A Five-Point “Systems Check” for Efficacy Studies of Programs under Development (Especially in the case of No Impact Findings)

Andrew P. Jaciw, Jenna Zacamy, Adam Schellinger and Li Lin
Empirical Education Inc.

Background

When a randomized control trial (RCT), designed to assess the efficacy of a program under development, demonstrates no impact, it provides an important opportunity to revisit the program blueprint and evaluate process limitations. Lack of impact may be from deficits in Fidelity of Implementation (FOImp) or Fidelity of Intervention (FOInt), (Hulleman and Cordray, 2009) or from other conditions not being met.

We propose a series of “post-experimental” exploratory analyses (e.g., methods developed by Unlu et al., 2009, applied in this work) as “systems checks” that diagnose where steps in the program logic model are obstructed. Opening the “black box” of the process to provide formative feedback and inform the next iteration of development, however, requires comprehensive data collection, to support the kinds of analyses described here.

Purpose

The purpose of the work is to demonstrate a 5-point diagnostic framework for identifying program processes limitations. Application of the framework informs developers about where potential weaknesses lie in their program logic model, allowing program remediation and further development. We illustrate the approach using data from an experimental impact evaluation.

Setting

The experiment was conducted in 27 rural high schools in Michigan and Pennsylvania, during the 2014/15 school year, and included 82 science teachers and 1468 students. It was conducted through the Investing in Innovation (i3) Fund.

Intervention

iRAISE (internet-Based Reading Apprenticeship Improving Science Education) is a 65-hour literacy professional development program delivered through an online format. It
is a year-long learning community in which high school science teachers practice and refine ways to improve their students’ ability to understand scientific texts. Elements of the program implementation are described below, under Methods.

**Research Design**

Impacts were assessed using a cluster randomized trial with 82 teachers randomized within high schools, to iRAISE, or Business as Usual (wait-listed) condition. Extensive baseline data were collected on teachers, and impacts were assessed after one year on measures of instructional practice and students’ reading literacy in grades 9–12.

**Data Collection and Analysis**

**The Systems Check**

To provide the developers with a “systems check” of their program process, under the condition where iRAISE had no impact on achievement, we investigated five hypotheses: *The program does not have impact because of:*

1. A weak treatment-control contrast,
2. Lack of fidelity of implementation (FOImp),
3. Lack of fidelity of intervention (FOInt) (Hulleman and Cordray, 2009),
4. Misalignment between skills tested, and skills targeted by the program,
5. Effect moderation (misalignment of population of individuals who would benefit from the program and the study sample.)

In the framework, these hypotheses are neither mutually exclusive nor exhaustive; however, they provide a sequential test of the process. Each is explored independently; however, failure at a given stage is seen as a potential process limiter along a sequence that is seen to operate “in series”.

Given the brevity of the proposal, we focus on methods for (2). Assessment of (1) and (4) relies on a review/analysis of what the counterfactual consist of, and comparison of assessment specifications and program blueprints. We discuss these, and (3), briefly in the results.

**Methods for assessing hypothesis (2)**

We assessed whether impact on achievement varies by level of FOImp, and is present under conditions of strong FOImp. If the investigation shows no impact, even with high FOImp, then we explore the next potential process limiter: a lack of FOInt.
To operationalize FOImp we used indicators, measured through the i3 evaluation, of teacher participation in, and receptiveness to, the program. There were four “Action” indicators quantifying participation in: (a) initial 5-day training, (b) monthly online training, (c) monthly online Professional Learning Communities (PLC’s), and (d) completion of assignments. There were three “Receptiveness” indicators assessing perceptions of (a) preparedness from training, (b) helpfulness of online training and (c) helpfulness of PLCs. Based in these indicators, FOImp was formulated four ways (allowing flexibility in operationalization):

MA(Multiplicative Action): “Action” indicators were weighted and multiplied; this “conjunctive” (stricter) approach was such that a low score on any indicator would limit the contribution of any other.

CA(Compensatory Action): “Action” indicators were averaged, allowing high scores on some indicators to compensate for low scores on others.

MAR (Multiplicative Action/Receptiveness): Receptiveness indicators were multiplicatively combined with the MA measure.

CAR (Compensatory Action/Receptiveness): Receptiveness indicators were averaged with the CA indicators.

To assess whether FOImp influences impact on achievement, we used a method developed by Unlu et al., (2009) based on work by Peck (2003) and Schochet and Burghard (2007). The steps of the process are as follows (displayed in Figure 1):

1) Regress the measure of FOImp against a series of baseline characteristics (BC) in the treatment group.

2) Use the regression model from (1) to obtained predicted levels of fidelity of implementation (FOImp*) in both treatment and control groups.

3) Estimate the moderating effect of FOImp* on impact on student achievement, and average impact in the upper range of FOImp*. (Because FOImp* is based entirely on baseline characteristics, it functions essentially as a moderator of the impact of iRAISE.)

