



From What Works to What Replicates? Methodological Foundations for a Replication Science

Vivian C Wong (University of Virginia) & Peter M Steiner (University of Maryland)

The methodological research presented here was supported by the Institute of Education Sciences R305D190043 and R305D220034. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.



28:17

+ Queue

Download

When Great Minds Think Unalike: Inside Science's 'Replication Crisis'

May 24, 2016 · 12:10 AM ET

NPR STAFF

SCIENCE

THE STATE OF THE UNIVERS

Sloppy Sci

Psychology's Replication Crisis Can't Be Wished Away

It has a real and heartbreaking cost.

ED YONG

MAR 4, 2016

SCIENCE

Are sketchy practices in the lab to blame for the replication crisis in psychology research?

By *Daniel Engber*



464



117



162

Efforts to Address Replication Crisis

- Proposed solution: **More funding and publication of replication efforts**
- NIH supports **training** of researchers and graduate students to promote replication of results
- Suggestion that **replication of results should be required** before publication, or inclusion in a registry for decision-making (e.g. What Works Clearinghouse)
- But ... replication itself is **not a well-established method**
 - No consensus on what replication is, what constitutes as a replication study, and how to analyze and interpret one.

Message from IES Director:

A More Systematic Approach to Replicating Research

IES' two research centers, the National Center for Education Research (NCER) and the National Center for Special Education Research (NCSEER), have funded around 450 projects testing whether interventions improve student outcomes. Most of the roughly 300 completed projects have found no impact, conforming to Peter Rossi's "*Iron Law of Evaluation*" that the expected value of any impact assessment of any large scale social program is zero. Despite this iron law, of these projects, some 1/3 show some evidence of success. While we certainly will work to improve that success rate, it's what we do with the projects that have evidence of impact that is a growing concern for us.

A central goal of the mission of IES is to identify what works for whom under what conditions. Unfortunately most of the studies that have found impact contribute little to helping us meet that goal. Many, if not the great majority, of these projects were carried out in a single location and/or tested using a relatively small number of settings, teachers, and learners. Given the limited scope of these projects, it is usually impossible to judge whether the tested interventions would work with different types of students or in different education venues.

In the last few years, IES has explicitly called for and supported replications, but these far too often replicate the problem of the original study: too few settings, too few subjects, too little variation. Our current approach to replication, in short, does not *systematically* test conditions that affect the impact interventions could have and accumulates knowledge very slowly, if at all.

We are considering a different approach to replication that, hopefully, will accelerate the accumulation of knowledge about which interventions might work for whom under what conditions. This approach revolves around the *systematic replication of interventions that already have strong evidence of impact*. We envision supporting *sets of replications* that will implement and evaluate interventions in carefully chosen venues that systematically vary in student demographics, geographic locations, implementation, or technology.

Building a Replication Science (Wong (PI), Steiner, Co-PI)

The Collaboratory Replication Lab is focused on how to conduct replication studies in field settings.

- Organized around the [Causal Replication Framework](#), which provides a common understanding what replication is from a potential outcomes perspective, and assumptions required for direct replication of results.
- Replication failure is not inherently a problem, as long as we have a systematic ways for understanding why failure occurred (sources of [effect heterogeneity!](#))
 - Current ad-hoc approaches to replication often fails to identify why replication failure has occurred.

Our goal is to present replication from a [design-based approach](#) to understand *why replication failure occurs*.

- How [can research designs be used](#) to address and test replication assumptions systematically?
- How can [diagnostic measures](#) can be used for assessing replication assumptions?
- How do we determine [replication success](#)?
- How to plan [multiple conceptual replication designs](#) to test [sources of effect variation](#) for [generalizing](#) (preview this)

What is Replication?

“Replication is a methodological tool based on a repetition of procedure that is involved in establishing a fact, truth or piece of knowledge”

(Schmidt, 2009)

What is Replication?

