



Making Every Study Count: Learning From Replication Failure to Improve Intervention Research

Citation

Kim, J. S. (2019). Making Every Study Count: Learning from Replication Failure to Improve Intervention Research. *Educational Researcher*, 48(9): 599-607.

Published version

<https://doi.org/10.3102/0013189x19891428>

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37378514>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

Making Every Study Count: Learning from Replication Failure to Improve Intervention Research

James S. Kim

Harvard University

Kim, J. S. (2019). Making Every Study Count: Learning from Replication Failure to Improve Intervention Research. *Educational Researcher*, 48(9): 599-607.

[\[DOI\]](#)

Abstract

Why, when so many educational interventions demonstrate positive impact in tightly controlled efficacy trials, are null results common in follow-up effectiveness trials? Using case studies from literacy, this article suggests that replication failure can surface hidden moderators—contextual differences between an efficacy and effectiveness trial—and generate new hypotheses and questions to guide future research. First, replication failure can reveal systemic barriers to program implementation. Second, it can highlight for whom and in what contexts a program theory of change works best. Third, it suggests that a fidelity first and adaptation second model of program implementation can enhance the effectiveness of evidence-based interventions and improve student outcomes. Ultimately, researchers can make every study count by learning from both replication success and failure to improve the rigor, relevance, and reproducibility of intervention research.

In recent years, a growing number of scholars in *Educational Researcher* and other outlets have highlighted the failure of researchers to replicate the causal effects of educational interventions in real world contexts (Bryk, 2015; Gilbert, King, Pettigrew, & Wilson, 2016; Ginsburg & Smith, 2016; Lareau, 2009; Lewis, 2015). Strictly speaking, replication refers to the “ability of a researcher to duplicate the results of a prior study if the same procedures are followed but new data are collected” (Bollen et al., 2015, p. 5). Direct replications that duplicate the methods of an original study are both rare and difficult to implement in the education sciences because the district, school, and classroom context for experimentation changes over time (Makel & Plucker, 2014; Nosek et al., 2015; Van Bavel, Mende-Siedlecki, Brady & Reinero, 2015). As a result, the methods and participants in a replication study are inevitably different from an original study in applied fields like education. In many ways, however, replication failure is a normal feature of science, and null results can contain useful information for improving the conduct of experiments.

Although there are numerous books, articles, and reports on conducting rigorous experimental and quasi-experimental studies in education (Angrist & Pischke, 2009; Bloom, 2005; Campbell & Stanley, 1963; Murnane & Willett, 2011; Schochet, Puma, & Deke, 2014), few texts contribute to an understanding of factors related to a successful or failed replication. I contend that one of the key factors to replication success or failure revolves around the role of intervention contexts. In fact, replication failure affords a unique opportunity for identifying contextual factors that influence the conduct of an experimental study.

My goal in this article is to focus on contextual factors that potentially interact with the fidelity of program implementation or the intervention program theory of change, to the success or detriment of student outcomes. Consequently, I address two specific and related questions. First, why is replication failure so common when a previously validated educational intervention is replicated under typical practice conditions? Second, what lessons can be learned from replication failure?

My approach to studying the role of context in replication failure and success is to focus on examples from literacy. Experimental (and quasi-experimental) research designs have played a particularly prominent role in evaluating the impact of several large-scale literacy interventions (Borman et al., 2007; Gamse et al., 2009; May et al., 2013; Quint et al., 2015) and research-based interventions supported by reform networks such as Reading for Understanding (Douglas & Albro, 2014). As a result, the field of literacy provides several programs of intervention research that have progressed from the development of an innovation, to a tightly controlled efficacy trial, and to a real-world effectiveness trial.

In this article, I also focus on the challenge of replicating results from an original efficacy trial in a later effectiveness trial. Essentially, the challenge for replicators is to maintain strong internal validity while enhancing the external validity of an experiment. An **efficacy trial** (Shavelson & Towne, 2002; Slavin, 2008) is a randomized controlled trial that tests a novel intervention under favorable implementation conditions and aims to enhance internal validity. As a result, researchers exercise greater control in selecting program sites and participants, training highly-skilled interventionists, and monitoring the quality of program implementation.

In an **effectiveness trial** (Flay, 1986; McDonald, Keesler, Kauffman, & Schneider, 2006; Starfield, 1977), researchers aim to follow-up on an efficacy trial by replicating the experimental design and program activities in typical practice conditions that reflect “the everyday practice occurring in classrooms, schools, and districts” (U. S. Department of Education, 2015, p. 45). In an effectiveness trial, researchers must shift the burden of implementation to front-line actors (e.g., superintendents, principals, practitioners) in a broader and more heterogeneous sample. As a result, researchers must delegate more control for delivering programs and practices to educators in a variety of district and school contexts and seek to understand the contextual factors that moderate program implementation and effectiveness (Francis, 2008).

In the following sections, I begin by describing the prevalence of null results in efficacy and effectiveness trials. Next, I use examples from effectiveness trials of previously validated literacy interventions to show how replication failure can surface hidden contextual moderators that affected program implementation and student outcomes. The discussion highlights lessons for improving the conduct of experiments and broader research implications.

Why is Replication Failure Common in Effectiveness Research?

Over the past 15 years, there has been a renaissance of experimentation in the education sciences. Before 2002, randomized controlled trials (RCTs) were a rare design in education, and numerous educational theorists raised philosophical and practical arguments against RCT designs (Mosteller & Boruch, 2002). Boruch, de Moya, and Snyder (2002) noted that the fiscal year 2000 budget for the U.S. Department of Education included only one randomized controlled trial, representing less than 1% of program evaluations and studies (p.

