Diagnosing bottlenecks in development-stage field experiments: Troubleshooting and finding opportunities for improvement

Andrew P Jaciw, Li Lin and Adam Schellinger

Abstract
Randomized trials of social programs yield internally valid estimates of causal impacts on key outcomes. While estimates of gross impact serve as useful summaries, program developers require deeper information to drive improvement efforts, especially when no impact is observed. The first main goal of this work is to present a seven-stage diagnostic method for assessing process bottlenecks in experiments. Designed for programs still in development, the troubleshooting sequence uses mixed methods to assess where in a program’s logic model the process is compromised. It includes post-experimental methods that are built into the design, to account for impact variation and test where effects are intensifying or diminishing. The second main purpose of this work is to demonstrate one such method in detail. The approach tests the relationship between fidelity of program implementation and impact. First, levels of achieved fidelity in the treatment group are modeled in terms of informative baseline covariates. The model is then used to index fidelity in both conditions. Informed only by pre-randomization characteristics of individuals, the model-based fidelity scores are unbiased by endogeneity, and allow assessment of whether impacts on key outcomes vary by levels of fidelity. Results can help program developers focus improvement efforts. We illustrate the seven-step diagnostic process through a randomized trial of the Internet-Based Reading Apprenticeship Improving Science Education (iRAISE) program. Eighty-two high school science teachers and 1468 students were randomly assigned to a literacy program or control. There was no overall impact on achievement. Applying the diagnostic process revealed this was not due to a weak program contrast between conditions, or an inadequate assessment; rather, lower-than-expected impact was likely due to weaker than intended implementation.

Keywords
Randomized experiments, innovative methods, improvement science, implementation fidelity, process troubleshooting, mixed methods

Introduction
Randomized control trials (RCTs) have assumed a prominent role in social science research in the past two decades, including in education (Spybrook, 2013) and their use and applications are increasing (Peck and Goldstein, 2016). This is especially true in the United States, where federal funding of educational research programs is often contingent on building-in a summative program evaluation that assesses program impact using methods that limit potential for selection bias in results. Randomizing participants to treatment or control rules out selection effects as a rival explanation of differences between the experimental groups in their outcomes. Because of this, randomized trials are considered the “gold standard” method for supporting valid causal inferences.

While growth in use of experiments has continued in specific policy environments, we are reminded of their limitations. It has long been recognized, and recently re-asserted, that

Empirical Education, Inc., Palo Alto, CA, USA

Corresponding author:
Andrew P Jaciw, 425 Sherman Ave. Unit 210, Palo Alto, CA, USA.
Email: ajaciw@empiricaleducation.com; ajaciw@hotmail.com
experiments by themselves do not warrant inferring results to populations beyond the study sample (Berliner, 2002; Bryk, 2014; Cronbach, 1982). That is, experiments conclusively demonstrate whether a program can have a positive causal effect; however, they do not support the extrapolation of a given impact finding beyond the study context (Bryk, 2014). Research that gives scant attention to how programs work in relation to specific contexts—focusing on estimation of gross impact quantities instead of understanding conditions conducive to those results—provides little support for necessary adaptations to allow programs to succeed in the field. This has led to an “adopt, attack and abandon” mind-set (Rohanna, 2017: 68), whereby programs are implemented in specific contexts, and then abandoned, if they do not produce results, before their potential to work is adequately understood.

This disconnect between results of experiments, and their uses and applications, holds when going from an experiment to the field; however, it is also problematic when program developers conduct experiments to inform program improvement. While there is often an imperative for early-stage demonstration of program promise and efficacy to satisfy accountability interests of the funder, and ultimately to secure more funds, simply showing that a program does or does not produce a positive marginal impact in its current stage of development—a “yes or no” determination—yields scant productive information for program improvement, and encourages a “fear of failure” (Lemire et al., 2017: 27) as a program’s survival hinges on one or a small number of marginal impact findings.

The purpose of this work is to expand understanding of potential uses of experiments, especially when conducted with programs that are under development. Recognizing that randomized experiments often may be implemented even in early stages of program development, we investigate how experimental designs and analysis can be re-purposed, with complementary methodology, to help developers assess, not just whether there is a positive marginal impact of the program, but the conditions for impact. More specifically, we are interested in presenting an innovative methodology that yields useful “improvement information” in the case where an experiment of a nascent program fails to produce a positive effect. In that case, with adequate planning, data collection, and use of mixed complementary methods, a diagnostic-remedial process may be integrated into the RCT itself, turning an implementation that fails to meet expectations into an opportunity to plan for success. We believe this work holds value not only in research and policy contexts where RCTs are commonplace, such as in the United States, but also in places where the potential value of RCTs is being weighed against benefits of alternative methods—we show that RCTs can be designed to be complementary with other methods, and together they can reveal conditions supporting program effectiveness.

This work has two main purposes. The first is to describe and illustrate, through an RCT, a seven-stage diagnostic-remedial “improvement process” for use with RCTs of programs under development, including building on principles from the experimentalist’s frame, as well as from improvement science. The sequential process is designed to work as a “systems check” or troubleshooting sequence, whereby process bottlenecks are hypothesized and tested. The second is to provide an in-depth demonstration of one of these diagnostic steps, which applies an innovative methodology that builds on the experiment to establish a correspondence between level of program implementation and impact. Understanding this relationship, central to program development, is usually handled through weaker correlational methods. The alternative we present leverages the randomized design, and informative baseline and implementation data, to produce internally valid inferences addressing the question. The overarching goal of this work is to demonstrate an expanded set of methods to support developers in program improvement.

This work proceeds as follows: First, we provide a short background on the benefits and limitations of randomized experiments. Second, we describe how the design and analysis of randomized trials can be expanded, in tandem with complementary mixed methods, to better address needs of program developers. Third, we illustrate the seven-step troubleshooting process using the example of an RCT designed to assess the efficacy of a literacy program in education. A no impact finding is the start point for hypothesizing and testing where the program process may be problematic, with a view to its improvement. Fourth, as part of the troubleshooting sequence, we demonstrate in detail the innovative approach noted above to assessing the relationship between program implementation and impact. Fifth, we draw conclusions about the diagnostic process.

**Background**

**Randomized experiments in the social sciences are popular and their use is expanding**

Social experiments are used widely to address questions of program efficacy. The Randomized Experiment eJournal counts 2000+ entries since its inception in 2007 (Peck and Goldstein, 2016). In the context of educational research in the United States, between 2002 and 2013, the Institute of Education Sciences funded more than 100 cluster randomized trials (Spybrook, 2013). Most recently, in American education, use of randomized experiments and quasi-experiments has been explicitly written into law through the Every Student Succeeds Act, with the goal of limiting selection bias in evaluating the efficacy of federally funded programs. Also, results of causal impact evaluations are being summarized through compendia such as the federally funded What Works Clearinghouse (Institute of Education Sciences, n.d., https://ies.ed.gov/ncee/wwc/), where designs with random assignment can meet the coveted ‘without reservations’ evidence standard. It is safe to say, that experiments are a
Randomized experiments in the social sciences: benefits and limitations

Experiments support internally valid inferences. The advantage and appeal of randomized experiments is that, by design, they allow accurate assessment of average causal impacts of programs. Random assignment of individuals to treatment or control ensures that the groups, thus formed, are on average balanced on all factors, whether known or unknown, and measured or not, that affect outcomes. Such factors would produce bias from confounding if distributed unevenly between conditions (Shadish et al., 2002). If outcomes are obtained for almost everyone randomized (i.e. provided attrition of participants is low) the mean difference in outcomes represents the average causal effect of being randomized to the program of interest, relative to the control condition (usually “business as usual”), for individuals in the study. (This causal quantity is referred to as the effect of Intent-To-Treat (ITT).)

More formally, random assignment assures that the difference between treatment and control in their average outcomes (a quantity we can observe), in expectation equals the average, across the study sample, of differences in outcomes between conditions for individuals (i.e. of the causal effect for individuals), which we are unable to observe because we cannot see how a person performs simultaneously in both conditions (Rubin, 1974). This is a profound result saying that while we cannot know the causal effect for any given individual, random assignment allows us to estimate this quantity on aggregate without bias. The experimental manipulation—random assignment—has been likened to a figurative “twist(ing) Nature by the tail” by Francis Bacon (Cook and Payne, 2002): while Nature may be complex, and it is hard to fathom its relations, we can coax out critical truths through experimental control.

