65.	Efficacy of the BrightStart! Program for Promoting Emergent Literacy Skills of PreKindergarten Children at Risk for Reading Difficulties	2016	RCT	- Fidelity (adherence, dosage, quality, participant responsiveness) - Adaptation	Observation rubric, questionnaire (qualitative)	Simple sum or average	No	(S. B. Piasta et al., 2021), (S. B. Piasta et al., 2023)
66.	Efficacy Evaluation of Zoology One: Kindergarten Research Labs	2016	RCT	- Fidelity (adherence, monitoring of control condition) - Adaptation	Teacher surveys, teacher logs, teacher interviews	Simple sum or average	Yes	(A. Gray et al., 2020), (A. Gray et al., 2022), (A. M. Gray et al., 2022)
67.	Efficacy of the Core Knowledge Language Arts Read Aloud Program	2016	RCT	Fidelity (adherence)	Not reported	Not reported	No	(Cabell, 2020)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
	in Kindergarten through Second Grade Classrooms							
68.	An Efficacy Study of Interleaved Mathematics Practice	2016	RCT	Fidelity (dosage)	Student artifacts	Not reported	No	(Rohrer et al., 2020)
69.	The Scale Up of Promoting Adolescents Comprehension of Text	2016	RCT	Fidelity (adherence, monitoring control condition, quality)	Observation rubric	Simple sum or average	No	(Roberts et al., 2023), (Scammacca et al., 2020)
70.	Middle School Matters: Promoting Research- and Evidence-Based Practices to Support Reading Comprehension (MSMPREP)	2017	RCT	- Fidelity (dosage)	Teacher surveys	Simple sum or average	No	(Stevens et al., 2022)
71.	Refinement of GlobalEd2 and Testing New Intervention Impact	2017	Quasi-experimental	Not reported	–	–	–	(Riel & Lawless, 2021), (Riel et al., 2022)
72.	SRSD+: Development of a Powerful Writing Program for Children in Grades 1 and 2	2017	RCT	Fidelity (adherence, quality)	Observation rubric	Simple sum or average	No	
73.	Efficacy of the TELL Curriculum for Preschool Children who are Economically Disadvantaged	2017	RCT	Fidelity (adherence, quality).	Observation rubric	Simple sum or average	Yes	(S. I. Gray et al., 2024)
74.	Efficacy Study of an Integrated Science and Literacy Curriculum for Young Learners	2018	RCT	Not reported	–	–	–	(C. J. Harris et al., 2023), (Rutstein et al., 2021)

	Title	Year Awar-ded	Study Design	Impl. Framework and Constructs	Instrument(s) to Measure Impl.	Method(s) to Aggregate Impl. Data	Associates Impl. Data with Outcomes	References
75.	Efficacy Replication Study of the Impact of MyTeachingPartner-Secondary (MTP-S)	2018	RCT	Not reported	–	–	–	(Wayne et al., 2023)
76.	A Regression Discontinuity Study of the Impact of ALFA Lab on 9th-Graders' Reading Achievement, Motivation, and Reading Frequency	2018	Quasi-experimental	Fidelity (not reported)	Not reported	Not reported	No	(Davis et al., 2024)
77.	Efficacy Trial of the We-Write Intervention with 4th- and 5th-Grade Students	2018	RCT	Fidelity (adherence)	Observation rubric	Simple sum or average	No	(McKeown et al., 2023)
78.	Developing Talkers: Building Effective Teachers of Academic Language Skills	2019	RCT	Fidelity (adherence, dosage, participant responsiveness)	Observation rubric	Simple sum or average	Yes	(Zucker et al., 2019), (Zucker, Jacbos, et al., 2021)
79.	Project Citizen Research Program	2019	RCT	Not reported	–	–	No	(Owen, 2024)

Appendix 2

Table A 2: Constructs and Subconstructs and Items in the School Achievement Snapshot

Construct	Subconstruct	Items	Classification	Programs
1. Challenging Individualized Instruction	Cooperative Learning	- Teachers use think-pair-share	IP	KC, RR, RW
		- Teachers provide time for team talk	IP	KC, RR, RW
		- Teachers facilitate team discussion	IP	KC, RR, RW
		- Following team talk, teachers discuss	IP	RW
		- Teachers use team scores to set goals	IP	RR, RW
	Cognitively Demanding Instruction	- Teachers restate and elaborate	IP	KC, RR, RW
		- Teachers summarize	IP	RW
		- Teachers ask students to share	IP	RW
	Pacing	- Active instruction appropriately paced	IP	KC, RR, RW
	Media Use	- Teachers use the basic lesson structure	IP	KC, RR, RW
	Grouping	- Cross grade regrouping	SS	
		- Multiple measures to determine placement	SS	
		- Placement is aggressive	SS	
	Tutoring	- Capacity exists to tutor	SS	
		- A certified teacher-tutor	SS	
		- Tutoring provided daily	SS	
		- Team Alphie	SS	
	Celebration	- Teachers calculate team scores	IP	RR, RW
		- Read and respond forms collected	IP	KC, RR, RW
2. Non-Instructional Issues that Affect Reading Instruction	Solutions Teams	- Solutions Coordinator	SS	
		- School wide solutions teams	SS	
		- Solutions coordinator supports solutions teams	SS	
		- School wide solutions teams set targets	SS	
	Parent/ Community Involvement	- Parent involvement	SS	
		- Volunteer listeners	SS	
		- Community supported vision program	SS	
	Attendance	- Attendance plans are complete	SS	
	Behavior	- Positive behavioral intervention support (PBIS) teams	SS	

Construct	Subconstruct	Items	Classification	Programs
		- Getting Along Together (GAT) structures classroom	SS	
		- GAT structures schoolwide	SS	
		- Intervention team	SS	
		- Teachers facilitate use of emotion control	SS	
3. Continuous Improvement of Students and Staff	Use of Data	- Accurate grade summary form	SS	
		- Formal reading level assessments	SS	
		- Classroom assessment summary	SS	
		- Member center data collection	SS	
		- Leadership team	SS	
		- Members of the leadership team know students meeting goals	SS	
		- Leading for Success quarterly meetings	SS	
		- Instructional component teams set targets	IP	KC, RR, RW
		- Leading for Success teams set targets	SS	
		- Intervention team	SS	
	Professional Development	- Instructional component teams	SS	
		- Facilitator uses the coaching process	SS	

Source: Quint (2016f).

Note. IP stands for Instructional Processes, SS for Schoolwide Structures, KC for Kinder Corner, RR for Reading Roots, and RW for Reading Wings.

Appendix 3

Table A 3: Teacher Survey Items

Section	Subsection	Item	Response Options
Background and Current Responsibilities		Q1. What grade level do you currently teach?	Kindergarten; 1st Grade; 2nd Grade 3rd Grade; 4th Grade; 5th Grade; Another role
		a. What is your role during the reading block?	Main teacher; Classroom aide; Reading interventionist; Reading specialist
		Q2. Do you currently have primary responsibility for teaching reading to students in any of grades 1-5?	Yes; No
		a. For how many years (including the current year) have you taught reading in any of grades 1-5?	Enter number (school years)
		b. **(SFA Only)** Which of the following reading levels do you currently teach?	Kinder Corner; Reading Roots; Reading Wings
		Q3. How many years during your career (including the 2013-14 school year) have you spent in the following roles?	
		a. Years spent as a classroom teacher in an elementary school.	Enter number (school years)
		b. Years spent as a reading specialist in an elementary school.	Enter number (school years)
		c. Years spent as a classroom aide in an elementary school.	Enter number (school years)
		d. Years spent in any role at schools (including your current school) involved in a whole-school reform effort before 2013- 14.	Enter number (school years)
		e. Years spent at your current school (including 2012-13).	Enter number (school years)
The Reading Program at Your School	The Current Program	Q4. On a typical day, what is the length of the reading block (excluding writing and grammar instruction) at your school?	Enter number (minutes per day)
		Q5. Since the beginning of the current school year, have students at your school been re-grouped for reading by ability level?	Yes (continue to Q5a); No (skip to Q6)
		a. How often are students re-grouped for reading by ability level?	Once every four months; Once every three months; Once every two months; Once a month; Twice a month or more; Never
		Q6. How many students are in your reading class currently?	Enter number (students)
		Q7. In the reading class you teach, are students divided into smaller groups by ability level?	Yes (continue to Q7a); No (skip to Q8)
		a. If students are divided into smaller groups by ability level for reading, what is the average number of students in these small groups? Please provide your best estimate.	Enter number (students)

Section	Subsection	Item	Response Options
		Q8. To what extent do you agree with the following statements?	
		a. Your reading class is small enough for individual students to receive adequate attention	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Reading groups are small enough in your reading class for individual students to receive adequate attention.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Re-grouping students for reading lessons is an effective strategy for improving students' reading skills.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q9. What percentage of students in your reading class are reading at grade level?	Enter number (percent of class reading on grade level)
		Q10. To what extent do you agree with the following statements about the reading program at your school during the 2013-14 school year?	
		a. You are satisfied with the overall quality of the reading program at your school.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. You have been given the support you need, in terms of additional resources, to implement your school's reading program.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. You are satisfied with the overall quality of the reading materials (including technology) that you use.	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. You find the reading program at your school too time- consuming or work-intensive.	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. You change the parts of your school's reading program that do not work for your students.	Strongly Disagree; Disagree; Agree; Strongly Agree
		f. The reading program at your school adequately serves most of your students.	Strongly Disagree; Disagree; Agree; Strongly Agree
		g. The reading program at your school adequately serves students who struggle the most with reading.	Strongly Disagree; Disagree; Agree; Strongly Agree
		h. The reading program at your school is too rigid or scripted.	Strongly Disagree; Disagree; Agree; Strongly Agree
		i. The reading program at your school promotes teacher collaboration.	Strongly Disagree; Disagree; Agree; Strongly Agree
		j. The reading program at your school involves students working together in pairs or small groups almost daily.	Strongly Disagree; Disagree; Agree; Strongly Agree
		k. Your school's reading program gets students excited about reading or learning how to read.	Strongly Disagree; Disagree; Agree; Strongly Agree
		l. Your school's reading program helps you to teach for mastery of concepts.	Strongly Disagree; Disagree; Agree; Strongly Agree
		m. You feel the pacing of your school's reading program allows most students in your class to learn critical concepts.	Strongly Disagree; Disagree; Agree; Strongly Agree

Section	Subsection	Item	Response Options
		n. You feel the pacing of your school's reading program allows you to get through almost all the material you need to cover in each class session	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q11. To what extent do you agree that the following aspects of the reading program help students in your class become better readers?	
		a. Cooperative Learning	Strongly Disagree; Disagree; Agree; Strongly Agree; Our school does not have this component
		b. Grouping	Strongly Disagree; Disagree; Agree; Strongly Agree; Our school does not have this component
		c. Tutoring	Strongly Disagree; Disagree; Agree; Strongly Agree; Our school does not have this component
		d. Length of the Reading Block	Strongly Disagree; Disagree; Agree; Strongly Agree; Our school does not have this component
		e. If another significant aspect to school reading program how much do you agree it helps students become better readers	Enter text – Strongly Disagree; Disagree; Agree; Strongly Agree; Our school does not have this component
		Q12. To what extent do you agree with the following statement about your reading class during the current school year?	
		a. You use educational media or technology as part of the reading program at your school	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q13. During your most recent reading block, for about how many minutes did you use educational technology as part of your active instruction?	Enter number (minutes)
		Q14. The following questions ask you about your own reading class during the 2013-14 school year. To what extent do you agree with the following statements?	
		a. Frequent student absence affects students' learning in your reading class	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. The number of students in your reading class interferes with your teaching	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Your students are well-behaved during your reading class	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. Your students are engaged during your reading class	Strongly Disagree; Disagree; Agree; Strongly Agree

Section	Subsection	Item	Response Options
		e. Your reading program provides you tools to help with classroom management	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q15. During the current 2012-13 school year, how frequently do students respond to questions in complete sentences?	Not at all; Some class sessions; Most class sessions; All class sessions
		Q16. To what extent do you agree with the following statements about your principal’s role in the reading program at your school in the 2013-14 school year?	
		a. Served as a knowledgeable source concerning reading standards and curriculum	Strongly Disagree; Disagree; Agree; Strongly Agree; Not Applicable
		b. Ensured that teachers have time for planning reading instruction	Strongly Disagree; Disagree; Agree; Strongly Agree; Not Applicable
		c. Provided teachers with adequate classroom materials to improve student reading proficiency	Strongly Disagree; Disagree; Agree; Strongly Agree; Not Applicable
		d. Ensured that teachers receive adequate professional development in reading.	Strongly Disagree; Disagree; Agree; Strongly Agree; Not Applicable
		e. Ensured that teachers receive regular feedback regarding their reading instruction.	Strongly Disagree; Disagree; Agree; Strongly Agree; Not Applicable
		f. Ensured outreach to parents to support reading practices at home.	Strongly Disagree; Disagree; Agree; Strongly Agree; Not Applicable
		Q17. Since the start of the 2013-14 school year, about how frequently has your principal…	
		a. Conducted walk- throughs of your class or briefly observed your reading instruction?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		b. Conducted unscheduled classroom observations to get a sense of your reading instruction?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		c. Provided informal feedback to you about how you could improve your reading instruction?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		d. Formally evaluated your reading instruction?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		e. Taught a demonstration class or modeled a reading class?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable

Section	Subsection	Item	Response Options
		f. Participated in a grade- level meeting with all reading teachers?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		g. Met in small groups with you and other teachers to discuss reading?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		Q18. Did you participate in any professional development concerning reading during the summer of 2013?	Yes (enter number of days); No
		Q19. Did you participate in any professional development concerning reading during the summer of 2013?	Yes (enter number of days); No
		Q20. Did you participate in any professional development concerning reading during the 2013-14 school year?	Yes; No
		Q21. Did you participate in any workshops concerning reading instruction during the 2013-14 school year?	Yes; No
		Q22. Since the start of the 2013-14 school year, your professional development in reading instruction has…	
		a. Helped you learn how to implement your school's reading program properly.	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area
		b. Helped you learn new techniques for reading instruction	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area
		c. Your PD in reading instruction has… Helped you learn how to teach students at different reading levels or in different reading groups	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area
		d. Helped you develop strategies to better meet the needs of the reading students who struggle the most	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area
		e. Helped you learn how to use classroom materials, including technology, to improve reading instruction	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area

Section	Subsection	Item	Response Options
		f. Helped you learn how to better use the time allocated to reading instruction.	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area
		g. Helped you learn how to implement cooperative learning techniques among students.	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area
		h. Helped you learn how to use reading assessment data to guide instruction.	Strongly Disagree; Disagree; Agree; Strongly Agree; Professional Development Was Not Received in This Area
The Reading Program at Your School	Use of Data	Q23. Did you receive any coaching in reading instruction during the 2013-14 school year?	Yes; No
		Q24. Since the start of the 2013-14 school year, how often have you…	
		a. Accessed data from the school's reading data system?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never
		b. Entered data in the school's reading data system?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never
		c. Used the school's reading data system to plan instruction?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never
		Q25. To what extent do you agree with the following statements?	
		a. Your reading program helps to prepare students to do well on state achievement tests	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. You are satisfied with the way that the reading assessments used during the 2013-14 school year measure your students' reading skills.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q26. To what extent do you agree with the following statements?	
		a. Evaluate the reading progress of students over time.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Communicate with and inform parents about student reading performance.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Identify students struggling with reading.	Strongly Disagree; Disagree; Agree; Strongly Agree

Section	Subsection	Item	Response Options
		d. Develop strategies to move students from the below basic and basic categories into proficient category on standardized tests of reading skills	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. Examine school-wide instructional issues related to reading.	Strongly Disagree; Disagree; Agree; Strongly Agree
		f. Identify reading teachers who need instructional improvement	Strongly Disagree; Disagree; Agree; Strongly Agree
General School Functioning	Climate	Q27. Considering your experiences during the 2013-14 school year, to what extent do you agree with the following statements?	
		a. You help administrators in school decision-making processes.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. You approach the principal when you have a problem with a colleague	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. You approach the principal when you have a concern or question related to general school functioning.	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. You approach other teachers when you have a concern or question related to instruction.	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. You believe that other teachers are teaching similar skills as you are to students at a given reading level	Strongly Disagree; Disagree; Agree; Strongly Agree
		f. You discuss student behavioral challenges with other teachers.	Strongly Disagree; Disagree; Agree; Strongly Agree
		g. You discuss how to improve instruction with other teachers.	Strongly Disagree; Disagree; Agree; Strongly Agree
		h. You discuss lesson plans that were not particularly successful with other teachers.	Strongly Disagree; Disagree; Agree; Strongly Agree
		i. You discuss lesson plans that were successful with other teachers.	Strongly Disagree; Disagree; Agree; Strongly Agree
		j. You share your students’ work with other teachers.	Strongly Disagree; Disagree; Agree; Strongly Agree
		k. You believe that your students come to school ready to learn.	Strongly Disagree; Disagree; Agree; Strongly Agree
		l. You build relationships with students’ parents/guardians.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q28. Considering your experiences during the 2013-14 school year, to what extent do you agree with the following statements?	
		a. You can help most students attain grade-level reading skills by the end of the year, regardless of their family or economic circumstances	Strongly Disagree; Disagree; Agree; Strongly Agree

Section	Subsection	Item	Response Options
		b. You can help most students improve their reading but not necessarily attain grade-level reading skills by the end of the year.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q29. To what extent do you agree your reading program can adequately serve students with the following characteristics:	
		a. English language learners	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Students with a reading IEP	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Students with behavioral challenges	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q30. Since the start of the current school year, teacher morale at your school has been high.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q31. To what extent do you agree with the following statements about your principal's leadership during the 2013-2014 school year?	
		a. Encourages team work among staff members	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Ensures that students are given academically demanding and challenging work.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Ensures that instruction follows the adopted curriculum	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. Makes expectations for meeting student learning goals clear to teachers.	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. Provides support for classroom discipline and order.	Strongly Disagree; Disagree; Agree; Strongly Agree
		f. Engages parents in school activities and student learning.	Strongly Disagree; Disagree; Agree; Strongly Agree
		g. Monitors your school's progress toward district and state standards.	Strongly Disagree; Disagree; Agree; Strongly Agree
		h. Manages and responds to issues related to the community outside the school.	Strongly Disagree; Disagree; Agree; Strongly Agree
		i. has a friendly and positive attitude.	Strongly Disagree; Disagree; Agree; Strongly Agree
		j. Is visible throughout your school.	Strongly Disagree; Disagree; Agree; Strongly Agree
		k. Has focused on teachers' sense of belonging and well- being	Strongly Disagree; Disagree; Agree; Strongly Agree

Section	Subsection	Item	Response Options
General School Functioning	The Success for All Program	Q32.	
		a. **(SFA Only)** How knowledgeable about SFA do you think your SFA facilitator is this year?	Not at all; Somewhat; Extremely
		b. **(SFA Only)** How knowledgeable about SFA do you think your principal is?	Not at all; Somewhat; Extremely
		c. **(SFA Only)** How adequate did you find the training on SFA prior to the start of the current school year?	Not at all; Somewhat; Extremely
		d. **(SFA Only)** How adequate did you find the feedback from the SFA point coach since the start of the 2013-14 school year?	Not at all; Somewhat; Extremely
		Q33. To what extent do you agree with the following statements about the 2013-14 school year?	
		a. **(SFA Only)** As a result of SFA, you received training in reading instruction that you had not received before.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. **(SFA Only)** The SFA facilitator provides you with useful feedback.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. **(SFA Only)** The SFA program doesn't provide you with enough autonomy in how you teach.	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. **(SFA Only)** As a result of SFA, you have changed your process for reviewing student reading data.	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. **(SFA Only)** As a result of SFA, you have changed your process for grouping students for reading.	Strongly Disagree; Disagree; Agree; Strongly Agree
		f. **(SFA Only)** Overall, your school has benefited from the SFA program.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Please use this space for additional comments about your school's instructional improvement efforts in reading. In particular, tell us what you think researchers and reformers need to know in order to better understand your school and its experiences	Enter text

Appendix 4

Table A 4: Principal Survey Items

Section	Subsection	Item	Response Options
Background and Broad Responsibilities	Professional Activities	Q1. How many years during your career (including the 2013-14 school year) have you spent doing the following activities?	
		a. Years spent as a classroom teacher of any subject in an elementary school	Enter number (school years)
		b. Years spent in an elementary school, but not as a classroom teacher	Enter number (school years)
		c. Years spent as a principal at your current elementary school	Enter number (school years)
		d. Years spent as a principal at other elementary school(s)	Enter number (school years)
		e. Years spent in any role at schools (including your current school) involved in a whole-school reform effort before 2013- 14	Enter number (school years)
		Q2. This question asks about the 2013-14 school year. How much does your district specify detailed job responsibilities for principals?	Not at all; A little bit; A lot; Completely
		Q3. Recognizing that all responsibilities are important and necessary, which two of the following items take priority above the others during the 2013-14 school year? Your responsibility to… (Check only TWO)	
		a. Encourage teamwork among staff members.	Select / Do not Select
		b. Set high learning standards for all students.	Select / Do not Select
		c. Make expectations for meeting student learning goals clear to teachers.	Select / Do not Select
		d. Provide support for classroom discipline and order.	Select / Do not Select
		e. Engage parents in school activities and student learning.	Select / Do not Select
		f. Monitor your school's progress toward district and state standards	Select / Do not Select
		g. Manage your school's finances.	Select / Do not Select
		h. Manage and respond to district policies and requirements.	Select / Do not Select
		i. Manage and respond to issues related to the community outside the school.	Select / Do not Select
		j. Ensure instruction is of high quality and follows the adopted curriculum	Select / Do not Select
		k. Other (please describe)	Select / Do not Select; Enter text

Section	Subsection	Item	Response Options
The Reading Program at Your School	The Current Program	Q4. What reading program is your school using in the 2013-2014 school year?	Enter text
		Q5. Do you use any other reading program in your school?	Yes (continue to Q5a); No (skip to Q6)
		a. What is the name?	Enter text.
		Q6. Does your school group students by ability or skill level for reading in any grade?	Yes (continue to Q6a); No (skip to Q7)
		a. Does your school group students who are in the same reading class into smaller groups according to their ability level?	Yes; No
		b. Does your school group students who are in the same grade into separate reading classes according to their ability level? (For example, third-grade students from three different homerooms who read at the same level might be in the same reading class.)	Yes; No
		c. Does your school group students who are in different grades into separate reading classes according to their ability level? (For example, a fifth-grade student and a third-grade student who read at the same level might be in the same reading class.)	Yes; No
		Q7. How much is your school trying to align its reading program with Common Core State Standards during this school year?	Not at all; To some extent; To a great extent; Unsure/Don't know
		Q8. On a typical day, what is the length of the reading block (excluding writing and grammar instruction) at your school?	Enter number (minutes)
		Q9. Which two of the following items related to improving reading instruction take priority above the others during the 2013-14 school year? Your responsibility to…	
		a. Serve as a knowledgeable source concerning reading standards and curriculum	Select / Do not Select
		b. Ensure that teachers have time for planning reading instruction	Select / Do not Select
		c. Provide teachers with adequate classroom materials to improve student reading proficiency	Select / Do not Select
		d. Ensure that teachers receive adequate professional development in reading	Select / Do not Select
		e. Reach out to parents to support reading practices at home	Select / Do not Select

Section	Subsection	Item	Response Options
		f. Ensure that teachers receive regular feedback regarding their reading instruction from you, a reading specialist or other instructional coach.	Select / Do not Select
		g. Other	Select / Do not Select; Enter text
The Reading Program at Your School	Instruction	Q10. The following items concern the frequency of your activities with teachers involved in reading instruction. Please select the answer that comes closest to the frequency with which you personally did the following during the first semester of the school year:	
		a. Conducted walk-throughs or briefly observed reading instruction.	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		b. Conducted unscheduled classroom observations to get a sense of reading instruction.	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		c. Provided informal feedback to teachers who you think need improvement in reading instruction.	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		d. Formally evaluated teachers in your school on their reading instruction	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		e. Taught a demonstration class or modeled a reading class.	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		f. Participated in a grade-level meeting with reading teachers.	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		g. Met in small groups with teachers to discuss reading.	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily; Never or Not Applicable
		Q11. Do you personally observe teacher instruction in reading?	Yes (continue to Q11a); No (skip to Q12)
		a. How many teachers do you observe during a typical *week to* get a sense of reading instruction?	Enter number (Teachers)
		b. About how often have you observed teacher instruction in reading in the past month?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily

Section	Subsection	Item	Response Options
		c. In the past month at your school, about how often have you looked for the following when you observed any individual teacher's instruction in reading?	
		i. Reading classes following a prescribed or recommended sequence of activities	Never; Some of the time; Most of the time; All of the time
		ii. Students working in pairs/teams.	Never; Some of the time; Most of the time; All of the time
		iii. Teachers asking questions that require students to think deeply about what they are reading.	Never; Some of the time; Most of the time; All of the time
		iv. Teachers communicating clearly to students the expectations for their assignment.	Never; Some of the time; Most of the time; All of the time
		v. Teachers engaging students in specific reading techniques when students are reading a challenging text.	Never; Some of the time; Most of the time; All of the time
The Reading Program at Your School	Tutoring/ Intervention	Q12. Does your school have a dedicated time for tutoring? [The following questions concern tutoring of students in reading at your school during the 2013-14 school year. 'Tutoring' refers to instruction that may occur before, during or after the school day. It may occur either one-on-one or in a group. It is sometimes called a reading intervention.]	Yes (continue to Q12a); No [skip pattern unclear]
		a. If yes, is this time in addition to the reading block?	Yes; No
		Q13. What is the length of the average tutoring session?	0-10 minutes; 11-20 minutes; 21-30 minutes; 31-40 minutes; 41 or more minutes
		Q14. About how many students in each grade at your school receive tutoring in reading?	
		a. # in Grade 1	Enter number
		b. # in Grade 2	Enter number
		c. # in Grade 3	Enter number
		d. # in Grade 4	Enter number
		e. # in Grade 5	Enter number
		Q15. Is tutoring in reading offered by school staff members at your school?	Yes (continue to Q15a); No (skip to Q16)
		a. How many tutors work as reading teachers at your school?	Enter number (Tutors)
		Q16. Is tutoring in reading offered by volunteers (non-school staff) at your school?	Yes (continue to Q16a); No (skip to Q17)