58). Today, the quantity and quality of RCT designs testing a range of interventions has increased dramatically in the education sciences. The clear majority of these RCTs are efficacy trials in which trained interventionists implement program activities and the chief study aim is to protect a study from threats to internal validity. For example, a recent review found that 705 RCTs meet What Works Clearinghouse (WWC) Standards without reservations—that is, studies with strong internal validity to infer cause-effect relations—and 295 (42%) of these studies have yielded positive impacts on one or more targeted student outcomes (Maynard, 2017).

Positive impacts in an efficacy trial, however, are often difficult to replicate in a follow-up effectiveness trial. During the transition from efficacy to effectiveness research, there is a shift in researchers' roles as well as the contexts for experimentation—from a more controlled lab-like setting to a larger number of diverse educational contexts. For example, the Coalition of Evidence-Based Policy (2013) reviewed 90 effectiveness trials that were designed to evaluate a previously validated educational intervention. Only 11 of 90 (12%) of these effectiveness trials—a substantially smaller percentage than in efficacy research—replicated positive effects on student outcomes. Replication failure in follow-up effectiveness studies is also common in medicine (Ioannidis, 2005).

Fundamentally, replication failure underscores the challenge of faithfully implementing an educational intervention and enacting a program theory of change in noisy contexts. In designing an effectiveness trial, researchers must overcome the challenge of getting complex interventions to work in noisy contexts where educators must respond to simultaneous and

competing pressures (Bryk, 2015). In particular, larger urban school districts that participate in experimental research are already involved in myriad education reforms and research projects. For example, Stuart et al. (2017) found that districts participating in RCTs are larger and more urban and also enroll more economically disadvantaged students than non-participating schools from the broader inference population. Similarly, an effectiveness study of supplemental elementary reading programs revealed that participating districts had higher poverty levels, were larger, and more urban than districts nationally (James-Burdumy et al., 2009). In recruiting districts for an effectiveness trial of reading (and math curricula), Tipton and colleagues (2016) reported that the main reason why district leaders declined to participate was that “district resources were tied up with other changes, be they competing programs, new standards of requirements, or other administrative changes” (p. 217). These descriptive data suggest that the density of reform activity and experimentation is greatest in disadvantaged urban school districts and schools, leading to a proliferation of RCT studies of educational interventions designed to help low-performing schools and students improve (Murnane & Nelson, 2007).

What Lessons Can Be Learned From Replication Failure?

Throughout the conduct of an effectiveness trial, researchers must balance tensions between research design and operational reality in real-world district, school, and classroom contexts. Although researchers typically focus on establishing strong preconditions for setting up an experimental study design, Gueron (2002) has argued that thinking only about preconditions is misguided and short-sighted, and “the key to success lies in how you till the soil

and do the hard work of planting and harvesting. You have to understand the context and clear away potential land mines (p. 17).” The broader implication is that researchers may need (a) to attend to relationships with educators across a school system in order to identify and remove *systemic* and unanticipated barriers to implementing an evidence-based intervention, (b) to collaborate with educators to learn for whom and in what contexts an intervention works best, and (c) to partner with practitioners to first implement an intervention with fidelity and then with structured adaptations.

Replication Failure Can Reveal Systemic Barriers to Program Implementation

Some district-level moderators of program effectiveness are visible—but others are not apparent to researchers at the beginning of a study. For example, “external validity bias” can arise when districts are drawn by convenience sampling and program effects systematically vary along (a) measured characteristics of districts—for example, the number of prior years of district experience with a reform or policy, urbanicity, size, student performance (Bell et al., 2016; Gamse et al., 2008; Stuart et al., 2017) or (b) the specificity and intensity of the program theory (Weiss, et al., 2017). In addition, there are also “hidden district moderators”—contextual differences between an earlier efficacy trial and later effectiveness trial—that may affect program implementation and effectiveness. For example, every RCT design requires a project home, a project lead and principal investigators, and a partnership between practitioners and researchers. As a result, the degree of shared agency, the extent to which research addresses practical problems, and the duration of the partnership can vary across the district contexts. The contextual supports can include district leaders’ political and normative

support for research, the ongoing effort to align research aims with district priorities, and attention to building practitioners' capacity to implement and sustain evidence-based interventions.

By design, effectiveness trials initiated by researchers target external validity but often fail to attend to systemic barriers to program implementation. To illustrate this point, it is useful to consider the replication failure of Collaborative Strategic Reading (CSR), an evidence-based program that combines strategy instruction with peer-led discussions about text in middle school subjects (Klingner, Vaughn, & Schumm, 1998). One systemic barrier is mobilizing district capacity, money, and time to ensure that professional development is sufficiently robust to reach acceptable levels of implementation fidelity. Although an early efficacy trial of CSR (Vaughn et al., 2011) produced positive impacts on a standardized test of reading comprehension ($ES = .12$) and suggested that "18 hours of professional development [was] an amount of time that most districts would consider feasible" (p. 958), a follow-up effectiveness trial of CSR failed to replicate these positive effects. Led by a consortium of research organizations involving the Regional Educational Laboratory Southwest, the Instructional Research Group (IRG), ICF International, and the American Institutes of Research (AIR), Hitchcock and colleagues (2011) describes researchers' efforts to examine whether the effects of CSR on student outcomes generalized to a more diverse sample of districts. In contrast to the previous efficacy trial of CSR, teachers in the effectiveness trial received two days or professional development before the study and four coaching sessions and research staff were not present in classrooms to provide ongoing feedback, modeling, and implementation

support. Observational data revealed lower procedural fidelity than in the earlier efficacy trial. For example, only 21.6% of CSR teachers implemented all core components. Hitchcock and colleagues (2011) report statistically insignificant effects ($ES = .05$) on standardized reading comprehension outcomes, suggesting that future “research on CSR might focus on enhancing the fidelity of CSR implementation within classrooms” (Hitchcock et al., 2011, p. 49).