Limitations to using experiments. Causal effect estimates are averages. That is, they address “gross impacts” of a program for a given study sample (Chatterji, 2007). While potentially unbiased for that sample, an ITT estimate applies to the specific conditions of a study, and does not necessarily, or even normally, generalize to other contexts. Referring to clinical trials in education Bryk (2014) notes,

They tell us whether some intervention can work. That is, if a field trial produces a significant effect size, it means that the intervention had to work somewhere for somebody. Such studies, however, are not primarily designed to tell us what it will take to make the intervention work for different subgroups of students and teachers or across varied contexts. (p. 469)

Additional and complementary methods are needed to understand conditions for impact, including the role of program components and their implementation, the influence of moderators of impact, possible compensatory effects of the counterfactual program(s), and intermediate processes and mechanisms that modulate impact. The average “marginal” impact estimate reflects the joint influence of these factors. Understanding their role focuses research both inward, toward a program’s composition, and outward, toward the broader context of a program’s implementation.

Using mixed methods to expand applications of experiments

Experiment-based methods. The toolkit of complementary methods that contextualize the gross impact findings include ones that are adjuncts to the experiment itself. Among such “post-experimental” approaches, are moderator and subgroup analyses (Bloom, 2005b), which address whether program effects vary across subsets of individuals, or hold for specific subgroups in the sample; mediator analyses that investigate whether impacts on important distal outcomes (such as achievement, in educational research) can be accounted for through prior impacts on mediating variables (such as instruction) (Baron and Kenny, 1986; Bloom, 2005a; Krull and MacKinnon, 2001), and analyses that investigate the role of implementation processes in producing impact (e.g. Hulleman and Cordray, 2009). These approaches either are experimental or involve further analysis of experimental outcomes.

More recent approaches that follow from the experimental paradigm and contextualize results to guide decision-making, include introducing planned variation into experiments to help with explanation (Mead, 2016), establishing criteria for deciding when, on the basis of implementation and impact studies, programs should be refined or abandoned (Epstein and Klerman, 2016), and methods that support generalization, by adjusting experimental impact estimates for differences between study sample and inference population, in the distribution of compositional and contextual variables that interact with treatment (Tipton, 2013).

Other mixed methods

Research purposes determine methods. In this work, we focus on needs and interests of program developers. They are concerned primarily with creating program prototypes as proofs of concept, in preparation for wider roll-out. A central purpose of evaluative methods is to identify points for improvement in the blueprint. Marginal impact estimates give a non-descript summary of the current prototype’s effects. Post-experimental methods help to identify process limitations in the blueprint, doing so by leveraging the experimental design. Still other methods can give deeper insight into why a program works, or has failed to work, and how it may be improved. This includes mixed- and qualitative methods. A useful way to think about this is that experimental and post-experimental methods indicate what of the conjectured program model worked or failed to
work, whereas qualitative methods help discover what failed to be conjectured in the program model in the first place. That is, qualitative approaches reveal blind spots in conceptualization of the program. Further, qualitative data, suitably coded, allow important integration with quantitative analysis, supporting a mixed-methods approach. A principal goal of this work is to illustrate one such innovative mixed method, which codifies qualitative data, and leverages the experimental design, to establish the causal connection between program implementation and impact. The diagnostic step proves critical to partially explaining why, in the case study that we present below, the marginal program impact was no different from zero.

Taking stock and turning to the method. We have made the case that randomized experiments in the social sciences are growing in popularity, and have important methodological advantages. However, they also have basic limitations. In particular, gross impact estimates from experiments must be contextualized to establish their reach or external validity. This may be accomplished in part using post-experimental approaches that experiments engender, which help us to understand the conditions for impact more broadly. Used with qualitative mixed methods, they are especially well-suited to addressing the interests of program developers whose goal is to understand the conditions for their program’s effectiveness and improvement. In particular, they allow a sequential process of diagnosis and remediation in cases where randomized experiments of a developing program fail to show an effect. Next, we turn to the first major aim of this work: we use the example of a program under development, which did not show a positive impact on key outcomes, to illustrate a seven-step diagnostic process that takes advantage of experimental, post-experimental and mixed (including qualitative) methods to inform the developer about where deficits lie in the process, and improvements can be made.

The second main goal, which we noted above, is to present one diagnostic step in detail. It uses an innovative methodology to identify facets of program implementation that have a bearing on the impact. The method is important in three respects: (1) it integrates the full program process, using baseline, implementation and outcomes data, (2) it shows how by leveraging an experimental design one can rigorously address important questions that normally are given scant attention because they are answered using weak correlational methods, and (3) it demonstrates the crucial role of qualitative data in interpreting average causal effects, and evaluating the conditions necessary for observing them.

Methods: a sequential approach to identifying process bottlenecks in experiments

The starting point: the program logic model

The first five steps in the diagnostic sequence target structures and processes within a blueprint of how the program of interest is hypothesized to work. (The last two steps focus on discovering important features of the program model not previously conjectured.) One such representation of a program blueprint is a logic model (LM) (Alter and Murty, 1997; Hernandez, 2000; Knowlton and Phillips, 2013; McLaughlin and Jordan, 1999). LM’s typically include program inputs, including key program components that are implemented and taken up by users and beneficiaries; outputs, which quantify this process; mediators, which are first-stage outcomes impacted by the program and that facilitate subsequent impacts; and short-, mid- and long-term outcomes. The LM also contains “external factors,” including baseline characteristics (BC) of participants that moderate—intensify or impede—impacts, and unintended consequences that may cycle back to affect the program process. Figure 1 shows a generic form of an LM.

The sequencing of processes in LMs makes them especially valuable for diagnostic efforts. LMs depict the input-to-outcome process in steps. Reasons a program fails to have impact may be conjectured in terms of obstructions in this process, and then empirically tested using various methods. A given test may focus on a specific section of the LM. For example, a test of overall impact—the first step for diagnosis—may be represented across the process, from deployment of inputs to points of impact. Alternatively, qualitative methods may zero-in on highly specific parts of the blueprint, such as on the role of external factors at certain junctures in the model, while mediator analyses may focus on specific linking processes along the chain of effects. Qualitative methods also may address the overall completeness of the blueprint, that is, if factors have been left out that the developer becomes aware of only after the prototype is tested.

The method proper: seven steps

The methodology we present consists of seven steps or conditions that must be satisfied before an overall impact can
be observed on important distal outcomes. It follows the processes of the LM. The conditions assessed may be obvious, including basic requirements of the implementation model—such as that an Internet-based program is delivered with adequate Internet connectivity—or less obvious, such as a positive relationship between a specific mediator and key outcome. The approach poses one or more hypotheses for each of the seven steps, and applies mixed methods to test them. The steps are illustrated in Figure 2.

A note about the organization of the Methods section: rather than present the methods first, then the example with results; we introduce the example first, then describe the method in stages with reference to the example. At each step, we provide background and citations, and describe the methodology used at that point of the troubleshooting process. This approach facilitates exposition. As noted above, a thorough description of Step 2 is a primary goal of this work.

We stress several points: (1) as often depicted in LMs, impact may be achieved along various paths, and in parallel. The troubleshooting process identifies points in the LM process that plausibly block the process and there may be several points of obstruction; (2) impact may be seen even with certain paths blocked, as processes may work in parallel; (3) the diagnostics are premised on a conjectured object (the LM)—if the LM is far afield from how the process actually works, the quantitative methods may not troubleshoot effectively; and (4) qualitative methods are essential, especially where the LM underrepresents the program model in actuality. The complementariness of quantitative and qualitative approaches is represented as a figure-ground picture in Figure 2: the “figure” lists possible weaknesses in the LM, while “the ground” leaves room for new propositions about the development process using feedback from alpha- or beta-implementations.

Data sources and results

The example of the Internet-Based Reading Apprenticeship Improving Science Education (iRAISE) intervention

Our example is an RCT used to investigate the efficacy of iRAISE. Its development, and accompanying process and impact studies were funded through the Investing in Innovation (i3) program established as part of the American Recovery and Reinvestment Act of 2009 (ARRA, 2009). Each i3 grantee is required to fund independently conducted implementation and impact evaluations. Our application of the seven-step diagnostic process was based on this evaluation.

iRAISE implements the Reading Apprenticeship Improving Secondary Education (RAISE) 65 hour face-to-face literacy professional development (PD) in an online format with science teachers. The adaptation to an online format was the focus of the grant. RAISE is an established program,
with its impacts evaluated in earlier studies (Kemple et al., 2008; Francsali et al., 2015). Reading apprenticeship (RA) theory postulates that students' literacy skills and content knowledge are impacted positively through an interactive learning culture that includes social, personal, cognitive, and knowledge-building dimensions. Teachers and students learn strategies for understanding complex texts, and a shift occurs from teacher instruction to teacher modeling and student-to-student learning in pairs and groups. (The LM for iRAISE is displayed in Figure S1 in Supplement 1.)

