

Moving from *What Works* to *What Replicates*: A New Framework for Evidence Based Policy Analysis

Vivian C. Wong

University of Virginia, Charlottesville

Peter M. Steiner

University of Wisconsin-Madison

Open Science, Trier, March 12-14, 2019

Supported by NSF grant #2015-0285-00

Replication in Field Settings

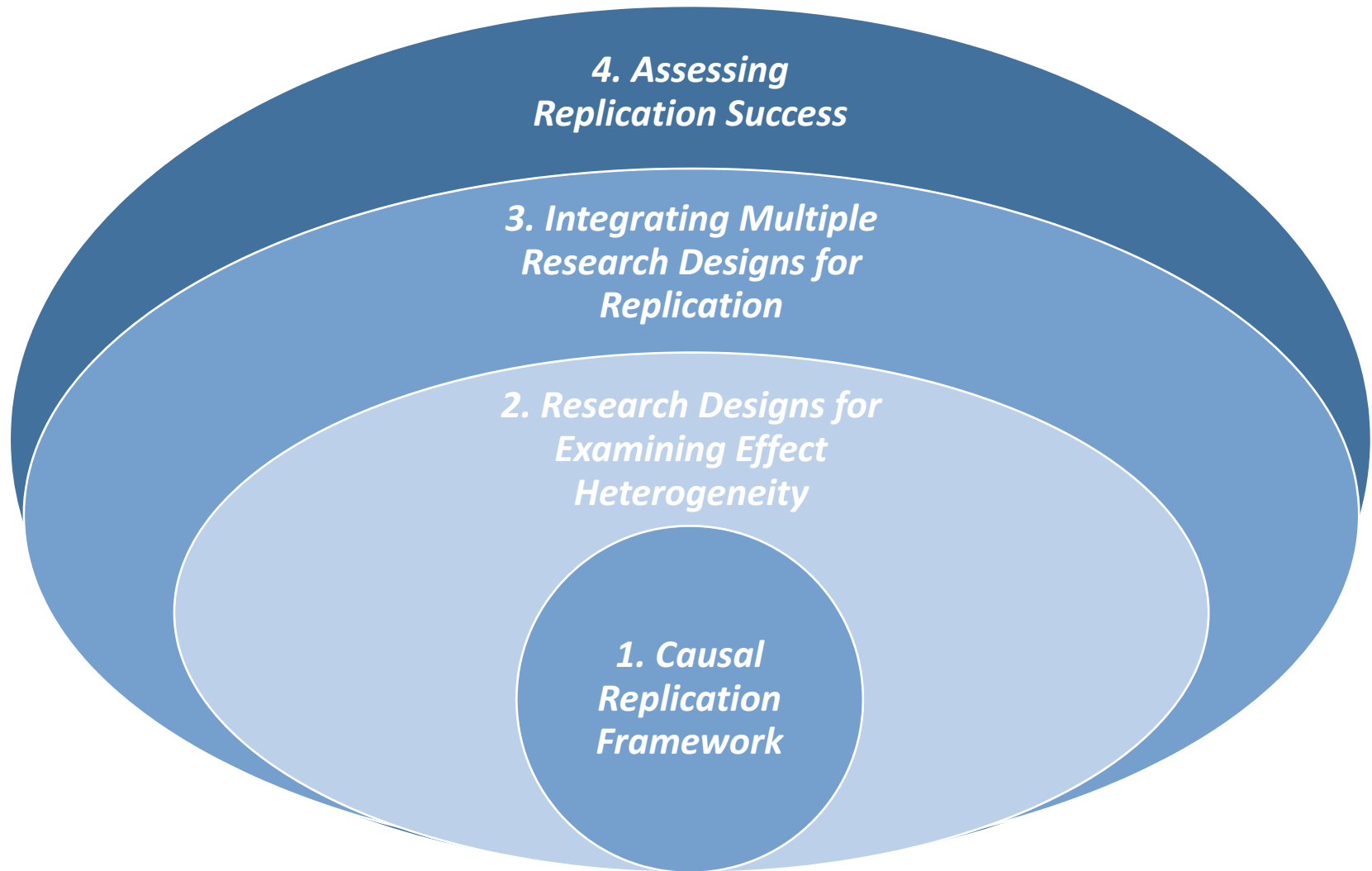
In light of the replication crisis in the social sciences, is replication **useful** and **feasible**?

