ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-414

Using Implementation Fidelity to Aid in Interpreting Program Impacts: A Brief Review

Heather C. Hill
Harvard University

Anna Erickson
University of Michigan

Poor program implementation constitutes one explanation for null results in trials of educational interventions. For this reason, researchers often collect data about implementation fidelity when conducting such trials. In this article, we document whether and how researchers report and measure program fidelity in recent cluster-randomized trials. We then create two measures—one describing the level of fidelity reported by authors and another describing whether the study reports null results—and examine the correspondence between the two. We also explore whether fidelity is influenced by study size, type of fidelity measured and reported, and features of the intervention. We find that as expected, fidelity level relates to student outcomes; we also find that the presence of new curriculum materials positively predicts fidelity level.

VERSION: May 2021

Suggested citation: Hill, Heather C., and Anna Erickson. (2021). Using Implementation Fidelity to Aid in Interpreting Program Impacts: A Brief Review. (EdWorkingPaper: 21-414). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/dt2s-9v59

Using Implementation Fidelity to Aid in Interpreting Program Impacts: A Brief Review

Heather C. Hill
Harvard Graduate School of Education
Anna Erickson
Youth Policy Lab, University of Michigan

Author's note: The research reported here was supported by the National Science Foundation through grant numbers 1348669 and 0918383 to Harvard University. We would like to thank Robin Jacob, Katherine Gonzalez, and Stacey Brockman for their assistance in preparing the manuscript. The opinions expressed are those of the authors.

Abstract

Poor program implementation constitutes one explanation for null results in trials of educational interventions. For this reason, researchers often collect data about implementation fidelity when conducting such trials. In this paper, we document whether and how researchers report and measure program fidelity in recent cluster-randomized trials (RCTs). We then create two measures – one describing the level of fidelity reported by authors, and another describing whether the study reports null results – and examine the correspondence between the two. We also explore whether fidelity is influenced by study size, type of fidelity measured and reported, and features of the intervention. We find that as expected, fidelity level relates to student outcomes; we also find that the presence of new curriculum materials positively predicts fidelity level.

Using Implementation Fidelity to Aid in Interpreting Program Impacts: A Brief Review

An examination of program fidelity, or "how well an intervention is implemented in comparison with the original program design" (O'Donnell, 2008, p. 33) is considered critical to modern program evaluation. Estimates of fidelity can help confirm that changes in outcomes are in fact attributable to the program, increasing the internal validity of experiments and bolstering claims made about program efficacy (Dane & Schneider, 1998; Mowbray, Holter, Teague, & Bybee, 2003). Beyond providing methodological support, reports on program fidelity can also provide substantive assistance to designers and practitioners in human service sectors, especially when scholars subject these reports to systematic review. Reviews that examine common reasons for implementation failure, for instance, can help program designers strengthen their product (e.g., Durlak & DuPre, 2008).

Estimates of implementation fidelity can also help explain null results, in particular distinguishing between the possibility that the program was not delivered as designed and other sources of failure, such as methodological problems, flaws in program theory, or lack of fit to local contexts (Dane & Schneider, 1998; Hohmann & Shear, 2002; Jacob, this issue; Mowbray et al., 2003). Yet while conventional wisdom in policy analysis often locates null results in implementation failure, we have no estimates of the extent to which this is true, particularly in recent, rigorous trials of educational interventions. To this end, we review the evidence regarding program fidelity in modern educational research by analyzing classroom-level intervention projects funded by seven Institutional Educational Sciences (IES) programs and a second set of studies identified during a meta-analysis of STEM curriculum and professional development programs, projects funded primarily by the National Science Foundation. Specifically, we ask:

1. How often is program fidelity reported, and how is it defined and measured in recent educational program evaluations? What proportion of evaluations report low, moderate, and high fidelity?
2. To what extent is implementation fidelity predictive of program success?
3. To what degree is the level of implementation fidelity related to study size, the type of fidelity measured and reported, and features of the intervention?

We also qualitatively explore how often authors connect null results to poor fidelity, and the explanations authors offer for a lack of fidelity. We describe our methods and results below.

Methods

Our analysis combines data from two samples of studies. The first involves IES-funded studies intended to change or improve K-12 classroom instruction. IES Requests for Applications (RFA) (e.g. U.S. Department of Education, 2009, p. 60) require that awardees collect implementation data, and thus we searched each major IES program (effective teachers and effective teaching, mathematics and science education, reading and writing, social and behavioral contexts for academic learning, social and character development, teacher quality in math and science, teacher quality in reading and writing) for grants awarded during the years 2002-2011. We chose these dates because we thought it unlikely that projects funded after 2011 would consistently have publicly available evidence on implementation fidelity and project outcomes at the time the search was originally conducted, in 2016. We restricted our search to efficacy and replication (Goal 3) and scale-up (Goal 4) studies because of our interest in fidelity under realistic school and classroom conditions. Because too much variation in program design and clientele would lead to difficult-to-interpret results, we excluded studies that were not based in K-12 classrooms

(e.g., tutoring or online learning; pre-school) and studies focused on special populations (e.g., ELLs). This screen reduced the number of eligible projects to 42.

