Ways Out of the Replication Crisis

Peter M. Steiner
University of Maryland, College Park
Department of Human Development and Quantitative Methodology

Vivian C. Wong
University of Virginia, Charlottesville

Trier, December 11, 2019

Supported by NSF grant #2015-0285-00 and IES grant #R305D190043
Psychology’s Replication Crisis Can’t Be Wished Away

It has a real and heartbreaking cost.

By Daniel Engber
The Replication Crisis

- Replication efforts failed in psychology, economics, medicine, education, prevention sciences, ... (e.g., OSC, 2015; Camerer et al., 2016; Duvendack, Palmer-Jones & Reed, 2017; Ioannidis, 2005; Makel & Plucker, 2014; Valentine et al., 2011)

- “Why most published research findings are false” (Ioannidis, 2005)

- “Selective Inference - the silent killer of replicability” (Benjamini, 2019)

This talk focuses not only on psychology but the social sciences in general (consider not necessarily only RCTs but also field experiments, quasi-experiments, and even observational studies)
Causes of the Replication Crisis

- Significance testing
- Power issues
- Publication bias
- Questionable research practices
- Researcher degrees of freedom / the garden of forking paths
- ...
- Effect heterogeneity across populations/sites/time
- Lack of a rigorous formal definition of replication
- Lack of (optimal & practical) replications designs
- Unclarity about how to assess replication success and failure
Ways Out of The Crisis

Need for “replication sciences” and “replication as a research design” (Wong & Steiner, 2018; Steiner, Wong, Anglin, 2019; Hedges & Schauer, 2019)

- *Clear definition* of replication and assumptions to better understand issues and difficulties

- *Strong designs* for replication
  - That allow us to learn from replication success and failure
  - That acknowledge effect heterogeneities and uncertainties
  - Power considerations

- *Appropriate analyses* of replication efforts (with respect to the replication questions)
  - Careful choice of correspondence measures/tests
  - Meta-analysis

- *Open science efforts*
Definition of Replication & Assumptions
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)
- $\rightarrow$ 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 – 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 the 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)
  - False positive finding (→ publication bias)
  - 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 *definition of replication is not precise*—there is no specific criterion what replication means (Hedges & Schauer, 2019), thus, it is difficult to determine the
  - optimal replication design
  - appropriate analysis
- 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?)
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-control contrast (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
- Prospective point of view – replication as a research design
- No (a priori) prioritization of the original study

→ Focus on causal effects rather than associations/correlations (which might also reflect bias due to attrition, confounding, or other validity threats; Shadish, Cook & Campbell, 2002)
→ Focus on 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)

\[
Y_i(0) \ldots \text{potential control outcome (} T_i = 0 \text{)} \\
Y_i(1) \ldots \text{potential treatment outcome (} T_i = 1 \text{)}
\]

Average treatment effect: \( \text{ATE} = E_{R}[Y_i(1) - Y_i(0)] \)

- 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**: $\text{ATE}_R = E_R[Y_i(1) - Y_i(0)]$

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

- **Intent-to-treat effect** (ITT):
  $\text{ITT}_R = E_R[Y_i(1) - Y_i(0) \mid Z_i = 1]$

- **Average treatment effect for the treated** (ATT; no-shows):
  $\text{ATT}_R = E_R[Y_i(1) - Y_i(0) \mid T_i = 1]$

- **Complier average treatment effect** (CATE or LATE; no-shows & cross-overs):
  $\text{CATE}_R = E_R[Y_i(1) - Y_i(0) \mid \text{Compliers}]$
Hedges & Schauer (2018, 2019)

Related definition from a meta-analytic point of view:

“Replication might be defined as a situation in which effects are ‘almost the same’ across studies, such that almost the same is defined precisely [based on scientific judgement]”

- “Effect:” refers to the meta-analytic effect (size) parameter ($\theta$) which may or may not be causal
- “almost the same:” Tolerable variation in effect (size) parameters ($\theta_j$), i.e., noncentrality parameter or variance of $\theta$s distribution
- Definition is sufficient for meta-analysis not necessarily satisfactory from a causal point of view and for assessing replication success (Mathur & VanderWeele, 2019)
The Causal Replication Framework: Assumptions

(Wong & Steiner, 2018; Steiner, Wong & Anglin, 2019)
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 identified and estimable without bias

- **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 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 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, same treatment-control contrast → 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 result in incorrect conclusions about:

- whether studies aim at the same causal estimand
- whether results successfully replicate
## Example: Two Perfectly Implemented RCTs

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Original Study: RCT</th>
<th>Replication I: RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1</strong> Treatment &amp; outcome stability</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects</td>
</tr>
<tr>
<td><strong>A2</strong> Equivalence of causal estimands</td>
<td>✓ ATE ✓ effect-generating process ✓ target population ( P = Q ) ✓ setting ( S_0 = S_1 )</td>
<td>✓ ATE ✓ effect-generating process ✓ target population ( Q = P ) ✓ setting ( S_1 = S_0 )</td>
</tr>
<tr>
<td><strong>A3</strong> Identification</td>
<td>✓ ATE is identified</td>
<td>✓ ATE is identified</td>
</tr>
<tr>
<td><strong>A4</strong> Estimation</td>
<td>✓ Unbiased (mean difference)</td>
<td>✓ Unbiased (mean difference)</td>
</tr>
<tr>
<td><strong>A5</strong> Reporting</td>
<td>✓ Correct reporting</td>
<td>✓ Correct reporting</td>
</tr>
</tbody>
</table>
### Example: Two Imperfect RCTs

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Original Study: RCT</th>
<th>Replication: RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1</strong> Treatment &amp; outcome stability</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✗ Participation incentives affect potential outcomes ✓ No peer-, spillover-, or carry-over effects</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects</td>
</tr>
<tr>
<td><strong>A2</strong> Equivalence of causal estimands</td>
<td>✓ ATE ✓ effect-generating process ✓ target population $P$ ✓ setting $S_0$</td>
<td>✓ ATE ✓ effect-generating process ✗ target population $Q \neq P$ ✗ setting $S_1$</td>
</tr>
<tr>
<td><strong>A3</strong> Identification</td>
<td>✗ $\text{ATE}_P$ is not identified (due to incentives’ effect)</td>
<td>✗ $\text{ATE}_P$ is not identified (due to above issues)</td>
</tr>
<tr>
<td><strong>A4</strong> Estimation</td>
<td>✓ Unbiased estimator (mean difference)</td>
<td>✓ Unbiased estimator (mean difference)</td>
</tr>
<tr>
<td><strong>A5</strong> Reporting</td>
<td>✓ Correct reporting</td>
<td>✓ Correct reporting</td>
</tr>
</tbody>
</table>
# Example: RCT and Observational Study

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Original Study: RCT (lab)</th>
<th>Replication: Observational (field)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1 Treatment &amp; outcome stability</strong></td>
<td>✓ High fidelity of treatment and control conditions&lt;br&gt;✓ Outcome measure, instruments &amp; timing&lt;br&gt;✓ No mode-of-study-selection effects&lt;br&gt;✓ No peer-, spillover-, or carry-over effects</td>
<td>× different control condition&lt;br&gt;× different timing of measurements&lt;br&gt;✓ No mode-of-study-selection effects&lt;br&gt;× carry-over effects</td>
</tr>
<tr>
<td><strong>A2 Equivalence of causal estimands</strong></td>
<td>✓ ATE&lt;br&gt;✓ effect-generating process&lt;br&gt;✓ target population $P$&lt;br&gt;✓ setting $S_0$</td>
<td>✓ ATE&lt;br&gt;× different effect-gener. process&lt;br&gt;✓ target population $P$&lt;br&gt;× setting $S_1$</td>
</tr>
<tr>
<td><strong>A3 Identification</strong></td>
<td>✓ ATE$_p$ is identified (mean difference)</td>
<td>× ATE$_p$ is not identified (due to above issues, and maybe violation of unconfoundedness)</td>
</tr>
<tr>
<td><strong>A4 Estimation</strong></td>
<td>✓ Unbiased (mean difference)</td>
<td>✓ Unbiased/consistent estimator (matching estimator)</td>
</tr>
<tr>
<td><strong>A5 Reporting</strong></td>
<td>✓ Correct reporting</td>
<td>✓ Correct reporting</td>
</tr>
</tbody>
</table>
Replication Designs
Design Aspects

- Design variants – designs for conceptual replication
- Prospective vs retrospective replication
- Power considerations (optimal design)
- Two large vs. multiple small replication studies

Efficient design of replication efforts requires
- Clearly defined replication goal
- Predetermined analytic methods (correspondence measure/meta-analysis)
- Subject-matter knowledge about underlying data-generating processes (effect heterogeneities)
Causal Replication Design Variants

Derivation of causal replication design variants (→ conceptual replication)

Instead of attempting to meet all replication assumptions (direct replication), researchers might

- systematically *relax* one (or more) assumptions
- while meeting all other assumptions, such that
- a meaningful interpretation of replication success of failure is maintained
## Effect Heterogeneity across Populations

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Original Study: RCT</th>
<th>Replication: RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1</strong> Treatment &amp; outcome stability</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects</td>
</tr>
<tr>
<td><strong>A2</strong> Equivalence of causal estimands</td>
<td>✓ ATE ✓ effect-generating process ✓ target population $P$ ✓ setting $S_0 = S_1$</td>
<td>✓ ATE ✓ effect-generating process ✓ target population $Q \neq P$ ✓ setting $S_1 = S_0$</td>
</tr>
<tr>
<td><strong>A3</strong> Identification</td>
<td>✓ $\text{ATE}_P$ is identified</td>
<td>✓ $\text{ATE}_Q$ is identified</td>
</tr>
<tr>
<td><strong>A4</strong> Estimation</td>
<td>✓ Unbiased (mean difference)</td>
<td>✓ Unbiased (mean difference)</td>
</tr>
<tr>
<td><strong>A5</strong> Reporting</td>
<td>✓ Correct reporting</td>
<td>✓ Correct reporting</td>
</tr>
</tbody>
</table>
# Effect Heterogeneity across Settings

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Original Study: RCT</th>
<th>Replication: RCT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A1</strong> Treatment &amp; outcome stability</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects</td>
<td>✓ High fidelity of treatment and control conditions ✓ Outcome measure, instruments &amp; timing ✓ No mode-of-study-selection effects ✓ No peer-, spillover-, or carry-over effects</td>
</tr>
<tr>
<td><strong>A2</strong> Equivalence of causal estimands</td>
<td>✓ ATE ✓ effect-generating process ✓ target population $P = Q$ ✓ setting $S_1$</td>
<td>✓ ATE ✓ effect-generating process ✓ target population $Q = P$ ✓ setting $S_1 \neq S_0$</td>
</tr>
<tr>
<td><strong>A3</strong> Identification</td>
<td>✓ ATE is identified</td>
<td>✓ ATE is identified</td>
</tr>
<tr>
<td><strong>A4</strong> Estimation</td>
<td>✓ Unbiased (mean difference)</td>
<td>✓ Unbiased (mean difference)</td>
</tr>
<tr>
<td><strong>A5</strong> Reporting</td>
<td>✓ Correct reporting</td>
<td>✓ Correct reporting</td>
</tr>
</tbody>
</table>
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.
Specific Replication Designs

- **Full factorial experimental designs** with one or multiple factors/blocks representing settings/sites/populations/time (Fisher, 1935)
- **Stepped-wedge designs**: participants are randomized to receive treatments in successive waves over time – replication over time (Hussey & Hughes, 2007)
- **Switching replication design**
- **Consecutive cohort design**
- **Multisite designs**
- **Doubly-randomized trials (design-replications, preference effects)**
Prospective vs. Retrospective Designs

**Prospective design**
- allow researchers to plan the entire replication effort in advance
- designed for a specific replication goal
- original study is not prioritized
- power calculations can be done for all studies together (optimal design; Schauer, 2018)

**Retrospective design**
- convenient for ad hoc replications of published results but often focus on different causal estimands (violations of A1 to A4) and lack power
Two vs Multiple Replication studies

Hedges & Schauer (2019) and Benjamini (2019) argue for *smaller but multiple replication studies*

- two studies have rarely sufficient power with respect to the replication question
- better cover effect heterogeneities across studies

However, it might be harder to systematically control or vary the causal estimand across studies (i.e., meeting assumptions A1 and A2)

- *Response surface modeling* (if sources of effect heterogeneity have mean measured)
Analysis of Replication Efforts
Pairwise Comparisons vs Meta-Analysis

Pairwise comparisons

- Directly related to *replication question*: can the results of one study be successfully replicated (Steiner & Wong, 2018; Mathur & VanderWeele, 2019)
- Different *correspondence measures*: Conclusion or distance-based measures

Meta-analysis

- Assessment criterion: *effect heterogeneity*
- More *power* with multiple studies
Choice of Correspondence Measures

- Conclusions drawn from a replication depend on the correspondence measure.

- In practice, correspondence measure are often chosen ad hoc, without careful considerations of the replication goal (Anderson & Maxwell, 2016a; Hedges & Schauer, 2019).

- The outcome of both studies need to be considered as stochastic (i.e., sampling uncertainty).
  - Do not consider measures that test whether the confidence interval covers the (fixed) effect of the original study (e.g., OSC, 2015).
Correspondence Measures

- Conclusion-based measures
  Would researchers and policy makers draw the same conclusion from each of the two studies? That is, has the intervention a meaningful effect or not?

- Distance-based measures
  Is the difference in effect estimates small / large enough to claim a significant equivalence / difference?

Conclusion-based Correspondence

- **Direction of effects**
  Is the sign of estimates identical?
  \[ C^D_C = 1[\text{sgn}(\hat{\tau}_i) = \text{sgn}(\hat{\tau}_{II})] \]

- **Magnitude of effects**
  Do the estimates exceed a certain magnitude?
  \[ C^M_C = 1[(\hat{\tau}_i \geq \lambda \& \hat{\tau}_{II} \geq \lambda) \text{ or } (\hat{\tau}_i < \lambda \& \hat{\tau}_{II} < \lambda)] \]

- **Statistical significance pattern**
  Is the significance of estimates identical?
  \[ C^S_C = 1[\text{sgn}(\hat{\tau}_i) = \text{sgn}(\hat{\tau}_{II}) \& p_i \leq \alpha \& p_{II} \leq \alpha \text{ or } (p_i > \alpha \& p_{II} > \alpha)] \]
Conclusion-based Correspondence

Correspondence of results depends on

- **Power** of study 1 & 2 to detect the *true but unknown effect*
  - Size of the true effect
  - Sample size
  - Error variance
- Direction and magnitude of *bias* (if any) in study 1 & 2

Note: the *minimum detectable effect size (MDES)* of study 1 & 2 only reflects sample size and error variance
Distance-based Correspondence

Investigates the estimated effect difference $\hat{t}_1 - \hat{t}_2$

- **Difference test**
  Is the difference in estimates insignificant? (two-sample $t$-test)

- **Equivalence test**
  Is the equivalence in estimates significant (with respect to a given equivalence threshold)? (Tryon, 2001; Tryon & Lewis, 2008)

- **Correspondence test**
  Combines the difference and equivalence test (Tryon & Lewis, 2008; Steiner & Wong, 2018)
Difference Test

- Standard null-hypothesis significance test (NHST) -> two-sample $t$-test
- Equivalence of effects is formulated as null hypothesis $H_0: \tau_1 - \tau_2 = 0$
- Correspondence requires insignificant $t$-test $C_D^D(\alpha) = 1[p_D > \alpha]$

Issues

- Incorrect interpretation of failure to reject the null
- Lack of power may result in failure to reject the false null hypothesis (no difference in effects), and thus an invalid correspondence conclusion
  - occurs even if both studies are sufficiently powered for a given MDES
Equivalence Test

- Uses NHST, but tries to overcome the difference test’s weakness

- **Equivalence** is formulated as alternative hypothesis with respect to an equivalence threshold $\delta_E$
  \[
  H_0: |\tau_1 - \tau_2| \geq \delta_E \quad H_1: |\tau_1 - \tau_2| < \delta_E
  \]

- The composite null hypothesis can be reformulated as two one-sided hypotheses
  \[
  H_{01}: \tau_1 - \tau_2 \geq \delta_E \\
  H_{02}: \tau_1 - \tau_2 \leq -\delta_E
  \]

- Equivalence can be tested with two one-sided $t$-tests

- Correspondence requires two significant $t$-tests
  \[
  C_D^E(\delta_E, \alpha) = 1[p_{E1}(\delta_E, \alpha) \leq \alpha \& p_{E2}(\delta_E, \alpha) \leq \alpha]
  \]
Equivalence Test

Issues

- Determination of equivalence threshold \( \delta_E \)
  - Effect difference that is substantively inconsequential or trivial
  - Small thresholds (.1 SD or smaller) require large sample sizes in each study
- Lack of power may result in a failure to reject the false null hypothesis (difference)
## Correspondence Test

- Combines the difference and equivalence test into a single test with four possible outcomes

<table>
<thead>
<tr>
<th>Difference ($C^D$)</th>
<th>Equivalence ($C^E$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C^E = 0$</td>
</tr>
<tr>
<td></td>
<td>insig. equivalence</td>
</tr>
<tr>
<td></td>
<td>$C^E = 1$</td>
</tr>
<tr>
<td></td>
<td>sig. equivalence</td>
</tr>
<tr>
<td>$C^D = 0$</td>
<td>Difference</td>
</tr>
<tr>
<td>sig. difference</td>
<td>Trivial Difference</td>
</tr>
<tr>
<td>$C^D = 1$</td>
<td>Indeterminacy</td>
</tr>
<tr>
<td>insig. difference</td>
<td>Equivalence</td>
</tr>
</tbody>
</table>
Correspondence Test

- explicitly deals with lack of power
- Correctly indicates equivalence or difference with high probability if both studies are sufficiently powered
- With insufficiently powered studies, indeterminacy is the most likely outcome
- Choice of equivalence threshold is crucial (affects power and test outcome)
Meta-Analysis

Hedges & Schauer (2018, 2019)
- **Fixed or random sample** of replication studies
- **Two vs multiple replications**
- If effect (size) estimates are “almost the same” conclude that effects are replicable

*Meaningful interpretation* ideally requires that the effects refer to the *same causal estimand* or that *heterogeneity can be modeled* (response surface modeling)
Conclusions:
Ways Out of The Replication Crisis
**Conclusions**

Important steps to overcome the replications crisis

- *Clear definition* of replication and assumptions
- Strong *prospective replication designs* (for direct and conceptual replication)
- Appropriate and sufficiently powered *analyses* for assessing replication success

- **Open Science**
  - Preregistration of replication efforts
  - Open access to data of replication studies

Acknowledge incomplete/partial knowledge in a changing world (Manski, 2013)

- Planning/decision making with partial knowledge
- Diversified treatment choices
Thank You!

psteiner@umd.edu