Classrooms, however, are nested in a broader school district. As such, systems-level supports such as supportive district leadership, connections to networks of other schools and teachers involved in the reform, and greater alignment between the district and research policies may foster stronger implementation of new classroom practices (Coburn, 2003, p. 6).

A second effectiveness trial of CSR suggests that stronger systemic supports for program implementation are needed to improve student outcomes. In the second effectiveness trial of CSR, the project home was a school district—the Denver Public Schools—and a researcher and practitioner were co-Principal Investigators. Political support for the project and the RCT in Denver was fostered by university and district leaders who “broadcasted” a vision statement to “(a) improve reading achievement for ELLs, students with LD, and struggling readers; and (b) use a whole-school strategy to do so” (Klingner, Boardman, & McMaster, 2013, p. 205). The final summative evaluation of CSR in the I3 funded randomized trial (Boardman et al., 2015) found a positive impact on students’ standardized reading comprehension scores that replicated results from earlier efficacy trials (Vaughn et al., 2011) and higher levels of procedural fidelity than in the previous multi-district effectiveness trial (Hitchcock et al., 2011). Although it is impossible to identify district factors that moderated program implementation in

a single-district RCT, some systemic factors may have helped rather than hindered the conduct of the second effectiveness trial. For example, it seems likely that efforts to connect researcher and practitioner goals, to embed CSR into existing district professional development structures and plans, and to align CSR with existing district reforms helped to foster shared agency over project goals and stronger implementation of CSR practices across a whole system (Slavin, 2013).

The effectiveness trials of CSR motivate several questions—and potential lessons—to inform the conduct of future intervention research. For example, to what extent are researchers systematically attending to relationships among educators across whole systems, particularly when an intervention requires buy-in among systems-leaders, principals, and teachers? To what extent do front-line actors—particularly school principals and classrooms teachers—perceive that the program was addressing a central problem of practice, aligned with existing programs, and likely to be sustained after the effectiveness trial? Are null effects more common in district contexts that are new to a program and where researchers are just starting to build systems-level support for an innovation and its implementation? The point of asking these questions is that there are likely to be many systemic barriers to program implementation that are not apparent to researchers at the beginning of a study. As a result, researchers may need to continually engage stakeholders across a school district to remove new barriers to program implementation.

Existing research suggests that both researchers' and practitioners' motivation and ability to overcome systems-level barriers to program implementation may affect project

outcomes. For example, in a multi-district effectiveness trial of KPALS, a classwide peer tutoring program that supplements core literacy instruction, researchers found evidence of larger effects in the original Tennessee district where KPALS was developed and smaller effects in the new Texas district. McMaster and Fuchs (2011) hypothesized that heterogeneity in effects across districts “may be attributed to the fact that Tennessee had the most K-PALS experience and resources” and “many teachers in the Texas site were not able to complete the full dose of K-PALS lessons due to competing district demands, such as testing that infringed upon instructional time” (McMaster & Fuchs, 2011). In other words, the failure to replicate the positive effects of previously validated literacy interventions in new contexts suggest that researchers may need to foster structured partnerships with district leaders to evaluate innovations and their implementations in noisy district contexts (Snow, 2015).

Replication Failure Can Highlight for Whom and in What Contexts a Program Theory of Change Works Best

With the rise of prominent schoolwide programs designed to improve outcomes for at-risk students (Berends, Bodilly, & Kirby, 2002; Comer et al., 1999; Cook et al., 2000; Quint et al., 2015), the education landscape is now replete with cluster randomized trials—experiments that assign whole schools to treatment and control conditions. In designing such trials, program developers and researchers must articulate a program theory of change that describes the causal levers that are likely to improve targeted student outcomes (Weiss, 1998). Although the program context frequently impacts the implementation and effectiveness of an educational intervention (Weiss, Bloom, & Brock, 2013), context is often overlooked in the development of

a program theory of change. For example, the W. K. Kellogg Foundation (2004) describes five components for developing a program theory of change (factors, activities, outputs, outcomes, impacts) but context is seen as one of many unknown “other factors” that influence program effectiveness.

In part, program effectiveness depends on the ability of school principals and teachers to translate a theory of change into typical practice conditions. The school organizational context can interact with a program theory of change because most, if not all, comprehensive schoolwide programs depend on the skill and will of school staff to coordinate program activities. Because staff readiness to implement complex school reforms matter, program developers and researchers frequently aim to measure staff readiness and sometimes to target programs in schools with higher levels of readiness. Proxies for staff readiness might include secret ballots designed to measure whether majority of school staff want to implement a reform (Borman et al., 2007; Cook et al., 1999), program-specific screener measures assessing the school organizational context (West, Morton, & Herlihy, 2016), or more general measures of school context such as teachers’ working conditions and the quality of teacher relationships (Kraft & Papay, 2014; McCormick et al., 2015).

A series of studies of Success for All (SFA) illustrates how replication failure can surface hidden moderators of the program theory of change and highlight for whom and in what contexts a program works best. For example, the SFA theory of change emphasizes early, intensive, high-quality, and ecologically pervasive intervention as a mechanism for preventing reading failure among low-income children in urban districts and schools (Ramey & Ramey,

1998). Early efficacy studies and follow-up effectiveness trials (Borman & Hewes, 2002; Borman et al., 2007) validated the program theory of change. The results of these trials suggested that Success for All could improve children's foundational early literacy skills if the broader school context was re-organized to facilitate the implementation of detailed lesson plans in reading, cooperative learning activities, cross-grade ability grouping, supplemental small group and 1-to-1 tutoring, and data-driven instructional decisions. To support school staff readiness to coordinate these instructional strategies, teachers receive

... 3 days of intensive training at the beginning of the first school year. Follow-up services over the first year of implementation consist of 16 days of on-site support provided by Success for All program staff as well as quarterly monitoring of student progress data. After the first year, approximately 15 days of additional training are provided each year. (Borman et al., 2007, p. 704)

SFA's intensive and sustained implementation plan is designed to facilitate the expansion of the program theory of change with integrity and to replicate success with new cohorts of schools.