We located as many publications as we could find from each project, as authors often distributed student outcomes and implementation data over several papers. We then contacted principal investigators from grants with no publications to learn about study results and implementation metrics. In two cases these investigators were able to provide information on main impacts, but had not completed implementation analyses. In seven cases, we either could not reach the principal investigator after repeated attempts or student impact results were not ready for release. Most reports focused on one intervention/program, but one described multiple treatment arms (Penuel, Gallagher & Moorthy, 2011). To accommodate this, we coded study design features (e.g., type of implementation reported) for each intervention, but null results and fidelity separately for each treatment arm. In total, IES-funded studies contributed 35 reports containing 37 treatments.

Our second sample arises from a meta-analysis of preK-12 STEM curriculum and professional development interventions (see Lynch, Hill, Gonzalez & Pollard, 2019). In the initial round of screening, we downloaded 1,698 studies and examined their abstracts. Of these, 477 papers and reports met basic criteria for relevance, and were advanced to the next round of screening. In this round, two raters examined each paper to determine whether it met the review's inclusion criteria.[1] After applying these inclusion/exclusion criteria, 42 papers and reports remained. As

[1] As reported in Lynch, Hill, Gonzalez & Pollard (2017), we required that projects have at least two teachers and 15 students in each treatment group (Slavin, Lake, & Groff, 2009). Projects also needed to provide student outcome data in at least one paper, and possess a randomized or strong quasi-

above, we identified all available reports from a project, then reviewed those reports for evidence about implementation. To prevent double-reporting, we excluded papers already included in the IES pool. One study (Heller et al., 2012) included three treatment arms. The STEM meta-analysis sample therefore contributes information on 37 reports and 39 treatments in total.

Our achieved sample of studies is thus a convenience sample, in some senses, derived from easily accessible IES reports and an existing pool of studies located for a STEM meta-analysis. Because the IES and STEM samples differed in the requirements regarding measuring implementation fidelity, we answer the first research question, on frequency of reporting, separately. Because many studies did report on implementation, we are able to answer research questions two and three with a moderate-sized dataset.

Scoring and Analysis

Our coding system was simple, designed mainly to categorize IES and STEM study results for descriptive analysis. Our first and simplest code was whether fidelity was reported at all. Second, we recorded the method(s) used to assess fidelity (e.g., teacher surveys, classroom observations), and whether project researchers designed their own fidelity measures or relied upon those designed by other research teams. Third, we assessed the type(s) of fidelity measured. How to do so was not immediately obvious; scholars have generated many ways to conceptualize fidelity and an equally large number of ways to measure it, with some offering as many as five different

experimental research design. We excluded papers that had no comparison group, where the comparison group was not measured at time periods commensurate with the treatment group, and where the comparison group was assembled through post-hoc matching. To provide comparability to the IES study pool, we then excluded projects that occurred exclusively or primarily in preK classrooms and that were conducted outside the United States.

categories for reporting (Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003). Several scholars, however, distinguish between what we will call structural fidelity (adherence to program design re: staffing levels, case load size, budget, procedures, frequency and intensity of contacts) and process fidelity (style, client-staff interactions, client-client interactions, individualization of treatment, climate) (Century, Rudnick & Freeman, 2010; Mowbray et al., 2003; O'Donnell, 2008). These scholars also call out dosage fidelity (Dane & Schneider, 1998 and Dusenbury, et al. 2003) which records whether a program was actually accessible to those meant to implement it. We adopted these categories and modified them to fit educational interventions. We coded positively for *structural* fidelity when the authors provided evidence on classroom-level compliance with program-specific elements, such as the use of project curriculum units, adherence to project-supplied lesson plans, and the deployment of program-specific instructional behaviors (e.g., worked examples featuring a particular sequence of teacher questions) with no attendant focus on quality. Typically, authors collected this data only from treatment-group classrooms. We coded positively for *process* fidelity when authors measured more complex and general classroom-level outcomes, such as teacher sensitivity to student learning needs, mathematical discussions, classroom climate and student behavior, and the cognitive challenge of student tasks. Typically, authors collected this data from both treatment and control classrooms. Finally, we coded yes for *dosage* fidelity when authors provided evidence on the extent to which teachers received a treatment (e.g., descriptions of attendance at professional development, whether curriculum materials were delivered to teachers in a timely manner). Structural fidelity typically measures adherence to program elements, process fidelity is a form of intermediate impact, and dosage fidelity measures teacher opportunity to learn or to use program materials.

