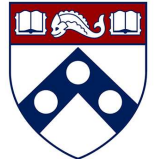# Toward a Science of Failure Analysis: A Narrative Review

Claire Allen-Platt, Clara-Christina Gerstner, Robert Boruch and Alan Ruby
University of Pennsylvania Graduate School of Education
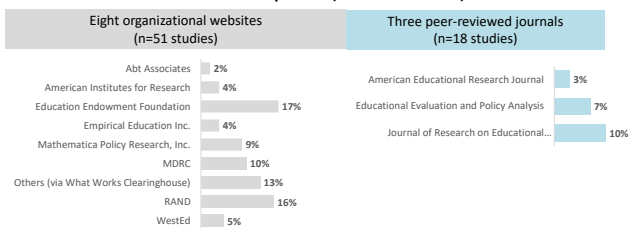
Background/purpose

- Review large-scale RCTs that yielded nonsignificant results and analyze reported reasons for null findings
- Develop Boruch and Ruby's (2015) proposition that the education and social sciences would benefit from a systematic approach to the study of failure
- Investigate the premise that accounts of null findings and nonsuccesses in education trials are an under-exploited source of evidence for practitioners and researchers
- Propose ways in which intervention and study designs can be informed by failure events in educational trials

Research questions

- What are the reported reasons for null findings in large-scale RCTs published in the last ten years?
- What assumptions do researchers make about possible failure mechanisms in educational trials, and how or when is failure addressed in research designs?

Methods

- Systematic search of organizational websites' databases and research journals for reports and peer-reviewed articles for school-based, K-12 randomized controlled trials (RCTs) conducted in the last ten years (2010 – 2019).

| Eight organizational websites (n=51 studies)		Three peer-reviewed journals (n=18 studies)	
Abt Associates	2%		
American Institutes for Research	4%	American Educational Research Journal	3%
Education Endowment Foundation	17%		
Empirical Education Inc.	4%	Educational Evaluation and Policy Analysis	7%
Mathematica Policy Research, Inc.	9%		
MDRC	10%	Journal of Research on Educational…	10%
Others (via What Works Clearinghouse)	13%		
RAND	16%		
WestEd	5%		

- RCTs with at least one major outcome that produced null or negative effects were included in the sample
- Null outcomes were those that failed to reach authors' prespecified probabilistic threshold, typically .05 or less, in a formal statistical test of a null hypothesis

Empirical findings
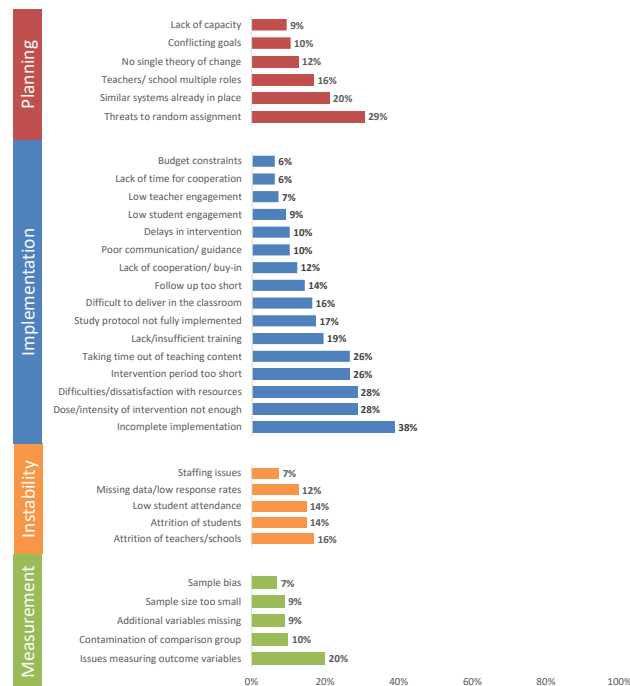
How do authors explain null findings in educational RCTs?

Figure 1. Percent of all studies citing reasons related to:

Planning	61%
Implementation	87%
Instability	41%
Measurement	43%
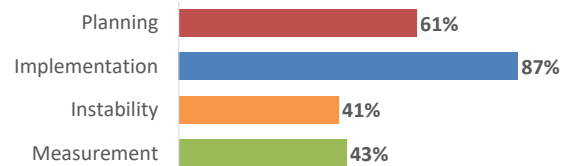
Reasons for null or negative findings were first coded in broad categories, determined a posteriori, as themes emerged from the sample of studies. Most authors provided reasons in more than one category, hence the graph does not sum to 100.

What mechanisms do authors specifically name when explaining null or negative findings in educational RCTs?

Figure 2. Percent of all studies citing reasons related to:

Planning
- Lack of capacity — 9%
- Conflicting goals — 10%
- No single theory of change — 12%
- Teachers/ school multiple roles — 16%
- Similar systems already in place — 20%
- Threats to random assignment — 29%

Implementation
- Budget constraints — 6%
- Lack of time for cooperation — 6%
- Low teacher engagement — 7%
- Low student engagement — 9%
- Delays in intervention — 10%
- Poor communication/ guidance — 10%
- Lack of cooperation/ buy-in — 12%
- Follow up too short — 14%
- Difficult to deliver in the classroom — 16%
- Study protocol not fully implemented — 17%
- Lack/insufficient training — 19%
- Taking time out of teaching content — 26%
- Intervention period too short — 26%
- Difficulties/dissatisfaction with resources — 28%
- Dose/intensity of intervention not enough — 28%
- Incomplete implementation — 38%

Instability
- Staffing issues — 7%
- Missing data/low response rates — 12%
- Low student attendance — 14%
- Attrition of students — 14%
- Attrition of teachers/schools — 16%

Measurement
- Sample bias — 7%
- Sample size too small — 9%
- Additional variables missing — 9%
- Contamination of comparison group — 10%
- Issues measuring outcome variables — 20%

Study characteristics

- Studies in the sample reported one of three types of effects: **null** (all major outcomes statistically non-significant), **mixed** (all outcomes a mix of positive, null and/or negative effects) and **negative** (all major outcomes statistically significantly negative)

Figure 3. Types of effects

Null (n=30)   Mixed (n=37)   Negative (n=8)

40%	49%	11%
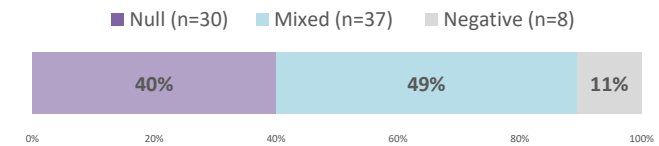
- Studies with null effects reported fewer outcome measures (3.5 outcomes, on average) than studies with mixed or negative effects in the sample (10.2 outcomes, on average)
- Most studies (69%) used reading and/or math assessments as the main outcome measure to assess intervention effect. Other outcomes included, for example, attendance, GPA, substance use and behavioral health indicators, socio-emotional learning measures, and grade retention and promotion

Preliminary conclusions

- Studies that do not yield expected effects seldom provide an orderly and transparent analysis of plausible reasons why the intervention did not work as anticipated
- Dependability of the evidence provided by authors to explain failed outcomes is highly variable, some is speculation while some is evidence-based
- Characteristics of strong post-mortems include the collection and use of implementation data and stakeholder exit interviews to accurately identify causes of failure
- 'Anticipating' failure in research designs may consist of, for example, including empirical benchmarks for student or teacher attrition in power calculations; utilizing more sensitive outcome measures; prioritizing fewer outcomes; ongoing collection of implementation data