Section	Subsection	Item	Response Options
		a. How many of these tutors are volunteers?	Enter number (Tutors)
		Q17. Please tell us more how your school's tutoring program operates.	
		a. Students at your school receive tutoring one-on-one.	Yes; No
		b. Students at your school receive tutoring in pull-out groups.	Yes; No
		c. Tutoring is scheduled to take place every day for all students assigned to tutoring	Yes; No
		d. The tutoring program at your school has changed quite a bit compared to the prior school year	Yes; No
		Q18. When students are being tutored, do they miss any of the following subjects? (Please check all that apply)	
		a. Reading	Select / Do not Select
		b. Math	Select / Do not Select
		c. Physical Education	Select / Do not Select
		d. Art	Select / Do not Select
		e. Another Subject	Select / Do not Select; Enter text
		Q19. Does your school use increasingly intensive interventions to provide help to students who are struggling with reading?	Yes (continue to Q19a); No (skip to Q20)
		a. How many levels of intervention (sometimes called tiers) does your school provide?	Enter number (Level(s) of intervention)
The Reading Program at Your School	Use of Data	Q20. Please select the answer that comes closest to the <u>frequency</u> for reviewing reading data at your school.	
		a. Since the start of the 2013-14 school year, how frequently has your school assessed the reading growth of students?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily
		b. Since the start of the 2013-14 school year, you have reviewed reading data…	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily
		c. How often have teachers reviewed data on their students' reading performance during the 2013-14 school year?	Once; Two to three times; Once or twice a month; Once a week; Daily or almost daily
		Q21. With whom do you personally review student reading assessment data? (Please check all that apply)	
		a. No one else	Select / Do not Select
		b. District representative	Select / Do not Select
		c. Assistant principal	Select / Do not Select

Section	Subsection	Item	Response Options
		d. Reading coach	Select / Do not Select
		e. Teachers who specifically teach reading	Select / Do not Select
		f. All teachers	Select / Do not Select
		g. **(SFA Only)** SFA Point Coach	Select / Do not Select
		h. **(SFA Only)** SFA Facilitator	Select / Do not Select
		i. **(SFA Only)** Other (please describe):	Select / Do not Select; Enter text
		Q22. How do you and the people mentioned above review student reading assessment data? Please check all that apply.	
		a. Review each student's score separately	Select / Do not Select
		b. Review a summary of all students by grade level	Select / Do not Select
		c. Review scores disaggregated by specific reading skill content	Select / Do not Select
		d. Review scores disaggregated by student demographic characteristics such as race and English learner status	Select / Do not Select
		e. Other (please describe):	Select / Do not Select; Enter text
		Q23. To what extent do you agree with the following statements?	
		a. Your reading program helps prepare students to do well on state achievement tests.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. You are satisfied with the way that reading assessments have been used during the 2013-14 school year to measure your students' reading skills.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q24. To what extent do you agree with the following statements? Since the start of the school year, your school has used data to…	
		a. Evaluate the reading progress of students over time.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Communicate with and inform parents about student reading performance.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Identify students struggling with reading.	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. Develop strategies to move students from the below basic and basic categories into the proficient category on standardized tests of reading skills.	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. Examine school-wide instructional issues related to reading.	Strongly Disagree; Disagree; Agree; Strongly Agree

Section	Subsection	Item	Response Options
		f. Identify reading teachers who need instructional improvement.	Strongly Disagree; Disagree; Agree; Strongly Agree
		g. Regroup students based on progress made during the last assessment period.	Strongly Disagree; Disagree; Agree; Strongly Agree
The Reading Program at Your School	Grouping	Q25. The following questions concern the use of reading groups at your school during the 2013-14 school year. To what extent do you agree with the following two statements?	
		a. Reading groups inside classrooms generally are small enough at your school for individual students to receive adequate attention	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Re-grouping students over the course of the year for reading classes is an effective strategy for improving students' reading outcomes in specific skill areas	Strongly Disagree; Disagree; Agree; Strongly Agree
General School Functioning	Teams/ Functions	Q26. Please indicate whether someone an individual or a group is responsible for the following at your school.	
		a. Developing school-wide solutions for individual students with behavioral challenges.	The school does not have someone with this responsibility; The school plans to have someone with this responsibility, but it is not in place now; The school has someone with this responsibility.
		b. Developing school-wide solutions for students with learning challenges.	The school does not have someone with this responsibility; The school plans to have someone with this responsibility, but it is not in place now; The school has someone with this responsibility.
		c. Helping teachers to improve their reading instruction of students.	The school does not have someone with this responsibility; The school plans to have someone with this responsibility, but it is not in place now; The school has someone with this responsibility.
		d. Implementing, monitoring, and improving a schoolwide program around social skills development and conflict resolution for all students.	The school does not have someone with this responsibility; The school plans to have someone with this responsibility, but it is not in place now; The school has someone with this responsibility.
		e. Developing school-wide solutions to improve student attendance	The school does not have someone with this responsibility; The school plans to

Section	Subsection	Item	Response Options
			have someone with this responsibility, but it is not in place now; The school has someone with this responsibility.
		f. Fostering closer relationships between the school and students' families.	The school does not have someone with this responsibility; The school plans to have someone with this responsibility, but it is not in place now; The school has someone with this responsibility.
		g. Building relationships with local businesses and institutions to increase community involvement.	The school does not have someone with this responsibility; The school plans to have someone with this responsibility, but it is not in place now; The school has someone with this responsibility.
General School Functioning		Q27. The following questions concern relationships among teachers (of any subject) and staff at your school. To what extent do you agree that the following activities take place at your school during the 2013-14 school year?	
		a. Teachers at your school help administrators in school decision-making processes	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Teachers approach you when they have a problem with a colleague	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Teachers approach each other when they have a concern or question related to classroom instruction	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. Teachers discuss student behavioral challenges with each other	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. Teachers discuss issues of improving instruction with each other	Strongly Disagree; Disagree; Agree; Strongly Agree
		f. Teachers discuss lesson plans that were not particularly successful with other reading instructors	Strongly Disagree; Disagree; Agree; Strongly Agree
		g. Teachers discuss lesson plans that were successful.	Strongly Disagree; Disagree; Agree; Strongly Agree
		h. Teachers share student work with other teachers.	Strongly Disagree; Disagree; Agree; Strongly Agree
		i. Teacher morale at your school during the current school year has been high.	Strongly Disagree; Disagree; Agree; Strongly Agree

Section	Subsection	Item	Response Options
		Q28. The following questions concern your perception of teachers' attitudes toward students at your school during the 2013-14 school year.	
		a. Most teachers at your school believe that their students come to school ready to learn.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. Most teachers at your school build relationships with students' parents/guardians.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. Most teachers at your school believe they can help most students attain grade-level reading skills by the end of the school year, regardless of their family or economic circumstances.	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. Most teachers at your school believe they can help most students improve their reading skills, but not necessarily attain grade-level reading skills by the end of the school year.	Strongly Disagree; Disagree; Agree; Strongly Agree
General School Functioning	Staffing	Q29. The following questions ask about what kinds of changes in staffing have taken place at your school:	
		a. How many reading teachers joined your school between Spring 2013 and Spring 2014?	Enter number (Total number between Spring 2013 and Spring 2014)
		b. How many reading teachers left your school between Spring 2013 and Spring 2014?	Enter number (Total number between Spring 2013 and Spring 2014)
		c. Of those reading teachers who *joined* your school between Spring 2013 and Spring 2014, how many are new to teaching?	Enter number (Total number between Spring 2013 and Spring 2014)
		d. **(SFA Only)** Of those reading teachers who joined your school during this period, how many received training from SFA?	Enter number (Total number between Spring 2013 and Spring 2014)
		e. **(SFA Only)** Of those reading teachers who joined your school during this period, how many did NOT receive training from SFA?	Enter number (Total number between Spring 2013 and Spring 2014)
		Q30. Please indicated the type and extent of the staff changes in your school's reading program in 2013-2014...	
		a. Your school increased total staff	Not at all; Yes, for some positions; Yes, for most positions
		b. Your school reassigned staff to new roles	Not at all; Yes, for some positions; Yes, for most positions

Section	Subsection	Item	Response Options
		c. Your school reduced total staff	Not at all; Yes, for some positions; Yes, for most positions
		Q31. To what extent do you agree with the following statement about school reform? Teachers at this school are frustrated by frequent school reform efforts	Strongly Disagree; Disagree; Agree; Strongly Agree
General School Functioning	Student Health Policies	Q32. Does your school screen students for health challenges?	Yes (continue to Q32a); No (skip to Q34)
		Q33. a. Please check to what extent your school screens students for vision challenges…	
		i. Grade 1	All Students; Some Students; No Students
		ii. Grade 2	All Students; Some Students; No Students
		iii. Grade 3	All Students; Some Students; No Students
		iv. Grade 4	All Students; Some Students; No Students
		v. Grade 5	All Students; Some Students; No Students
		Q33. b. Please check to what extent your school screens students for hearing challenges…	
		i. Grade 1	All Students; Some Students; No Students
		ii. Grade 2	All Students; Some Students; No Students
		iii. Grade 3	All Students; Some Students; No Students
		iv. Grade 4	All Students; Some Students; No Students
		v. Grade 5	All Students; Some Students; No Students
		Q34. Does your school help students diagnosed with hearing or vision problems obtain appropriate solutions?	Yes; No
General School Functioning	Attendance, Parental Involvement	Q35. Student behavior is a problem at your school.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Q36. How frequently do your staff do the following related to student attendance?...	
		a. Ask guardians of frequently tardy or absent students to meet with the school administration to discuss student progress and behavior	Never; Sometimes; Most of the time; Almost always
		b. Provide rewards for students who regularly arrive at school on time.	Never; Sometimes; Most of the time; Almost always
		c. Provide positive recognition to parents whose children attend school regularly.	Never; Sometimes; Most of the time; Almost always
		Q37. The following questions concern your perception of parent engagement at your school regarding: Since the start of the 2013-14 school year:	

Section	Subsection	Item	Response Options
		a. School activities, such as attending parent-teacher conferences, fundraising for school needs, or volunteering at the school.	Some of the parents are engaged, but not the majority; The majority of the parents are engaged; Almost all of the parents are engaged.
		b. Their child's reading practices outside of school	Some of the parents are engaged, but not the majority; The majority of the parents are engaged; Almost all of the parents are engaged.
		c. Teachers calling regarding a child's academic performance.	Some of the parents are engaged, but not the majority; The majority of the parents are engaged; Almost all of the parents are engaged.
General School Functioning	Funding Support	Q38. Does your school or district use outside funds and/or grants to help implement the reading program?	Yes; No
		Q39. To what extent do federal Title I funds support the reading program at your school?	Not at all; 1-25%; 26%-50%; 51%-75%; 76%-100%; Don't Know
		Q40. Are federal Title I funds sufficient to support the reading program at your school?	Not at all; Somewhat sufficient; Mostly sufficient; Unable to answer
		Q41. a. Are there elements of the reading program you are not able to implement at all because of insufficient <u>funding</u>? If yes, mark all that apply.	
		i. Tutoring Students	Select / Do not Select
		ii. Grouping or re-grouping students	Select / Do not Select
		iii. Parent outreach	Select / Do not Select
		iv. Frequent progress monitoring of students	Select / Do not Select
		v. Other (description)	Select / Do not Select; Enter text
		Q41. b. Are there elements of the reading program you are not able to implement at all because of insufficient staffing? If yes, mark all that apply.	
		i. Tutoring students	Select / Do not Select
		ii. Grouping or re-grouping students	Select / Do not Select
		iii. Parent outreach	Select / Do not Select
		iv. Frequent progress monitoring of students	Select / Do not Select
		v. Other (description)	Select / Do not Select; Enter text
General School Functioning		Q42. **(SFA Only)** Is your SFA facilitator currently funded by Title I funds?	Yes; No; Don't Know

Section	Subsection	Item	Response Options
	The Success For All Program	Q43. **(SFA Only)** Does your SFA facilitator split time between SFA and other responsibilities?)	Yes (continue to Q43a); No (skip to Q44)
		a. **(SFA Only)** If yes, what percent of your facilitator's time is spent on SFA?	1%-25 % of the time; 26%-50% of the time; 51%-75% of the time; 76%-100% of the time
		Q44. **(SFA Only)** Does implementing SFA (including parent outreach, tutoring, entering data, etc.) require teachers to spend time outside of official school hours?	Yes; No
		Q45. The following questions concern the changes in instructional practice at your school as a result of the school's involvement with SFA in the 2013-14 school year. Please indicate to what extent you agree that the following occurred at your school	
		a. **(SFA Only)** As a result of SFA, teachers received training in reading instruction that they had not received before.	Strongly Disagree; Disagree; Agree; Strongly Agree
		b. **(SFA Only)** The SFA facilitator and coach provided you with useful feedback.	Strongly Disagree; Disagree; Agree; Strongly Agree
		c. **(SFA Only)** The SFA facilitator has provided teachers with useful feedback.	Strongly Disagree; Disagree; Agree; Strongly Agree
		d. **(SFA Only)** Teachers at your school say the SFA program doesn't provide them enough autonomy in how they teach.	Strongly Disagree; Disagree; Agree; Strongly Agree
		e. **(SFA Only)** As a result of SFA, you have changed your process for observing classroom instruction.	Strongly Disagree; Disagree; Agree; Strongly Agree
		f. **(SFA Only)** As a result of SFA, your school has changed its process for reviewing student reading data.	Strongly Disagree; Disagree; Agree; Strongly Agree
		g. **(SFA Only)** As a result of SFA, your school has changed its process for grouping students for reading.	Strongly Disagree; Disagree; Agree; Strongly Agree
		h. **(SFA Only)** As a result of SFA, teachers at your school collaborate more with each other	Strongly Disagree; Disagree; Agree; Strongly Agree
		i. **(SFA Only)** As a result of SFA, students at your school use more cooperative learning strategies	Strongly Disagree; Disagree; Agree; Strongly Agree
		j. **(SFA Only)** Overall, your school has benefited from the SFA program.	Strongly Disagree; Disagree; Agree; Strongly Agree
		Please use this space for additional comments about your school's instructional improvement efforts in reading. In particular, tell us what you think researchers and reformers need to know in order to better understand your school and its experiences	Enter text