- Yes!
- Replication failure is not inherently a problem, as long as we have a systematic way for understanding why failure occurred

But unclear how we should

- **define** what **replication** is
- **determine** whether a replication is **high quality**
- **test** and **interpret replication success**

A Framework for Systematic Replication



What is Replication?

Methods & Procedures, Causal Estimands

What is Replication?

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

(Schmidt, 2009)

- ▣ “Most [replication] definitions pronounce the action of **repeating an experimental procedure**”
(Schmidt, 2009)
→ **direct replication** (exact or close replication)

Direct/Close Replication – *Definitions*

- “Replication is independently **repeating the methodology** of a previous study and obtaining the same results” (Nosek & Errington, 2017)
- “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)

Direct/Close Replication – *Issues*

Issues with repetition of *methods & procedures* (M&P)

- Repetition of M&P *prioritizes the original study* over the replication study
 - estimated effects of original study are implicitly assumed to reflect a true (causal) effect
- M&P of the original study might have been *imperfectly implemented or flawed* (e.g., noncompliance, attrition, low treatment fidelity, treatment contamination)
 - replicating a potentially flawed study might not be meaningful

Direct/Close Replication – *Issues (cont.)*

- M&P are rarely *fully documented* in the original study
 - exact or close replication is impossible/difficult
- M&P are *not the primary goal* of a replication
 - But the question about whether the treatment/intervention has an impact on the outcome is of interest

Replication of Causal Estimands

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

(Schmidt, 2009)

- aim at replicating the ***causal effect of a well-defined treatment-control contrast (-> causal estimand)***
- repeating methods and procedure might help in achieving the goal but it is no longer necessary

Replication of Causal Estimands

- ▣ Causal point of view instead of a procedural one
 - ▣ Prospective point of view – replication as a research design
 - ▣ No prioritization of the original study
- Focus is on the causal assumptions required for a successful replication of a causal estimand
- Formalized in potential outcomes notation

Causal Estimand (Target of Inference)

Causal estimand: A population parameter quantifying the causal effect of a treatment relative to a control condition

- the “true” but unknown causal effect in a well-defined inference population (R)
- defined in terms of *potential outcomes*
(Rubin Causal Model: Rubin, 1974; Holland, 1986)

$Y_i(0)$... potential control outcome ($T_i = 0$)

$Y_i(1)$... potential treatment outcome ($T_i = 1$)

Average treatment effect: $ATE = E_R[Y_i(1) - Y_i(0)]$

Causal Estimand (Target of Inference)

Examples of causal estimands for an RCT:

- *Average treatment effect*: $ATE_R = E_R[Y_i(1) - Y_i(0)]$

In case of noncompliance (no-shows & cross-overs)

- *Intent-to-treat effect* (ITT):

$$ITT_R = E_R[Y_i(1) - Y_i(0) \mid Z_i = 1]$$

- *Average treatment effect for the treated* (ATT; no-shows):

$$ATT_R = E_R[Y_i(1) - Y_i(0) \mid T_i = 1]$$

- *Complier average treatment effect* (CATE or LATE; no-shows & cross-overs):

$$CATE_R = E_R[Y_i(1) - Y_i(0) \mid \text{Compliers}]$$

The Causal Replication Framework: Assumptions

Causal Replication Assumptions

The valid replication of a causal estimand rests on five major assumptions

Across studies:

A1 Treatment & Outcome Stability

A2 Equivalence of Causal Estimands

Within studies:

A3 Causal Estimand is Identified in Both Studies

A4 Causal Estimand is Estimable without Bias
in Both Studies

A5 Treatments, Outcomes, Estimands, Estimators,
and Estimates are Correctly Reported in
Both Studies

A1 Treatment & Outcome Stability

A1.1 No variation in treatment and control conditions

- ▣ Identical treatment procedures, no unobserved variation in treatment dosage
- ▣ Identical control conditions

A1.2 No variation in outcome measures

- ▣ Identical outcome constructs and valid measurement
- ▣ Identical measurement setting and timing

A1 Treatment & Outcome Stability (cont.)

A1.3 No mode-of-study-selection effects

- Selection into studies has no effect on potential outcomes (e.g., random or self-selection, with or without incentives)

A1.4 No peer, spillover, or carryover effects

- The potential outcomes in the replication study are unaffected by researchers, participants, and characteristics of the original study

A2 Equivalence of Causal Estimands

A2.1 *Same causal quantity of interest*

- Both studies need to focus on the same causal quantity, e.g., ATE

A2.2 *Identical effect-generating processes*

- The process generating the *causal effects* must be identical in both studies
→ effect moderators have the same effect in both studies—across sites or time)

A2 Equivalence of Causal Estimands (cont.)

A2.3 Identical distribution of population characteristics

- ❑ target populations must be identical with respect to the joint distribution of individual characteristics
(→ same inference population, R)
- ❑ observed and unobserved population characteristics that moderate the causal effect

A2.4 Identical distribution of setting variables

- ❑ both studies must be implemented in the same setting

A3 Identification of Causal Estimands

In both studies, the causal estimand (ATE) must be *identified*

Example:

- ▣ RCTs with identical target populations and settings
→ perfect implementation
- ▣ RCTs with different target populations (P, Q) and settings (S_0, S_1)
→ perfect implementation
→ reweighting or matching with respect to inference population R and setting variables S

A4 Unbiased Estimation of Causal Estimands

In both studies, the causal estimand (ATE) is *estimable without bias*

- ▣ Unbiased or consistent estimator for ATE (correct model specification)
- ▣ Technical assumptions must be met (e.g., no perfect collinearity, sufficient degrees of freedom)

A5 Correct Reporting in Both Studies

In both studies, treatments, outcomes, estimands, estimators, and estimates need to be correctly reported

Mistakes in reporting may results in incorrect conclusions about

- ▣ whether studies aim at same causal estimand
- ▣ whether results successfully replicate

Example: Two Perfectly Implemented RCTs

Assumption	Original Study: RCT	Replication I: RCT
A1 Treatment & outcome stability	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects 	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects
A2 Equivalence of causal estimands	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ✓ target population $P = Q$ ✓ setting $S_0 = S_1$ 	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ✓ target population $Q = P$ ✓ setting $S_1 = S_0$
A3 Identification	✓ ATE is identified	✓ ATE is identified
A4 Estimation	✓ Unbiased (mean difference)	✓ Unbiased (mean difference)
A5 Reporting	✓ Correct reporting	✓ Correct reporting

Example: Two Imperfect RCTs

Assumption	Original Study: RCT	Replication: RCT
A1 Treatment & outcome stability	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✗ Participation incentives affect potential outcomes ✓ No peer-, spillover-, or carry-over effects 	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects
A2 Equivalence of causal estimands	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ✓ target population P ✓ setting S_0 	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ✗ target population $Q \neq P$ ✗ setting S_1
A3 Identification	✗ ATE_p is not identified (due to incentives' effect)	✗ ATE_p is not identified (due to above issues)
A4 Estimation	✓ Unbiased estimator (mean difference)	✓ Unbiased estimator (mean difference)
A5 Reporting	✓ Correct reporting	✓ Correct reporting