We measured the extent of implementation fidelity using two purely quantitative indicators as well as a holistic, more qualitative metric. Our quantitative measures consisted of:

- The proportion of positive structural fidelity results reported by authors. We considered structural fidelity metrics positive when authors observed 50% compliance with project activities.
- The proportion of positive process fidelity results reported by authors. Because process fidelity was typically reported as a treatment-control contrast, we considered process fidelity metrics positive when authors observed positive and significant results of this T/C test.

Our holistic measure took into account outcomes from these quantitative fidelity metrics, but also relied upon other sources of information about overall fidelity. Specifically, we considered both dosage fidelity, typically reported descriptively, and authors' comments about fidelity of implementation. We also weighed process fidelity results over structural fidelity results when the two conflicted. We assigned a score of "low fidelity" when less than half of the structural and process fidelity codes were positive, a score of "high fidelity" when more than 80% of the structural and process fidelity codes were positive, and "medium fidelity" for those in between. Both authors coded each report included in the review, then met to reconcile disagreements. We recognize that this coding system requires a fair amount of judgment, but a more deterministic coding system was impossible in light of the different fidelity measures used by study authors.

To allow us to qualitatively understand the relationship between implementation fidelity and null results, we developed a rudimentary metric for assessing whether results from a study were null. For each project we calculated the fraction of total impact estimates, aggregated across all available papers and reports, that were positive at significance levels of at least $p < 0.05$. We used this estimate of the proportion of positive impacts estimate in some of our models. We also created a binary measure by categorizing studies with less than 50% positive effects as null results studies – an arbitrary threshold, but one reflective of current hopes for consistent and positive results in the field. Authors and two research assistants double-coded each study, discussing and resolving discrepancies where they arose. One potential issue with this method for determining null results is that it does not distinguish between more and less central program outcomes – for instance, when a program expects a strong impact on executive function and weaker impacts on student achievement. In practice, however, few studies prioritized outcomes in this way, with many reports containing two outcomes (e.g., a researcher-developed and standardized measure) without information about which researchers valued more.

We also coded for a number of program and study design features that might impact fidelity. Our first study design feature was sample size; because we expected fidelity may be lower in studies with larger sample sizes, we recorded the number of teachers in the treatment and control groups combined. As noted above, we coded for the method for collecting fidelity evidence, structural vs. process fidelity, and whether the researchers designed their own fidelity metric. Among program characteristics, we coded for whether the program featured curriculum materials, professional development, and/or coaching, and noted the maximum number of hours teachers could have experienced the coaching and professional development.

To answer our first research question, regarding how often program fidelity is measured and reported, we calculated simple descriptives and crosstabs. To answer our second research question, regarding the relationship between implementation fidelity and program success, we generated both crosstabs and a regression of the fraction of null results over fidelity level, using study characteristics as controls. Finally, to answer research question three, we generated two regression models linking study and program characteristics to fidelity levels. To link study characteristics to fidelity level, we used a multi-level model; the multilevel model accounts for the nesting of multiple fidelity assessments within treatments. Because the model regressing our holistic variable as an outcome did not converge in this multilevel model, we used the ratio of positive fidelity outcomes in its place. In our second model linking fidelity to program features, we used OLS regression because program features were not nested within treatments.

Results

Program fidelity measurement, reporting and results

Program fidelity was reported in 97% of projects arising from IES grants, and 74% of projects in the STEM pool (Table 1). One project in the IES pool suggested that implementation data was collected, but did not provide results. Across both study pools, structural fidelity was measured in 54% of projects and process fidelity was measured in 50% of projects. Eighteen projects (24%) across both sources measured both process and structural fidelity. Dosage fidelity was reported in 26% of projects. For studies reporting fidelity, the most frequently used method for gauging fidelity was classroom observation, with 46% of projects reporting this data collection technique; teacher self-reports (typically logs and surveys) followed behind, with 29% of projects using this technique; 18% of studies used both (Table 2). Our read of the studies also

suggested scattered use of other methods, such as teacher interviews, student surveys, and periodic check-ins between teachers and study staff. Most projects evaluated interventions against original program design; in fact, we found no project that directly measured users' adaptations of the program, despite scholarly interest in this topic (Blakely et al., 1987; McMaster et al., 2014; O'Donnell, 2008; Quinn & Kim, 2017).