The RCT was conducted in the project’s second year in 25 districts in Michigan and Pennsylvania. After taking an extensive baseline survey, 82 high school teachers were randomized to iRAISE or Business-As-Usual. The program was implemented through initial summer training, with monthly booster trainings, and participation in online Professional Learning Communities (PLCs). Implementation focused on advancing literacy as “inquiry in science.” Monthly surveys assessed teachers’ use of instructional practices aligned to iRAISE theory, and their perceptions of helpfulness and usefulness of aspects of the program. At the end of the year impacts were measured on students’ reading literacy and comprehension using an assessment developed by the Educational Testing Service (ETS; O’Reilly et al., 2014).

The study met the criteria for the diagnostic process described in the current essay: (1) the program is under development, (2) an experiment is used to assess the average impact of the program, and (3) no impact is observed on a key outcome, in this case, on student reading literacy, (the standardized effect size was .001 standard deviation units, \(p = 0.93\)). Importantly, intensive data collection from multiple sources allowed use of post-experimental and qualitative methods in the diagnostic process as described earlier. Next, we consider, in sequence, the seven steps of the diagnostic process.

**Step 1. Assess the treatment-control contrast**

We hypothesize that impact on student achievement will not be observed if the treatment, as offered in the iRAISE group, is similar to what is made available to the control group.

**Background.** The control group, in field settings, often experiences one or a constellation of programs (Cronbach, 1982; Holland, 1986), described generically as “business as usual” (Hulleman and Cordray, 2009). Reduced contrasts on specific treatment components, because controls are implementing program components similar to treatment, may produce “positive infidelity” (Hulleman and Cordray, 2009). The result is reduced experimental impact attributable to those components.

**Observation.** iRAISE is a supplementary literacy program used in science classes, and we did not observe the offering or use of any systematic approaches to literacy instruction among controls. Science classes normally implicitly provide some opportunity to practice literacy; however, we did not observe use of explicit inquiry-based approaches.

**Conclusion.** Lack of a contrast between conditions on the key components of iRAISE is not a likely source of the bottleneck leading to the no impact finding.

**Step 2. Assess fidelity of implementation**

We address this step in considerable length given that it provides demonstration of an innovative mixed method for establishing the connection between fidelity of implementation (FOI) and impact. A central goal of this essay is to present this approach.

**Hypothesis.** We hypothesize that positive impact of the program on student achievement will not be observed if program implementation is substandard.

**Rationale.** We are interested, first, in whether FOI was achieved on average. If not, then it may explain lack of overall impact on achievement. If average implementation is adequate, and impact still is not observed, it may signify that what is considered adequate FOI, is unrealistic for impact, and even stronger implementation is required across the board. In that case, consistent with Improvement Science approaches, we are interested in what Bryk (2014) refers to as “positive deviants”: to assess whether there is variation in implementation, and identify the successful cases where stronger levels of fidelity are associated with greater impact. These cases may reveal the conditions necessary to achieve these successes more-broadly. On the other hand, if there is no impact even with increasing implementation, then even maximum achieved implementation in the study may be insufficient to produce impact. This signals that the impact process was wholly stymied elsewhere, drawing us to look further along the LM to identify potential points of obstruction to the program process.

**Background.** A challenge to assessing FOI for a given study is that there are many ways to conceptualize it (Hulleman and Cordray, 2009). These approaches may be differentiated in terms of the depth or extent to which they account for processes represented in the LM. For example, they may be limited to assessing whether the program was delivered as intended (provision of inputs). Alternatively, they may also account for uptake by the beneficiaries based on whether they responded to provision of materials and services (e.g. attended PD), or whether they subjectively reported those services as having value. Fidelity also may be assessed by different stake-holders and participants: study subjects may self-report perceived value of the program, study evaluators may survey adherence by program developer to program principles during PD, and program
facilitators may rate teachers’ adherence to program principles during PD interactions.

**Method: conceptualizing FOI.** We conceptualized FOI three ways. Participation as Action (“Action” for short): The fidelity study for i3 required identifying key components of the program, quantifying their levels of use, and setting a numeric threshold above which each component is considered implemented with fidelity. The thresholds were determined by the researchers in consultation with the program developer. The Action conceptualization is based on this information. Four indicators were used to assess teachers’ formal participation in the program through specific actions: attending foundational training, attending synchronous meetings (i.e. live online meetings), participating in PLCs, and completing and posting monthly assignments online. A value was assigned to each treatment teacher for each of these indicators. The data sources were program logs that recorded participants’ attendance at meetings and homework completion. Receptiveness: Beyond what was required by i3, we posited that greater receptiveness reflected more buy-in into the program, and that in addition to participation in the program (as captured through the Action indicators of implementation), buy-in and commitment are necessary for a program to take hold. Three Receptiveness indicators consisted of teachers’ self-reported levels of: preparation from training on several dimensions, helpfulness of monthly meetings, and helpfulness of the PLC. A value was assigned to each teacher for each of these three measures. The data sources were teacher surveys, with responses averaged over nine survey occasions for each teacher. Adherence: We considered also the rich reflections of the program facilitators. Four facilitators led iRAISE teachers through various aspects of the program over the course of the year. Specifically, the facilitators led online synchronous (all participants present online) and asynchronous (working online but at different times) activities, conferred with individual teachers, and provided feedback on homework. This frequent involvement with teachers allowed the facilitators, at the end of the study, to rate each teacher they worked with on four main dimensions of iRAISE: use of the core practices of the program, level of attention to student thinking, persistence in problem solving as related to the program, and use of a variety of texts in ways consistent with the principles of the program. The facilitators also were asked to provide rationales for their answers (which support the qualitative aspect of the diagnostic process discussed below). The facilitators’ scores were averaged across the four adherence dimensions for each teacher (“AVERAGE” in Table 1), and facilitators also produced a holistic rating (“HOLISTIC” in Table 1) for each teacher.

One way to think of the three measures of FOI—Action, Receptiveness and Adherence—is in terms of the depth to which they capture teachers’ commitment and involvement in the program, and their internalization and demonstration of program principles in practice. The Action metric is the most superficial, capturing only teachers’ participation in PD opportunities. Receptiveness goes further by measuring teachers’ dispositions toward the program. Adherence goes even further by using program experts’ ratings of individual teachers’ commitment to the program and fidelity to core principles in its use.

**Method: selecting metrics for FOI.** Next, we considered several approaches to scaling the implementation data, and combining the measures, to represent the logic of the process. The third measure, Adherence, was sufficiently different that it was considered separately from the other two. Specifically, it was based on more-objective ratings of PLC facilitators, rather than on teacher self-report, and second, as noted above, it tapped adherence to program principles as intended by the developer, and therefore yielded a deeper assessment of fidelity to the program.

### Table 1. Alternative measures and metrics for Fidelity of Implementation.

| Measure of FOI                      | Formulas (Metric)                                                                 |
|------------------------------------|----------------------------------------------------------------------------------|
| Adherence (AVERAGE)                | D(AVG) = D1 + D2 + D3 + D4                                                       |
| Adherence (HOLISTIC)               | D(HOL)                                                                          |
| Action                             | Compensatory                                                                     |
| Action and receptiveness            | CAR = (A1 + A2 + A3 + A4) + (R1 + R2 + R3)                                       |
|                                    | Multiplicative (conjunctive)                                                     |
|                                    | MA = A1 × A2 × A3 × A4                                                           |
|                                    | MAR = (A1 + A2 + A3 + A4) × (R1 + R2 + R3)                                        |

A1 “Action 1”: Days attended foundational training; A2 “Action 2”: Days attended monthly synchronous meetings; A3 “Action 3”: Days attended monthly Professional Learning Communities; A4 “Action 4”: Completion of monthly asynchronous assignments; R1 “Receptiveness 1”: Sense of level of preparation from training on various dimensions; R2 “Receptiveness 2”: Perceived levels of helpfulness of the monthly meetings; R3 “Receptiveness 3”: Perceived levels of helpfulness of the monthly Professional Learning Communities; D1 “Adherence 1”: level of use of the core practices of the program; D2 “Adherence 2”: level of teacher’s attention to student thinking; D3 “Adherence 3”: level of teacher’s persistence in problem solving as related to the program; D4 “Adherence 4”: level of teachers’ use of a variety of text in ways consistent with the principles of the program; D(AVG): “Adherence Average”; D(HOL): “Adherence Holistic.”; CA: “Compensatory(Action)”; MA: “Multiplicative(Action)”; CAR: “Compensatory(Action-Receptiveness)”; MAR: “Multiplicative (Action-Receptiveness).”

A1 was given 2.5 times the weight of each of A2—A4. (A1+A2+A3+A4) was weighted the same as (R1+R2+R3).
For the Action and Receptiveness measures, we followed the standards-setting literature in education (Haladyna and Hess, 2000) to specify the formulas for FOI. With a “conjunctive” (or “multiplicative”) approach, a low score on one indicator of FOI suppresses the contributions of another (like Christmas lights wired in-series—when one burns out, it stops the current). With a “compensatory” (additive) formula, a high score on one indicator can make up for a low score on another (like lights wired in parallel—if one burns out, the remainder stays lit). Table 1 summarizes the six metrics used to calculate FOI at the teacher level.

**Method: assessing average FOI.** To assess whether FOI was achieved on average, we focused separately on the Action and Adherence measures. Fidelity scores based on the Action metric were set according to developer-identified thresholds for what was considered adequate implementation. For Adherence, we report median levels of achieved fidelity, but no discrete threshold was identified a-priori. For both types of measures, in addition to assessing average levels of FOI, we estimated the variances, and determined whether greater FOI co-occurred with greater impact.

**Method: investigating changes of impact with increasing fidelity.**

The problem of a correlational analysis. One alternative for gauging the relationship between fidelity and impact is to examine the correlation between FOI and student achievement for the treatment (*iRAISE*) group only. This indicates whether levels of program participation, receptiveness to the program, and adherence to program principles by teachers are related to their students’ achievement. However, the validity of this approach is dubious because it does not deconfound levels of fidelity from characteristics inherent to individuals—formally, endogenous factors—that also influence outcomes. For example, a positive correlation between level of FOI and student achievement, may simply reflect that teachers with higher levels of incoming motivation achieve higher levels of fidelity, where motivation is causally related to achievement, but fidelity is not.

An alternative approach. A superior method is to assess performance of students of *iRAISE* teachers relative to that of students of similar control teachers—ones who would have achieved the same levels of FOI with *iRAISE*, had they been randomly assigned to treatment. While we cannot observe actual *iRAISE* fidelity outcomes for control teachers, the experimental design, supported by collection of strong baseline data, allows us to infer levels of FOI that plausibly would have been achieved among controls if they had been randomly assigned to *iRAISE*. The approach used, developed by Unlu et al. (2010) and based on work by Peck (2003) and Schochet and Burghard (2007), models the relationship between baseline data and observed FOI scores in the treatment group, then applies this model to infer values of FOI for teachers in the control group. Importantly, the method differentiates between high and low implementers using an index informed only by pre-randomization characteristics. That is, the index is a function of only exogenous variables measured at baseline (Unlu et al., 2010). Due to random assignment, within each stratum of the model-based implementation index, teachers are balanced, on average, between treatment and control conditions, on observable and unobservable traits. This allows an internally valid assessment of the causal impact of the program on student achievement at a given level of model-predicted FOI. The method, which combines aspects of implementation and moderation analysis, strategically builds on the randomized design to address a question important to the diagnostic effort. It exemplifies a post-experimental approach by retaining benefits of the experimental design while addressing important questions that give context to the marginal impact finding. The steps of the process, represented in Figure 3, are as follows:

1. Regress the measure of FOI against a series of BC in the treatment group.
2. Use the regression model from (1) to estimate predicted levels of fidelity (FOI*) in both treatment and control groups.
3. Assess the moderating effect of FOI* on impact on student achievement. (Moderators are attributes of individuals and settings, that exist before random assignment (and therefore could not be affected by that assignment), and that are associated with differences in impact. Because FOI* is derived entirely from baseline covariates, it functions essentially as a moderator of the impact of *iRAISE*).

We used linear regression to estimate the relationships between teacher baseline covariates and observed values of FOI in the treatment group. We selected from among

![Diagram](image-url)
the following teacher baseline covariates: (a) basic setting variables, such as which region a teacher’s school belongs to; (b) teacher averages of student baseline covariates, including average achievement and proportion of students eligible for free- or reduced-price lunch; (c) basic background characteristics of the teacher, including number of years in the profession, and having a master’s degree; and (d) measures of teacher attitudes and beliefs about literacy. The last of these was addressed through a survey on three dimensions: teachers’ perceived levels of confidence using different literacy strategies, self-reported levels of prior use of specific literacy strategies important to the iRAISE theory, and four items pertaining to teachers’ beliefs about the role of literacy instruction in teaching science.

With many teacher-level baseline covariates that could be used to predict FOI*, and only 34 teachers yielding outcomes in the iRAISE condition, it was necessary to reduce the number of covariates in the prediction model. We used three approaches: (a) Model A: We first used a forward selection algorithm to identify, from among variables pertaining to teacher attitudes to literacy, the eight that were most strongly related to the outcome. We forced these eight variables into the model, then forward selected eight more from among the set of general demographics variables, (b) Model B: We forward selected, from among all available covariates, the 16 covariates most strongly related to the outcome, (c) theoretical model: We chose items that either theory or common sense dictated would have a bearing on implementation, including scores for two subscales pertaining to instructional practices reported at baseline, four items asking teachers about levels of confidence with different instructional practices, and two items assessing teachers’ beliefs in the importance of reading literacy to the learning of science concepts. Added to this were four items reflecting classroom composition, including average pre-tests in reading and science, and three teacher variables, including years of science teaching.

Ordinary least squares regression was used to model the six FOI outcomes listed in Table 1: Compensatory Action (CA), Compensatory Action and Receptiveness (CAR), Multiplicative Action (MA), Multiplicative Action and Receptiveness (MAR), Adherence Average (DAVG), Adherence Holistic (DHOL), in terms of baseline covariates for each of the three models (Model A, Model B, Theoretical). This yielded six measures of FOI, crossed with three sets of variables for obtaining predicted FOI* values, resulting in 18 unique estimation models. Each model’s success at predicting FOI was assessed based on the \( R^2 \) value, which is the proportion of variability in FOI accounted by covariates in the treatment group, and the adjusted \( R^2 \) value.5

On obtaining FOI* scores, the primary objective was to assess their moderating effect on the impact of iRAISE on student achievement. To accomplish this, we expanded the hierarchical linear model (Raudenbush and Bryk, 2002) used to analyze impact on achievement, to include a term for the interaction between treatment status and FOI*, with the different measures of FOI* assessed one at a time. We focused on the statistical significance and magnitude of each estimate of the difference, across levels of FOI*, in the impact of iRAISE on student achievement.

Results. As noted earlier, there was no marginal impact of the program on student achievement (Standardized Effect Size = .001, \( p = .93 \)). As a whole, the level of delivery of the program by the developers, and their fidelity to iRAISE principles in that delivery, reached numeric thresholds representing adequate implementation; however, components of FOI used for the accountability study and the Action metric showed that teachers’ attendance at PD and doing assignments varied widely across teachers, and on average was just below the threshold set by the developers. For Adherence, the median holistic rating of teachers by program facilitators was 1 on an integer scale ranging from 0 to 5. In addition to assessing average FOI, we applied the method described above to examine the moderating effect of model-predicted FOI on the impact of iRAISE on student achievement. The results are summarized in Figures 4 and 5 (with detailed results in Table S1 in Supplement 1.)