Appendix 5

Table A 5: Mean Scores for Implementation Items in the Teacher and Principal Questionnaires by Treatment Status

Construct	Survey	Item	Scale	Mean (Standard deviation) Treatment	Mean (Standard deviation) Control	Difference Between T and C Schools
Length of the reading block	Teacher	- Average length of the reading block (in minutes) on a typical day (excluding grammar and writing), according to teacher reports.†	1 = 0-5 min 0.75 = 6-10 min 0.5 = 11-15 min 0.25 = 16-20 min 0 = > 20 min	0.893 (0.189)	0.673 (0.329)	0.220** (*p*=0.041)
Small class size	Teacher	- Average number of students in reading class, according to teacher report.	1 = <=20 students 0 = > 20 students	0.786 (0.426)	0.231 (0.439)	0.555*** (*p*=0.003)
Grouping	Principal	- Percentage of principals reporting that students in the same reading class are divided into smaller groups	1 = Yes 0 = No	0.571 (0.513)	0.462 (0.519)	0.110 (*p*=0.585)
	Principal	- Percentage of principals reporting that students in the same grade are grouped into different reading classes by ability level	1 = Yes 0 = No	0.986 (0.053)	0.385 (0.506)	0.601*** (*p*<0.001)
	Principal	- Percentage of principals reporting that: Students who are in different grades, but at the same ability level, are sometimes grouped together in the same reading class	1 = Yes 0 = No	1 (0)	0 (0)	
	Teacher	- Percentage of teachers reporting that students are periodically regrouped for reading by ability level	1 = Yes 0 = No	0.973 (0.08)	0.797 (0.211)	0.176*** (*p*=0.007)
Cooperative learning	Teacher	- Percentage of teachers who agree that students work in pairs or small groups daily or almost daily.	1 = strongly agree 0.66 = agree 0.33= disagree 0=strongly disagree	0.861 (0.066)	0.622 (0.075)	0.239*** (*p*<0.001)

Construct	Survey	Item	Scale	Mean (Standard deviation)		Difference Between T and C Schools
				Treatment	Control	
	Principal	- Percentage of classrooms in which students were observed working in small groups.	1 = all of the time 0.66 = most of the time 0.33= some of the time 0=never	0.857 (0.171)	0.538 (0.256)	0.319*** (*p*=0.001)
Tutoring	Principal	- Percentage of principals reporting that students are tutored one on one	1 = Yes 0 = No	0.357 (0.497)	0.385 (0.506)	-0.027 (*p*=0.888)
	Principal	- Percentage of principals reporting that students receive tutoring in pull-out groups	1 = Yes 0 = No	0.857 (0.363)	0.769 (0.439)	0.088 (*p*=0.574)
	Principal	- Percentage of principals reporting that tutoring is scheduled every day for all students assigned to tutoring	1 = Yes 0 = No	0.571 (0.514)	0.231 (0.439)	0.340* (*p*=0.077)
	Principal	- Average length of a tutoring session (in minutes)	1 = > 40 min 0.75 = 31-40 min 0.5 = 21-30 min 0.25 = 11-20 min 0 = 0-10 min	0.518 (0.268)	0.692 (0.309)	-0.174 (*p*=0.129)
	Principal	- Percentage of principals reporting that their school uses a system of increasingly intensive interventions for students who are struggling with reading	1 = Yes 0 = No	0.857 (0.363)	0.846 (0.376)	0.011 (*p*=0.939)
Use of educational media/technology	Teacher	- Teachers agree that they use educational media/technology as part of the reading program at their school	1 = strongly agree 0.66 = agree 0.33= disagree 0=strongly disagree	0.749 (0.086)	0.673 (0.067)	0.076** (*p*=0.017)

Construct	Survey	Item	Scale	Mean (Standard deviation)		Difference Between T and C Schools
				Treatment	Control	
	Teacher	- Average amount of time teachers report using educational media/technology in their most recent reading class.	1 > 50 min 0.8 = 41-50 min 0.6 = 31-40 min 0.4 = 21-30 min 0.2 = 11-20 min 0 = 0-10 min	0.743 (0.241)	0.369 (0.160)	0.374*** (*p*<0.001)

Notes. * p < 0.1, ** p < 0.05, *** p < 0.01.

† The scale for this variable represents the distance from the ideal lesson length, as defined by the program (90 minutes). For example, lessons that lasted 78 or 102 minutes would be assigned a score of 0.5, as they are 12 minutes under and 12 minutes over 90 minutes, respectively.

Bibliography

Abry, T., Rimm-Kaufman, S. E., Larsen, R. A., & Brewer, A. J. (2013). The influence of fidelity of implementation on teacher–student interaction quality in the context of a randomized controlled trial of the Responsive Classroom approach. *Journal of School Psychology*, *51*(4), 437–453. https://doi.org/10.1016/j.jsp.2013.03.001

American Institutes for Research. (n.d.). *HighScope Preschool Curriculum and Professional Development Efficacy Study*. American Institutes for Research. Retrieved July 21, 2024, from https://www.air.org/project/highscope-preschool-curriculum-and-professional-development-efficacy-study

American Psychological Association. (2002). *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists*. American Psychological Association.

Apthorp, H., Randel, B., Cherasaro, T., Clark, T., McKeown, M., & Beck, I. (2012). Effects of a Supplemental Vocabulary Program on Word Knowledge and Passage Comprehension. *Journal of Research on Educational Effectiveness*, *5*(2), 160–188. https://doi.org/10.1080/19345747.2012.660240

August, D., Barr, C., Carlson, C., Cárdenas-Hagan, E., Johnston, W. T., & Marken, A. (2023). A promising intervention designed to improve EAL learners' mathematics skills and associated academic language. *Bilingual Research Journal*, *46*(1–2), 117–141. https://doi.org/10.1080/15235882.2023.2225459

Autio, E., Nelsestuen, K., & Northwest, E. (2014). *What Is It Like to Be a Guinea Pig? Teacher Experiences in a Randomized Controlled Trial*. Annual Meeting of the American Evaluation Research Association.

Babendure, J., Thompson, L., Peterman, K., Teiper, L., Gastil, H., Liwanag, H., & Glenn-Lee, S. (2011). BioBridge Professional Development: Bringing Innovative Science into the Classroom. *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness Annual Conference. https://eric.ed.gov/?id=ED528926

Babinski, L. M., Amendum, S. J., Knotek, S. E., Sánchez, M., & Malone, P. (2018). Improving Young English Learners' Language and Literacy Skills Through Teacher Professional Development: A Randomized Controlled Trial. *American Educational Research Journal*, *55*(1), 117–143. https://doi.org/10.3102/0002831217732335

Bae, Y., Hand, B. M., & Fulmer, G. W. (2022). A generative professional development program for the development of science teacher epistemic orientations and teaching practices. *Instructional Science*, *50*(1), 143–167. https://doi.org/10.1007/s11251-021-09569-y

Balu, R., & Quint, J. (2015). Measuring Implementation Fidelity in Success for All. In C. V. Meyers & W. C. Brandt (Eds.), *Implementation Fidelity in Education Research: Designer and Evaluator Considerations* (pp. 154–175). Routledge.

Barber, A. T., Buehl, M. M., Kidd, J. K., Sturtevant, E. G., Richey Nuland, L., & Beck, J. (2015). Reading Engagement in Social Studies: Exploring the Role of a Social Studies Literacy Intervention on Reading Comprehension, Reading Self-Efficacy, and Engagement in Middle School Students with Different Language Backgrounds. *Reading Psychology*, *36*(1), 31–85. https://doi.org/10.1080/02702711.2013.815140

Bardach, E. (1980). *The implementation game: What happens after a bill becomes a law*. MIT Press.

Barr, C. D., Reutebuch, C. K., Carlson, C. D., Vaughn, S., & Francis, D. J. (2019). Explaining Variation in Findings From Efficacy and Effectiveness Studies for English Reading

Interventions for English Learners. *Journal of Research on Educational Effectiveness*, *12*(1), 116–134. https://doi.org/10.1080/19345747.2018.1529211

Barton, E. A., Whittaker, J. V., Kinzie, M. B., DeCoster, J., & Furnari, E. (2017). Understanding the relationship between teachers' use of online demonstration videos and fidelity of implementation in *MyTeachingPartner-Math/Science*. *Teaching and Teacher Education*, *67*, 189–201. https://doi.org/10.1016/j.tate.2017.06.011

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29. https://doi.org/10.1080/09243453.2018.1539014

Berman, P., & McLaughlin, M. W. (1974). *Federal Programs Supporting Educational Change: A Model of Educational Change. Volume I* (No. R-1589/1-HEW). RAND Corporation.

Berman, P., & McLaughlin, M. W. (1976). Implementation of Educational Innovation. *The Educational Forum*, *40*(3), 345–370. https://doi.org/10.1080/00131727609336469

Berman, P., & McLaughlin, M. W. (1978). *Federal Programs Supporting Educational Change, Vol. VIII: Implementing and Sustaining Innovations* (No. R-1589/8-HEW). RAND Corporation.

Billman, A. K., Pearson, P. D., & Barber, J. (2018). *First Grade Second Language: Uniting Science Knowledge and Literacy Development for English Learners* [IES Final Report].

Blakely, C. H., Mayer, J. P., Gottschaik, R. G., Sehmitt, N., Davidson, W. S., Roitman, D. B., & Emshoff, J. G. (1987). The fidelity-adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology*, *15*(3), 253–268.

Blanton, M., Stroud, R., Stephens, A., Gardiner, A. M., Stylianou, D. A., Knuth, E., Isler-Baykal, I., & Strachota, S. (2019). Does Early Algebra Matter? The Effectiveness of an Early Algebra Intervention in Grades 3 to 5. *American Educational Research Journal*, *56*(5), 1930–1972. https://doi.org/10.3102/0002831219832301

Burkhauser, M. A., & Lesaux, N. K. (2017). Exercising a bounded autonomy: Novice and experienced teachers' adaptations to curriculum materials in an age of accountability. *Journal of Curriculum Studies*, *49*(3), 291–312. https://doi.org/10.1080/00220272.2015.1088065

Burte, H., Gardony, A. L., Hutton, A., & Taylor, H. A. (2017). Think3d!: Improving mathematics learning through embodied spatial training. *Cognitive Research: Principles and Implications*, *2*(1), 13. https://doi.org/10.1186/s41235-017-0052-9

Burte, H., Gardony, A. L., Hutton, A., & Taylor, H. A. (2020). Elementary teachers' attitudes and beliefs about spatial thinking and mathematics. *Cognitive Research: Principles and Implications*, *5*, 17. https://doi.org/10.1186/s41235-020-00221-w

Buxton, C. A., Allexsaht-Snider, M., Kayumova, S., Aghasaleh, R., Choi, Y.-J., & Cohen, A. (2015). Teacher agency and professional learning: Rethinking fidelity of implementation as multiplicities of enactment. *Journal of Research in Science Teaching*, *52*(4), 489–502. https://doi.org/10.1002/tea.21223

Cabell, S. Q. (2020). *Impact of the Core Knowledge Language Arts' Read-Aloud Program on Kindergarteners' Vocabulary, Listening Comprehension, and General Knowledge*. Marvalene Hughes Research in Education Conference. https://cehhs.fsu.edu/sites/g/files/upcbnu4516/files/research/Sonia%20Cabell%2C%202020.pdf

Campbell, J. M., & Dommestrup, A. K. (2010). Evidence-based Assessment of Cognitive Functioning in Pediatric Psychology. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini Encyclopedia of Psychology*. https://doi.org/10.1002/9780470479216.corpsy0649

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, *2*(1), 40. https://doi.org/10.1186/1748-5908-2-40

Century, J., & Cassata, A. (2016). Implementation Research: Finding Common Ground on What, How, Why, Where, and Who. *Review of Research in Education*, *40*(1), 169–215. https://doi.org/10.3102/0091732X16665332

Chaparro, E. A., Smolkowski, K., Gunn, B., Dennis, C., & Vadasy, P. (2022). Evaluating the Efficacy of an English Language Development Program for Middle School English Learners. *Journal of Education for Students Placed at Risk (JESPAR)*, *27*(4), 322–352. https://doi.org/10.1080/10824669.2022.2045993

Cheung, A. C. K., Xie, C., Zhuang, T., Neitzel, A. J., & Slavin, R. E. (2021). Success for All: A Quantitative Synthesis of U.S. Evaluations. *Journal of Research on Educational Effectiveness*, *14*(1), 90–115. https://doi.org/10.1080/19345747.2020.1868031

Chhin, C. S., Taylor, K. A., & Wei, W. S. (2018). Supporting a Culture of Replication: An Examination of Education and Special Education Research Grants Funded by the Institute of Education Sciences. *Educational Researcher*, *47*(9), 594–605. https://doi.org/10.3102/0013189X18788047

Cho, J. (1998, April). *Rethinking Curriculum Implementation: Paradigms, Models, and Teachers' Work*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Clark, T. F., Arens, S. A., & Stewart, J. (2015). Efficacy Study of a Pre-Algebra Supplemental Program in Rural Mississippi: Preliminary Findings. *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED562175

Clements, D. H., Sarama, J., Layzer, C., Unlu, F., & Fesler, L. (2020). Effects on Mathematics and Executive Function of a Mathematics and Play Intervention Versus Mathematics Alone. *Journal for Research in Mathematics Education*, *51*(3), 301–333. https://doi.org/10.5951/jresemtheduc-2019-0069

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics Learned by Young Children in an Intervention Based on Learning Trajectories: A Large-Scale Cluster Randomized Trial. *Journal for Research in Mathematics Education*, *42*(2), 127–166. https://doi.org/10.5951/jresematheduc.42.2.0127

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal Evaluation of a Scale-Up Model for Teaching Mathematics With Trajectories and Technologies: Persistence of Effects in the Third Year. *American Educational Research Journal*, *50*(4), 812–850. https://doi.org/10.3102/0002831212469270

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine*, *30*(1), 65–73. https://doi.org/10.1207/s15324796abm3001_8

Cook, T. D. (2005). Emergent Principles for the Design, Implementation, and Analysis of Cluster-Based Experiments in Social Science. *The ANNALS of the American Academy of*

Political and Social Science, *599*(1), 176–198. https://doi.org/10.1177/0002716205275738

Corbin, C. M., Downer, J. T., Ruzek, E. A., Lowenstein, A. E., & Brown, J. L. (2020). Correlates of change in elementary students' perceptions of interactions with their teacher. *Journal of Applied Developmental Psychology*, *69*, 101144. https://doi.org/10.1016/j.appdev.2020.101144

Correnti, R., Matsumura, L. C., Walsh, M., Zook-Howell, D., Bickel, D. D., & Yu, B. (2021). Effects of Online Content-Focused Coaching on Discussion Quality and Reading Achievement: Building Theory for How Coaching Develops Teachers' Adaptive Expertise. *Reading Research Quarterly*, *56*(3), 519–558. https://doi.org/10.1002/rrq.317

Crosson, A. C., Correnti, R., Matsumura, L. C., & McKeown, M. G. (2024). Effects of the Triple Q Intervention on Argument Writing: Findings from a Small-Scale Cluster-Randomized Controlled Trial. *Journal of Research on Educational Effectiveness*, *17*(3), 590–613. https://doi.org/10.1080/19345747.2023.2231941

Crosson, A. C., McKeown, M. G., Lei, P., Zhao, H., Li, X., Patrick, K., Brown, K., & Shen, Y. (2021). Morphological analysis skill and academic vocabulary knowledge are malleable through intervention and may contribute to reading comprehension for multilingual adolescents. *Journal of Research in Reading*, *44*(1), 154–174. https://doi.org/10.1111/1467-9817.12323

Crosson, A. C., McKeown, M. G., Moore, D. W., & Ye, F. (2019). Extending the bounds of morphology instruction: Teaching Latin roots facilitates academic word learning for English Learner adolescents. *Reading and Writing*, *32*(3), 689–727. https://doi.org/10.1007/s11145-018-9885-y

Crosson, A. C., & Moore, D. (2017). When to Take Up Roots: The Effects of Morphology Instruction for Middle School and High School English Learners. *Reading Psychology*, *38*(3), 262–288. https://doi.org/10.1080/02702711.2016.1263699

Dane, A. V., & Schneider, B. H. (1998). Program Integrity in Primary and Early Secondary Prevention: Are Implementation Effects out of Control? *Clinical Psychology Review*, *18*(1), 23–45. https://doi.org/10.1016/S0272-7358(97)00043-3

Darrow, C. L. (2013). The Effectiveness and Precision of Intervention Fidelity Measures in Preschool Intervention Research. *Early Education & Development*, *24*(8), 1137–1160. https://doi.org/10.1080/10409289.2013.765786

Daugherty, L., Phillips, A., Pane, J. F., & Karam, R. T. (2012). *Analysis of Costs in an Algebra I Curriculum Effectiveness Study*. RAND Corporation. https://www.rand.org/pubs/technical_reports/TR1171-1.html

Davenport, J. L., Kao, Y. S., Johannes, K. N., Hornburg, C. B., & McNeil, N. M. (2022). Improving Children's Understanding of Mathematical Equivalence: An Efficacy Study. *Journal of Research on Educational Effectiveness*, *0*(0), 1–28. https://doi.org/10.1080/19345747.2022.2144787

Davis, M. H., Schoeneberger, J., Rhoads, C., Mac Iver, D. J., Zhang, X., Mac Iver, M., & Spinney, S. (2024). Accelerating Literacy for Adolescents (ALFA): Evaluating ALFA Lab Using a Regression Discontinuity Study. *Journal of Research on Educational Effectiveness*, *0*(0), 1–27. https://doi.org/10.1080/19345747.2024.2338129

De La Paz, S., Felton, M., Monte-Sano, C., Croninger, R., Jackson, C., Deogracias, J. S., & Hoffman, B. P. (2014). Developing Historical Reading and Writing With Adolescent

Readers: Effects on Student Learning. *Theory & Research in Social Education*, *42*(2), 228–274. https://doi.org/10.1080/00933104.2014.908754

De La Paz, S., Monte-Sano, C., Felton, M., Croninger, R., Jackson, C., & Piantedosi, K. W. (2017). A Historical Writing Apprenticeship for Adolescents: Integrating Disciplinary Learning With Cognitive Strategies. *Reading Research Quarterly*, *52*(1), 31–52. https://doi.org/10.1002/rrq.147

Deussen, T., Autio, E., Roccograndi, A., & Hanita, M. (2014). *The Impact of Project GLAD on Students' Literacy and Science Learning: Year 1 Results from a Cluster-Randomized Trial of Sheltered Instruction*. Annual Meeting of the Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED562842

Deussen, T., Roccograndi, A., Hanita, M., Autio, E., Rodriguez-Mojica, C., Rodriguez, C., & Northwest, E. (2015). *The impact of Project GLAD on fifth-grade literacy: Sheltered instruction and English learners in the mainstream classroom*. Annual Meeting of the American Education Research Association.

Dewey, J. (1933). *How We Think*. D.C. Heath and Company.

Dhillon, S., Darrow, C., & Meyers, C. V. (2014). Introduction to Implementation Fidelity. In *Implementation Fidelity in Education Research*. Routledge.

Diamond, B. S., Maerten-Rivera, J., Rohrer, R. E., & Lee, O. (2014). Effectiveness of a curricular and professional development intervention at improving elementary teachers' science content knowledge and student achievement outcomes: Year 1 results. *Journal of Research in Science Teaching*, *51*(5), 635–658. https://doi.org/10.1002/tea.21148

Doabler, C. T., Clarke, B., Kosty, D. B., Baker, S. K., Smolkowski, K., & Fien, H. (2016). Effects of a Core Kindergarten Mathematics Curriculum on the Mathematics

Achievement of Spanish-Speaking English Learners. *School Psychology Review*, *45*(3), 343–361.

Doabler, C. T., Nelson, N. J., Kosty, D. B., Fien, H., Baker, S. K., Smolkowski, K., & Clarke, B. (2014). Examining Teachers' Use of Evidence-Based Practices During Core Mathematics Instruction. *Assessment for Effective Intervention*, *39*(2), 99–111. https://doi.org/10.1177/1534508413511848

Dobson, D., & Cook, T. J. (1980). Avoiding type III error in program evaluation. *Evaluation and Program Planning*, *3*(4), 269–276. https://doi.org/10.1016/0149-7189(80)90042-7

Domitrovich, C. E., Bradshaw, C. P., Poduska, J. M., Hoagwood, K., Buckley, J. A., Olin, S., Romanelli, L. H., Leaf, P. J., Greenberg, M. T., & Ialongo, N. S. (2008). Maximizing the Implementation Quality of Evidence-Based Preventive Interventions in Schools: A Conceptual Framework. *Advances in School Mental Health Promotion*, *1*(3), 6–28. https://doi.org/10.1080/1754730X.2008.9715730

Doyle, N. B., Gomez Varon, J. A., Downer, J. T., & Brown, J. L. (2023). Testing the integration of a teacher coaching model and a social-emotional learning and literacy intervention in urban elementary schools. *Teaching and Teacher Education*, *132*, 104232. https://doi.org/10.1016/j.tate.2023.104232

Duncan Seraphin, K., Harrison, G. M., Philippoff, J., Brandon, P. R., Nguyen, T. T. T., Lawton, B. E., & Vallin, L. M. (2017). Teaching aquatic science as inquiry through professional development: Teacher characteristics and student outcomes. *Journal of Research in Science Teaching*, *54*(9), 1219–1245. https://doi.org/10.1002/tea.21403

Dunst, C. J., Trivette, C. M., & Raab, M. (2013). An Implementation Science Framework for Conceptualizing and Operationalizing Fidelity in Early Childhood Intervention Studies. *Journal of Early Intervention*, *35*(2), 85–101. https://doi.org/10.1177/1053815113502235

Durlak, J. A. (2015). Studying Program Implementation Is Not Easy but It Is Essential. *Prevention Science*, *16*(8), 1123–1127. https://doi.org/10.1007/s11121-015-0606-3

Durlak, J. A., & DuPre, E. P. (2008). Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation. *American Journal of Community Psychology*, *41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Dusenbury, L. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Dusenbury, L., Brannigan, R., Hansen, W. B., Walsh, J., & Falco, M. (2005). Quality of implementation: Developing measures crucial to understanding the diffusion of preventive interventions. *Health Education Research*, *20*(3), 308–313. https://doi.org/10.1093/her/cyg134

Emshoff, J. G., Blakely, C., Gottschalk, R., Mayer, J., Davidson, W. S., & Erickson, S. (1987). *Innovation in Education and Criminal Justice: Measuring Fidelity of Implementation and Program Effectiveness*. *9*(4), 300–311.

Enders, C. K. (2022). *Applied missing data analysis, 2nd ed* (pp. ix, 546). The Guilford Press.

Every Student Succeeds Act, 14th Congress (2015-2016) S.1177 (2015).

Farran, D. C., Wilson, S. J., Meador, D., Norvell, J., & Nesbitt, K. (2015). Experimental Evaluation of the Tools of the Mind Pre-K Curriculum. Technical Report. Working Paper.

In *Peabody Research Institute*. Peabody Research Institute. https://eric.ed.gov/?id=ED574842

Fernández, M. P., & Martínez, J. F. (2022). Evaluating Teacher Performance and Teaching Effectiveness: Conceptual and Methodological Considerations. In J. Manzi, Y. Sun, & M. R. García (Eds.), *Teacher Evaluation Around the World: Experiences, Dilemmas and Future Challenges* (pp. 39–70). Springer International Publishing. https://doi.org/10.1007/978-3-031-13639-9_3

Firetto, C. M., Murphy, P. K., Greene, J. A., Li, M., Wei, L., Montalbano, C., Hendrick, B., & Croninger, R. M. V. (2019). Bolstering students' written argumentation by refining an effective discourse intervention: Negotiating the fine line between flexibility and fidelity. *Instructional Science*, *47*(2), 181–214.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. A., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature* (No. FHMI Publication #231). University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network.

Fullan, M., & Pomfret, A. (1977). Research on Curriculum and Instruction Implementation. *Review of Educational Research*, *47*(2), 335–397.

Germeroth, C., Bodrova, E., Day-Hess, C., Barker, J. E., Sarama, J., Clements, D., & Layzer, C. (2019). Play It High, Play It Low: Examining the Reliability and Validity of a New Observation Tool to Measure Children's Make-Believe Play. *American Journal of Play*. https://www.semanticscholar.org/paper/Play-It-High%2C-Play-It-Low%3A-Examining-the-and-of-a-Germeroth-Bodrova/ebd6c362f7b166acf1073f15da43b87596d2bc7d

Gómez, J. A., Brown, J. L., & Downer, J. T. (2023). High quality implementation of 4Rs + MTP increases classroom emotional support and reduces absenteeism. *Frontiers in Psychology*, *14*, 1065749. https://doi.org/10.3389/fpsyg.2023.1065749

Gonzalez Canche, M. S., Portes, P. R., Mellon, P. J., Stollberg, R. A., & Turk, J. M. (2014). Instrumental Iteration toward FOI Evaluation of Pedagogical Methods. *Society for Research on Educational Effectiveness*. https://eric.ed.gov/?id=ED562767

Gonzalez, J. E., Kim, H., Anderson, J., & Pollard-Durodola, S. (2024). The Effects of a Science and Social Studies Content Rich Shared Reading Intervention on the Vocabulary Learning of Preschool Dual Language Learners. *Early Childhood Research Quarterly*, *66*, 34–47. https://doi.org/10.1016/j.ecresq.2023.08.011

Gray, A. M., Sirinides, P. M., Fink, R. E., & Bowden, A. B. (2022). Integrating Literacy and Science Instruction in Kindergarten: Results From the Efficacy Study of Zoology One. *Journal of Research on Educational Effectiveness*, *15*(1), 1–27. https://doi.org/10.1080/19345747.2021.1938313

Gray, A., Sirinides, P., Bowden, B., Fink, R., DuBois, T., Suwak, K., & Castillo, W. (2022). *Efficacy Evaluation of Zoology One: Kindergarten Research Labs Online Appendix of Measures and Tools for Data Collection* (CPRE Policy Briefs). Consortium for Policy Research in Education (CPRE). https://repository.upenn.edu/cpre_researchreports/116

Gray, A., Sirinides, P., Fink, R., & Bowden, B. (2020). *Zoology One Efficacy Evaluation Summary of Findings (April 2020)* (CPRE Policy Briefs). Consortium for Policy Research in Education (CPRE). https://repository.upenn.edu/cpre_policybriefs/87

Gray, S. I., Wilcox, M. J., & Reiser, M. (2024). Efficacy of the Teaching Early Literacy and Language Curriculum With Preschoolers From Low-Income Families. *Language, Speech, and Hearing Services in Schools*. https://doi.org/10.1044/2024_LSHSS-23-00140

Gropen, J., Clark-Chiarelli, N., Ehrlich, S., & Thieu, Y. (2011). Examining the Efficacy of "Foundations of Science Literacy": Exploring Contextual Factors. *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness Annual Conference. https://eric.ed.gov/?id=ED518137

Gropen, J., Kook, J. F., Hoisington, C., & Clark-Chiarelli, N. (2017). Foundations of Science Literacy: Efficacy of a Preschool Professional Development Program in Science on Classroom Instruction, Teachers' Pedagogical Content Knowledge, and Children's Observations and Predictions. *Early Education and Development*, *28*(5), 607–631. https://doi.org/10.1080/10409289.2017.1279527

Gunn, B., Smolkowski, K., Strycker, L. A., & Dennis, C. (2021). Measuring Explicit Instruction Using Classroom Observations of Student–Teacher Interactions (COSTI). *Perspectives on Behavior Science*, *44*(2–3), 267–283. https://doi.org/10.1007/s40614-021-00291-1

Hadley, E. B., Scott, M., Foster, M. E., Dickinson, D. K., Hirsh-Pasek, K., & Golinkoff, R. M. (2022). Preschool Teachers' Fidelity in Implementing a Vocabulary Intervention: Variation Across Settings and Strategies. *Topics in Language Disorders*, *42*(4), 319. https://doi.org/10.1097/TLD.0000000000000294

Hall, G. E., & Loucks, S. F. (1978). *Innovation Configurations: Analyzing the Adaptations of Innovations*. Texas University, Austin. Research and Development Center for Teacher, Education.

Hand, B., Park, S., & Suh, J. K. (2018). Examining Teachers' Shifting Epistemic Orientations in Improving Students' Scientific Literacy Through Adoption of the Science Writing Heuristic Approach. In K.-S. Tang & K. Danielsson (Eds.), *Global Developments in Literacy Research for Science Education* (pp. 339–355). Springer International Publishing. https://doi.org/10.1007/978-3-319-69197-8_20

Hand, B., Shelley, M. C., Laugerman, M., Fostvedt, L., & Therrien, W. (2018). Improving critical thinking growth for disadvantaged groups within elementary school science: A randomized controlled trial using the Science Writing Heuristic approach. *Science Education*, *102*(4), 693–710. https://doi.org/10.1002/sce.21341

Hand, B., Therrien, W., & Shelley, M. (2013). *Examining the Impact of Using the Science Writing Heuristic Approach in Learning Science: A Cluster Randomized Study*. Society for Research on Educational Effectiveness Conference. https://eric.ed.gov/?id=ED563211

Hansen, W. B. (2014). Measuring Fidelity. In Z. Sloboda & H. Petras (Eds.), *Defining Prevention Science* (pp. 335–359). Springer.

Hansen, W. B., Graham, J. W., Wolkenstein, B. H., & Rohrbach, L. A. (1991). Program integrity as a moderator of prevention program effectiveness: Results for fifth-grade students in the adolescent alcohol prevention trial. *Journal of Studies on Alcohol*, *52*(6), 568–579. https://doi.org/10.15288/jsa.1991.52.568

Hansen, W. B., Pankratz, M. M., Dusenbury, L., Giles, S. M., Bishop, D. C., Albritton, J., Albritton, L. P., & Strack, J. (2013). Styles of adaptation: The impact of frequency and valence of adaptation on preventing substance use. *Health Education*, *113*(4), 345–363. https://doi.org/10.1108/09654281311329268

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the Black Box: Using Process Evaluation Measures to Assess Implementation and Theory Building. *American Journal of Community Psychology*, *27*(5), 711–731. https://doi.org/10.1023/A:1022194005511

Harris, C. J., Murphy, R., Feng, M., & Rutstein, D. W. (2023). Supporting Science Learning and Literacy Development Together: Initial Results from a Curriculum Study in 1st Grade Classrooms. In *WestEd*. WestEd. https://eric.ed.gov/?id=ED638494

Harris, K. R., Kim, Y.-S., Yim, S., Camping, A., & Graham, S. (2023). Yes, they can: Developing transcription skills and oral language in tandem with SRSD instruction on close reading of science text to write informative essays at grades 1 and 2. *Contemporary Educational Psychology*, *73*, 102150. https://doi.org/10.1016/j.cedpsych.2023.102150

Hassinger-Das, B., Ridge, K., Parker, A., Golinkoff, R. M., Hirsh-Pasek, K., & Dickinson, D. K. (2016). Building Vocabulary Knowledge in Preschoolers Through Shared Book Reading and Gameplay. *Mind, Brain, and Education*, *10*(2), 71–80. https://doi.org/10.1111/mbe.12103

Hendy, E., & Cuevas, J. (2020). The Effects on Instructional Conversations on English Language Learners. *Georgia Educational Researcher*, *17*(2). https://eric.ed.gov/?id=EJ1262458

Herrmann-Abell, C. F., Hardcastle, J., & Roseman, J. E. (2019, April). Evaluating a Unit Aimed at Helping Students Understand Matter and Energy for Growth and Activity. *Grantee Submission*. Annual Meeting of the American Educational Research Association, Toronto, Canada. https://eric.ed.gov/?id=ED598354

Herrmann-Abell, C. F., Koppal, M., & Roseman, J. E. (2016). Toward High School Biology: Helping Middle School Students Understand Chemical Reactions and Conservation of

Mass in Nonliving and Living Systems. *CBE Life Sciences Education*, *15*(4), ar74. https://doi.org/10.1187/cbe.16-03-0112

Hill, H. C., & Erickson, A. (2019). Using Implementation Fidelity to Aid in Interpreting Program Impacts: A Brief Review. *Educational Researcher*, *48*(9), 590–598. https://doi.org/10.3102/0013189X19891436

Hmelo-Silver, C. E., Jordan, R., Eberbach, C., & Sinha, S. (2017). Systems learning with a conceptual representation: A quasi-experimental study. *Instructional Science*, *45*(1), 53–72. https://doi.org/10.1007/s11251-016-9392-y

Hoewe, J. (2017). Manipulation Check. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The International Encyclopedia of Communication Research Methods* (1st ed., pp. 1–5). Wiley. https://doi.org/10.1002/9781118901731.iecrm0135

Honig, M. (2006). Introduction: Complexity and policy implementation. *Confronting Complexity*, 1–23.

Howard, E., Yee, D., Weinberg, E., Ogut, B., & Lee, D. H. (2020, November 12). *Highscope Preschool Curriculum and Professional Development Efficacy Study*. 2020 APPAM Fall Research Conference. https://appam.confex.com/appam/2020/meetingapp.cgi/Paper/37914

Howard, E., Yee, D., Weinberg, E., Ogut, B., & Lee, D. H. (2021). *HighScope Preschool Curriculum and Professional Development Efficacy Study: Results in Brief*. https://www.air.org/project/highscope-preschool-curriculum-and-professional-development-efficacy-study

Hulleman, C. S., & Cordray, D. S. (2009). Moving From the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88–110. https://doi.org/10.1080/19345740802539325

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Institute of Education Sciences. (2015). *Request for Applications: Education Research Grants. CFDA Number: 84.305A*. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences. (2016). *Request for Applications: Education Research Grants. CFDA Number: 84.305A*. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences. (2017). *Request for Applications: Education Research Grants. CFDA Number: 84.305A*. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences. (2018). *Request for Applications: Education Research Grants. CFDA Number: 84.305A*. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences. (2019). *Request for Applications: Education Research Grants. CFDA Number: 84.305A*. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences. (2020). *Request for Applications: Education Research Grants. CFDA Number: 84.305A*. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences. (2022a). *Request for Proposals: Education Research Grants Program* [Assistance Listing Number (ALN): 84.305A]. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences. (2022b). *Request for Proposals: Research Grants Focused on Systematic Replication* [Assistance Listing Number (ALN): 84.305R]. U.S. Department of Education, Institute of Education Sciences.

Institute of Education Sciences & National Science Foundation. (2013). *Common Guidelines for Education Research and Development*. https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf13126

Jitendra, A. K., Harwell, M. R., Dupuis, D. N., Karl, S. R., Lein, A. E., Simonson, G., & Slater, S. C. (2015). Effects of a research-based intervention to improve seventh-grade students' proportional problem solving: A cluster randomized trial. *Journal of Educational Psychology*, *107*(4), 1019–1034. https://doi.org/10.1037/edu0000039

Karam, R., Pane, J. F., Griffin, B. A., Robyn, A., Phillips, A., & Daugherty, L. (2017). Examining the implementation of technology-based blended algebra I curriculum at scale. *Educational Technology Research and Development*, *65*(2), 399–425. https://doi.org/10.1007/s11423-016-9498-6

Klein, A., Starkey, P., Clements, D., Sarama, J., & Iyer, R. (2008). Effects of a Pre-Kindergarten Mathematics Intervention: A Randomized Experiment. *Journal of Research on Educational Effectiveness*, *1*(3), 155–178. https://doi.org/10.1080/19345740802114533

Lawless, K. A., Brown, S. W., Rhoads, C., Lynn, L., Newton, S. D., Brodowiksa, K., Oren, J., Riel, J., Song, S., & Wang, M. (2018). Promoting students' science literacy skills through a simulation of international negotiations: The GlobalEd 2 Project. *Computers in Human Behavior*, *78*, 389–396. https://doi.org/10.1016/j.chb.2017.08.027

Lawrence, J. F., Crosson, A. C., Paré-Blagoev, E. J., & Snow, C. E. (2015). Word Generation Randomized Trial: Discussion Mediates the Impact of Program Treatment on Academic Word Learning. *American Educational Research Journal*, *52*(4), 750–786. https://doi.org/10.3102/0002831215579485

Lawrence, J. F., Francis, D., Paré-Blagoev, J., & Snow, C. E. (2017). The Poor Get Richer: Heterogeneity in the Efficacy of a School-Level Intervention for Academic Language. *Journal of Research on Educational Effectiveness*, *10*(4), 767–793. https://doi.org/10.1080/19345747.2016.1237596

Lemire, S., Kwako, A., Nielsen, S. B., Christie, C. A., Donaldson, S. I., & Leeuw, F. L. (2020). What Is This Thing Called a Mechanism? Findings From a Review of Realist Evaluations. *New Directions for Evaluation*, *2020*(167), 73–86. https://doi.org/10.1002/ev.20428

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*(5), 635–652. https://doi.org/10.1080/03054985.2012.734800

Lesaux, N. K., Kieffer, M. J., Kelley, J. G., & Harris, J. R. (2014). Effects of Academic Vocabulary Instruction for Linguistically Diverse Adolescents: Evidence From a Randomized Field Trial. *American Educational Research Journal*, *51*(6), 1159–1194. https://doi.org/10.3102/0002831214532165

Li, L., Ringstaff, C., Tripathy, R. G., Flynn, K., & Thomas, L. (2019). Improving Elementary School Students' Vocabulary Skills and Reading Comprehension through a Word Learning Strategies Program. *Grantee Submission*. Annual Meeting of the American Educational Research Association (AERA. https://eric.ed.gov/?id=ED604594

Lin, A. R., Lawrence, J. F., Snow, C. E., & Taylor, K. S. (2016). Assessing Adolescents' Communicative Self-Efficacy to Discuss Controversial Issues: Findings From a Randomized Study of the Word Generation Program. *Theory & Research in Social Education*, *44*(3), 316–343. https://doi.org/10.1080/00933104.2016.1203852

Lindblom, C. E. (1959). The Science of "Muddling Through." *Public Administration Review*, *19*(2), 79. https://doi.org/10.2307/973677

LoCasale-Crouch, J., Williford, A., Whittaker, J., DeCoster, J., & Alamos, P. (2018). Does Fidelity of Implementation Account for Changes in Teacher–Child Interactions in a Randomized Controlled Trial of Banking Time? *Journal of Research on Educational Effectiveness*, *11*(1), 35–55. https://doi.org/10.1080/19345747.2017.1329365

MacDonald, B., & Walker, R. (1976). *Chaging the Curriculum*. Open Books Publishing.

Maerten-Rivera, J., Ahn, S., Lanier, K., Diaz, J., & Lee, O. (2016). Effect of a Multiyear Intervention on Science Achievement of All Students including English Language Learners. *The Elementary School Journal*, *116*(4), 600–624. https://doi.org/10.1086/686250

Marsh, C. J., & Willis, G. (2007). *Curriculum: Alternative Approaches, Ongoing Issues* (4th ed.). Pearson.

Marti, M., Melvin, S., Noble, K. G., & Duch, H. (2018). Intervention fidelity of Getting Ready for School: Associations with classroom and teacher characteristics and preschooler's school readiness skills. *Early Childhood Research Quarterly*, *44*, 55–71. https://doi.org/10.1016/j.ecresq.2018.02.010

Marti, M., Merz, E. C., Repka, K. R., Landers, C., Noble, K. G., & Duch, H. (2018). Parent Involvement in the Getting Ready for School Intervention Is Associated With Changes in School Readiness Skills. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00759

Mashburn, A., Justice, L. M., McGinty, A., & Slocum, L. (2016). The Impacts of a Scalable Intervention on the Language and Literacy Development of Rural Pre-Kindergartners.