Of 65 projects that presented quantitative fidelity data, 26 (40%) included evidence of strong implementation (Table 3, row 3). For instance, Lara-Alacio et al. (2012) reported that treatment teachers earned 108 of 124 points on a structural fidelity metric, and Schwartz-Bloom and Halpin (2003) reported that most teachers used at least three of the four curriculum modules developed as part of their project. Twenty-five (38.4%) projects included evidence of moderate implementation. For instance, Star et al. (2015) reported that many treatment-group teachers used the newly developed curriculum materials with some aspects of structural fidelity (e.g., when using the curriculum, teacher displayed learning objective; teacher summarized major points from student discussion), but also reported that roughly one-fifth of teachers did not use those materials at all. Finally, 14 (21.5%) projects included evidence of weak or non-existent implementation. For example, Lang, Schoen, LaVenia & Oberlin (2014) report no treatment/control contrast in the use of formative assessment practice as recorded in classroom observations.

Does fidelity predict program success?

Table 4 shows results from our holistic fidelity level metric by whether we categorized the study as having null results. Overall, we classified 27 (35.5%) treatments as producing null results, a

figure much lower than the 91% null-result rate found by the Coalition for Evidence-Based Policy (CEBP, 2013). Studies coded as moderate or high fidelity had more than double the chance of yielding positive results than null results; for studies coded as low fidelity, the odds of a positive and null categorization were about even. There appeared little difference between moderate and high fidelity ratings, in terms of the likelihood of positive results. Table 4 also shows that fidelity, at least as we have defined it, is not deterministic of program outcomes. Six studies had majority-positive impacts yet low fidelity ratings, and another eight had strong fidelity ratings but majority-null impacts.

To further understand the relationship between fidelity and null results, we regressed the fraction of positive student impacts over both our holistic fidelity rating and controls, including the teacher sample size and the type of assessment used to measure student learning. Because neither Table 4 nor exploratory regressions revealed a difference between moderate and high-fidelity studies in terms of the likelihood of positive outcomes, we simplified our fidelity measure to a dummy variable representing low fidelity (Table 5). We found a significant relationship between the dummy variable representing low-fidelity implementation and student outcomes; treatments with low fidelity averaged 24% fewer positive outcomes than those with moderate or strong fidelity. We also observed that teacher sample size had a small but statistically significant negative relationship to the fraction of positive results; a treatment one SD above average in teacher sample size (442 teachers) had, on average, 6.6% fewer positive impacts than a program with an average-sized sample (167 teachers). In line with Hill, Bloom, Black and Lipsey (2008), researcher-designed assessments were also more likely to post positive impacts as compared to standardized assessments (shown) and studies that used both standardized and researcher-

designed measures (the referent variable). Separately, we also examined the likelihood of null results by content area, including STEM, reading/writing, and social/behavioral interventions, and found no relationship (not shown).

To complement our quantitative analysis, we examined reports from null-report studies to see the extent to which authors indicate that implementation may have contributed to a lack of impacts. We saw that five of the 27 null-result treatments (Gersten, Dimino, Jayanthi, Kim, & Santoro, 2010; Jacob, Hill & Corey, 2017; Matsumura, Garnier, & Spybrook, 2012; Santagata, Kersting, Givvin, & Stigler, 2011; Schneider & Meyer, 2012) identify dosage fidelity – teachers' receipt of an appropriate amount of professional development – as problematic, although one of those studies (Gersten et al., 2010) reports generally strong dosage and classroom fidelity. Only seven of 27 null-result treatments (Borman, Gamoran & Bowdon, 2008; Cavalluzzo et al., 2014; Dominguez, Nicholls, & Storandt, 2006; Hurtig, 2009; Santagata et al., 2011; Star et al., 2015; Thompson, Senk, & Yu, 2012) contain evidence suggesting that lack of structural or process implementation fidelity may have led to an absence of impacts on student outcomes. One of those (Cavalluzzo et al., 2014) reported moderate fidelity based on teacher and student surveys but commented on a lack of fidelity in its discussion, while another (Grigg, Kelly, Gamoran, & Borman, 2013) noted that while treatment teachers were nearly twice as likely to use an inquiry science teaching method than control teachers, the quality of those instructional elements was questionable.

Next we turn to models that predict implementation fidelity by study characteristics and program features. In Table 6, we used a two-level model (fidelity measures nested within studies) to

regress the fraction of positive fidelity impacts on study design characteristics. We find that using a process (vs. structural) fidelity measure is associated with a lower fidelity rating; unexpectedly, we find the use of classroom observations (vs. teacher self-reports) associated with stronger fidelity in our final model. Researcher-designed (vs. third-party designed) fidelity measures have a positive relationship when entered into the models alone, but none in the final model. Finally, sample size was not related to the fraction of positive fidelity impacts.

