Title: Toward a Science of Failure Analysis: A narrative review of null findings in large-scale randomized school interventions

Authors: Clara-Christina Gerstner, Claire Allen-Platt, Robert Boruch
Affiliation: University of Pennsylvania Graduate School of Education

Background:

When a researcher tests an educational program, product or policy in a randomized controlled trial and detects a significant effect on an outcome, the intervention is usually classified as something that "works" and is worthy of replication or emulation. When the expected effects are not found, the intervention is often labelled as a failure and there is seldom an orderly and transparent analysis of plausible reasons why the intervention did not work as anticipated. Accumulating and learning from possible failure mechanisms are not standard practice in education research, nor is it common to design educational interventions with common causes of failure in mind. This paper develops Boruch and Ruby's (2015) proposition that the education and social sciences would benefit from a systematic approach to the study of failure. It draws inspiration partly from the emerging field of implementation science, which itself evolved through the study of failure in the health sector (Kelly & Perkins, 2012), as well as the practice of postmortems in medicine and time-to-failure and safety factor calculations in engineering. However, "while a bridge collapse is usually plain and spectacular, failures of education innovations … are often quieter, not spectacular, and often occur for no transparent reasons" (Boruch & Ruby, 2015, p.2), with experimental null findings likely underreported (e.g., Dawson & Dawson, 2018; Pigott et al, 2013) but, by current estimates, also prevalent (e.g., Boulay et al., 2018). The premise here is that null findings are an underutilized source of evidence. This paper reviews and taxonomizes researchers' accounts of failure in educational experiments, including the nature of a null event and reasons (if provided) for why it occurred. Our systematic analysis of educational interventions with a failed major outcome can help researchers, policymakers and practitioners anticipate failure mechanisms common to real-world settings and build a culture of learning from missteps.

Objective:

Our purpose is to introduce a broad framework for thinking about educational interventions that do not produce expected effects in controlled trials and to seed a cumulative knowledge base on when, how and why interventions do not reach expectations. The expectations are based on assumptions in the analyses that precede the experiment's design and execution. The paper addresses three research questions: 1) What are the reported reasons for null findings in educational evaluations published in the last 10 years?, 2) How do evaluators assess effectiveness?, 3) How do evaluators communicate what can be learned from null findings? The analysis identifies four sources of null findings in school-based experiments: challenges with the planning and logic of an intervention; challenges with implementation; instability in the system; and measurement error. We conclude with recommendations for evaluators and practitioners to improve the design and implementation of large-scale educational interventions in ways that anticipate failure to reach success criteria.

Research Design:

We confine our review to recent, large-scale randomized controlled trials (RCTs) in K-12 schooling with at least one non-significant or negative major outcome. We conducted a systematic search of the reports of eight United States- and United Kingdom-based evaluation firms, as well as the U.S. What Works Clearinghouse, to identify academic papers and evaluation reports within these confines. The search yielded 57 papers published in 2010-2019. Researchers' accounts of null findings were coded to classify the nature of null events and why they occurred.

Findings:

The analysis has surfaced four broad types of events that led to null findings in school-based interventions, including issues with planning and theory of change (65 percent of studies), implementation constraints and errors (86 percent), instability/attrition in the population of participants (47 percent), and measurement issues (44 percent). We present selected examples of educational interventions that illustrate common challenges, such as adapting a theory of change to the study context; recruiting and retaining students and schools in the study sample; or providing sufficient training to teachers. In addition, we find that 44 percent of studies in our sample yielded mixed outcomes (positive, nonsignificant and/or negative), inviting development of a more systematic approach to judging relative effectiveness or interpreting heterogeneous effects (Datta, 1976). Finally, we observe that most evaluators present recommendations on how to improve the implementation of a research project, yet, few reports address issues with attrition or measurement.

Conclusions:

Teachers in schools know that learners can greatly benefit from making mistakes ("productive failure," in classroom parlance). Similarly, researchers can learn a lot from challenges and missteps in their research projects. The framework for analyzing program failure introduced in this paper augments researchers' understanding of the conditions in which educational interventions succeed or fail. It allows us to study more closely the mechanisms that lead to outcomes and determine when these mechanisms might fail, as suggested by Weiss (2002). We propose that program failure be considered in the early design stages of a program, rather than retrospectively analyzed. For example, 47 percent of studies in this review related a failed outcome to participant attrition, but almost none suggested statistical designs that might anticipate or correct for known instabilities in school environments or intervention designs that predict and plan for student or teacher departures. More knowledge about a study context prior to implementation, for example, obtaining the average student or teacher mobility rates in a district, could increase the likelihood that the study reveals what works, under what conditions. RCTs with nonsignificant or negative major outcomes are an untapped source of high-quality evidence, which suggest the presence of major sources of failure that can be preempted through thoughtful design, statistics, or both. In the absence of a systematic study of failure in education, valuable insights might be overlooked from the rich research evidence generated by unsuccessful interventions.

References:

Boulay, B., Goodson, B., Olsen, R., McCormick, R., Darrow, C., Frye, M., Gan, K., Harvill, H., & Sarna, M. (2018). *The Investing in Innovation Fund: Summary of 67 Evaluations: Final Report* (NCEE 2018-4013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Boruch, R., & Ruby, A. (2015). To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure. *In Emerging Trends in the Social and Behavioral Sciences*, 1–16.

Datta, L.-E. (1976). Does It Work When It Has Been Tried? and Half Full or Half Empty? *Journal of Career Education*, *2*(3), 38–55.

Dawson, P. & Dawson, S.L. (2018). Sharing successes and hiding failures: 'Reporting bias' in learning and teaching research. *Studies in Higher Education, 43*(8), 1405-1416.

Kelly, B., & Perkins, D. F. (Eds.). (2012). *Handbook of implementation science for psychology in educatio*n. Cambridge: Cambridge University Press.

Pigott, T.D., Valentine, J.C., Polanin, J.R., Williams, R.T., & Canada, D.D. (2013). Outcome-reporting bias in education research. *Educational Researcher, 42*(8), 424-432.

Weiss, C. (2002). What to do until the random assigner comes, in F. Mosteller and R. Boruch (eds) *Evidence Matters: Randomized Trials in Education Research.* Washington: Brookings Institution.