First, we consider the adequacy of the model-based FOI measure. In Figure 4 we see that in most cases the covariates did a very good job of predicting observed FOI in the iRAISE group: Close to 80% of the variance in FOI scores is accounted for by baseline covariates (except in the Theoretical model). These results give us confidence that with most of the FOI metric-covariate combinations the prediction model imputes fidelity values close to those that would have actually been observed among the control teachers, had they been assigned to the iRAISE program.6

Next, we consider the result of main interest, which is the moderating effect of FOI* on the impact of iRAISE; that is, whether impact on student achievement increases for teachers demonstrating stronger fidelity. Results are summarized in Figure 5. Several findings are noteworthy: first, all except four of the \( t \) values for the estimates of the interaction between treatment status and FOI* are positive. Second, Adherence measures of FOI have higher \( t \) values, and two effects reach marginal statistical significance in their moderation of impact on achievement (\( p < .10 \)). The differential impacts for these two scales—the gain in impact associated with a 1-unit increase on the FOI scale—are .04 (Average, Model B) and .03 (Holistic, Model A) scale score units (.16 and .10 standardized effect size units, respectively). Because each of the scale scores are normalized to range from 0 to 1, the moderating effect represent impact gains associated with going from 0 implementation to the highest level achievable. The added gains, as standardized effect sizes, would be considered small by some standards (Cohen, 1988) but substantively important when translated into a metric of additional days of instruction (e.g. Newman et al. (2012) found that a standardized effect size of .05 translates to 28 additional days of schooling in
impacts on reading achievement). Reduced power to detect interaction effects (Jaciw et al., 2016a; Bloom, 2005b) may limit differences from reaching conventional levels of statistical significance among the Adherence measures.

A further noteworthy result is that the correlation between FOI as measured through participation-related teacher self-report (i.e. Action, Receptiveness) and as assessed through the facilitator’s ratings of Adherence (i.e. the average of ratings of Adherence to iRAISE principles) is low: for instance, the correlation in the treatment group between observed “Holistic” and “Compensatory Action” is .424 (p=0.022, J=29) and between “Holistic” and “Compensatory Action Receptiveness” is .627 (p<0.001, J=29). These alternative measures are addressing different and weakly related aspects of FOI, and one is moderately associated with greater impact, and the other is not.

**Conclusions, and revisiting hypothesis.** For this second step in the diagnostic process, we see first that FOI is multidimensional. PLC-leaders’ assessments of teacher adherence to program
principles are not strongly correlated to the Action and Receptiveness metrics; also, only the facilitators’ Adherence ratings to some extent account for variation in impact: the relationship between FOI and impact is moderate compared to the expected marginal impact on achievement the study was powered to detect (.20 standard deviations). It is important to reiterate that the moderating effect of model-based FOI differentiates impact between groups of teachers identified solely in terms of their baseline (pre-randomization) characteristics, but on an index highly predictive of actual FOI. The inference is stronger than one supported by a mere correlation between fidelity and achievement. Overall, we conclude that the lack of an average impact of iRAISE on achievement is in part due to low FOI among some teachers (it accounts for differences in the impact of approximately .03 to .04 scale score units); however, this deficit does not fully account for the less-than-expected effect, and we must look to the next stages in the pipeline of the LM to further troubleshoot the process and understand where the bottleneck(s) to impact occur.

**Step 3. Assessing the role of mediators of program impact**

We hypothesize that impact on student achievement will not be observed if (1) the program has no impact on posited mediating processes and (2) there is no association between those processes and achievement.

**Background.** Mediation analysis assesses whether impact on a more distal outcome, such as achievement, depends entirely or in part on prior impacts on intermediate outcomes. There are several approaches to assessing mediation, including use of instrumental variables, (Bloom 2005a), and regression-based approaches that assess the relationship...
Table 2. Impacts of iRAISE on twelve dimensions of instruction.

| Construct (M) | Teacher N | Effect Size | df | t  | p value |
|---------------|-----------|-------------|----|----|---------|
| 1. Teacher uses a variety of text types | 60 | 0.393** | 52 | 2.18 | 0.033 |
| 2. Teacher fosters student independence | 62 | 0.382*** | 53 | 2.18 | 0.034 |
| 3. Teacher uses traditional instructional strategies | 62 | 0.329* | 53 | 1.88 | 0.066 |
| 4. Teacher uses metacognitive inquiry in instruction | 62 | 0.236* | 52 | 1.7 | 0.095 |
| 5. Teacher models metacognitive inquiry in instruction | 62 | 0.250* | 52 | 1.75 | 0.086 |
| 6. Students practice metacognitive inquiry | 62 | 0.457*** | 52 | 3.14 | 0.003 |
| 7. Teacher uses reading comprehension strategies | 62 | 0.316** | 52 | 2.1 | 0.040 |
| 8. Teacher models reading comprehension strategies | 62 | 0.243* | 52 | 1.67 | 0.100 |
| 9. Students practice reading comprehension strategies | 62 | 0.516*** | 52 | 3.46 | 0.001 |
| 10. Students collaborate in classroom activities | 63 | 0.285 | 53 | 1.54 | 0.129 |
| 11. Students are engaged and participate in class activities | 62 | -0.136 | 53 | -0.7 | 0.487 |
| 12. Teacher is confident in providing literacy instruction | 63 | 0.619*** | 49 | 3.03 | 0.004 |

The degrees of freedom reflect the number of teachers in the analytic sample and the number of covariates used in the analysis. *p < 0.10 **p < 0.05, ***p < 0.01.

Traditional instructional strategies are not excluded from the iRAISE approach, but the focus is on how they are used.

Observations and results. Estimates of impact on the 12 posited mediators (stage 1) are displayed in Table 2 (expressed as standardized effect sizes). We observe positive impacts of iRAISE on 6 of 12 dimensions (p < 0.05) (or on 10 of 12, using a significance level of .10). The conditional associations between the mediating variables and student achievement in reading literacy (stage 2 of the mediation process) are displayed in Table 3. No effects reach statistical significance.

Conclusions and revisiting the hypothesis. We observed impacts on mediators (stage 1), but no relationship between mediators and student achievement (stage 2). Given the centrality of these mediating practices to the program, it is premature to conclude that instructional strategies impacted by iRAISE are unrelated to achievement outcomes. The mediating steps should be scrutinized further, especially their role in past studies where impacts on achievement were observed (e.g. Kemple et al., 2008) as well as in future studies. Stage 2 results may have limited accuracy in this study. For example, relationships between teacher practices and student outcomes may be more complex than represented in the linear model used for estimation, teacher self-report may not accurately represent what teachers actually do in the classroom, and lack of random assignment to the mediator could bias the results.

Step 4. Assessing the level of alignment between the outcome measure and the intervention

Given the scope and space of the current article, we summarize Steps 4 and 5 briefly.

Hypothesis. We hypothesize that a marginal impact of the program will not be observed if the outcome measures do not assess the domain targeted by the program.
Results and conclusion. The outcome measure was carefully selected to be aligned (but not over-aligned) to the program. It was independently developed by ETS, and its suitability was discussed frequently and in-depth (a representative from ETS was a member of Technical Working Group for a prior efficacy study of Reading Apprenticeship). The assessment had strong psychometric properties. (The instrument is described in detail in Jaciw et al., (2016b); Francsali et al. (2015)). We conclude that under-alignment of the assessment to the program is unlikely to be a source of the bottleneck limiting impact on achievement.

Step 5. Assessing influence of external factors and effect moderators

Hypothesis. We hypothesize that impact will not be observed if characteristics of students, setting, or context interact with the program so as to suppress its effect(s).

Results and conclusions. This is a two-part problem of external validity. First there is the question of whether the setting as a whole precludes impact: this would be the case, if, for example, the rural nature of the schools prevented impact because of issues with online connectivity. We address questions of overall accessibility through qualitative inquiry described in the next section. Second, there is the question of whether there are differential impacts within the sample (e.g. if impact for females is positive and impact for males is negative, then we may expect zero gross impact in spite of a differential effect). Moderator analyses showed a marginal differential impact of the program favoring students with lower incoming reading achievement, (p = 0.076), and no differential impacts across levels of student socioeconomic status (p = 0.180), level of English proficiency (p = 0.311), or science subject area (p = 0.168). Overall, we do not observe large and statistically significant deviations in impact across levels of a moderator that, when averaged over, result in zero gross impact. However, the finding of greater impact for students with lower incoming achievement is noteworthy, given that Reading Apprenticeship has previously been shown efficacious among low-achieving students (Kemple et al., 2008).

Steps 6 and 7: assessing implementation experiences reported by individual participants, including teachers, and iRAISE PLC facilitators

Hypothesis. Impacts will not be observed if there are unanticipated obstacles to implementation.