Example: RCT and Observational Study

Assumption	Original Study: RCT (lab)	Replication: Observational (field)
A1 Treatment & outcome stability	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects 	<ul style="list-style-type: none"> ✗ different control condition ✗ different timing of measurements ✓ No mode-of-study-selection effects ✗ carry-over effects
A2 Equivalence of causal estimands	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ✓ target population P ✓ setting S_0 	<ul style="list-style-type: none"> ✓ ATE ✗ different effect-gener. process ✓ target population P ✗ setting S_1
A3 Identification	<ul style="list-style-type: none"> ✓ ATE_p is identified (mean difference) 	<ul style="list-style-type: none"> ✗ ATE_p is not identified (due to above issues, and maybe violation of unconfoundedness)
A4 Estimation	<ul style="list-style-type: none"> ✓ Unbiased (mean difference) 	<ul style="list-style-type: none"> ✓ Unbiased/consistent estimator (matching estimator)
A5 Reporting	<ul style="list-style-type: none"> ✓ Correct reporting 	<ul style="list-style-type: none"> ✓ Correct reporting

The Causal Replication Framework: Design Variants

Causal Replication Design Variants

Derivation of causal replication design variants

(→ conceptual replication)

Instead of attempting to meet all replication assumptions, researchers might

- ▣ systematically *relax* one (or more) assumptions
- ▣ while meeting all other assumptions

Effect Heterogeneity across Populations

Assumption	Original Study: RCT	Replication: RCT
A1 Treatment & outcome stability	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects 	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects
A2 Equivalence of causal estimands	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ☑ target population P ✓ setting $S_0 = S_1$ 	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ☑ target population $Q \neq P$ ✓ setting $S_1 = S_0$
A3 Identification	<ul style="list-style-type: none"> ✓ ATE_P is identified 	<ul style="list-style-type: none"> ✓ ATE_Q is identified
A4 Estimation	<ul style="list-style-type: none"> ✓ Unbiased (mean difference) 	<ul style="list-style-type: none"> ✓ Unbiased (mean difference)
A5 Reporting	<ul style="list-style-type: none"> ✓ Correct reporting 	<ul style="list-style-type: none"> ✓ Correct reporting

Effect Heterogeneity across Settings

Assumption	Original Study: RCT	Replication: RCT
A1 Treatment & outcome stability	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects 	<ul style="list-style-type: none"> ✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments & timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects
A2 Equivalence of causal estimands	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ✓ target population $P = Q$ ☑ setting S_1 	<ul style="list-style-type: none"> ✓ ATE ✓ effect-generating process ✓ target population $Q = P$ ☑ setting $S_1 \neq S_0$
A3 Identification	<ul style="list-style-type: none"> ✓ ATE is identified 	<ul style="list-style-type: none"> ✓ ATE is identified
A4 Estimation	<ul style="list-style-type: none"> ✓ Unbiased (mean difference) 	<ul style="list-style-type: none"> ✓ Unbiased (mean difference)
A5 Reporting	<ul style="list-style-type: none"> ✓ Correct reporting 	<ul style="list-style-type: none"> ✓ Correct reporting

Summary / Implications

Summary

Focus of the causal replication framework is on

- ▣ *Causal estimand* instead of methods & procedures
- ▣ *Assumptions for a successful replication*
 - Assumptions are strong
 - Replication success is not very likely (easier in lab than field conditions)
- ▣ *Replication variants* for probing effect heterogeneities, generalizability, research designs & methods -> replication designs

Implications for Replication Practice

- ❑ *Plan replications prospectively* (two or multiple studies together)
- ❑ *Make data available* such that studies can be reanalyzed with respect to different target populations (R) → weighting or matching
- ❑ When using design variants, *systematically vary only one factor at a time* (i.e., relax only one assumption)
- ❑ Clearly report causal estimands, estimators and estimates
- ❑ Do replications! Learn from replication failure.

Thank You!

Wong & Steiner (2018). *Replication Designs for Causal Inference*. Working paper.

https://curry.virginia.edu/sites/default/files/uploads/epw/62_Replication_Designs.pdf