Pilot studies of educational interventions: How to conduct them and what we can learn from them

Organizer and Chair

Robert Olsen
George Washington University, GW Institute of Public Policy
robolsen@gwu.edu

Discussant

Anne Ricciuti
U.S. Department of Education, Institute of Education Sciences
anne.ricciuti@ed.gov

Symposium justification

The term "pilot study" seems to mean different things to different people. It can be used to describe an early test of study *procedures* to ensure that the research team is ready to implement them. Alternatively, it can be used to describe an early test of *an intervention* to assess its implementation, estimate impacts, or both.

Pilot studies play an important role in evaluations of educational interventions. The U.S. Department of Education, the largest sponsor of evaluations in education, expects its grant recipients to pilot test their interventions before requesting additional funding for broader implementation and testing. This expectation is reflected in the requirements of the Education Innovation and Research (EIR) program and the requirements of the Education Research Grant program. Such requirements are typical of tiered evidence programs, which provide more financial support for interventions with a more extensive base of supporting evidence.

Despite the importance of pilot studies in educational evaluations, there has been relatively little research or discussion at conferences like SREE on the potential value of pilot studies and how to conduct them. Obvious questions include:

1. Should pilot studies focus on implementation fidelity—leaving impacts for later studies if fidelity proves to be adequate? Or should they assess both implementation and impacts?
2. Should pilot studies be large enough to detect impacts and rigorous enough to meet the What Works Clearinghouse standards? Or should they just be minimally adequate to establish whether the intervention is "promising"?
3. How should the results from pilot studies be used? Should we use them to justify a more rigorous test and broader implementation? Or perhaps to set sample size targets for future studies?

4. What challenges do organizations face in conducting pilot studies? And how can these challenges be addressed?

This symposium will address many of these questions and facilitate a question and answer period focused on questions like these. The goal of the session is to take an initial step toward consensus about how to conduct pilot tests and what we can learn from them.

The first presentation, by Elizabeth Stuart from the Johns Hopkins Bloomberg School of Public Health will explain how pilot studies of the impact of an intervention should be used; it also highlights the pitfalls from putting too much weight on the results from a single pilot study based on a small sample. The second presentation, by Alexandra Resch of Mathematica Policy Research, will discuss the obstacles that school districts face in conducting pilot studies and how these challenges can be addressed based on her experience leading Ed Tech RCE Coach, a free online platform to support the conduct of pilot studies of technology-based educational interventions.

Presentation #1: The nonuse, misuse, and proper use of pilot studies in education research

Presenter's name, affiliation, and email:

Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health
estuart@jhu.edu

Co-author

Erik Westlund
University of Iowa
erik-westlund@uiowa.edu

Abstract

Background/Context:

When designing and conducting experimental research, pilot studies—small-scale studies conducted before a full trial—are considered a best practice, and represent a key step in the Department of Education Institute of Education Sciences grants structure. In education research, where experiments are the gold standard of research design, pilot studies should be an integral part of the research design

process. Yet, there is little explicit guidance in the literature to guide researchers who seek to incorporate pilot studies into experimental evaluation research.

Objective:

This article has three aims. First, we document the dearth of guidance available to researchers conducting pilot studies as part of experimental research. Using program evaluation in education research as a case study, we also show that although pilot studies are required by grant protocols of major education research grant-making bodies, methods for their appropriate use are rarely discussed or provided. As such, we show how researchers seeking guidance on how to use pilot studies in experimental evaluation research will find little guidance from either theory or practice. Second, we discuss how pilot studies can be misused in experimental evaluation research.

We focus on two particular errors: using pilot studies to decide whether to conduct a full trial of an intervention and using pilot studies to determine how many cases to sample in such a full trial. Third, we draw on the literature, our experience in reviewing grant protocols, and the information gleaned from these simulations to offer practical advice on how to responsibly use pilot studies in experimental research.

Research Design:

We did a systematic review of textbooks on program evaluation and papers in the education research literature to identify whether any papers or books give concrete guidance regarding pilot studies. We use mathematical simulations to illustrate how each of the above practices can result in researchers making poor research

design choices. To run our simulations, we developed an open source, publicly available web application using the R programming language and the shiny web framework (http://pilotpower.table1.org/). The program allows researchers to set various assumptions about the true effect size of an intervention. It also allows them to choose decision rules for interpreting pilot study results. This framework is also used to clearly illustrate the results described below.