In a multi-district effectiveness trial, SFA demonstrated positive effects on early literacy outcomes for children who participated in the intervention from Kindergarten to Grade 2 (Borman et al., 2007).

As part of the 21st Century Community Learning Center program, the SFA program theory was adapted for after school programs serving older students (Black et al., 2009). When SFA was implemented in after school learning centers with older students in grades 2 to 5, the program impacts on student outcomes did not replicate the results of the school year model.

As SFA was implemented in an entirely new context, the SFA implementation theory became more complex; there were several new challenges in delivering the program in afterschool settings. These challenges included the need to align afterschool with school year curricula and to enhance both students' and teachers' motivation to participate in the program at the end of the school day. Put simply, the replication failure suggested that the SFA program theory did not work in afterschool contexts and with older students. Given the null results of the effectiveness trial of SFA in afterschool programs, Black et al. (2015) note that the program developers "decided after this study to leave the field of afterschool instruction and instead focus on improving reading achievement through school day reading instruction" (p. 2).

The replication failure of SFA surfaced a potentially important hidden moderator—the time of day when the program is delivered—that was not obvious at the beginning of the study. In other words, the afterschool context was missing several enabling conditions—for example, strong principal and teacher support, cross grade grouping in reading, and schoolwide ownership for the program. These contextual supports fostered the coordination and implementation of SFA program activities during the regular school day, particularly in early grade literacy instruction. In other words, the replication failure of SFA suggests that comprehensive schoolwide programs may not work well in afterschool contexts that have a fundamentally different organizational structure from the regular school day. Researchers and program developers, then, must grapple with questions like: (a) To what extent can program developers successfully adapt programs for entirely new contexts? (b) How can effectiveness

trials be designed to understand whether evidence-based interventions are context-bound—that is, effective only in specific contexts and for specific subgroups of students?

Shedding light on these questions will depend on strategically replicating studies of the same intervention over time and in different contexts. For example, the Building Educated Leaders for Life (BELL) intervention is a 5-week, full-day summer program focused on improving students' academic and leadership skills and produced positive effects on reading when implemented with elementary grade students (Chaplin & Capizzano, 2006). An effectiveness trial involving middle school students in grades 6 to 8 failed to replicate these positive effects in reading (Somers et al., 2015). Although there are differences between the elementary and middle school evaluation of BELL, Somers and colleagues speculated that “BELL’s model may not have been engaging enough to keep middle school students engaged” with the program activities (Somers et al., 2015, p. 5).

The replication failure of SFA and BELL suggest that researchers and program developers cannot simply adapt program activities for new contexts and for new subgroups of students and expect to replicate positive impacts on student outcomes. Instead, researchers and practitioners may need to learn for whom and in what contexts an evidence-based intervention is most effective—and least effective. Lessons from replication success and failure might then inform the design of effectiveness trials in a more narrowly and precisely defined target population of students and school contexts.

Replication Failure Suggests the Value of a Fidelity First and Adaptation Second Approach to Program Implementation

A replication failure can highlight the need for teachers to first learn how to implement evidence-based practices with fidelity and then to make structured adaptations to help make programs work better for their students and their specific contexts. A hidden moderator of program effectiveness may be the extent to which teachers enact “structured adaptations” — that is, the extent to which teachers receive guidance on how to make acceptable program adaptations that enhance rather than undermine program implementation and effectiveness (Frank, Zhao, Penuel, Ellefson, & Porter, 2011; Quinn & Kim, 2017).

A series of experimental studies of READS (Reading Enhances Achievement During Summer), a low-cost and large-scale summer reading intervention for elementary grade students, suggest the importance of a fidelity first and adaptation second approach to program implementation. Two early efficacy trials of READS provided evidence of positive impacts on students’ reading comprehension outcomes (Kim, 2006; Kim & White, 2008), but these results were not replicated in follow-up effectiveness trials (Kim & Guryan, 2010; White, Kim, Kingston, & Foster, 2014). The replication failure of READS provided useful information for designing future studies. To test the impact of the program when fidelity of implementation was emphasized, the project team (Kim et al., 2016) conducted a year 1 effectiveness trial in 59 schools that revealed a positive main effect on a delayed measure of reading comprehension ($ES = .04$) and larger effects in high poverty schools ($ES = .05$) than moderate poverty schools ($ES = .01$). In the year 2 study, 27 high poverty schools were randomly assigned to a program implementation that emphasized (a) faithfully replicating procedures that worked in the year 1 effectiveness trial, or (b) to make structured teacher adaptations (Kim et al., 2017). In latter

condition, teachers participated in grade-level meetings throughout the school year, completed a series of online modules designed to deepen their understanding of the research-based principles underlying READS, and were given the flexibility to make adaptations aligned with those principles. The results indicated that students taught by teachers in the structured adaptations condition enjoyed a significant advantage in reading comprehension ($ES = .12$) relative to the control condition emphasizing fidelity of program implementation.