In Table 7, we regress our holistic measure of fidelity level on program features. We find that when entered singly, the program's provision of professional development and curriculum materials are each positively associated with implementation fidelity; the number of hours of professional development has a slight negative relationship with fidelity level. However, all but two programs provided some professional development, leading us to both concern about making strong inferences from this variable, and also leading us to omit this variable from the model with multiple predictors. In that model, curriculum materials remained a positive and significant predictor of fidelity level but professional development hours did not.

To again complement our quantitative analysis, we examined project reports for factors linked to fidelity. Some projects measured factors thought to affect implementation fidelity and formally tested them as part of their analyses. For instance, Matsumura, Garnier, and Resnick (2010) evaluated the extent to which coach background, coach orientation toward their role, school-level professional community, teacher experience, and principal support explained teacher take-up of coaching (dosage fidelity). Wanless, Rimm-Kaufman and colleagues (2014) conducted both qualitative and quantitative analyses and identified principal buy-in, coach attributes, and

teachers' perceptions of validation for their efforts as critical to teacher take-up and classroom implementation.

In addition to formal testing of the factors linked to implementation fidelity, other projects offered more impressionist accounts of such factors; these accounts are especially common among studies that found a lack of fidelity. Authors of one study that had low fidelity as reported on a process metric (Murray et al., 2014) commented that its measurement of fidelity may have been problematic – that the observational metric used to capture classroom processes (the Classroom Assessment Scoring System, or CLASS) was not sufficiently aligned to the intervention's outcomes. Thus fidelity itself may not have been an issue. Santagata, Givvin and colleagues (Santagata et al., 2011; Givven & Santagata, 2011) discussed a wide array of reasons for lack of implementation, from inadequate principal support for the program, competing programs that absorbed teacher time, and, for some teachers, insufficient content knowledge to fully understand and implement the program. Hill, Jacob & Corey (2018) identify a similar set of reasons for a teacher professional development program's failure to impact practice. Santagata and colleagues (2011) also noted that teachers often came unprepared to meetings, an observation echoed by Gersten and colleagues (2010). In some cases, the difficulty teachers experienced when implementing novel instructional practices seemed at issue. For instance, Cavalluzzo and colleagues (2014) surmise that although their teachers did engage in routines around data use, the focus of her intervention, they were not able to translate what they learned from data into classroom practice. Borman, Gamoran & Bowdon (2008) surmise an "implementation dip," in which instructional quality declines as teachers encounter new curricula and instructional routines. Three other authors (Hill, Santagata, and Hurtig) also speculate that

their intervention may have not been sufficiently strong to overcome obstacles to implementation.

Conclusion

The field of implementation fidelity research has come far from the days when interventions were black boxes converting inputs to outputs. Much of the promise of implementation fidelity noted by scholars – validating experimental designs, understanding mechanisms, and explaining null results – has been realized in recent studies. Four-fifths of studies reported on some measure of fidelity. Many included measures of more than one type of fidelity – structural, process, or dosage – and many used classroom observations to examine impacts on practice. Low fidelity increases the likelihood of weak student outcomes, and fidelity itself is predicted by both study design characteristics and program features. We offer several interpretations of our findings and suggestions in this conclusion.

On average, better fidelity correlated with better program outcomes, confirming an assumption made by many scholars, and aligning with empirical evidence from studies in which implementation fidelity observably mediates program impact (e.g., Penuel, Gallagher, Moorthy, 2011; Rimm-Kaufmann et al., 2014). Interestingly, our results suggest that moderate and strong fidelity yield the same likelihood of on-average positive impacts on student outcomes, leading to the intriguing hypothesis that moderate fidelity may be enough to yield positive program outcomes. Understanding better what level of fidelity is "enough" is a key task for future researchers. Nevertheless, this evidence suggests that intervenors should continue to place bets, as they have done, on improving teacher take-up of key program practices.

Our results also imply that implementation fidelity is a partial but not complete explanation for null-result studies. Eight studies had null results but high fidelity, suggesting that the program design and contextual factors outlined by Jacob and Kim (this issue) may play a role in producing program outcomes. We also found six studies with low fidelity but with positive program impacts, suggesting that authors' fidelity measures may not have been sensitive to key changes in classroom practice, or that teachers may not have implemented the intervention "by the book" yet nevertheless saw positive results.