Findings:

None of the books or papers identified offered practical advice directed specifically at experimental evaluation researchers, although a number of existing studies provide relevant insights. There are also a number of findings with respect to the dangers of using pilot study results to explicitly plan future studies. First, using unbiased effect sizes from underpowered pilot studies to decide whether to conduct a full trial is considerably error-prone and volatile practice. The typical sample size of a pilot study (in our simulations, 50 cases) is often too small to detect true effects when they exist, resulting in premature abandonment of trials. Even applying lenient rules for proceeding to a full trial leads to prematurely abandoning effective interventions in a high number of cases. Conversely, applying these lenient rules to an ineffective intervention results in mistakenly conducting full trials in a substantial number of cases. Both of these results are a consequence of the uncertainty inherent to conducting studies with small sample sizes. Second, using observed pilot study effect sizes as the basis for power calculations in full trials results in significant loss in statistical power in comparison to basing power calculations upon a predetermined practically significant effect size. Third, using observed pilot study effect sizes as the basis for proceeding to trial and then using the same effect size as a basis for power calculations for those full trials results in systematically overestimating the effect size in cases where statistically significant effect sizes are detected in those full trials. The closer in magnitude the true effect size is to the practically significant effect size, the more likely the true effect size will be overestimated in the full trial. Put succinctly, using pilot study effect sizes to power full trials leads to incorrectly judging effective interventions to be ineffectual, and judging effective interventions to be more effective than they are.

Conclusions:

Pilot study results should be used as one data point among many, possibly including data from past non-experimental research and solid theoretical models, to determine whether a full trial is worth conducting. To avoid Type 2 errors and overestimating the true effectiveness of interventions, full trials should always be powered to detect the practically significant effect size. While it may be alluring to save resources by sampling fewer cases in light of a large observed pilot study effect size, this always results in conducting an underpowered (vis-à-vis the predetermined level of practical significance) full trial. However, it is important to note that pilot studies can be a key tool to help determine the feasibility of carrying out a larger trial of an intervention. For example, pilot studies allow researchers to judge the adequacy of recruitment and consent procedures, to evaluate the acceptability of randomization procedures, to gauge the quality of measurement instruments, and to assess whether compliance with the intervention in the field will be sufficient to achieve an intervention's goals.

Presentation #2: What can we learn from small pilots conducted by school districts? Lessons from the Ed Tech RCE Coach

Presenter's name, affiliation, and email

Alexandra Resch
Mathematica Policy Research
aresch@mathematica-mpr.com

Abstract

Background/Context:

As school districts across the nation invest in technology, they need timely, reliable evidence on the effectiveness of educational technology (ed tech) products and on the strategies for implementing them. Many educators feel that traditional program evaluation does not meet this need. New ed tech products are being released every day, but traditional, large-scale evaluation can cost more than the development of the product itself. At the same time, educators want to know what works in the classroom: what tools and strategies are successful at moving the needle for their students? Locally implemented pilot studies provide an opportunity for educators and district leaders to get results more quickly and to assess whether ed tech tools and other interventions work in their local context.

The Ed Tech RCE Coach (the Coach) is a free online platform where any user can conduct a rigorous evaluation with their own data using a matched comparison group or randomized pilot. The Coach was designed to help districts rigorously (and quickly) evaluate the ed tech tools that they are using. The Coach was also designed to meet districts where they are without assuming that they have done research studies or data analysis in the past. The Coach's tools build capacity for and comfort with evaluation so that educators and district leaders can slowly progress toward more rigorous, and therefore convincing, evaluations over time. This focus on capacity building and local control was designed in part to bridge the gap between research and practice – if practitioners are empowered to conduct the research, the resulting evaluations will be more likely to answer their questions and address their needs.

Objective:

This paper has two objectives. First, we document lessons learned about whether it is feasible for school districts to conduct rigorous and useful pilot studies on their own. Second, we propose approaches for aggregating and analyzing results from these local pilots to begin to build knowledge that will be useful to the broader community.

Research Design:

As of September 2017, users from school districts and charter networks have completed and made publicly available approximately twenty evaluations of local ed tech initiatives and other interventions. These evaluations, and additional evaluations currently in process, provide an opportunity to examine what can be learned from small, localized pilot studies. The shared findings contain details on the products or interventions studies, characteristics of the districts and participants, the analysis choices made by the user, and the ultimate findings of the analysis. We also sought feedback from Coach users and received additional information about their use of the Coach through emails, user interviews, and debrief phone calls. In supporting districts in using the Coach, we also observed several examples of districts that started evaluations, but did not complete them, and we recorded observations from these examples as well.

We draw from the evaluation briefs and feedback to summarize the types of interventions tested, the research designs employed, and characteristics of the participants. We use the experiences of these districts to describe challenges and opportunities observed in the evaluations completed to date.

Findings:

Our examination of the completed evaluations to date suggest that local evaluation is desired and valued, but that users struggle with many parts of the process. Rigorous evaluation is new to many district-based users, so there is a learning curve to conducting local studies and using the Coach. Inconsistency in the accessibility, format, and content of usage data for ed tech products poses a challenge for many users. Districts with some evaluation capacity see the Coach as a way to expand their capacity and to get more staff involved in the practice of evaluating programs and practices and using evidence in decision-making. We also identified three factors that caused districts to fail to complete evaluations: (1) technology implementation problems, (2) lack of a champion with capacity to follow through, and (3) reluctance to use random assignment. We will update these findings as additional evaluations are completed in the next several months.