The effectiveness trials of READS suggest that practitioners must first learn what works on average and then learn how to make programs work better in their specific context. As noted by Easton (2012), the challenge is “understanding how to apply research findings in new situations” given enormous variability in “context, needs, and capacity across different learning environments” (p. 20). Thus, the important question for intervention researchers is this: can a customized approach to program adaptations help teachers use research-based knowledge and their local knowledge to make programs work better for their students? At minimum, the multi-year effectiveness trials of READS suggest the feasibility of testing a “scaffolded sequence of program implementation” (Quinn & Kim, 2017, p. 1194) in which teachers first learn how to implement programs with fidelity before making structured adaptations tailored for their local contexts. Consistent with this idea, McMaster and colleagues (2014) report the results of a second year study of KPALS in which teachers chose (a) to continue implementing KPALS “by-the-book” as prescribed by researchers or (b) to “customize PALS” by adapting noncore program activities to meet the needs of their particular students and contexts (McMaster et al., 2014). Students taught by “customized PALS” teachers outperformed “by-the-book KPALS”

students by .25 to .60 standard deviations across reading measures (Lemons et al., 2014). The multi-year effectiveness trials of READS and KPALS study suggests that an emphasis on fidelity in the first year program implementation and then structured teacher adaptations in a second year of program implementation may enhance program effectiveness.

Conclusion

There is growing evidence that null effects on targeted student outcomes are most common in follow-up effectiveness trials of previously validated interventions (Coalition for Evidence-Based Policy, 2013; Ioannidis, 2005). In many ways, however, the ubiquitous challenge of bridging the divide between a controlled efficacy trial and real-world effectiveness trial compels scholars to re-think the role of practitioners in experimental research. Viewing district administrators, school leaders, and classroom teachers as true collaborators rather than mere study participants might encourage researchers (a) to form structured partnerships with district leaders to overcome systemic barriers to implementing an evidence-based intervention with fidelity across a large numbers of schools and classrooms, (b) to collaborate with principals and teachers to understand for whom and in what contexts an educational intervention is most effective, and (c) to work with practitioners to first implement an intervention with fidelity and then with structured adaptations. In closing, this article elaborates on each of these lessons for improving the conduct of experiments and highlights broader research implications.

First, replication failure can surface hidden moderators that are not apparent to researchers and thus help generate a priori hypotheses to test using a variety of research methods. The chief lesson from the CSR effectiveness trial is not that replication success

depends on housing projects in school districts, recruiting more practitioners to serve as principal investigators of research studies, or cultivating long-term partnerships that address practical educational problems. Rather, replication failure can generate new hypotheses and questions about district-level moderating variables that could be explored more systematically in future research. Currently, for example, little is known about the relationship between the nature of the research-practice partnership (RPP) and project outcomes. Coburn and Penuel (2016) have suggested more “comparative studies that investigate how RPPs of different designs interact with their contexts to impact various outcomes of interest” (pp. 51-52). Comparative case studies of interventions that are tested in diverse district contexts might detail the level of commitment across a school system that is needed to implement an experimental design and an intervention with fidelity and quality. In addition to case study research, it is also clear that research syntheses are needed to improve the statistical power to detect district- and school-level moderators of program outcomes. For example, a recent meta-analysis of 100 replication studies in social psychology found that macro-level contextual factors like time, location, and culture affected replication success (Van Bavel et al. 2016). A similar design could be used to examine whether both visible and hidden district-level moderator variables—for example, the size, urbanicity, and student demographic characteristics as well as the nature and duration of the partnership between researchers and district leaders—affect the replicability of experimental findings in the education sciences (Stuart et al., 2017; Tipton et al., 2016).

Second, replication failure can also surface hidden moderators that can interact with a program theory of change and thus help to improve the match between an educational intervention and the broader school context. The replication failure of Success for All in afterschool contexts surfaced a hidden moderator—the time of day when program activities are delivered to children—that was not obvious to program developers and researchers. Yet, it also sharpened the program theory of change by highlighting for whom and in what contexts Success for All is most effective. For example, recent replication failures and successes of Success for All (Black et al., 2009; Hanselman & Borman, 2013; Quint et al., 2015) suggest that the schoolwide program may work best in urban district and school contexts, when core literacy activities are implemented longitudinally from Kindergarten to second-grade for students at-risk of later school failure, and when school organizational supports are in place to coordinate program activities across grades and among teachers. A precise answer to the question—what works, for whom, and under what conditions—provides useful guidance for policymakers and practitioners seeking to cost-effectively target scarce resources and maximize impacts on student outcomes.

Finally, replication failure highlights the need to strike the right balance between fidelity and adaptation and for stronger research designs that establish causal links between the quality of program implementation and student outcomes. Although fidelity and adaptation inevitably co-occur during any program implementation, teachers ultimately choose the extent to which they adhere to researcher-developed program procedures or adapt them, making it difficult for researchers to infer whether and to what extent the quality of implementation caused student

outcomes to improve. At minimum, future research designs might test whether multi-year interventions that enable teachers to adhere to the core program components and then adapt noncore components can improve program implementation and student outcomes. In real world contexts where teachers must implement and integrate evidence-based programs and practices into typical practice without special implementation supports from researchers, they must have a deep understanding of foundational principles to make “productive adaptations” (McLaughlin & Mitra, 2001) and the agility to adapt noncore program components for their local contexts.

More broadly, multi-year effectiveness trials might help researchers transcend and reconcile tensions between experimental and improvement science paradigms (Lewis, 2015) and the fidelity-adaptation debate (Dane & Schneider, 1998). That is, strategic replications that are designed to test the causal effects of a fidelity-focused program implementation first and then structured adaptations second may indicate that teachers and schools need more time to learn. In other words, they may need more time to learn how to faithfully implement evidence-based interventions that worked “on average” then and there and enact structured adaptations to make the intervention work better here and now.