Method. For these steps, our focus is on identifying, not what of the planned LM failed, but what failed to be incorporated into the model. That is, we seek “blindspots” in the conceptualization of the current LM, and seek additional insights from close participants. We report iRAISE facilitators’ median ranks of teachers’ implementation levels on four ordinal adherence scales (the ones used also in Step 2: teachers’ use of the core practices of the program, teachers’ attention to student thinking, teachers’ persistence in problem solving as related to the program, and teachers’ use of a variety of text in ways consistent with the principles of the program). The scales ranged from 0 (no to limited implementation) through 3 (high to full implementation), and a separate holistic scale was used, ranging from 0 to 5. We summarized PLC facilitators’ reflections on each teacher and adherence scale using a grid with 16 cells (4 implementation dimensions × 4 levels of implementation). We then holistically profiled teacher responses at each implementation level. We also tabulated themes in teachers’ responses to open ended questions about challenges to implementation.7

Table 3. Estimates of the conditional association between the teacher-level mediator and student achievement in the iRAISE experiment.

| Construct                                                                 | Student N | Estimate | SE   | df  | t    | p value |
|--------------------------------------------------------------------------|-----------|----------|------|-----|------|---------|
| 1. Teacher uses a variety of text types                                  | 1435      | .008     | .010 | 1368| .81  | 0.416   |
| 2. Teacher fosters student independence                                 | 1435      | .003     | .003 | 1369| 1.02 | 0.310   |
| 3. Teacher uses traditional instructional strategies                      | 1435      | .001     | .003 | 1369| 0.46 | 0.644   |
| 4. Teacher uses metacognitive inquiry in instruction                     | 1116      | −.002    | .014 | 1052| −0.13| 0.898   |
| 5. Teacher models metacognitive inquiry in instruction                   | 1186      | −.018    | .013 | 1120| −1.40| 0.163   |
| 6. Students practice metacognitive inquiry                               | 1250      | −.007    | .013 | 1184| −0.52| 0.605   |
| 7. Teachers use reading comprehension strategies                          | 1273      | .014     | .010 | 1208| 1.44 | 0.151   |
| 8. Teachers model reading comprehension strategies                         | 1284      | .009     | .010 | 1218| 0.88 | 0.379   |
| 9. Students practice reading comprehension strategies                      | 1309      | −.004    | .007 | 1242| −0.50| 0.615   |
| 10. Students collaborate in classroom activities                           | 1435      | −.002    | .003 | 1369| −0.82| 0.411   |
| 11. Students are engaged and participate in class activities              | 1435      | .003     | .003 | 1369| 0.94 | 0.348   |
| 12. Teachers are confident in providing literacy instruction              | 1401      | .001     | .002 | 1334| 0.58 | 0.562   |

The mediators are averaged across nine survey occasions. The estimates are of the change in student achievement for each unit-increase in the mediator. The estimation model adjusts for effects of a series of student and teacher baseline characteristics. The degrees of freedom reflect the number of students per analysis.
Results. We address the PLC facilitators’ comments first. The facilitators rated 30 teachers. The median rating for each of the four 4-level integer scales (0–3) and the one holistic 6-level scale (0–5) was 1, indicating a regular pattern of implementation at the low end of the scale. The grid summarizing facilitators’ responses across levels for each dimension is displayed in Table S2 in Supplement 1. Profiling responses across the four dimensions, at level 0 we observe that teachers provide either no evidence of their implementation or that implementation involves adapting iRAISE to their existing programs or approaches. At this level, teachers focus on students as passive recipients of information. At Level 1, teachers implement sufficiently to experience challenges; however, they hold back, and find reasons to not implement fully, offering up excuses (e.g. they report struggling with learning themselves, that students do not have the ability, that students already know how to handle difficult text, or that it is not the science teacher’s role to teach literacy). At Level 2, teachers seem to break through into consistent use of core routines and putting student experiences at the center. They obtain genuine insight into the process, value the opportunity to dialogue with peers about the iRAISE process, and are not afraid to fail. Level-3 is characterized by an even more “learner-focused” approach, by a holistic appreciation of the program, a focus on the role of community including help and leadership roles, and nuance in the use of different text types. Interestingly, as the level of implementation increases, self-reported experiences of challenge also go up, reflecting that iRAISE is not just about resolution, but about being comfortable with the discomfort of challenge. Most frequent teacher responses to open-ended questions about challenges to implementation are as follows (with the number of instances they are mentioned across the 30 responding teachers indicated in parentheses): finding time and outside priorities (25), difficulty of getting student buy-in (9), finding appropriate text (7), lack of knowledge (7), difficulty changing their teaching routine (6), and issues with the on-line nature of program delivery (1).

Conclusions. Several unexpected themes emerged. First, the online nature of the program was not a major impediment; however, several teachers expressed preference for an in-school (as opposed to virtual) community of iRAISE practitioners to collaborate with. Several teachers noted a lack of subject-specific examples and text. Customization of implementation (including PD) by science subject may have added value. The iRAISE study underscores the challenge that unless teachers have high levels of buy-in, they will regard the program as just supplementary, to be done in addition to the work that really matters. In the case of iRAISE, additional integration of the program with science standards may help reluctant teachers to see that literacy and science instruction are not at cross-purposes. Developers may focus some of their next-stage efforts on Level-1 teachers who are at the tipping point of buying into the program, stressing growth through individual risk-taking in instruction, and on challenging teachers’ assumptions about the value of teacher-versus student-centered instruction.

Qualitative, including anecdotal, information reveals potential factors that contribute to the bottleneck that obstructs an overall causal effect. Some were not anticipated in the preliminary LM and may be remedied through adjusting the blueprint for the program (e.g. providing additional subject-specific material and instruction around modeling, and complementing the virtual community of users with a school-based network) while other “external factors” may be difficult to counter programmatically (e.g. the predominant sense that iRAISE is competing with other priorities).

Conclusions about iRAISE

What does the diagnostic remedial effort reveal in the case of iRAISE? What recommendations for remediation do we have to offer, given a no impact finding on student achievement?

It is an oversimplification to assume that the program process leading to impact is impeded at a single point. Rather, several factors in the LM lead to a gradually diminishing signal. The mixed methods converged to show a limitation of implementation: it was low (a median value of 1 on a scale of 0–3) on the Adherence dimension, which we consider the most-valid. Yet, even across the full range of implementation on this metric (from worst to best) impact on achievement varied, at most, by only .03—.04 scale score units (.10 and .16 standard deviation units, respectively), less than the .20 standard deviation average impact envisioned at the start of the study. Therefore, additional factors other than weak adherence must explain the failure to reach this target.

Teacher feedback indicated that, while the online delivery system was successfully implemented, and an online community was established, some teachers sought on-site collaboration. More generally, certain teachers requested a more customized solution, including materials that were geared to the specific science content area of instruction.

A further obstruction may be that instructional practices impacted by iRAISE fail to mediate the effect on student achievement. The two-stage mediation analysis suggests disconnects between specific instructional strategies and achievement. However, given that differences in adherence to the iRAISE model account for some of the variation in impact, the effect must move along some mediating channels. Given this and the fact that positive impacts of Reading Apprenticeship are observed in previous studies with specific populations, and given the importance of the instructional practices to the program, we consider the mediator results inconclusive.

As with many new supplementary programs, there is a dual challenge (discussed in footnote 4) to make the program effective among teachers who are ready to implement, and to motivate sluggish adopters, first, to see the value of the program, including that it can increase instructional efficiency, and second, once buy-in occurs, to maintain
fidelity in implementation. Separate mechanisms are driving the impact process for quick and slow adopters. The developer may consider different approaches to encouraging involvement. We noted that a breakthrough in implementation appeared to happen among Level-2 implementers (on the 0-3 Adherence scale), therefore encouragement and support efforts may be focused on Level-1 participants, to get them over the hurdle to implement successfully.

**Conclusion and discussion**

The focus of this work has been on how to conceive of and plan RCTs to be more than a method that yields merely a summative impact result. An average impact finding can be seen as a starting point, but by itself, provides little value for program improvement. With adequate planning and data collection, an RCT can be designed to support a diagnostic process that informs program remediation. In particular, in the event of a no impact finding, it allows tracing the program process and checking it in sequence to identify where the program model fails to represent what occurs in reality.

