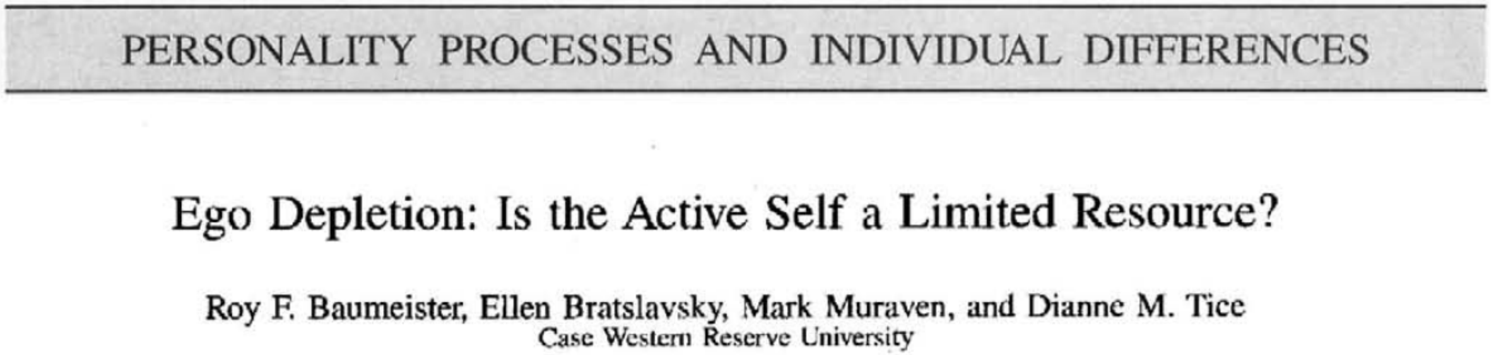
Moving From What Works to What Replicates: Promoting the Systematic Replication of Results

A Few Reflections and Two Questions

James E. Pustejovsky

American Educational Research Association 2023 Convention
April 15, 2023
Chicago, IL

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

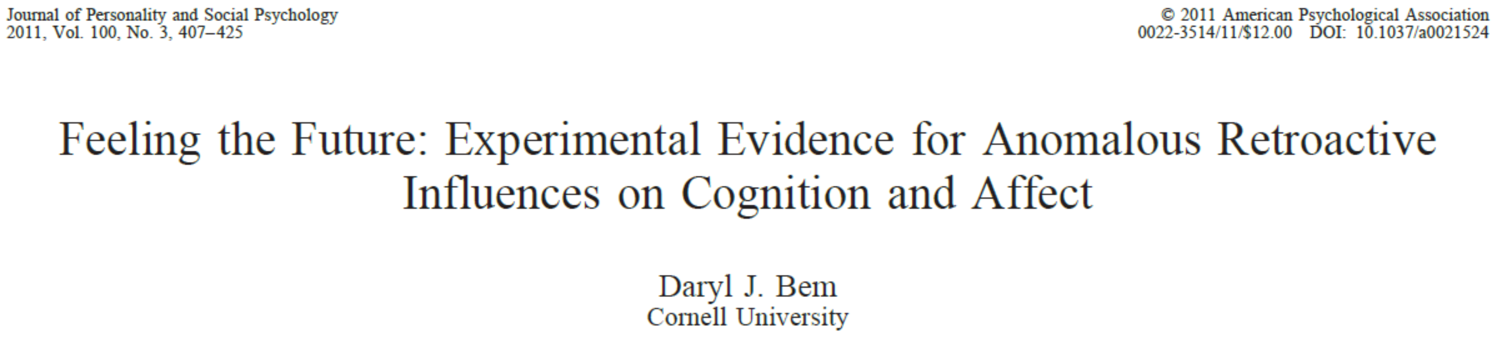


PERSONALITY PROCESSES AND INDIVIDUAL DIFFERENCES

Ego Depletion: Is the Active Self a Limited Resource?

Roy F. Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice
Case Western Reserve University

> Choice, active response, self-regulation, and other volition may all draw on a common inner resource. In Experiment 1, people who forced themselves to eat radishes instead of tempting chocolates subsequently quit faster on unsolvable puzzles than people who had not had to exert self-control over eating. In Experiment 2, making a meaningful personal choice to perform attitude-relevant behavior caused a similar decrement in persistence. In Experiment 3, suppressing emotion led to a subsequent drop in performance of solvable anagrams. In Experiment 4, an initial task requiring high self-regulation made people more passive (i.e., more prone to favor the passive-response option). These results suggest that the self's capacity for active volition is limited and that a range of seemingly different, unrelated acts share a common resource.



Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

© 2011 American Psychological Association
0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

> The term psi denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are precognition (conscious cognitive awareness) and premonition (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process.... This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (d) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

Open science is essential

- Absolutely vital that planned replication studies are ***fully and completely reported***.

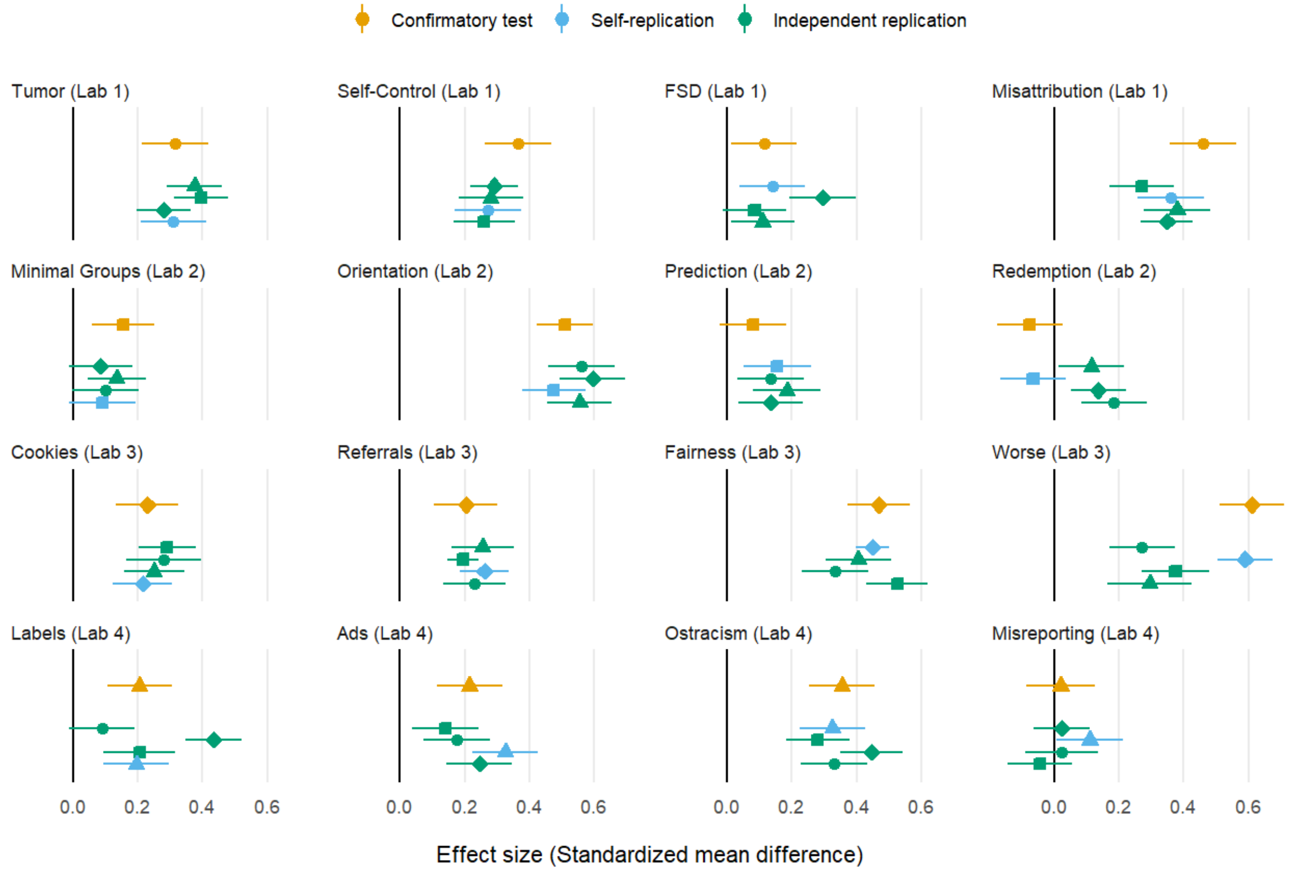
PREREGISTERED
OPEN DATA
OPEN MATERIALS

Expect some imperfections

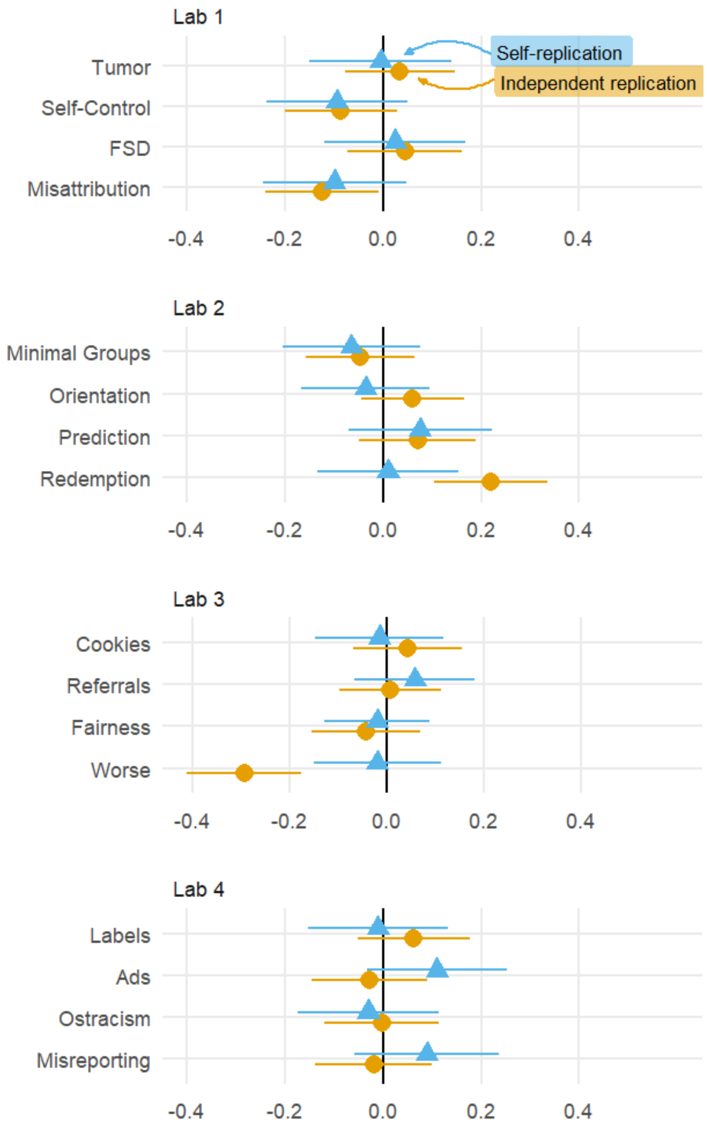