A Strategic and Principled Approach to Replication

The increasing number of null results in effectiveness trials of previously validated education interventions has fueled the perception that the cost of doing RCTs outweighs the benefits (Berwick, 2008; Bryk, 2015; Lareau, 2008) and that science is in the midst of a replication crisis (Fiske, Schacter, & Taylor, 2016). This article concludes with a more optimistic

message: replication failure should be viewed as an unusually rich opportunity to learn how to improve intervention research and to explore new questions and hypotheses.

Ultimately, a strategic and principled approach to replication requires a long-term view of science. Focusing on a single topic or question over the long-term affords scientists opportunities to conduct replications of important findings and to determine whether novel findings are an anomaly or robust across different contexts and under different implementation conditions. To build usable knowledge, researchers, program developers, and practitioners will need to work together (a) to identify and remove systemic and unanticipated barriers to the conduct of experiments and program implementation in noisy district contexts, (b) to recognize that a program theory of change must reveal for whom and in what contexts evidence-based interventions work best, and (c) to design multi-year experiments that test whether a fidelity first and structured adaptations second approach to program implementation can enhance student outcomes. Ideally, replication failure should inspire the scientific community to make every study count by illuminating what works, for whom, and in what contexts.

NOTES

I gratefully acknowledge the thoughtful suggestions of three anonymous reviewers and the editors of this special edition, Rebecca Maynard and Carolyn Herrington. I also thank Robin Jacob, Heather Hill, Mary Burkhauser, Jill Fitzgerald, David Quinn, Catherine Armstrong, and Thomas Kelley-Kemple for helpful comments on earlier versions of this paper. The research reported in this article was made possible by an Investing in Innovation Fund (i3) grant from the U.S. Department of Education (PR/Award U396B100195); however, the contents of this article do not represent the policy of the U.S. Department of Education.

References

- Angrist, J. D., & Pischke, J. S. (2009). *Mostly harmless econometrics*. Princeton, NJ: Princeton University Press.
- Bell, S. H., Olson, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, 38(2), 318-335.
- Berends, M., Bodilly, S., & Kirby, S. N. (2002). *Facing the challenges of whole-school reform: New American Schools after a decade*. Santa Monica, CA: RAND.
- Berwick, D. M. (2008). The science of improvement. *JAMA*, 299(10), 1182-1184.
- Black, A. R., Somers, M.-A., Doolittle, F., Unterman, R., & Grossman, J. B. (2009). *The Evaluation of Enhanced Academic Instruction in After-School Programs: Final Report* (NCEE 2009-4077). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Bloom, H. (2005). *Learning more from social experiments*. New York, NY: Russell Sage Foundation.
- Boardman, A. G., Klinger, J. K., Buckley, P., Annamma, S., & Lasser, C. J. (2015). The efficacy of Collaborative Strategic Reading in middle school science and social studies classes. *Reading and Writing*, 28, 1257-1283.
- Bollen, K., Cacioppo, J. T., Kapan, R. M., Krosnick, J. A., & Olds, J. L. (2015). *Social, behavioral,*

economic sciences perspectives on robust and reliable science. Report of the subcommittee on replicability in science advisory committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.

Borman, G. D., & Hewes, G. M. (2002). The long-term effects and cost-effectiveness of Success for All. *Educational Evaluation and Policy Analysis*, 24(4), 243-266.

Borman, G. D., Slavin, R. E., Cheung, A. C. K., Chamberlain, A. M., Madden, N. A., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, 44, 701-731.

Boruch, R., de Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. M. Mosteller & R. Boruch (Eds.), *Evidence matters*, (pp. 50-79). Washington, DC: The Brookings Institution.

Bryk, A. S. (2015). Accelerating how we learn to improve. *Educational Researcher*, 44, 467-477.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, 33(3), 378-398.

Coalition for Evidence-Based Policy (2013). *Randomized controlled trials commissioned by the Institute of Education Sciences since 2002: How many found positive versus weak or no effects.* Retrieved from: <http://coalition4evidence.org/wp->

[content/uploads/2013/06/IES-Commissioned-RCTs-positive-vs-weak-or-null-findings-7-2013.pdf](#)

Coburn, C. E., & Penuel, W. R. (2016). Research-practice partnerships in education: Outcomes, dynamics, and open questions. *Educational Researcher*, 45(1), 48-54.

Comer, J., Haynes, N., Joyner, E. T., & Ben Avie, M. (1996). *Rallying the whole village: The Comer process for reforming education*. New York: Teachers College Press.

Cook, T. D., Habib, F. N., Phillips, M., Settersten, R. A., Shagle, S. C., & Degirmencioglu, S. M. (1999). Comer's School Development Program in Prince George's County, Maryland: A theory-based evaluation. *American Educational Research Journal*, 36, (3), 543-597.

Cook, T. D., Murphy, R. F., & Hunt, H. D. (2000). Comer's School Development Program in Chicago: A Theory-based Evaluation, *American Educational Research Journal*, 37, (2), 535-597.

Coyne, M. D., Little, M., Rawlinson, D., Simmons, D., Oi-man, K., Kim, M., Simmons, L., Hagan-Burke, S., & Civetelli, C. (2013). Replicating the impact of a supplemental beginning reading intervention: The role of instructional context. *Journal of Research on Educational Effectiveness*, 6(1), 1-23.

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.

Douglas, K. M., & Albrow, E. R. (2014). The progress and promise of the Reading for Understanding research initiative. *Educational Psychological Review*, 26, 341-355.

Easton, J. (2012). Commentary on the uses of research in policy and practice. *Social Policy Report, 26*, 19-20.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist, 48*(2), 90.

Fiske, S. T., Schacter, D. L., & Taylor, S. E. (2016). Introduction. *Annual Review of Psychology*, vol. 67.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine, 15*, 451-474.

Francis, D. F. (2008). Reading First Impact Study: What have we learned and where do we go from here? Presented at the IES Research Conference, Washington, DC.

Frank, K. A., Zhao, Y., Penuel, W. R., Ellefson, N., & Porter, S. (2011). Focus, fiddle, and friends: Experiences that transform knowledge for the implementation of innovations. *Sociology of Education, 84*(2), 138–156.

Fuchs, D., McMaster, K. L., Saenz, L., Kearns, D., Fuchs, L. S., Yen, L., . . . & Schatschneider, C. (2010, June). *Scaling up PALS: A comparison of top down versus bottom up approaches*. Poster presented at the IES Research Conference, Washington, DC.

Fryer, R. G. (2016). The production of human capital in developed countries: Evidence from 196 randomized field experiments. Cambridge, MA: National Bureau of Economic Research.

- Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (2009). Reading First Impact Study final report. NCEE 2009-4039. Washington, DC: National Center for Education Evaluation.
- Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016). More on "Estimating the reproducibility of psychological science," *Science*, 351(6277), 1037a-1037b.
- Ginsburg, A., & Smith, M. S. (2016). Do randomized controlled trials meet the "gold standard"? Washington, DC: *American Enterprise Institute*.
- Retrieved from: <https://www.aei.org/events/do-randomized-control-trials-in-education-meet-the-gold-standard/>
- Gawande, A. (2007). *Better*. New York NY: Picador.
- Gueron, J. M. (2002). The politics of random assignment: Implementing studies and affecting policy. In F. M. Mosteller & R. Boruch (Eds.), *Evidence matters*, (pp. 15-49). Washington, DC: The Brookings Institution.
- Hanselman, P., & Borman, G.D. (2013). The impacts of Success for All on reading achievement in grades 3-5: Does intervening during the later elementary grades produce the same benefits as intervening early? *Educational Evaluation and Policy Analysis*, 35, 237-251.
- Herman, J. L., Epstein, S., Leon, S., Dai, Y., La Torre Matrundola, D., Reber, S., & Choi, K. (2015). The implementation and effects of the literacy design collaborative (LDC): Early findings in sixth-grade advanced reading courses. CRESST/University of California, Los Angeles.
- Hitchcock, J., Dimino, J., Kurki, A., Wilkins, C., & Gersten, R. (2011). *The Impact of Collaborative Strategic Reading on the Reading Comprehension of Grade 5 Students in Linguistically Diverse Schools*. (NCEE 2011-4001). Washington, DC: National Center for

Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S.
Department of Education.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, 2(8),
e124.

Jacob, R. T., Armstrong, C., & Willard, J. A. (2016). Leveraging *volunteers: An experimental evaluation of a tutoring program for struggling readers*. *Journal of Research on Educational Effectiveness*, 9(S1), 67-92.

James-Burdumy, S., Mansfield, W., Deke, J., Carey, N., Lugo-Gil, J., Hershey, A., Douglas, A.,
Gersten, R., Newman-Gonchar, R., Dimino, J., and Faddis, B. (2009). *Effectiveness of Selected Supplemental Reading Comprehension Interventions: Impacts on a First Cohort of Fifth-Grade Students* (NCEE 2009-4032). Washington, DC: National Center for
Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S.
Department of Education.

Kim, J. S. (2006). "The Effects of a Voluntary Summer Reading Intervention on Reading
Achievement: Results from a Randomized Field Trial." *Educational Evaluation and Policy Analysis* 28(4): 335-355.

Kim, J. S., Burkhauser, M. B., Quinn, D. M., Guryan, J., Kingston, H. C., & Aleman, K. (2017).
"Effectiveness of Structured Teacher Adaptations to an Evidence-Based Summer Literacy
Program." *Reading Research Quarterly*, 52(4): 443-468.

Kim, J. S., & Guryan, J. (2010). "The Efficacy of a Voluntary Summer Book Reading Intervention

- for Low- Income Latino Children from Language Minority Families.” *Journal of Educational Psychology* 99(3):505-515.
- Kim, J. S., Guryan, J., White, T. G., Quinn, D. M., Capotosto, L., & Kingston, H. C. (2016). “Delayed Effects of a Low-Cost and Large-Scale Summer Reading Intervention on Elementary School Children’s Reading Comprehension.” *Journal of Research on Educational Effectiveness*, 9(S1): 1-22. DOI: 10.1080/19345747.2016.1164780
- Kim, J. S. & White, T. G. (2008). “Scaffolding Voluntary Summary Reading for Children in Grades 3 to 5: An Experimental Study.” *Scientific Studies of Reading* 12(1): 1-23.
- Klingner, J. K., Boardman, A. G., & McMaster, K. L. (2013). What does it take to scale up and sustain evidence-based practices? *Exceptional Children*, 79(2), 195-211.
- Klingner, J. K., Vaughn, S., & Schumm, J. S. (1998). Collaborative Strategic Reading during social studies in heterogeneous fourth-grade classrooms. *The elementary school journal*, 99(1), 3–22.
- Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, 36(4), 476-500.
- Lareau, A. (2009). Narrow questions, narrow answers: The limited value of randomized controlled trials for education research. In P. Walters, A. Lareau, & S. H. Ranis (Eds.). *Education Research on Trial: Policy Reform and the Call for Scientific Rigor*, (145-161). New York, NY: Routledge.
- Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational*