We presented a seven-step innovative diagnostic methodology for troubleshooting and identifying program limitations, focused specifically on development-stage RCTs. It includes “post-experimental” approaches built into randomized trials that retain specific advantages of experiments, including strong internal validity. The diagnostic process also draws on a range of mixed methods, designed to appraise the currently held LM, and identify potential blindspots that obscure both catalysts and impeders of program success. Direct feedback from individuals most involved in program development and implementation can challenge the program structures and processes as currently framed, even leading to a full overhaul of the program, which is appropriate especially for programs under development. We conclude with several ideas that are important for conceiving of and planning an RCT as a key component of a broader mixed-methods approach to program development and remediation.

**Strong and timely data collection are critical**

In diagnostic efforts where improvement cycles take time, quality of data and its collection are critical to getting the greatest possible yield from the effort. This includes, gathering strong data at baseline, mid-study and at posttest. At baseline, this consists of thorough measurement of factors potentially affecting outcomes, specifically attributes of students, teacher and contexts that may influence quality of implementation and impact. At mid-study, this includes assessing both potential mediators of impact and indicators of FOI. A central goal of this work was to illustrate a rigorous post-experimental method, with strong internal validity, which quantifies the relationship between program implementation and impact. The method was successful because various baseline data were collected to predict implementation levels, and strong and diverse measures of implementation were used (based on Action and Adherence conceptualizations) to investigate possible differences in their connections to impact on achievement. At posttest, the measure must be sufficiently representative of the construct that the program is designed to have impact on: even if all prior processes are intact and work as conjectured, if the outcome measure is mis- or under-aligned, impact will not register.

**Program impact is the results of an accumulation of factors and conditions; understanding their roles is revealing**

The mixed-methods diagnostic approach that we presented identifies process limitations gradually and in stages. An overall no impact finding does nothing to reveal the process of a gradually weakening signal. A no-impact results is not likely due to a single bottleneck that is amenable to a “magic bullet” solution. Furthermore, due diligence in identifying and measuring baseline factors that in theory moderate impact reflects cognizance that a program does not function as a panacea—it will probably work under specific conditions, for example, with certain teachers and students. If planned and designed properly, RCTs can yield information that helps to identify boundary conditions for observing impact, which may inform further development and adaptation of the program to expand such boundaries. Increased external validity is an important result of looking into the program (its mechanism) and outside of the program (to the conditions of implementation) to understand conditions necessary for impact, and thereby process improvement.

In the case of iRAISE, we identified several points where the program and its effects are expected to be variable or degraded: lack of customization of the program for specific science content areas, low adoption among hesitant teachers, and trends in moderation of impact favoring only certain subgroups, with greater impact for lower achieving students. Addressing any of these requires understanding why they are happening, and whether the program can address the reasons. Concerning program customization, a physics program emphasizing unification of content through core principles summarized through mathematical formulae may integrate a literacy program differently than a biology course with broad content coverage and that develops a different kind of knowledge base among students. This adaptation is asked for, but can the program expand to cover it? Concerning the challenge of overcoming certain teachers’ reluctance to use the program, the questions is how to assess and address deficits in interest and motivation. As we noted in Footnote 4, a single solution may not be adequate for everyone, if the goal is to improve implementation generally. Concerning the finding of positive impact for low achievers, the question is “why?” Additional theory of iRAISE, and concerning literacy skills acquisition generally, may help to explain and point to leverageable aspects of the program that further increase benefits for low achievers. The mechanism to accomplishing
this may not be the same as for increasing achievement of mid- to high-achievers who appear to receive less benefits from the current version.

**There are some tradeoffs to validity among the mixed methods used in the diagnostic-remedial process:** in combination they allow triangulation in producing evidence and a comprehensive account

The seven-step diagnostic process applies mixed methods in a social science context, and therefore shares the limitations of those methods. These are often expressed as tradeoffs to validity. A marginal impact result, while internally valid, reveals little about where to focus program improvement efforts. Opening the “black box” to examine program processes through post-experimental methods tells more about the possible “why” behind effects, or lack of them, but with less assurance of the causal role or connection among the parts of the system. Analyzing wider swaths of the LM with quantitative methods identifies trends on a broader scale, and limitations at one level of granularity, but this analytic lens has a specific focus limited by available statistical power; qualitative approaches give insight into the processes, and their failure, on a granular scale, suggesting heretofore not considered alternatives; however, sample sizes necessary for adequately powered inferential claims may be hard to achieve.

**Planning for what you know is only half the response, the other is knowing what you have not planned for**

In the multi-stage method Steps 1–5 are designed to take advantage of what we know about a program. Steps 6 and 7 focus on discovering what we do not yet know. We concur with Lee Cronbach’s (1975) point that with empirical generalizations “exceptions are taken as seriously as the rule.” Troubleshooting a program process and finding areas for improvement, especially for programs under development, means gathering data that challenge basic assumptions. In the iRAISE study, variability in and unexpected challenges to implementation were revealed, potentially focusing improvement efforts in new directions not yet identified in the LM.

**Acknowledgements**

We would like to thanks Robin Means for her contribution to this manuscript. We would also like to thank our colleagues in the Strategic Literacy Initiative at WestEd, and all study participants.

**Declaration of conflicting interests**

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

**Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The study reported in this article was funded by the Investing in Innovation (i3) program through the U.S. Department of Education.

**Supplemental Material**

Supplementary material for this article is available online.

**Notes**

1. http://edworkforce.house.gov/uploadedfiles/every_student_succeeds_act_-_conference_report.pdf
2. More formally, we are interested in the expected value of the causal effect of T = 1 relative to T = 0 across individuals randomly assigned to treatment, E(Y_i(T = 1) - Y_i(T = 0)|T = 1). We can rewrite this simply as: E(Y_i(T = 1)|T = 1) - E(Y_i(T = 0)|T = 1). It is possible to observe the former of these quantities, but not the latter (which is the expected value of performance of individuals assigned to treatment, had they been randomly assigned to control). However, because of random assignment, the latter quantity is equal to average performance of actual controls: E(Y_i(T = 0)|T = 0). This means we can estimate the average causal impact of the program as the difference between conditions in their average outcomes, an observed quantity: E(Y_i(T = 1)|T = 1) - E(Y_i(T = 0)|T = 0).
3. A full description of the program, and the impact study on which our example is based can be found in Jaciw et al. (2016b).
4. We briefly note an interesting paradox: It is reasonable to expect similar outcomes among strongly implementing teachers in both conditions because the “strong implementers” deliver their respective programs well (i.e. achieve strong implementation generally). And it is reasonable to expect similar outcomes among “weak implementers” in both conditions because they implement their respective programs poorly. If this is the case, how does impact arise, even in theory? It must be because weak implementers become stronger implementers under assignment to treatment, while among strong implementers, the treatment itself must equip teachers with methods and materials that provide advantage over business-as-usual (there must be material or service contrasts between treatment and control programs that influence achievement). Put another way, among the weak implementers, the program must both induce motivation to implement and establish a strong achieved treatment-control contrast, while for strong implementers the effect must depend largely on this contrast (assuming motivation is already present). This suggests two mechanisms at play that may be targeted separately by program developers.
5. It is beyond the scope of this work to discuss the technical aspects of the method being used to arrive at predicted values for fidelity. However, it is important to note that we are alert to the possibility of bias in the results, and addressed it. First, there is a concern that R^2 values may be biased from overfitting (the number of predictors is large, given the number of teachers in the analytic sample). Austin and Steyerberg (2015) show that reducing the number of subjects per variable leads to increased bias in the estimate of R^2; however, the adjusted R^2 value—which changes only if including a new term in the model improves model fit more than would be expected by chance—has much lower bias. We report both. Austin and Steyerberg (2015) also
demonstrate that regression coefficients remained unbiased even with as few as two subjects per variable. It is the accuracy of the regression coefficient that is most important for predicting fidelity in our application. As part of robustness checks of results reported in this article, we explored effects of reducing the number of covariates to 8 and then to 4. The predictive power of the models was reduced (as would be expected), but the patterns of results remained the same. To further address the possibility of overfitting, we also examined the effect of modeling fidelity with a random half-sample of treatment cases (Step 1 of the process) and applying it to the remaining half-sample and controls (Step 3), and repeating the procedure multiple times. The median values of these repeated analyses also revealed patterns of results like the ones we report in this article.

6. In Table S1 in Supplement 1, we also report the correlation between observed levels of fidelity (FOI) and predicted levels of fidelity (FOI*) which is close to .90 (lower for the Theoretical Model). This gives confidence in the correspondence between observed and model-predicted FOI. However, this result may to some extent reflect a degree of bias from over-specification, and we interpret it cautiously.