“Replication is a methodological tool based on a *repetition of procedure* that is involved in *establishing a fact, truth or piece of knowledge*”

(Schmidt, 2009)

Most [replication] definitions emphasize *repeating an experimental procedure* (Schmidt, 2009)

→ *direct replication* (exact or close replication)

“Close replications refer to those replications that are based on *methods and procedures as close as possible to the original study*”
(Brandt et al., 2014)

Challenges with Current Definition

Replication quality is judged by how closely the replication is able to repeat methods and procedures from original study (Brandt et al., 2014, Kahneman, 2014)

But,

1. **Original study may fail to report** all relevant and necessary methods, making direct replication difficult if not impossible (Hansen, 2011)
2. **Privileges methods and procedures in original study** ... but original study may be flawed or not perfectly implemented
3. **Replication of methods is not the primary goal** of replication. We want *replication of effects*.

What is Replication? ... Revisited

“Replication is a methodological tool based on a repetition of procedure that is involved in establishing a fact, truth or piece of knowledge”

(Schmidt, 2009)

What is Replication? ... Revisited

“Replication is a methodological tool based on a repetition of procedure that is involved in establishing a fact, truth or piece of knowledge”

(Schmidt, 2009)

Focus on the “fact, truth, or piece of knowledge” that we want to establish

What is Replication? ... Revisited

“Replication is a methodological tool based on a repetition of procedure that is involved in establishing a fact, truth or piece of knowledge”

(Schmidt, 2009)

Focus on the “fact, truth, or piece of knowledge” that we want to establish

- In program evaluation, we say the goal is to replicate the causal effect of a well-defined treatment effect → *Causal estimand*
 - Causal estimand is defined as the causal effect of a well-defined treatment-control contrast for a clearly defined target population and setting.
- This means that repeating methods and procedures will help in achieving goal, *but it is not required*

Successful Replication of the Causal Estimand

Successful replication can be expected only if the causal estimand in the original and replication study is identical.

In most current replication studies, the causal estimand is not well-defined. The reader must surmise it from the study description.

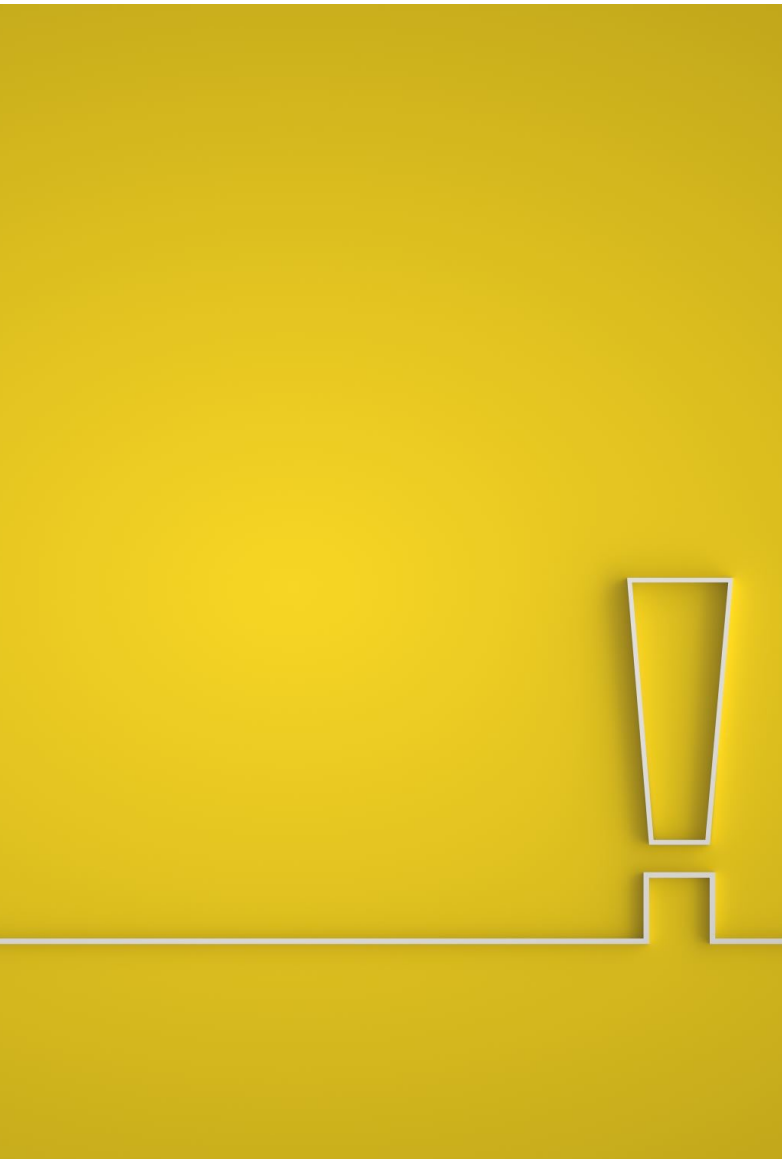
Introduce the Causal Replication Framework to address challenges:

- Derived using potential outcomes notation (Rubin, 1977)
- Implies a research design perspective for determining “high quality” replication
- Focus mostly on causal identification and estimation (point estimation) but ignore efficiency and power issues [for now]

Assumption	Source of Variation	Replication Design
R1. Treatment and Outcome Stability	Is there variation in treatment and control conditions, and in the outcome measures used?	Multi-Arm treatment designs / Replication of proximal and distal measures
R2. Equivalence of the Causal Estimand	Is there variation in contexts, study settings, and sample characteristics across studies?	Multi-site designs / Robustness checks / switching replication designs / multiple cohort designs
S1. & S2. Identification and Estimation Assumptions	Do different research designs and estimation approaches produce the same result?	Design-replication studies / Robustness checks with alternative model specifications
S3. Correct Reporting	Given the same data and syntax files, are multiple investigators able to reproduce the same results?	Reproducibility / Reanalysis Designs

	Assumption	Source of Variation	Replication Design
	R1. Treatment and Outcome Stability	Is there variation in treatment and control conditions, and in the outcome measures used?	Multi-Arm treatment designs / Replication of proximal and distal measures
	R2. Equivalence of the Causal Estimand	Is there variation in contexts, study settings, and sample characteristics across studies?	Multi-site designs / Robustness checks / switching replication designs / multiple cohort designs
Direct Replication	S1. & S2. Identification and Estimation Assumptions	Do different research designs and estimation approaches produce the same result?	Design-replication studies / Robustness checks with alternative model specifications
	S3. Correct Reporting	Given the same data and syntax files, are multiple investigators able to reproduce the same results?	Reproducibility / Reanalysis Designs

	Assumption	Source of Variation	Replication Design
Conceptual Replication	R1. Treatment and Outcome Stability	Is there variation in treatment and control conditions, and in the outcome measures used?	Multi-Arm treatment designs / Replication of proximal and distal measures
	R2. Equivalence of the Causal Estimand	Is there variation in contexts, study settings, and sample characteristics across studies?	Multi-site designs / Robustness checks / switching replication designs / multiple cohort designs
Direct Replication	S1. & S2. Identification and Estimation Assumptions	Do different research designs and estimation approaches produce the same result?	Design-replication studies / Robustness checks with alternative model specifications
	S3. Correct Reporting	Given the same data and syntax files, are multiple investigators able to reproduce the same results?	Reproducibility / Reanalysis Designs



Replication failure is not inherently a problem, *as long as the researcher understand why.*

Replication failure occurs when one or more assumption is violated.