Implementation fidelity appeared shaped by several factors, including how scholars measured fidelity. Structural fidelity measures – often checklists but almost always surface-level indicators of implementation – tended to show stronger fidelity than process metrics, which often required more substantial changes in classroom climate or teacher practices. Classroom observations tended to see more positive fidelity outcomes than teacher self-reports, perhaps because of issues with response bias in the latter; it is not unusual for treated teachers to report enacting fewer practices once they gain more specific information about what the survey items intend to capture (see Jacob, Hill & Corey, 2017). Program characteristics, including the presence of professional development and curriculum materials, also positively predicted fidelity outcomes. Against expectation, fidelity was not influenced by the size of the teacher sample, suggesting that high-quality implementation can occur at scale. We also found that the maximum hours of professional development was either negatively related (when considered alone) or unrelated (in models with multiple predictors) to fidelity. These results suggest specific pathways through which fidelity can be intentionally supported by intervenors, and also that strong fidelity is not out of reach for large programs and/or programs with limited resources for teacher professional

development.

Descriptive evidence from our reading of these studies highlights other themes, themes which align with Kennedy's (2005) analysis of teaching and efforts to improve teaching. Support from principals and peers is critical to implementation success, and programs placed in complex environments are likely to have more difficulty seeing their key components carried out. Programs that ask teachers to complete more 'difficult' tasks – for instance, introducing higher cognitive demand tasks into classrooms, or using data to inform instruction – may simply be more difficult for teachers to enact, and less likely to be implemented with fidelity.

In reading project reports, we also noticed the systematic absence of information we argue should be collected to advance our knowledge of implementation. To start, projects that studied teacher adaptation of interventions were rare (see also Durlak & DuPre, 2008). Given the long history of debates over viewing implementation from a fidelity versus adaptation perspective (O'Donnell, 2008), and the notion that adaptation is likely and even desireable in some settings, the field should do much more to both qualitatively understand adaptation and to perhaps systematically test whether planned teacher adaptation can lead to improved program outcomes (see, e.g., DeBarger et al. 2017; Kim et al. 2015; McMaster et al. 2014). Second and relatedly, few reports presented teachers' perspectives on program implementation. For instance, we rarely found teachers' insights into typical barriers to implementation, typical difficulties working with ideas from professional development or instructional materials, and typical reasons in which implementation varied, qualitatively, from what the authors of the interventions intended. Improving program implementation at scale cannot occur without a more nuanced understanding

of these issues. Finally, the reports we reviewed often contained basic information about district contexts (size, student demographics) but rarely contained insights into other factors that might condition implementation fidelity, such as the number of competing programs, alternative sources of instructional guidance, teacher capacity, and school and district organizational characteristics (see also Lynch et al., 2019). Without such information, it is difficult to develop a field-wide sense for what level of implementation is realistic in a given context, and which contextual factors need to be recognized and navigated by program developers.

Finally, for the field to continue to grow, we will need more rigorous studies of implementation, as well as better understanding of how program features lead to or help mitigate against implementation challenges. A good start appears in the small number of studies that predict implementation fidelity from teacher and school characteristics (e.g., Matsumura et al., 2010). Advancing the field likely means encouraging such studies, perhaps using the framework described by Durlak and DuPre (2008), to structure the systematic measurement and testing of factors related to classroom implementation.

References

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, 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.

Borman, G. D., Gamoran, A., & Bowdon, J. (2008). A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness*, *1*(4), 237-264.

Cavalluzzo, L., Geraghty, T. M., Steele, J. L., Holian, L., Jenkins, F., Alexander, J. M., & Yamasaki, K. Y. (2014). Using data to inform decisions: How teachers use data to inform practice and improve student performance in mathematics. Arlington, VA: CNA

Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation: A foundation for shared language and accumulation of knowledge. *American Journal of Evaluation*, *31*(2), 199-218.

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

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.

Debarger, A. H., Penuel, W. R., Moorthy, S., Beauvineau, Y., Kennedy, C. A., & Boscardin, C. K. (2017). Investigating purposeful science curriculum adaptation as a strategy to improve teaching and learning. Science Education, 101(1), 66-98.

Dominguez, P. S., Nicholls, C., & Storandt, B. (2006). Experimental Methods and Results in a Study of PBS TeacherLine Math Courses. *Hezel Associates (NJ1)*.

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.

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health Education Research*, *18*(2), 237-256.

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal, 47*(3), 694–739.

Givvin, K. B., & Santagata, R. (2011). Toward a common language for discussing the features of effective professional development: The case of a US mathematics program. *Professional Development in Education*, *37*(3), 439-451.

Grigg, J., Kelly, K. A., Gamoran, A., & Borman, G. D. (2013). Effects of two scientific inquiry professional development interventions on teaching practice. *Educational Evaluation and Policy Analysis*, *35*(1), 38-56.

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*(3), 333-362.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172-177.

