Impact Evaluations in Real-World Contexts: Prudence, Adaptation, Transparency

Andrew P. Jaciw & Thanh Nguyen, Empirical Education Inc.

Presented April 14, 2018 at the annual meeting of the American Educational Research Association

Over the past decade, education policy has sharpened its focus on ensuring that decisions about education programs, products, and services are evidence-based. Studies are scrutinized under different sets of review standards, including but not limited to the What Works Clearinghouse (WWC), Evidence for ESSA, and other criteria used by the larger research community. We consider three cases arising from two cluster randomized trials where real-world contexts present challenges to reaching conclusions. For each case, we briefly describe the issue at hand, discuss the implications, and suggest a possible response. We argue that while the official charge and evidence standards guiding impact evaluations are important, evaluators and researchers must report the full details to allow the broader research community and direct beneficiaries to judge the relevance and validity of the results to advance knowledge and transparency.

Case #1: When Education Standards and Assessments Change

Background

We are conducting a multi-year, multi-site Investing In Innovation (i3) evaluation of a science teacher professional development program that connects hands-on science with integrated teaching and literacy supports.

Challenges

- The evaluation is being conducted while science education standards are being dramatically reframed. The Next Generation Science Standards (NGSS) are being adopted by states across the country, prompting modifications in instruction and assessment of student science achievement.
- Even though the study design called for a measure of student science achievement as a confirmatory outcome, by the time of data collection, there was not yet an established, valid, and reliable instrument available for the study's use.

Actions taken

The lack of an NGSS-aligned instrument compelled researchers to develop an instrument that requires strong reliability and construct and face validity, while ensuring that the instrument would be accepted by independent review and not considered over-aligned to the intervention.

Lessons learned and recommendations

- Clear division of labor between researcher and program developer to provide evidence of independence.
- Detailed documentation of item development, including source, permission to use, and other characteristics (content domain, grade, DOK level, etc.) for full transparency and posterity.
- Report item and form-level statistics from pilot rounds to provide rationale for researchers' decisions during the process of instrument construction.
- Provide contextual information and rationale for critical decisions as they are made to help preserve collective memory.

Case #2: Internally Valid Result Conflicts with Real-World Causal Quantity

Background

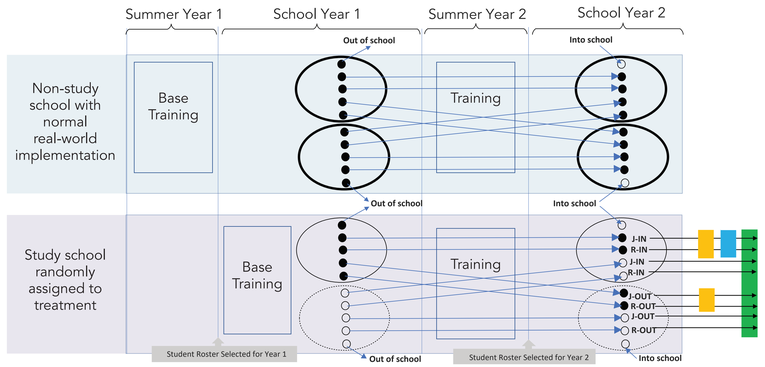
Same as case #1

Challenges

- A study design aimed at achieving internal validity is sometimes at odds with the realities of the school context (priorities, schedules, resources, etc.).
- Evidence standards were being modified as the evaluation was occurring, and the study design would be reviewed under the new (and unknown to the evaluator) standards.

Lessons learned and recommendations

- Use a design that accommodates anticipated changes in evidence standards, if it is practical and sensible.
- Empirically test and report the assumptions concerning internal validity.
- Pursue balanced treatment of validity. Assess and discuss the tradeoffs among the different types of validity given certain design constraints.



	Teacher in non-study school implementing the program
	Teacher in study treatment school who joined study before randomization
	Teacher in study treatment school who did not join study over the course of the trial
●	In study school: grade 4 student in Y1 who is on roster of study teacher in grade 4
	In non-study school: grade 4 student who is present in school at start of implementation who gets at least some to full exposure to program
○	In study school: grade 4 student in Y1 NOT on roster of study teacher in grade 4
	In non-study school: student not present in school at start of its implementation but who joins implementing school within 2 years
J-IN	Student who jockeys to be IN the program in year 2
J-OUT	Student who jockeys to be OUT of the program in year 2
R-IN	Student who is randomly IN the program in year 2
R-OUT	Student who is randomly OUT of the program in year 2
	We include as many of these students as possible (i.e., from baseline sample) in the impact analysis to limit attrition so that result meets standards w/o reservations
	These are students for whom we can assess impact of receiving full implementation
	Students who would receive a full two years of the program if all teachers in both grades participated in the study (it will include students in J-IN, J-OUT, R-IN, R-OUT but where there is no opportunity to Jockey-OUT.)

Case # 3: Transparency in Reporting Confirmatory and Exploratory Findings

Background

The result in question was an exploratory finding from a randomized control trial of a statistically significant differential impact of a math, science, and technology program, such that a traditionally privileged subgroup received a modest positive impact, while an underprivileged group received zero benefit.

Challenges

- The exploratory research question may be consequential, but if not declared confirmatory in advance, requires replication for corroboration (to avoid "fishing" and reaching a false-positive conclusion.)
- A replication effort may never be funded.

Lessons learned and recommendations

- Better to provide more information rather than less, to make explicit the standards being applied and to find venues to presents results from different perspectives.
- Conduct a multi-armed study that incorporates the standard version of treatment and an "improvement version" that is specifically designed to address the deficits identified through secondary analysis. The goal is to replicate the secondary result (as recommended) while working on improvement, because we judge the potential cost of no impact for the minority group to be large (if it is real.)

One Year Impact of a Math Science and Technology Intervention on Achievement in Multiple Subject Areas in Elementary and Middle Schools