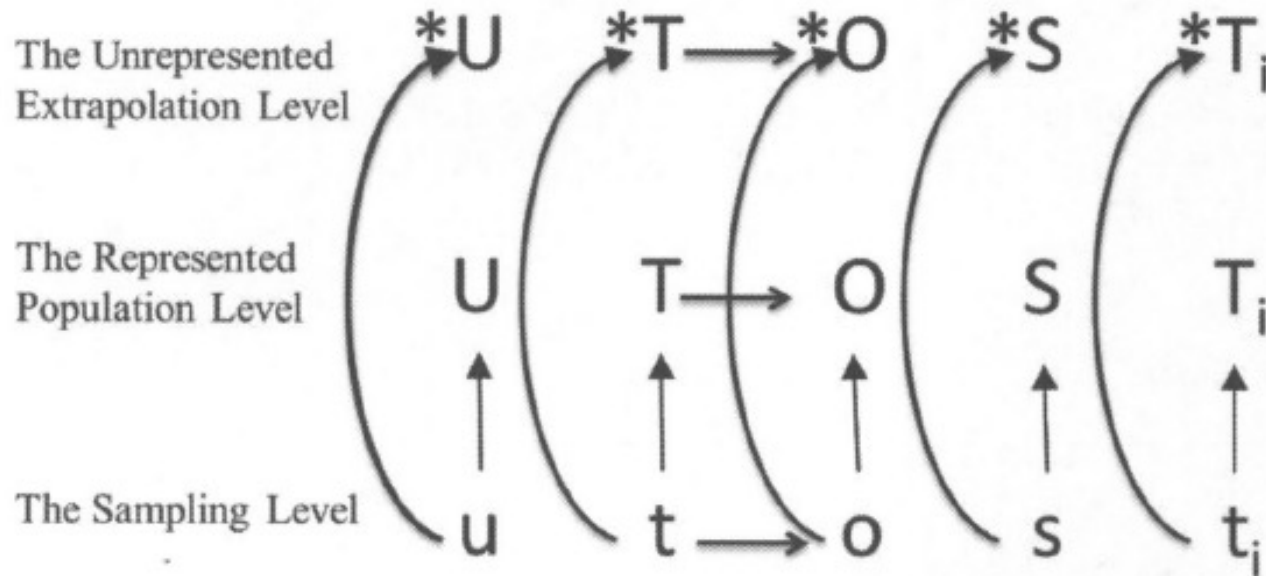
When the researcher is able to test one or more assumptions systematically -- while addressing all other assumptions -- the researcher may identify *why* replication failure occurred.

→ Achieved through [prospective research designs](#)

Extension: From Systematic Replication Designs to Generalizing Effects

- Field evaluators are often interested in conducting **systematic conceptual replication** studies for understanding “**what works, for whom, under what conditions.**”
- The goal is to understand the **replicability and generalizability of effects**, or at least to identify generalizability boundaries over a **response space of interest**.
- The presumption here is that in field settings, **how treatment effects vary over unit, treatment, outcome, setting, and time characteristics** are as important for *understanding as an average treatment effect*

Cronbach's View (1982)



Meta-Analysis

Meta-analysis involves the analysis of multiple study effects with ideally the same treatments and outcomes.

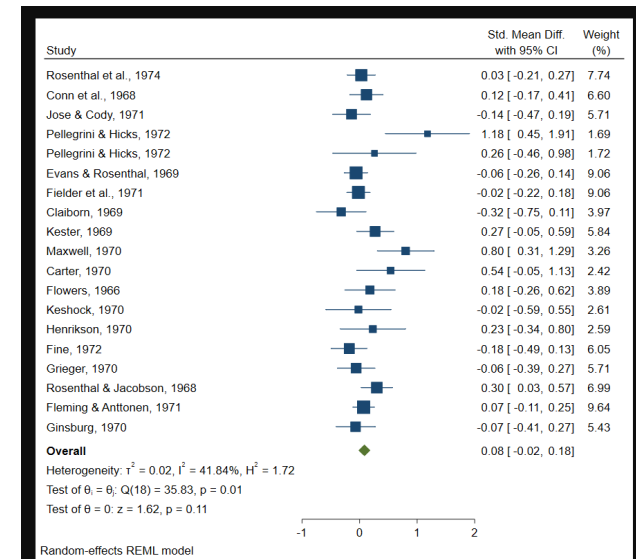
Challenges

Requires synthesizing effects from *obtainable samples*.