Hill, H. C., Corey, D. L., & Jacob, R. T. (2018). Dividing by Zero: Exploring Null Results in a Mathematics Professional Development Program. *Teachers College Record*, *120*(6), n6.

Hohmann, A. & Shear, M. (2002). Community-based intervention research: Coping with the "noise" of real life in study design. *American Journal of Psychiatry, 159*(2), 201-207.

Hurtig, R. (2009). IES Annual Performance Report CFDA #84.305G (Grant Award #R305G04145). Iowa City, IA: The University of Iowa.

Jacob, R., Doolittle, F., Kemple, J., & Somers, M.-A. (2020). A framework for null results. *Educational Researcher.*

Jacob, R., Hill, H., & Corey, D. (2017). The Impact of a Professional Development Program on Teachers' Mathematical Knowledge for Teaching, Instruction, and Student Achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 379-407.

Kennedy, M. M. (2005). *Inside teaching: How classroom life undermines reform.* Cambridge, MA: Harvard University Press.

Kim, J. (2020). Making every study count: Learning from replication failure to improve intervention research. *Educational Researcher*.

Kim, J. S., Burkhauser, M. A., 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), 443-467.

Lang, L. B., Schoen, R. R., LaVenia, M., & Oberlin, M. (2014). Mathematics Formative Assessment System--Common Core State Standards: A Randomized Field Trial in Kindergarten and First Grade. *Society for Research on Educational Effectiveness*.

Lara-Alecio, R., Tong, F., Irby, B. J., Guerrero, C., Huerta, M., & Fan, Y. (2012). The effect of an instructional intervention on middle school english learners' science and english reading achievement. *Journal of Research in Science Teaching*, *49*(8), 987-1011.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the Research Base that Informs STEM Instructional Improvement Efforts: A Meta-Analysis. *Educational Evaluation and Policy Analysis*, 0162373719849044.

Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2012). The effect of content-focused coaching on the quality of classroom text discussions. *Journal of Teacher Education*, *63*(3), 214-228.

Matsumura, L. C., Garnier, H. E., & Resnick, L.B. (2010). Implementing literacy coaching: The role of school social resources. *Educational Evaluation and Policy Analysis, 32*(2), 249-272.

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.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315-340.

Murray, D.W., Rabiner, D.L. & Carrig, M. M. (2014) *Grade level effects of the Incredible Years Teacher Training Program on emotion regulation and attention.* Paper presented at the annual conference of the Society for Research on Educational Effectiveness, Washington D.C.

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.

Quinn, D. M., & Kim, J. S. (2017). Scaffolding fidelity and adaptation in educational program implementation: Experimental evidence from a literacy intervention. *American Educational Research Journal*, 0002831217717692.

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programs. *American Educational Research Journal*, *48*(4), 996-1025.

Rimm-Kaufman, S. E., Larsen, R. A., Baroody, A. E., Curby, T. W., Ko, M., Thomas, J. B., ... & 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.

Santagata, R., Kersting, N., Givvin, K. B., & Stigler, J. W. (2011). Problem implementation as a lever for change: An experimental study of the effects of a professional development program on students' mathematics learning. *Journal of Research on Educational Effectiveness, 4*(1), 1–24.

Schneider, M. C., & Meyer, J. P. (2012). Investigating the efficacy of a professional development program in formative classroom assessment in middle school English language arts and mathematics. *Journal of Multidisciplinary Evaluation*, *8*(17), 1-24.

Schwartz-Bloom, R. D., & Halpin, M. J. (2003). Integrating pharmacology topics in high school biology and chemistry classes improves performance. *Journal of Research in Science Teaching*, *40*(9), 922-938.

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*(2), 839-911.

Star, J. R., Pollack, C., Durkin, K., Rittle-Johnson, B., Lynch, K., Newton, K., & Gogolen, C. (2015). Learning from comparison in algebra. *Contemporary Educational Psychology*, *40*, 41-54.

Thompson, D. R., Senk, S. L., & Yu, Y. (2012). An evaluation of the third edition of the University of Chicago school mathematics project: Transition mathematics. *Chicago, IL: University of Chicago School Mathematics Project*.

U.S. Department of Education. (2009). Education research grant request for application: CFDA Number 84.305A. Washington, DC: Institute for Education Sciences. Retrieved August 4, 2016 from http://ies.ed.gov/funding/pdf/2009_84305A.pdf

Wanless, S. B., Patton, C. L., Rimm-Kaufman, S. E., & Deutsch, N. L. (2013). Setting-level influences on implementation of the Responsive Classroom approach. *Prevention Science*, *14*(1), 40-51.

Table 1

Types of fidelity reported by source

	IES-funded studies (N=37)		STEM Studies (N=39)		Total (N=76)	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
Fidelity reported	36	97%	29	74%	65	86%
Structural fidelity	23	62%	18	46%	41	54%
Process fidelity	21	57%	17	44%	38	50%
Structural & process	9	24%	9	23%	18	24%
Dosage fidelity	8	22%	12	31%	20	26%

Note. Percentages reported are of the total set of studies and do not sum to 100 because many studies reported more than one type of fidelity. All overlapping studies are included in the IES-funded count here because of the IES requirement to report fidelity.

Table 2

Method of gauging fidelity

Method	Frequency	Percent
Observational	30	46%
Self-report	19	29%
Both	12	18%

Note. Total N=65. Four of the 65 studies that measured fidelity only measured dosage fidelity so percentages do not sum to 100. Classroom observation includes in-person, video-taped, and one instance of audio recorded observations. Teacher self-report includes daily activity logs or post-intervention surveys.

Table 3

Fidelity ratings by type of fidelity measured

	Structural		Process		Dosage	
Fidelity Rating	Frequency	Percent	Frequency	Percent	Frequency	Percent
Low fidelity (N=14)	5	12%	12	32%	6	30%
Moderate fidelity (N=25)	18	44%	15	39%	6	30%
High fidelity (N=26)	18	44%	11	29%	8	40%
Total	41	100%	38	100%	20	100%

Note. N=65. Some studies reported more than one type of fidelity so row frequencies do not sum to the total N. A study was coded low when less than half of the structural and process fidelity codes were positive, a score of "high fidelity" when more than 80% of the structural and process fidelity codes were positive, and "medium fidelity" for those in between. Ratings also factored in dosage fidelity, when reported, and any additional author comments about fidelity.

Table 4

Fidelity rating by student impacts

	Null impacts		Positive impacts	
	Frequency	Percent	Frequency	Percent
Low fidelity	8	10.5%	6	7.9%
Moderate fidelity	8	10.5%	17	22.4%
High fidelity	8	10.5%	18	23.7%
Fidelity not reported	3	3.9%	8	10.5%
Total	27	35.5%	49	64.5%

Note. N=76. Studies were coded null if they had less than 50% student impacts that were positive with significance levels of at $p < 0.05$.

Table 5

Results of OLS regression analysis for variables predicting percent positive student achievement outcomes (N=76)

	Model 1	Model 2	Model 3	Model 4	Model 5
Fidelity rating = 1 (Low)	-0.19				-0.24**
	(0.13)				(0.11)
Sample size - teacher sample		-0.00032***			-0.00024**
		(0.000089)			(0.000095)
Standardized assessment only			-0.27**		-0.085
			(0.088)		(0.094)
Researcher-developed assessment only				0.36***	0.29**
				(0.095)	(0.11)
Intercept	0.57***	0.58***	0.66***	0.43***	0.58***
	(0.049)	(0.050)	(0.058)	(0.050)	(0.064)
R-squared	0.033	0.048	0.115	0.160	0.247

Standard errors in parentheses.
* p<0.1, ** p<0.05, *** p<0.001

Table 6

Results of multilevel regression analysis for variables predicting percent positive fidelity impacts (N=61)

	Model 1	Model 2	Model 3	Model 4	Model 5
Process fidelity	-0.34***				-0.36***
	(0.069)				(0.071)
Classroom observation		0.058			0.14**
		(0.08)			(0.072)
Researcher-designed instrument			0.27*		0.13
			(0.14)		(0.13)
Teacher sample size				0.00008	0.00013
				(0.00013)	(0.00011)
Intercept	0.85***	0.66***	0.45***	0.68***	0.63***
	(0.049)	(0.065)	(0.13)	(0.049)	(0.14)

* p<0.1, ** p<0.05, *** p<0.001

Note: Table displays results from a multilevel model in which impact estimates are nested within programs. Four of the 65 studies that measured fidelity only measured dosage fidelity. Standard errors in parentheses.

Table 7

Results of OLS regression analysis for variables predicting fidelity rating (N=61)

	Model 1	Model 2	Model 3	Model 4	Model 5
Professional development	1.22***				
	(0.095)				
New curriculum		0.37*			0.39*
		(0.19)			(0.22)
Coaching			0.13		0.15
			(0.20)		(0.21)
Duration of professional development				-0.0070**	-0.0045
				(0.0024)	(0.0030)
Intercept	1.00***	1.96***	2.14***	2.48***	2.10***
	(0.000000042)	(0.15)	(0.12)	(0.14)	(0.28)
R-squared		0.054	0.007	0.099	0.153

* p<0.1, ** p<0.05, *** p<0.001

Note: Four of the 65 studies that report fidelity did not include PD or were missing a duration. Standard errors in parentheses.