- Researcher*, 43(5), 242-252.
- Lewis, C. (2015). What is improvement science? Do we need it in education? *Educational Researcher*, 44(1), 54-61.
- Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, 43(6), 304-316.
- May, H., Gray, A., Gillespie, J. N., Sirinides, P., Sam, C., Goldsworthy, H., Armijo, M., & Tognatta, N. (2013). *Evaluation of the i3 scale-up of Reading Recovery*. University of Delaware: Consortium for Policy Research in Education.
- Maynard, R. A. (2017). Studies Examined by the WWC, by Quality of Evidence on Effectiveness and Whether Any Impact Estimates Relevant to the Topic Area Were "Favorable" and Statistically Significant. Source: <https://ies.ed.gov/ncee/wwc/ReviewedStudies#/>
- McCormick, M. P., Cappella, E., O'Connor, E. E., & McClowry, S. G., (2015). Context matters for social-emotional learning: Examining variation in program impact by dimensions of school climate. *American Journal of Community Psychology*, 56(1-2), 101-119.
- McDonald, S., Keesler, V., Kauffman, N., & Schneider, B. (2006). Scaling-up exemplary interventions. *Educational Researcher*, 35(3), 15-24.
- McLaughlin (1990). The Rand Change Agent Study Revisited: Macro Perspectives and Micro Realities. *Educational Researcher*, 19(9), 11-16.
- McLaughlin, M. W., & Mitra, D. (2001). Theory-based change and change-based theory: Going deeper, going broader. *Journal of Educational Change*, 2(4), 301-323.
- McMaster, K. L., & Fuchs, D. F. (2011). Scaling-up: From the laboratory to the field site to

multiple sites. Paper presented at the 2011 Annual Meeting of the Society for Research on Educational Effectiveness, Washington, DC.

McMaster, K.L., Jung, P.G., Brandes, D., Pinto, V., Fuchs, D., Kearns, D., ... Yen, L. (2014).

Customizing a research-based reading practice. *The Reading Teacher*, 68(3), 173–183.

doi:10.1002/trtr.1301

Mosteller, F., Boruch, R.F. (Eds.). (2002). *Evidence matters: Randomized trials in education research*. Washington, D.C.: Brookings Institution Press.

Murnane, R., & Nelson, R. (2007). Improving the performance of the education sector: The valuable, challenging, and limited role of random assignment evaluations. *Economics of Innovation and New Technology*, 16(5), 307-322.

Murnane, R. & Willett, J. B. (2011). *Methods matter*. New York, NY: Oxford Press.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S.J., . . . & Contestabile, M. (2015). Promoting an open research culture. *Science*, 348(6242), 1422-1425.

Quinn, D. M., & Kim, J. S. (2017). Scaffolding Fidelity and Flexibility in Educational Program Implementation: Experimental Evidence from a Literacy Intervention. *American Educational Research Journal*, 54(6): 1187-1120.

DOI: <https://doi.org/10.3102/0002831217717692>

Quint, J., Zhu, P., Balu, R., Rappaport, S., DeLaurentis, M., Alterman, E., Bottles, C., & Pramik, E. (2015). *Scaling up the Success for All Model of School Reform*. New York: MDRC.

Ramey, C. T., & Ramey, S. L. (1998). Early intervention and early experience. *American Psychologist*, 53, 109–120.

- Schochet, P. Z., Puma, M., & Deke, J. (2014). *Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods* (NCEE 2014–4017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from <http://ies.ed.gov/ncee/edlabs>.
- Shavelson, R. J., & Towne, L. (2002). *Scientific research in education*. Washington, DC: National Academy Press
- Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, 37(1), 5-14.
- Slavin, R. E. (2013). Lessons from innovators: Collaborative Strategic Reading. [Blog post]. Retrieved from http://www.huffingtonpost.com/robert-e-slavin/lessons-from-innovators-c_b_3988028.html
- Snow, C. E. (2015). 2014 Wallace Foundation Distinguished Lecture: Rigor and Realism: Doing Education Science in the Real World. *Educational Researcher*, 44(9), 460-466.
- Somers, M-A., Welbeck, R., Grossman, J. B., & Gooden, S. (2015). An analysis of the effects of an academic summer program for middle school students. New York, NY: MDRC.
- Starfield, B. (1977). Efficacy and effectiveness of primary medical care for children. In Harvard Health Project, Volume 2. Children's medical care needs and treatment: Report of the Harvard Child Health Project. Cambridge, MA: Ballinger.
- Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Saenz, 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. DOI: 10.3102/0162373708322738
- Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness*, 10(1), 168-206.
- Tipton, E., Fellers, L., Caverly, S., Vaden-Kiernan, M., Borman, G. D., Sullivan, S., & Ruiz de Castilla, V. (2016). Site selection in experiments: An assessment of site recruitment and generalizability in two scale-up studies. *Journal of Research on Educational Effectiveness*, 9(S1), 209-228.
- Van Bavel, J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2015). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, 113, 6454-6459.
- Vaughn, S., et al. (2011). Efficacy of Collaborative Strategic Reading with middle school students. *American Educational Research Journal*, 48, 938–964.
- Weiss, C. H. (1998). *Evaluation*. Upper Saddle, NJ: Prentice-Hall, Inc.
- Weiss, M. J., Bloom, H. S., & Brock, T. (2013). *A conceptual framework for studying the sources of variation in program effects*. New York: Manpower Development Research Company.
- Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. J. (2017):

How Much Do the Effects of Education and Training Programs Vary Across Sites?

Evidence From Past Multisite Randomized Trials, *Journal of Research on Educational Effectiveness*, DOI: 10.1080/19345747.2017.1300719

West, M. R., Morton, B. A., & Herlihy, C. (2016). Achievement Network's Investing in Innovation Expansion: Impacts on Educator Practice and Student Achievement. Harvard University, Center for Education Policy Research.

White, T. G., Kim, J. S., Kingston, H. C., & Foster, L. F. (2014). "Replicating the Effects of a Teacher-Scaffolded Voluntary Summer Reading Program: The Role of Poverty." *Reading Research Quarterly* 49(1): 5-30.

Yin, R. (2003). *Case study research*. Thousand Oaks, CA: Sage Publications.