“What is this target population, and, even if I could conceptualize it, why should I care about average effects in it?” (Rubin, 1992, pg. 365)

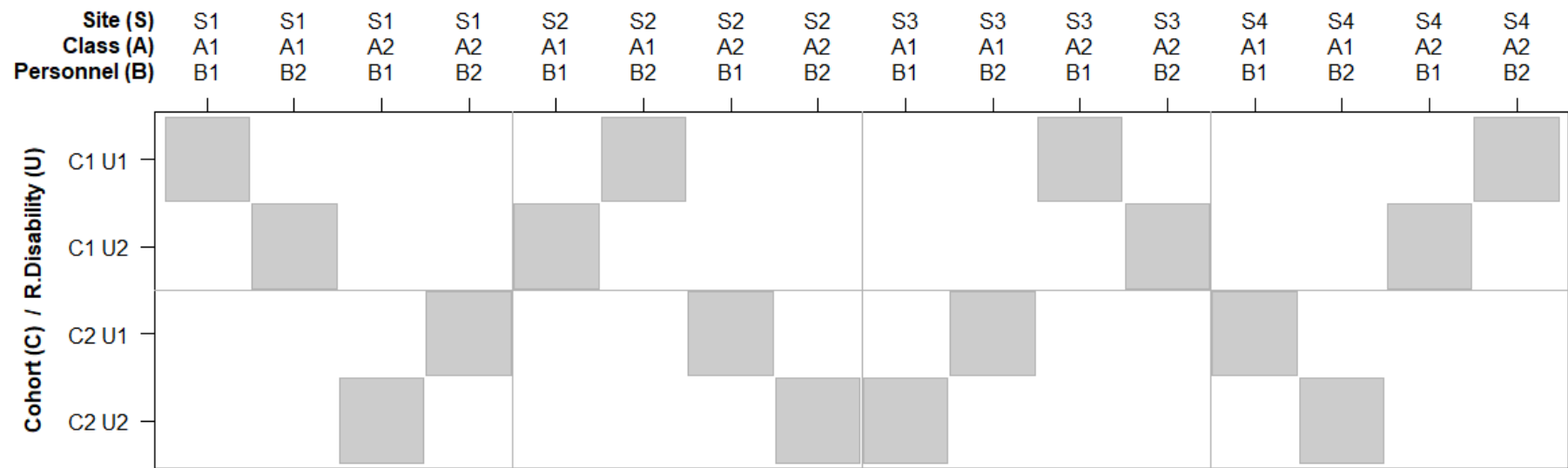
Effect heterogeneity may be correlated with study characteristics and features, but causal interpretations of effect heterogeneity remains unclear