---------------------------------------------
INSERT FIGURE 1 HERE
---------------------------------------------

Importantly, because FOImp* is a function of pre-randomization characteristics, and treatment status is determined through random assignment, the FOImp* × Treatment interaction is exogenous, and its estimate is unbiased (as described in Unlu et al., 2009).
With an extensive set of informative baseline measures pertaining to teachers’ levels of confidence in use of reform based methods in instruction, and their beliefs about the role of literacy in science instruction, we predicted FOImp* three ways: two regression models where covariates were chosen through algorithmic forward selection (S1, S2), and one where covariates were selected using program theory (PT).

**Results:**

Summarized in Table 1, we observe no differential impact across levels of FOImp* for any of the 4 FOImp metrics × 3 FOImp* prediction models = 12 analyses. Further, there was no impact at the 75th percentile of FOImp* with any of these approaches.

---

**Table 1**

In the case of iRAISE, the treatment-control contrast was strong, the test of reading literacy assessed skills targeted by the program, and in spite of this, even with increasing (and strong) FOImp*, no impact was observed. In the sequence of systems checks, we next turned our attention to possible limitations in Fidelity of Intervention (FOInt). As we describe in the full paper, we observed positive impacts of iRAISE on posited instructional mediators, but weak relationships between mediators and achievement. These analyses of FOInt were non-experimental; however, in the case of iRAISE, they point to a potential process limitation in the second stage of mediation – impacts on mediators did not translate into impacts on achievement (i.e., described as a “conceptual theory failure” as opposed to an “action theory failure” (Chen, 1990; Krull and MacKinnon, 2001)).

**Conclusion**

The five-stage systems check applied to the iRAISE experiment helped us to point the developer to program processes where improvement efforts can be focused. While a lack of impact may be attributed to other process deficits as well, the framework we propose provides a systematic approach to identifying program limitations and areas for improvement.
Table 1. Differential Impacts across the Range of Model-Predicted Fidelity Implementation

<table>
<thead>
<tr>
<th>Implementation Metric</th>
<th>R-squared (Proportion of Variance in FOImp* explained by teacher-level baseline covariates)</th>
<th>Correlation between FOImp (i.e., measured) and FOImp* (i.e., predicted) in the treatment condition</th>
<th>Differential impact across levels of predicted implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>Model S1: .807, J=35, R-squared = .903 (p&lt;.001), J=35</td>
<td>-0.018 (SE=.041), DF=1396, t=-.45, p=.653, J=68, J(T)=35, J(C)=33, n=1462,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model S2: .843, J=33, R-squared = .918 (p&lt;.001), J=33</td>
<td>0.022 (SE=.029), DF=1348, t=.75, p=.450, J=66, J(T)=33, J(C)=33, n=1414,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model PT: .416, J=34, R-squared = .645 (p&lt;.001), J=34</td>
<td>-0.009 (SE=.047), DF=1368, t=-.20, p=.844, J=67, J(T)=34, J(C)=33, n=1434,</td>
<td></td>
</tr>
<tr>
<td>MA</td>
<td>Model S1: .844, J=34, R-squared = .919 (p&lt;.0001), J=34</td>
<td>0.024 (SE=.030), DF=1368, t=.78, p=.435, J=67, J(T)=34, J(C)=33, n=1434,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model S2: .758, J=34, R-squared = .872 (p&lt;.0001), J=34</td>
<td>0.015 (SE=.026), DF=1368, t=.78, p=.563, J=68, J(T)=34, J(C)=34, n=1462,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model PT: .384, J=34, R-squared = .620 (p&lt;.0001), J=34</td>
<td>0.032 (SE=.040), DF=1368, t=.81, p=.420, J=67, J(T)=34, J(C)=33, n=1434,</td>
<td></td>
</tr>
<tr>
<td>CAR</td>
<td>Model S1: .844, J=34, R-squared = .919 (p&lt;.0001), J=34</td>
<td>0.024 (SE=.030), DF=1368, t=.78, p=.435, J=67, J(T)=34, J(C)=33, n=1434,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model S2: .758, J=34, R-squared = .872 (p&lt;.0001), J=34</td>
<td>0.015 (SE=.027), DF=1368, t=.58, p=.563, J=68, J(T)=34, J(C)=34, n=1462,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model PT: .384, J=34, R-squared = .620 (p&lt;.0001), J=34</td>
<td>0.032 (SE=.040), DF=1368, t=.81, p=.420, J=67, J(T)=34, J(C)=33, n=1434,</td>
<td></td>
</tr>
<tr>
<td>MAR</td>
<td>Model S1: .912, J=34, R-squared = .955 (p&lt;.0001), J=34</td>
<td>0.016 (SE=.036), DF=1396, t=.45, p=.656, J=68, J(T)=34, J(C)=34, n=1462,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model S2: .685, J=34, R-squared = .838 (p&lt;.0001), J=34</td>
<td>0.017 (SE=.049), DF=1396, t=.34, p=.733, J=68, J(T)=34, J(C)=34, n=1462,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model PT: .425, J=34, R-squared = .652 (p&lt;.0001), J=34</td>
<td>0.070 (SE=.057), DF=1396, t=1.24, p=.214, J=68, J(T)=34, J(C)=34, n=1434,</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1:
The Steps in the Process of Estimating Differential Impact Across Model-Predicted Level of Fidelity of Implementation
References:


