What is a Replication?

Replication Summer School 2026 — Day 1 of 3

Branden Fitelson

18 May 2026

The Three R's

(of Trustworthy Science)

- **Replicability**. Obtaining *consistent results* across studies aimed at answering the same scientific question, each of which has obtained its own data.
- **Reproducibility**. Obtaining *consistent computational results* using the *same input data, methods, code,* and conditions of analysis [13].
- **Robustness**. Obtaining *consistent conclusions* across legitimate alternative analytic/modeling choices and assumptions — *e.g.*, multiverse analysis [18], sensitivity to the "garden of forking paths" [4].

☞ All three concern the *trustworthiness* of scientific findings. This lecture is about the first one. (Margherita will be discussing the second R this afternoon.)

Defining Replication

Replicability asks: do two studies of *the same scientific question* give consistent results?

But what makes two studies "the same"? Two experiments are *never* literally identical (different participants, different day, different lab, …).

There is a large *gray area*: studies that use *similar but not identical* methodology.

☞ When two such studies disagree, we need a principled answer to: *which differences should count as a failure to replicate, and which should not?*

Framework: **Cronbach**'s UTOS + **Machery**'s resampling.

Cronbach & Shapiro [3]: UTOS. A framework for the constituents of an experiment.

- **U**nits — the elements of the population from which the study sample has been drawn.
- **T**reatments — the independent variables (*e.g.* experimental interventions, population subgroups).
- **O**utcomes — the dependent variables on which the study is focused.
- **S**etting — the remainder of factors that can affect the outcome and its relation to the treatments.

☞ UTOS gives us a vocabulary, but does not tell us *which* of the four components a replication is permitted to vary. For that, we turn to Machery …

Machery [10]: the Resampling Account of Replication.

Each component of an experiment is either *fixed* or *random*:

- *random* components are sampled from a population (*e.g.*, the participants in a study);
- *fixed* components are not (*e.g.*, the experimental manipulation under test).

Resampling Account. *E' counts as a replication of E* iff E' holds E's fixed components *fixed* and *resamples* (some of) E's random components.

☞ A *successful* replication E' of E confirms that the original result R was not an outlier of the (re)sampling process.

Example. A study tests a depression treatment on a sample of participants drawn from a population.

A **replication** (in Machery's sense): another study using *the same protocol, definitions, and measures*, but collecting *another sample* of participants from the same population.

- Units (participants — random component, resampled),
- Treatment, protocol, outcome measure — fixed components (held fixed).

☞ Machery relates E's *replicability* to its *reliability* — whether E's outcome was a stable feature of the setup, or an artifact of who happened to be sampled.

Replication *vs.* Extension

Replications in Machery's strict sense probe *reliability*. But scientists also speak of **conceptual replication/extension**.

Conceptual replication: *"a replication experiment with similar aims and methodology **but** important differences."*

Example. A different study tests the same depression treatment, but uses a *different theoretical definition* or a *different measure* of depression.

☞ In Machery's terms, a conceptual replication $E^\star$ *changes some fixed components*. So, strictly speaking, $E^\star$ is a *different experiment* targeting the same phenomenon P.

From this POV, $E^\star$ probes the (external) *validity* (*viz.*, *generalizability*) of E, not the *reliability* of E *per se*.

Rosenthal [16] introduces a graded notion: the **precision** of a replication — *the degree of similarity between the replication and the original*.

A maximally precise replication is a strict (Machery-style) replication: only random components vary. As precision drops, more fixed components may also be varied.

☞ *Reliability/Validity Precision Tradeoff*

- higher precision ⇒ more probative test of E's *reliability*, but less informative about its *validity*;
- lower precision ⇒ potentially more informative about E's *validity*; but, less probative of E's *reliability*.

Should the field prioritize *strict* replications or *extensions* (conceptual replications)? An ongoing debate:

- **Simons [17]**: "The value of direct replication." Without strict replications, we never establish that a phenomenon P is even present in the first place.
- **Crandall & Sherman [2]**: extensions deliver more *theoretical* information per study: they test the phenomenon *and* a hypothesis about its scope.

☞ In a sense, replication is *more fundamental* than extension. If an extension $E^\star$ of E *itself* fails to replicate, this undermines $E^\star$'s evidential value (even wrt E's validity).

But, if E is successfully (and independently) replicated, then this supports E's reliability — *even if* E does not generalize.

Phenomena differ in how **context-sensitive** they are [19, 7, 6].

Context-sensitive effects — social-psychological priming, behavioral interventions, educational treatments — depend strongly on cultural, historical, and demographic factors.

"Portable" effects — low-level perception, sensorimotor learning, classical conditioning — look similar across settings and populations.

☞ Replication (and extension) expectations should be calibrated to the kind of effect. A failure to replicate (or extend) a context-sensitive effect may be less informative than a failure to replicate (or extend) a portable one.

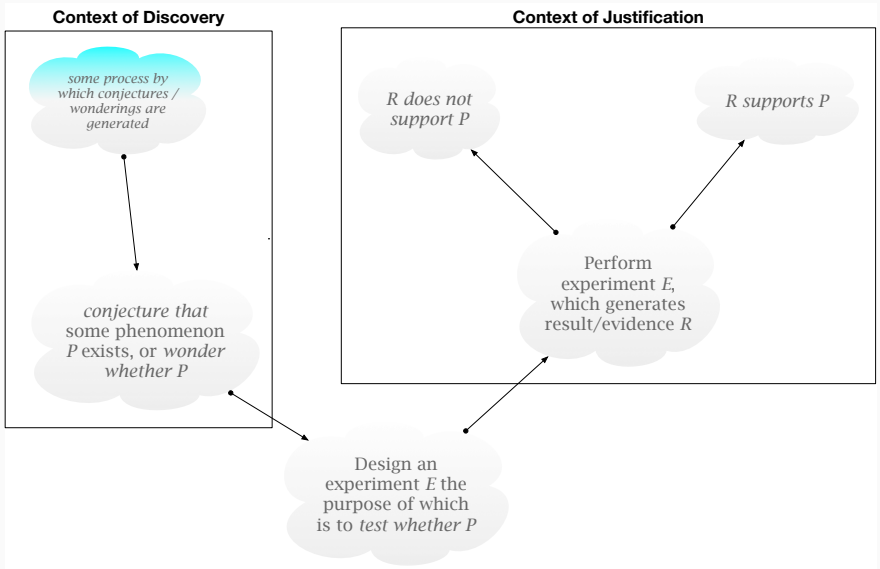
Rosenthal [16] also flagged the problem of **correlated replicators**.

Replications carried out by researchers who share theoretical commitments, training, networks, or methodological habits tend to produce *similar* results.

Original authors replicating their *own* work are an extreme case. "Many labs" multi-site projects [8] are an attempt to ameliorate these sorts of effects.

☞ Strong evidence for E's reliability requires not just *many* replications, but replications by *independent, skeptical* replicators.

Replication & Evidential Support



Suppose an experiment E produces a result R, which is claimed to *support* a phenomenon-claim P (think of $\neg P$ as the null; R as a "significant" result).

R counts as supporting P *only if* replications E' of E tend to produce results R' that also support P.

On the other hand, if

(D) replications E' of E tend to produce results R' that do *not* support P,

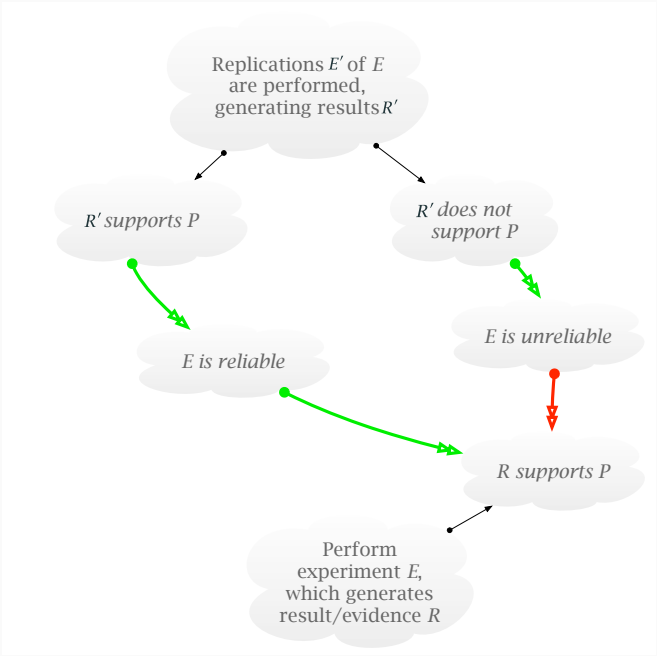
☞ then D is a *defeater* [11] of R's support for P (*via supporting E's unreliability*). *Failures of replication serve as defeaters of the evidential support of (the results of) experiments.*

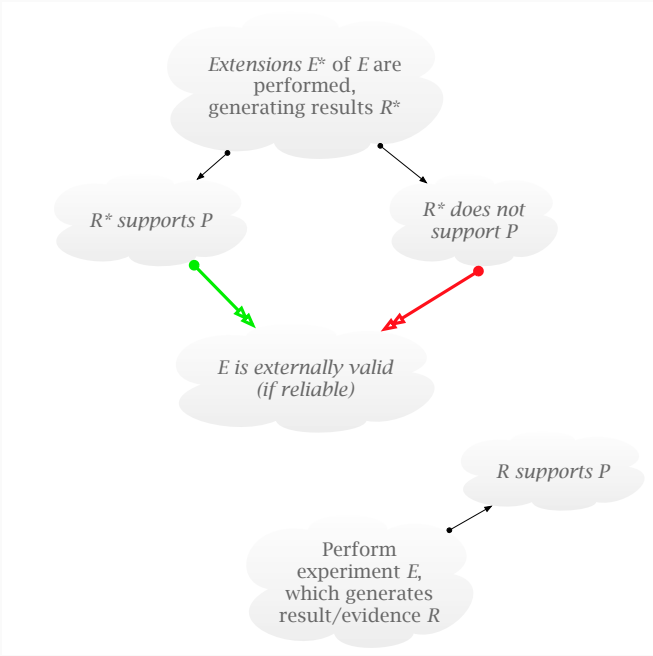
There are 2 ways D might disrupt R's support for P [9, 11].

- D **rebuts** R's support for P if D is a reason to believe $\neg P$, *e.g.*, D is the result of a different, more powerful experiment $E^\star$ yielding the opposite conclusion as E.
- D **undercuts** R's support for P if D disrupts the inference from R to P *without telling against P itself — e.g., D disconfirms/rebuts E's reliability.*

☞ Failures to replicate E generally **undercut** the support R provides for P — they don't rebut/disconfirm P itself.

Failed *extensions* ($E^\star$) are not defeaters of R's support of P (in either sense). Rather, they call into questions E's *validity*.





Fun Example: “perytons” at Parkes Observatory [15, 14, 5].

A *fast radio burst* (FRB) is a transient radio pulse caused by some high-energy astrophysical process (not yet fully understood); first detected in 2007.

In 2015, Parkes believed it had discovered a new kind of FRB. Other observatories failed to replicate the detection.

The replication failure *undercut* the FRB *inference*, and led to a deeper investigation into the cause of the “bursts.”

☞ Diagnosis: the “bursts” were caused by the observatory’s *microwave ovens* (which tended to operate at the same time each day). Such man-made bursts are now called **perytons**.

References

[1] `http://filippogambarota.github.io/replicability-book/`
[2] C. Crandall and J. Sherman, *On the scientific superiority of conceptual replications…*, 2016.
[3] L. Cronbach and K. Shapiro, *Designing Evaluations of Educational and Social Programs*, 1982.
[4] A. Gelman and E. Loken, *The Garden of Forking Paths…*, 2013.
[5] Wikipedia, *Fast Radio Burst*, 2023.
[6] M. Gollwitzer and U. Schwabe, *Context dependence as a predictor of replicability*, 2022.
[7] Y. Inbar, *Association between contextual dependence and replicability in psychology…*, 2016.
[8] R. Klein et al, *Many Labs: Investigating variation in replicability…*, 2014.
[9] M. Kotzen, *A Formal Account of Epistemic Defeat*, 2019.
[10] E. Machery, *What is a replication?*, 2020.
[11] L. Moretti and T. Piazza, *Defeaters in current epistemology*, 2018.
[12] NASEM, *Reproducibility and Replicability in Science*, 2019.
[13] R. Peng, *Reproducible research in computational science*, 2011.
[14] Wikipedia, *Peryton (astronomy)*, 2023.
[15] R. Petroff *et al*, *Identifying the source of perytons at the Parkes radio telescope*, 2015.
[16] R. Rosenthal, *Replication in behavioral research*, 1990.
[17] D. Simons, *The value of direct replication*, 2014.
[18] S. Steegen et al, *Increasing transparency through a multiverse analysis*, 2016.
[19] J. Van Bavel et al, *Contextual sensitivity in scientific reproducibility*, 2016.