

# Replication Designs for Causal Inference

***Peter M. Steiner***

University of Wisconsin-Madison &  
Freie Universitaet Berlin

***Vivian C. Wong***

University of Virginia, Charlottesville

FU Berlin, June 15, 2018

Supported by NSF grant #2015-0285-00

# Shift to Alternative Perspective on Replication

From

*Why might an experiment not replicate?* (West's talk)

to

***When can we expect an experiment to replicate?***

and from

*Replicating methods & procedures and results*

to

***Replicating the (unknown) causal effect  
of a treatment***

- Focus on “ideal” replication design for causal effects
- Derive different causal replication designs

# 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)
- ❑ “Direct replication is the attempt to **recreate the conditions** believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data” (OSC, 2015)
- ❑ “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 – *Success*

- ❑ “Replication is independently repeating the methodology of a previous study and **obtaining the same results**” (Nosek & Errington, 2017)
- ❑ “Direct replication is the attempt to recreate the conditions believed sufficient for **obtaining a previously observed finding** and is the means of establishing reproducibility of a finding with new data” (OSC, 2015)
- ❑ “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) of original study

- M&P are rarely *fully documented*  
→ exact or close replication is impossible/difficult
- M&P might be *imperfectly implemented or flawed* (e.g., noncompliance, attrition, low treatment fidelity, treatment contamination)  
→ replicating a potentially flawed study might be not meaningful
- M&P are *not the primary goal* of a replication  
→ establishing a fact, truth or piece of knowledge (i.e., does the treatment/intervention have an impact on the outcome?)



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

- Repetition of M&P *prioritizes the original study* over the replication study
  - estimated effects of the original study are implicitly assumed to reflect the true (causal) effect
  - methods & procedures of the original study must be replicated

# 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)

→ Focus on the fact, truth or piece of knowledge we want to establish (target of inference)

- aim at replicating the *causal effect of a well-defined treatment condition (causal estimand)*
- repeating (some) 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  
→ could also be an association instead of a causal effect
- ❑ *No (a priori) prioritization of the original study*  
→ prospective point of view – replication as a research design
- ❑ Derive *assumptions* required for a successful replication of a causal estimand

# 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*:  $ATE = E_R[Y_i(1) - Y_i(0)]$

- in general, the causal estimand is not the same as a model parameter (e.g., of a regression model)

# 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):

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

# When does an Experiment Replicate?

*Successful replication* can be expected only if the *causal estimands* of the original and replication study are *identical*

In deriving a *causal replication framework* we

- use a *prospective point of view*, i.e., we do not address issues like publication bias, *p*-hacking or HARKing
- focus mostly on *causal identification and estimation* (point estimation) but ignore efficiency and power issues

# Identification vs. Estimation

**Identification:** Can the average causal effect be estimated *nonparametrically* from a hypothetically *infinite sample/population*?

- ignoring random fluctuations allows us to focus on systematic biases due to confounding, selection, and measurement
- do not require any modeling assumptions → nonparametric

**Estimation:** Can the average causal effect be *uniquely* estimated from the *finite sample*?

- does a unique solution exist?
- focus is on bias, consistency, efficiency of an estimator

# The Causal Replication Framework: Assumptions



# Assumptions for Replicating Causal Effects

Require two sets of assumptions

- ▣ *Causal assumptions* for the identification & estimation of a causal effect in *each study* (original and replication study)  
→ make sure that a *causal* quantity is estimated
- ▣ *Causal replication assumptions* for the valid replication of a causal estimand *across studies*  
→ make sure that the causal quantities are identical for both studies

# Causal Assumptions for Single Studies

*Identification assumptions* for an RCT

- Unconfoundedness (independence)
- Positivity
- Stable Unit Treatment Value Assumption (SUTVA)
  - uniquely defined and implemented treatment and control conditions
  - assignment procedure itself has no direct effect on outcome
  - no interference among subjects

→ perfect implementation of RCT (high treatment fidelity, no treatment contamination, no attrition, no non-compliance, no missing data, etc.)

# Causal Assumptions for Single Studies

## *Identification assumptions (cont.)*

- ❑ Other study designs like regression discontinuity, nonrandomized pretest-posttest, or instrumental variable designs
  - require more and stronger assumptions
  - might identify other causal estimands

## *Estimation assumptions*

- ❑ unbiasedness / consistency of the estimator with respect to the causal estimand (correctly specified model)
- ❑ technical assumptions (no perfect collinearity, sufficient degrees of freedom)

# Causal Replication Assumptions

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

**A1** Treatment & Outcome Stability (→ procedures)

**A2** Equivalence of Causal Estimands

**A3** Causal Estimand is Identified in Both Studies

**A4** Causal Estimand is Estimable without Bias  
in Both Studies

**A5** 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 measures and instruments
- ▣ Identical 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 Estimands, Estimators, and Estimates are Correctly Reported in Both Studies**

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

# Example: Two Imperfect RCTs

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

# Example: RCT and Observational Study

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

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

# Effect Heterogeneity across Settings

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

# Evaluation of Variation in Treatments

Assumption	Original Study: RCT	Replication: RCT
<b>A1</b> Treatment & outcome stability	<p>☑ Treatment variant A (unique control condition)</p> <ul style="list-style-type: none"> <li>✓ Outcome measure, instruments &amp; timing</li> <li>✓ No mode-of-study-selection effects</li> <li>✓ No peer-, spillover-, or carry-over effects</li> </ul>	<p>☑ Treatment variant B (same control condition)</p> <ul style="list-style-type: none"> <li>✓ Outcome measure, instruments &amp; timing</li> <li>✓ No mode-of-study-selection effects</li> <li>✓ No peer-, spillover-, or carry-over effects</li> </ul>
<b>A2</b> Equivalence of causal estimands	<ul style="list-style-type: none"> <li>✓ ATE</li> <li>✓ effect-generating process</li> <li>✓ target population <math>P</math></li> <li>✓ setting <math>S_0 = S_1</math></li> </ul>	<ul style="list-style-type: none"> <li>✓ ATE</li> <li>✓ effect-generating process</li> <li>✓ target population <math>Q = P</math></li> <li>✓ setting <math>S_1 = S_0</math></li> </ul>
<b>A3</b> Identification	<ul style="list-style-type: none"> <li>✓ ATE is identified</li> </ul>	<ul style="list-style-type: none"> <li>✓ ATE is identified</li> </ul>
<b>A4</b> Estimation	<ul style="list-style-type: none"> <li>✓ Unbiased (mean difference)</li> </ul>	<ul style="list-style-type: none"> <li>✓ Unbiased (mean difference)</li> </ul>
<b>A5</b> Reporting	<ul style="list-style-type: none"> <li>✓ Correct reporting</li> </ul>	<ul style="list-style-type: none"> <li>✓ Correct reporting</li> </ul>

# Evaluation of Research Designs

Assumption	Original Study: <b>RCT</b>	Replication: <b>Observational Study</b>
<b>A1</b> Treatment & outcome stability	<ul style="list-style-type: none"> <li>✓ High fidelity of treatment and control conditions</li> <li>✓ Outcome measure, instruments &amp; timing</li> <li>✓ No mode-of-study-selection effects</li> <li>✓ No peer-, spillover-, or carry-over effects</li> </ul>	<ul style="list-style-type: none"> <li>✓ High fidelity of treatment and control conditions</li> <li>✓ Outcome measure, instruments &amp; timing</li> <li>✓ No mode-of-study-selection effects</li> <li>✓ No peer-, spillover-, or carry-over effects</li> </ul>
<b>A2</b> Equivalence of causal estimands	<ul style="list-style-type: none"> <li>✓ ATE</li> <li>✓ effect-generating process</li> <li>✓ target population <math>P</math></li> <li>✓ setting <math>S_0 = S_1</math></li> </ul>	<ul style="list-style-type: none"> <li>✓ ATE</li> <li>✓ effect-generating process</li> <li>✓ target population <math>Q = P</math></li> <li>✓ setting <math>S_1 = S_0</math></li> </ul>
<b>A3</b> Identification	☑ ATE is identified (under RCT assumptions)	☑ ATE is identified (if confounding variables are reliably measured)
<b>A4</b> Estimation	✓ Unbiased (mean difference)	✓ Unbiased/consistent (matching estimator)
<b>A5</b> Reporting	✓ Correct reporting	✓ Correct reporting

# Goals of Replication Design Variants

- *Probe generalizability / effect heterogeneity*  
across units, treatments, observations, settings,  
time (Cronbach's UTOST)
- *Evaluate research designs* against an RCT  
benchmark estimate (within-study comparisons)  
(Lalonde, 1986; Shadish et al., 2008)  
Regression discontinuity design, nonrandomized  
pretest-posttest design (DiD), interrupted time  
series designs, matching designs  
→ implemented at the same time
- *Evaluate different estimators*  
→ can be done with the same data

# Interpretation of Replication Designs

- ▣ *Meaningful (causal) interpretation* requires that no or only one assumption is relaxed at a time
- ▣ If more assumptions do not hold simultaneously, any effect difference cannot be attributed to single sources
- ▣ *Repeating methods & procedures* may help in meeting the other replication assumptions

How to Assess Replication Success?

# How to Assess Replication Success

How one should assess replication success depends on the main goal of the replication:

- Replication of an effect's *significance*
  - > compare pattern of significance
- Replication of an effect's *size*
  - > compare and test difference in estimates
    - significance test
    - equivalence test
    - *correspondence test* with four possible outcomes
      - (1) equivalence
      - (2) difference
      - (3) trivial difference
      - (4) indeterminacy



Summary / Implications

# Summary

- ❑ Focus on *causal estimand* instead of methods & procedures
- ❑ Laid out *assumptions for a successful replication*
  - Assumptions are strong
  - Replication success is not very likely (easier in lab than field conditions)
- ❑ Derived different *replication variants* for probing effect heterogeneities, generalizability, research designs & methods

# Implications for Replication Practice

- ❑ *Plan replications prospectively* (two or multiple studies together)
- ❑ *Make data available* such that studies can reanalyze data 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!

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