Applied Developmental Science, *20*(1), 61–78. https://doi.org/10.1080/10888691.2015.1051622

Matsumura, L. C., Correnti, R., Walsh, M., Bickel, D. D., & Zook-Howell, D. (2019). Online content-focused coaching to improve classroom discussion quality. *Technology, Pedagogy and Education*, *28*(2), 191–215. https://doi.org/10.1080/1475939X.2019.1577748

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward Measuring Instructional Interactions "At-Scale." *Educational Assessment*, *13*(4), 267–300. https://doi.org/10.1080/10627190802602541

McKeown, D., Wijekumar, K., Owens, J., Harris, K., Graham, S., Lei, P., & FitzPatrick, E. (2023). Professional development for evidence-based SRSD writing instruction: Elevating fourth grade outcomes. *Contemporary Educational Psychology*, *73*, 102152. https://doi.org/10.1016/j.cedpsych.2023.102152

McLaughlin, M. W. (1984). *Implementation Realities and Evaluation Design* (Program Report Nos. 84-B1). Institute for Research on Educational Finance and Governance.

McLaughlin, M. W. (1990). The Rand Change Agent Study Revisited: Macro Perspectives and Micro Realities. *Educational Researcher*, *19*(9), 11–16.

Meador, D., Nesbitt, K., & Farran, D. C. (2015). Experimental Evaluation of the Tools of the Mind Pre-K Curriculum. Fidelity of Implementation Technical Report. In *Peabody Research Institute*. Peabody Research Institute.

Mellom, P. J., Straubhaar, R., Balderas, C., Ariail, M., & Portes, P. R. (2018). "They come with nothing:" How professional development in a culturally responsive pedagogy shapes teacher attitudes towards Latino/a English language learners. *Teaching and Teacher Education*, *71*, 98–107. https://doi.org/10.1016/j.tate.2017.12.013

Monte-Sano, C., De La Paz, S., & Felton, M. (2014). Implementing a disciplinary-literacy curriculum for US history: Learning from expert middle school teachers in diverse classrooms. *Journal of Curriculum Studies*, *46*(4), 540–575. https://doi.org/10.1080/00220272.2014.904444

Moore, J. E., Bumbarger, B. K., & Cooper, B. R. (2013). Examining Adaptations of Evidence-Based Programs in Natural Contexts. *The Journal of Primary Prevention*, *34*(3), 147–161. https://doi.org/10.1007/s10935-013-0303-6

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity Criteria: Development, Measurement, and Validation. *American Journal of Evaluation*, 26.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, *12*(1), 53–74. https://doi.org/10.1080/13803610500392236

Murnane, R. J., & Willett, J. B. (2011). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Murphy, P. K., Greene, J. A., Firetto, C. M., M. V. Croninger, R., Duke, R. F., Li, M., & Lobczowski, N. G. (2022). Examining the effects of quality talk discussions on 4th- and 5th-grade students' high-level comprehension of text. *Contemporary Educational Psychology*, *71*, 102099. https://doi.org/10.1016/j.cedpsych.2022.102099

No Child Left Behind Act of 2001, H.R.1 107th Congress (2001-2002) (2002).

O'Donnell, C. L. (2008). Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K–12 Curriculum Intervention Research. *Review of Educational Research*, *78*(1), 33–84. https://doi.org/10.3102/0034654307313793

Owen, D. (2024). *Project Citizen Research Program Final Report*. Civic Education Research Lab. https://doi.org/10.13140/RG.2.2.21416.38407

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n160

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014a). Effectiveness of Cognitive Tutor Algebra I at Scale. *Educational Evaluation and Policy Analysis*, *36*(2), 127–144. https://doi.org/10.3102/0162373713507480

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. T. (2014b). *Addendum to Effectiveness of Cognitive Tutor Algebra I at Scale*. RAND Corporation. https://www.rand.org/pubs/working_papers/WR1050.html

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. Sage.

Peterman, K., Pan, Y., Robertson, J., & Lee, S. G. (2014). Self-Report and Academic Factors in Relation to High School Students' Success in an Innovative Biotechnology Program. *Journal of Technology Education*, *25*(2), 35–51.

Piasta, S. B., Logan, J. A. R., Thomas, L. J. G., Zettler-Greeley, C. M., Bailet, L. L., & Lewis, K. (2021). Implementation of a small-group emergent literacy intervention by preschool teachers and community aides. *Early Childhood Research Quarterly*, *54*, 31–43. https://doi.org/10.1016/j.ecresq.2020.08.002

Piasta, S. B., Logan, J. A. R., Zettler-Greeley, C. M., Bailet, L. L., Lewis, K., & Thomas, L. J. G. (2023). Small-Group, Emergent Literacy Intervention Under Two Implementation Models: Intent-to-Treat and Dosage Effects for Preschoolers at Risk for Reading Difficulties. *Journal of Learning Disabilities*, *56*(3), 225–240. https://doi.org/10.1177/00222194221079355

Piasta, S., Justice, L., McGinty, A., Mashburn, A., & Slocum, L. (2015). A Comprehensive Examination of Preschool Teachers' Implementation Fidelity When Using a Supplemental Language and Literacy Curriculum. *Child & Youth Care Forum*, *44*(5), 731–755. https://doi.org/10.1007/s10566-015-9305-2

Pollard-Durodola, S. D., Gonzalez, J. E., Saenz, L., Resendez, N., Kwok, O., Zhu, L., & Davis, H. (2018). The Effects of Content-Enriched Shared Book Reading Versus Vocabulary-Only Discussions on the Vocabulary Outcomes of Preschool Dual Language Learners. *Early Education and Development*, *29*(2), 245–265. https://doi.org/10.1080/10409289.2017.1393738

Portes, P. R., González Canché, M., Boada, D., & Whatley, M. E. (2018). Early Evaluation Findings From the Instructional Conversation Study: Culturally Responsive Teaching

Outcomes for Diverse Learners in Elementary School. *American Educational Research Journal*, *55*(3), 488–531. https://doi.org/10.3102/0002831217741089

Price, R. H., Friedland, D. S., Choi, J. N., & Caplan, R. D. (1998). Job-Loss and Work Transitions in a Time of Global Economic Change. In X. Arriaga & S. Oskamp (Eds.), *Addressing Community Problems* (pp. 195–222). SAGE.

Proctor, C. P., Silverman, R. D., Harring, J. R., Jones, R. L., & Hartranft, A. M. (2020). Teaching Bilingual Learners: Effects of a Language-Based Reading Intervention on Academic Language and Reading Comprehension in Grades 4 and 5. *Reading Research Quarterly*, *55*(1), 95–122. https://doi.org/10.1002/rrq.258

Proctor, C. P., Silverman, R. D., & Jones, R. L. (2021). Centering Language and Student Voice in Multilingual Literacy Instruction. *The Reading Teacher*, *75*(3), 255–267. https://doi.org/10.1002/trtr.2051

Quint, J. (2016a). *Impacts and Implementation of the i3-Funded Scale-Up of Success for All – Primary Student Sample Data* (Version 1) [Dataset]. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36387.v1

Quint, J. (2016b). *Impacts and Implementation of the i3-Funded Scale-Up of Success for All – Principal Survey Data* (Version 1) [Dataset]. Inter-university Consortium for Political and Social Research [distributor].

Quint, J. (2016c). *Impacts and Implementation of the i3-Funded Scale-Up of Success for All – School Achievement Snapshot Data* (Version 1) [Dataset]. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36387.v1

Quint, J. (2016d). *Impacts and Implementation of the i3-Funded Scale-Up of Success for All – Teacher Log Data* (Version 1) [Dataset]. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36387.v1

Quint, J. (2016e). *Impacts and Implementation of the i3-Funded Scale-Up of Success for All – Teacher Survey Data* (Version 1) [Dataset]. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36387.v1

Quint, J. (2016f). *Impacts and Implementation of the i3-Funded Scale-Up of Success for All – User Guide* (No. ICPSR 36387). MDRC, Inter-university Consortium for Political and Social Research. 10.3886/ICPSR36387.v1

Quint, J., Zhu, P., Balu, R., Rappaport, S., & DeLaurentis, M. (2015). *Scaling Up the Success for All Model of School Reform: Final Report from the Investing in Innovation (i3) Evaluation*. MDRC.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning About and From a Distribution of Program Impacts Using Multisite Trials. *American Journal of Evaluation*, *36*(4), 475–499. https://doi.org/10.1177/1098214015600515

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical-Linear Models: Applications and Data Analysis Methods* (2nd edition). Sage Publications.

Riel, J., & Lawless, K. A. (2021). Enhancing Student Affect From Multi-Classroom Simulation Games via Teacher Professional Development: Supporting Game Implementation With the ROPD Model. *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, *13*(1), 34–54. https://doi.org/10.4018/IJGCMS.20210101.oa3

Riel, J., Lawless, K. A., & Oren, J. B. (2022). Comparisons of Synchronous and Asynchronous Discussions in an Online Roleplaying Simulation to Teach Middle School Written

Argumentation Skills. *Online Learning*, *26*(4), Article 4. https://doi.org/10.24059/olj.v26i4.3468

Riel, J., Lawless, K., & Brown, S. (2016). Listening to the Teachers: Using Weekly Online Teacher Logs for ROPD to Identify Teachers' Persistent Challenges When Implementing a Blended Learning Curriculum. *Journal of Online Learning Research*, *2*(2), 169–200.

Rimm-Kaufman, S. E., Larsen, R. A. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., Merritt, E. G., Abry, T., & DeCoster, J. (2014). Efficacy of the Responsive Classroom Approach: Results From a 3-Year, Longitudinal Randomized Controlled Trial. *American Educational Research Journal*, *51*(3), 567–603. https://doi.org/10.3102/0002831214523821

Rimm-Kaufman, S. E., Merritt, E. G., Lapan, C., DeCoster, J., Hunt, A., & Bowers, N. (2021). Can service-learning boost science achievement, civic engagement, and social skills? A randomized controlled trial of Connect Science. *Journal of Applied Developmental Psychology*, *74*, 101236. https://doi.org/10.1016/j.appdev.2020.101236

Roberts, G., Vaughn, S., Wanzek, J., Furman, G., Martinez, L., & Sargent, K. (2023). Promoting adolescents' comprehension of text: A randomized control trial of its effectiveness. *Journal of Educational Psychology*, *115*(5), 665–682. https://doi.org/10.1037/edu0000794

Rogers, E. M. (1983). *Diffusion of innovations* (3rd ed). Free Press ; Collier Macmillan.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., & Cheung, C.-N. (2020). A randomized controlled trial of interleaved mathematics practice. *Journal of Educational Psychology*, *112*(1), 40–52. https://doi.org/10.1037/edu0000367

Roseman, J. E., & Herrmann-Abell, C. F. (2016, April). *Integrating NGSS Core Ideas and Practices: Supporting and Studying Teachers' Implementation*. Poster presented at the American Educational Research Association meeting, Washington, DC.

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2019). *Evaluation: A Systematic Approach* (8th ed.). SAGE.

Ruiz-Primo, M. A. (2006). *A Multi-Method and Multi-Source Approach for Studying Fidelity of Implementation* (No. CSE Report 677; p. 51). National Center for Research on Evaluation, Standards, and Student Testing (CRESST) Center for the Study of Evaluation (CSE).

Rutstein, D., Alozie, N., Fujii, R., & Fried, R. (2021, June). Addressing Challenges When Designing NGSS Aligned 3-Dimensional Assessments for Young Learners. *Proceedings of the 15th International Conference of the Learning Sciences*. International Society of the Learning Sciences. https://repository.isls.org//handle/1/7432

Sanetti, L. M. H., Gritter, K. L., & Dobey, L. M. (2011). Treatment integrity of interventions with children in the school psychology literature from 1995 to 2008. *School Psychology Review*, *40*, 72–84.

Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2016). Professional development in early mathematics: Effects of an intervention based on learning trajectories on teachers' practices. *Nordic Studies in Mathematics Education*, *21*(4), 29–55.

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*(3), 489–502. https://doi.org/10.1016/j.ecresq.2011.12.002

Scammacca, N., Swanson, E., Vaughn, S., & Roberts, G. (2020). Cost-Effectiveness of a Grade 8 Intensive Reading and Content Learning Intervention. *School Psychology Review*, *49*(4), 374–385. https://doi.org/10.1080/2372966X.2020.1760691

Scanlon, J. W., Horst, P., Nay, J. N., Schmidt, R. E., & Waller, J. D. (1977). Evaluability assessment: Avoiding Type III and Type IV errors. In G. R. Gilbert & P. J. Conklin (Eds.), *Evaluation management: A sourcebook of readings*. United States Civil Service Commission - Bureau of Training.

Scheirer, M. A., & Rezmovic, E. L. (1983). Measuring the Degree of Program Implementation: A Methodological Review. *Evaluation Review*, *7*(5), 599–633. https://doi.org/10.1177/0193841X8300700502

Schoen, R. C., Lewis, C. C., Rhoads, C., Lai, K., & Riddell, C. M. (2024). Impact of Lesson Study and Fractions Resources on Instruction and Student Learning. *The Journal of Experimental Education*, *92*(2), 225–246. https://doi.org/10.1080/00220973.2023.2183374

Schöpfel, J., & Farace, D. J. (2017). Grey Literature. In J. D. McDonald & M. Levine-Clark (Eds.), *Encyclopedia of Library and Information Science* (4th Edition, pp. 1746–1756). CRC Press. https://doi.org/10.1081/E-ELIS4

Scriven, M. (1994). The Fine Line between Evaluation and Explanation. *Evaluation Practice*, *15*(1), 75–77.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Sibuma, B., Wunnava, S., John, M.-S., Anggoro, F., & Dubosarsky, M. (2018). The impact of an integrated Pre-K STEM curriculum on teachers' engineering content knowledge, self-efficacy, and teaching practices. *2018 IEEE Integrated STEM Education Conference (ISEC)*, 224–227. https://doi.org/10.1109/ISECon.2018.8340489

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, *69*(1), 99–118.