7. The approaches described here were feasible with the grant resources available. Ideally, more resources and additional data collection methods, such as use of classroom observations, would allow a more thorough assessment of implementation patterns by individual teachers.

References

Alter C and Murty S (1997) Logic modeling: A tool for teaching practice evaluation. Journal of Social Work Education 33: 103–117.

American Recovery and Reinvestment Act of 2009 (ARRA) (2009) Public Law no. 111–5§ 14007. Available at: https://www.gpo.gov/fdsys/pkg/PLAW-111publ5/pdf/PLAW-111publ5.pdf

Austin PC and Steyering EW (2015) The number of subjects per variable required in linear regression analyses. Journal of Clinical Epidemiology 68: 627–636.

Baron RM and Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 51: 1173–1182.

Berliner DC (2002) Educational research: The hardest science of all. Educational Researcher 31: 18–20.

Bloom HS (2005a) Constructing instrumental variables from experimental data to explore how treatments produce effects. In: Bloom HS (ed.) Learning More from Social Experiments. New York: Russell Sage Foundation, pp. 75–114.

Bloom HS (2005b) Randomizing groups to evaluate place-based programs. In: Bloom HS (ed.) Learning More from Social Experiments. New York: Russell Sage Foundation, pp. 115–172.

Bryk AS (2014) 2014 AERA distinguished lecture: Accelerating how we learn to improve. Educational Researcher 44: 467–477.

Chatterji M (2007) Grades of evidence: Variability in quality of findings in effectiveness studies of complex field interventions. American Journal of Evaluation 28: 239–255.

Chen HT (1990) Theory-Driven Evaluations. Newbury Park, CA: SAGE.

Cohen J (1988) Statistical Power Analysis for the Behavioral Sciences. New York: Taylor & Francis.

Cook TD and Payne MR (2002) Objecting to the objections to using randomized assignment in educational research. In: Mosteller F and Boruch R (eds) Evidence Matters: Randomized Trials in Education Research. Washington, DC: Brookings Institution Press, pp. 150–178.

Cronbach LJ (1975) Beyond the two disciplines of scientific psychology. American Psychologist 116–127.

Cronbach’s LJ (1982) Designing Evaluations of Educational and Social Programs. San Francisco, CA: Jossey-Bass.

Epstein D and Klerman JA (2016) On the “when” of social experiments. The tension between program refinement and abandonment. In: Peck LR (ed.) Social Experiments in Practice: The What, Why, When and How of Experimental Design and Analysis (Series: New directions for evaluation, no. 152). Malden, MA: Wiley Periodicals, Inc., pp. 33–45.

Fancsali C, Abe Y, Pyatigorsky M, et al. (2015) The Impact of the Reading Apprenticeship Improving Secondary Education (RAISE) Project on Academic Literacy in High School: A Report of a Randomized Experiment in Pennsylvania and California Schools. (Empirical Education Rep. No. Empirical_RAISE-7019-FR1-O.2), Palo Alto, CA: Empirical Education Inc. Retrievable from http://empiricaleducation.com/pdfs/RAISEFr.pdf.

Haladyna T and Hess RK (2000) An evaluation of conjunctive and compensatory standard-setting strategies for test decisions. Educational Assessment 6: 129–153.

Hernandez M (2000) Using logic models and program theory to build outcome accountability. Education and Treatment of Children 23: 24–41.

Holland PW (1986) Statistics and causal inference. Journal of the American Statistical Association 81: 945–960.

Hulleman CS and Cordray DS (2009) Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. Journal of Research on Educational Effectiveness 2: 88–110.

Institute of Education Sciences (n.d.) What Works Clearinghouse. Available at: https://ies.ed.gov/ncee/wwc/

Jaciw A, Lin L and Ma B (2016a) An empirical study of design parameters for assessing differential impacts for students in group randomized trials. Evaluation Review 40: 410–443.

Jaciw AP, Schellinger A, Lin L, et al. (2016b) Effectiveness of Internet-Based Reading Apprenticeship Improving Science Education (iRAISE): A Report of a Randomized Experiment in Michigan and Pennsylvania. (Empirical Education Rep.No. Empirical_iRAISE-7027-FR1-Y2-O.3), Palo Alto, CA: Empirical Education Inc. Retrievable at http://empiricaleducation.com/pdfs/iRAISEFr.pdf.

Kemple J, Corrin W, Nelson E, et al. (2008) The enhanced reading opportunities study: Early impact and implementation findings. NCEE 2008–4015, January. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Knovalton LW and Phillips CC (2013) The Logic Model Guidebook: Better Strategies for Great Results. Thousand Oaks, CA: SAGE.

Krull JL and MacKinnon DP (2001) Multilevel modeling of individual and group level mediated effects. Multivariate Behavioral Research 36: 249–277.
Lemire S, Christie CA and Inkelas M (2017) The methods and tools of improvement science. In: Christie CA, Inkelas M and Lemire S (eds) Improvement Science in Evaluation: Methods and Uses (Series: New directions for evaluation, no. 153). Hoboken, NJ: Wiley Subscription Services, Inc., pp. 23–34.

McLaughlin JA and Jordan GB (1999) Logic models: A tool for telling your programs performance story. Evaluation and Program Planning 22: 65–72.

Mead LM (2016) On the “how” of social experiments: Using implementation research to get inside the black box. In: Peck LR (ed.) Social Experiments in Practice: The What, Why, When, Where and How of Experimental Design and Analysis (Series: New directions for evaluation, no. 152). Malden, MA: Wiley Periodicals, Inc., pp. 73–84.

Newman D, Finney PB, Bell S, et al. (2012) Evaluation of the Effectiveness of the Alabama Math, Science, and Technology Initiative (AMSTI). (NCEE 2012–4008). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

O’Reilly T, Weeks J, Sabatini J, et al. (2014) Designing reading comprehension assessments for reading interventions: How a theoretically motivated assessment can serve as an outcome measure. Educational Psychology Review 26: 403–424.

Peck LR (2003) Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice. American Journal of Evaluation 24(2): 157–187.

Peck LR and Goldstein N (2016) Social experiments in practice: Introduction, framing and context. In: Peck LR (ed.) Social Experiments in Practice: The What, Why, When, Where and How of Experimental Design and Analysis (Series: New directions for evaluation, no. 152). Malden, MA: Wiley Periodicals, Inc., pp. 9–17.

Raudenbush SW and Bryk AS (2002) Hierarchical Linear Models (2nd edn). Thousand Oaks, CA: SAGE.

Rohanna K (2017) Breaking the “Adopt, Attack, Abandon” cycle: A case for improvement science in K-12 education. In: Christie CA, Inkelas M and Lemire S (eds) Improvement Science in Evaluation: Methods and Uses (Series: New directions for evaluation, no. 153). Hoboken, NJ: Wiley Subscription Services, Inc., pp. 65–77.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66: 688–701.

Schochet PZ (2009) Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? NCEE 2009-4065, 13 October. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Schochet PZ and Burghard J (2007) Using propensity scoring to estimate program-related subgroup impacts in experimental program evaluations. Evaluation Review 31: 95–120.

Shadish WR, Cook TD and Campbell DT (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin.

Spybrook J (2013) Detecting intervention effects across context: An examination of the precision of cluster randomized trials. Journal of Experimental Education 82: 334–357.

Tipton E (2013) Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. Journal of Educational and Behavioral Statistics 38: 239–266.

Unlu F, Bozzi L, Layzer C, et al. (2010) Linking implementation fidelity to impacts in an RCT: A matching approach. In: Paper presented in symposium: Using matching methods to analyze RCT impacts on program-related subgroups, Association for Public Policy Analysis and Management, 13 October. Available at: http://www.aepweb.org/sites/default/files/webform/Unlu_InPress.pdf

Author biographies

Andrew P Jaciw is the chief scientist for Empirical Education Inc. He specializes in the evaluation of programs in education, including the design and analysis of group randomized trials. His interests include research on programs in mathematics education, socio-emotional and behavioral interventions, and alternative approaches to generalizing results from impact evaluations.

Li Lin is a chief statistician at Empirical Education Inc. She specializes in modeling and analysis of data from experiments, quasi-experiments and observational studies. Her focus includes statistics, economics, business administration and computational modeling.

Adam Schelling is a research analyst at Empirical Education Inc. He works to collect, manage, and analyze data for evaluations of education technology products.