Teacher Expectancy & Pupil IQ

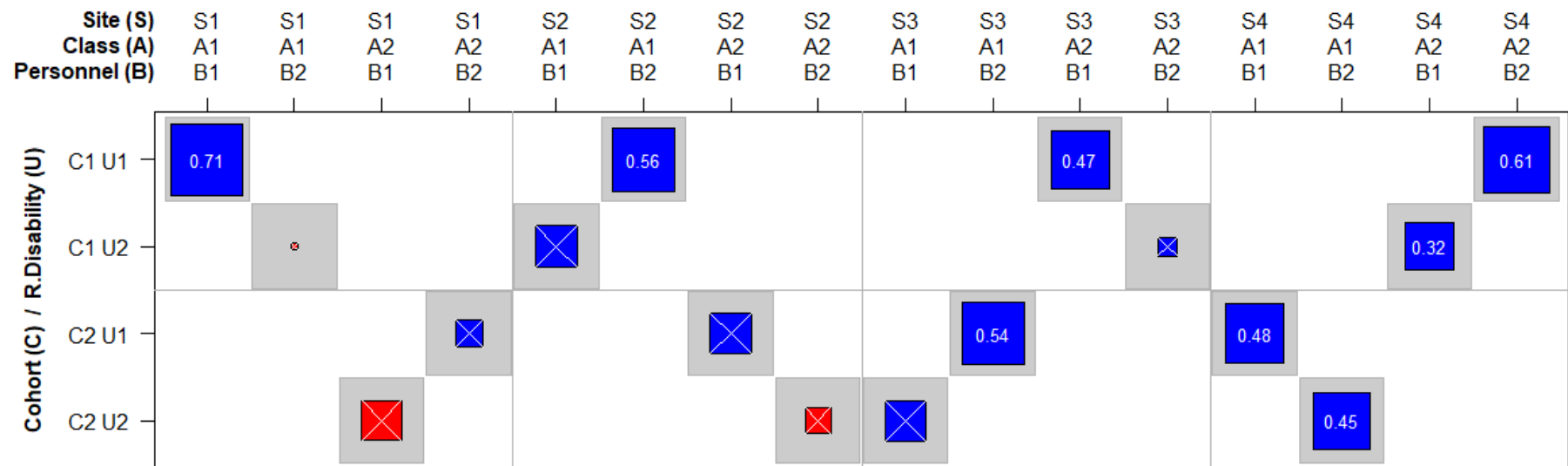


Raudenbush (1984)

Factional Design Plan with Planned Effect Estimates for the Grey Cells



Illustrative Example: Observed Effect Estimates of Response Effect Grid



Illustrative Example: Observed and Predicted Effect Estimates for the Entire Response Grid

	Site (S)				Site (S)				Site (S)				Site (S)				
	S1	S1	S1	S1	S2	S2	S2	S2	S3	S3	S3	S3	S4	S4	S4	S4	
Class (A)	A1	A1	A2	A2	A1	A1	A2	A2	A1	A1	A2	A2	A1	A1	A2	A2	
Personnel (B)	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	
Cohort (C) / R.Disability (U)	C1 U1	0.71	0.53	0.41	×	0.55	0.56	0.33	0.3	0.66	0.63	0.47	0.39	0.57	0.72	0.51	0.61
	C1 U2	×	0	×	-0.35	×	×	×		0.32	×	×	×	0.38	0.54	0.32	0.42
	C2 U1	0.62	0.45	0.32	×	0.46	0.47	×	×	0.58	0.54	0.38	0.3	0.48	0.63	0.42	0.52
	C2 U2	×	×	×	-0.44	×	×	×	×	×	×	×	×	0.3	0.45	×	0.34

What is Needed

- Well-articulated theory about what **characteristics will affect (moderate) the magnitude of the intervention effect.**
 - Consider scientifically important moderators
 - And boundary conditions that examine the robustness of results
- The moderators are factors that are hypothesized to generate gradations in the “response surface”/effect grid
- Since its not feasible to estimate effects for every cell on the response surface, we **need design-based approaches to select cells** for observing effects
- A **series of conceptual replication studies** to estimate observed effects in selected cells -- and moderator effects -- for the purpose of generalizing effects to cells that cannot be observed.
- Implementing with the Special Education Research Accelerator (SERA).

Conclusion: Design-based Approaches to Replication

- Conceptualize replication as a *prospective research design* of organizing multiple studies.
 - When results do not replicate in *direct replication* studies, the researcher concludes that individual study results were *biased or incorrectly reported*.
 - When results fail to replicate in *conceptual replication studies*, the researcher concludes *the presence of effect variation*.
- The Causal Replication Framework provides assumptions for determining “high quality” direct and conceptual replication designs for the purpose of *identifying sources of effect heterogeneity*.
 - Understanding effect heterogeneity is critical for *generalization of effects*
- Metric for determining *replication success* must also be determined in advance.
 - Different statistical properties for different measures (and different sample size requirements)
 - Metric for determining replication success should match replication hypotheses under investigation.



Thank you

Email: vcw2n@virginia.edu

Systematic Replication Resources:
Website: <http://edreplication.org/>