Slavin, R. E. (2002). Evidence-Based Education Policies: Transforming Educational Practice and Research. *Educational Researcher*, *31*(7), 15–21. https://doi.org/10.3102/0013189X031007015

Slavin, R. E., & Madden, N. A. (2001). *One million children: Success for All* (pp. xii, 339). Corwin Press.

Slavin, R. E., & Madden, N. A. (2012). *Success for All: Summary of Research on Achievement Outcomes* (p. 77). Success for All Foundation.

Slavin, R. E., Madden, N. A., Chambers, B., & Haxby, B. (2009). *2 million children: Success for all, 2nd ed* (2nd edition, pp. xii, 392). Corwin Press.

Snyder, J., Bolin, F., & Zumwalt, K. (1992). Curriculum Implementation. In P. W. Jackson (Ed.), *Handbook of Research on Curriculum* (pp. 402–435). Macmillan.

Solomon, T., Dupuis, A., O'Hara, A., Hockenberry, M.-N., Lam, J., Goco, G., Ferguson, B., & Tannock, R. (2019). A cluster-randomized controlled trial of the effectiveness of the JUMP Math program of math instruction for improving elementary math achievement. *PLoS ONE*, *14*(10), e0223049. https://doi.org/10.1371/journal.pone.0223049

Spencer, T. D., Moran, M., Thompson, M. S., Petersen, D. B., & Restrepo, M. A. (2020). Early Efficacy of Multitiered Dual-Language Instruction: Promoting Preschoolers' Spanish and English Oral Language. *AERA Open*, *6*(1), 2332858419897886. https://doi.org/10.1177/2332858419897886

Starkey, P., Klein, A., Clarke, B., Baker, S., & Thomas, J. (2022). Effects of early mathematics intervention for low-SES pre-kindergarten and kindergarten students: A replication study. *Educational Research and Evaluation*, *27*(1–2), 61–82. https://doi.org/10.1080/13803611.2021.2022316

Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., Fuchs, L. S., & Compton, D. L. (2008). Scaling Up an Early Reading Program: Relationships Among Teacher Support, Fidelity of Implementation, and Student Performance Across Different Sites and Years. *Educational Evaluation and Policy Analysis*, *30*(4), 368–388. https://doi.org/10.3102/0162373708322738

Stevens, E. A., Murray, C. S., Scammacca, N., Haager, D., & Vaughn, S. (2022). Middle school matters: Examining the effects of a schoolwide professional development model to improve reading comprehension. *Reading and Writing*, *35*(8), 1839–1864. https://doi.org/10.1007/s11145-022-10271-9

Stieff, M. (2011). Improving representational competence using molecular simulations embedded in inquiry activities. *Journal of Research in Science Teaching*, *48*(10), 1137–1158. https://doi.org/10.1002/tea.20438

Stieff, M. (2019). Improving Learning Outcomes in Secondary Chemistry with Visualization-Supported Inquiry Activities. *Journal of Chemical Education*, *96*(7), 1300–1307. https://doi.org/10.1021/acs.jchemed.9b00205

Stylianou, D. A., Stroud, R., Cassidy, M., Knuth, E., Stephens, A., Gardiner, A., & Demers, L. (2019). Putting early algebra in the hands of elementary school teachers: Examining fidelity of implementation and its relation to student performance. *Journal for the Study* *of Education and Development*, *42*(3), 523–569. https://doi.org/10.1080/02103702.2019.1604021

Sullivan, K., Bell, N., Jones, D. H., Caverly, S., & Vaden-Kiernan, M. (2016). Implementation Work at Scale: An Examination of the Fidelity of Implementation Study of the Scale-Up Effectiveness Trial of Open Court Reading. *Society for Research on Educational Effectiveness*. Annual Meeting of the Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED567238

Swanson, E., Stewart, A. A., Stevens, E. A., Scammacca, N. K., Capin, P., Bhat, B. H., Roberts, G., & Vaughn, S. (2024). The Efficacy of Two Models of Professional Development Mediated by Fidelity on Fourth Grade Student Reading Outcomes. *Journal of Research on Educational Effectiveness*, *17*(2), 288–317. https://doi.org/10.1080/19345747.2023.2181897

Swanson, E., Vaughn, S., Fall, A.-M., Stevens, E. A., Stewart, A. A., Capin, P., & Roberts, G. (2021). The Differential Efficacy of a Professional Development Model on Reading Outcomes for Students With and Without Disabilities. *Exceptional Children*, *87*(4), 497–516. https://doi.org/10.1177/00144029211007149

Tarar, J. M., Meisinger, E. B., & Dickens, R. H. (2015). Test Review: Test of Word Reading Efficiency–Second Edition (TOWRE-2) by Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. *Canadian Journal of School Psychology*, *30*(4), 320–326. https://doi.org/10.1177/0829573515594334

Teague, G. B., Drake, R. E., & Ackerson, T. H. (1995). Evaluating Use of Continuous Treatment Teams for Persons With Mental Illness and Substance Abuse. *Psychiatric Services*, *46*(7), 689–695.

Turner, J. L. (2014). The Non-Parametric Spearman's Rho and Kendall's Tau Statistics. In *Using Statistics in Small-Scale Language Education Research*. Routledge.

Unlu, F., Bozzi, L., Layzer, C., Smith, A., Price, C., & Hurtig, R. (2016). Linking Implementation Fidelity to Impacts in an RCT. In *Treatment Fidelity in Studies of Educational Intervention*. Routledge.

U.S. Department of Education. (2024, February 2). *RE: FOIA Request No. 24-00781-F* [Personal communication].

Vadasy, P. F., Sanders, E. A., & Logan Herrera, B. (2015). Efficacy of Rich Vocabulary Instruction in Fourth- and Fifth-Grade Classrooms. *Journal of Research on Educational Effectiveness*, *8*(3), 325–365. https://doi.org/10.1080/19345747.2014.933495

Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., Ruiz de Castilla, V., Sullivan, K., & Rodriguez, D. (2015). Findings from a Multi-Year Scale-up Effectiveness Trial of Everyday Mathematics. *Society for Research on Educational Effectiveness*. Annual Meetinf of the Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED567630

Vaden-Kiernan, M., Borman, G., Caverly, S., Bell, N., Sullivan, K., Ruiz de Castilla, V., Fleming, G., Rodriguez, D., Henry, C., Long, T., & Hughes Jones, D. (2018). Findings From a Multiyear Scale-Up Effectiveness Trial of Open Court Reading. *Journal of Research on Educational Effectiveness*, *11*(1), 109–132. https://doi.org/10.1080/19345747.2017.1342886

Vaughn, S., Klingner, J. K., Swanson, E. A., Boardman, A. G., Roberts, G., Mohammed, S. S., & Stillman-Spisak, S. J. (2011). Efficacy of Collaborative Strategic Reading With Middle

School Students. *American Educational Research Journal*, *48*(4), 938–964. https://doi.org/10.3102/0002831211410305

Vaughn, S., Roberts, G., Klingner, J. K., Swanson, E. A., Boardman, A., Stillman-Spisak, S. J., Mohammed, S. S., & Leroux, A. J. (2013). Collaborative Strategic Reading: Findings From Experienced Implementers. *Journal of Research on Educational Effectiveness*, *6*(2), 137–163. https://doi.org/10.1080/19345747.2012.741661

Vaughn, S., Swanson, E., Fall, A.-M., Roberts, G., Capin, P., Stevens, E. A., & Stewart, A. A. (2022). The efficacy of comprehension and vocabulary focused professional development on English learners' literacy. *Journal of Educational Psychology*, *114*(2), 257–272. https://doi.org/10.1037/edu0000684

Vitale, M. R., & Romance, N. R. (2013a). Accelerating Vocabulary Development and Reading Comprehension in Grades 3-4-5 through an Inductive Vocabulary Model. *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness Conference. https://eric.ed.gov/?id=ED563063

Vitale, M. R., & Romance, N. R. (2013b). Inductive Use of Semantic Word-Families to Accelerate Vocabulary Development and Reading Comprehension in Grades 3-4-5. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35).

Wakabayashi, T., Andrade-Adaniya, F., Schweinhart, L. J., Xiang, Z., Marshall, B. A., & Markley, C. A. (2020). The impact of a supplementary preschool mathematics curriculum on children's early mathematics learning. *Early Childhood Research Quarterly*, *53*, 329–342. https://doi.org/10.1016/j.ecresq.2020.04.002

Wasik, B. A., & Hindman, A. H. (2020). Increasing preschoolers' vocabulary development through a streamlined teacher professional development intervention. *Early Childhood Research Quarterly*, *50*, 101–113. https://doi.org/10.1016/j.ecresq.2018.11.001

Wasik, B. A., & Hindman, A. H. (2023). Story Talk: Using Strategies from an Evidence-Based Program to Improve Young Children's Vocabulary. *The Reading Teacher*, *76*(4), 429–438. https://doi.org/10.1002/trtr.2174

Wayne, A. J., Song, M., Bishop, A., Graczewski, C., Kitmitto, S., & Lally, H. (2023). *Evaluation of MyTeachingPartner-Secondary Delivered Using Local Coaches during the COVID-19 Pandemic: Evidence from a Randomized Experiment*. American Institutes for Research.

Weiland, C., Eidelman, H., & Yoshikawa, H. (2011). A Regression Discontinuity Analysis of the Impact of "Building Blocks" in an Urban Public Prekindergarten Program and Associations between Fidelity-to-Curriculum and Child Outcomes. *Society for Research on Educational Effectiveness*. Society for Research on Educational Effectiveness. https://eric.ed.gov/?id=ED528500

Weiland, C., & Yoshikawa, H. (2013). Impacts of a Prekindergarten Program on Children's Mathematics, Language, Literacy, Executive Function, and Emotional Skills. *Child Development*, *84*(6), 2112–2130. https://doi.org/10.1111/cdev.12099

Weiss, C. H. (1995). Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families. In J. P. Connell, A. C. Kubisch, L. B. Schorr, & C. H. Weiss (Eds.), *New Approaches to Evaluating Community Initiatives* (Vol. 1). The Aspen Institute.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A Conceptual Framework for Studying the Sources of Variation in Program Effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808. https://doi.org/10.1002/pam.21760

Wendling, B. J., Schrank, F. A., & Schmitt, A. J. (2007). *Educational Interventions Related to the Woodcock-Johnson III Tests of Achievement* (No. 8; Assessment Service Bulletin). Riverside Publishing.

What Works Clearinghouse. (2011). *What Works Clearinghouse Procedures and Standards Handbook* (No. Version 2.1). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

What Works Clearinghouse. (2022). *What Works Clearinghouse Procedures and Standards Handbook* (No. Version 5.0). U.S. Department of Education, Institute of Education Sciences.

Whitehurst, G. J. (Russ). (2018). The Institute of Education Sciences: A Model for Federal Research Offices. *The ANNALS of the American Academy of Political and Social Science*, *678*(1), 124–133. https://doi.org/10.1177/0002716218768243

Whittaker, J. V., Kinzie, M. B., Vitiello, V., DeCoster, J., Mulcahy, C., & Barton, E. A. (2020). Impacts of an Early Childhood Mathematics and Science Intervention on Teaching Practices and Child Outcomes. *Journal of Research on Educational Effectiveness*, *13*(2), 177–212. https://doi.org/10.1080/19345747.2019.1710884

Whittaker, J. V., Kinzie, M. B., Williford, A., & DeCoster, J. (2016). Effects of MyTeachingPartner–Math/Science on Teacher–Child Interactions in Prekindergarten Classrooms. *Early Education and Development*, *27*(1), 110–127. https://doi.org/10.1080/10409289.2015.1047711

Wills, M. C., & Wolf, J. M. (2021). Woodcock-Johnson Cognitive and Achievement Batteries. In F. R. Volkmar (Ed.), *Encyclopedia of Autism Spectrum Disorders* (pp. 5206–5213). Springer International Publishing. https://doi.org/10.1007/978-3-319-91280-6_760

W.K. Kellogg Foundation. (2004). *W.K. Kellogg Foundation Evaluation Handbook*. W.K. Kellogg Foundation.

Wolery, M. (2011). Intervention Research: The Importance of Fidelity Measurement. *Topics in Early Childhood Special Education*, *31*(3), 155–157. https://doi.org/10.1177/0271121411408621

Zucker, T. A., Cabell, S. Q., Petscher, Y., Mui, H., Landry, S. H., & Tock, J. (2021). Teaching Together: Pilot study of a tiered language and literacy intervention with Head Start teachers and linguistically diverse families. *Early Childhood Research Quarterly*, *54*, 136–152. https://doi.org/10.1016/j.ecresq.2020.09.001

Zucker, T. A., Carlo, M. S., Landry, S. H., Masood-Saleem, S. S., Williams, J. M., & Bhavsar, V. (2019). Iterative Design and Pilot Testing of the Developing Talkers Tiered Academic Language Curriculum for Pre-Kindergarten and Kindergarten. *Journal of Research on Educational Effectiveness*, *12*(2), 274–306. https://doi.org/10.1080/19345747.2018.1519623

Zucker, T. A., Jacbos, E., & Cabell, S. Q. (2021). Exploring Barriers to Early Childhood Teachers' Implementation of a Supplemental Academic Language Curriculum. *Early Education and Development*, *32*(8), 1194–1219. https://doi.org/10.1080/10409289.2020.1839288