- Causal Replication Framework (Wong & Steiner, 2018) is framed in terms of **exact equivalence of causal estimands**.
    - Slough and Tyson (2022) describe conditions of "***target equivalence***" for interpretability of meta-analyses.
- But exact equivalence is very stringent. We should expect some imperfection, even in very close replications.

Close replications of online experiments

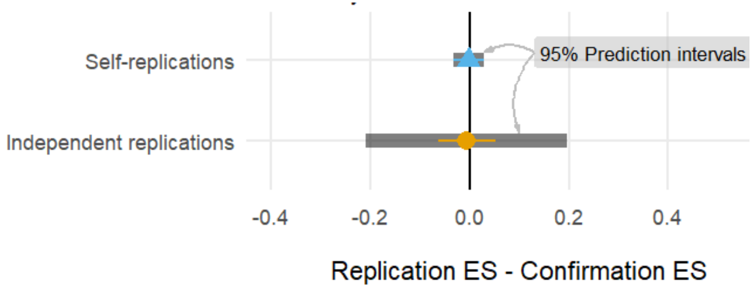
- Protzko and colleagues (2022) conducted multi-lab replications of 16 pre-registered online experiments of basic social/behavioral effects.



Differences between replication effect sizes and original effect sizes



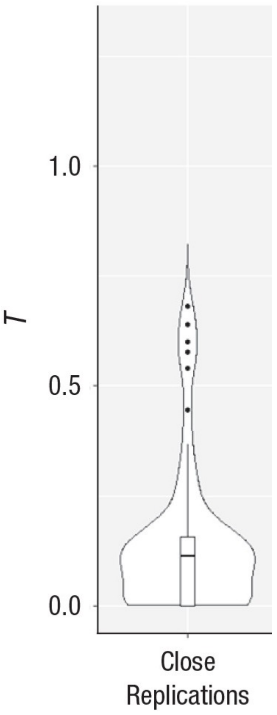
Meta-analysis of differences between replication effects and original effects



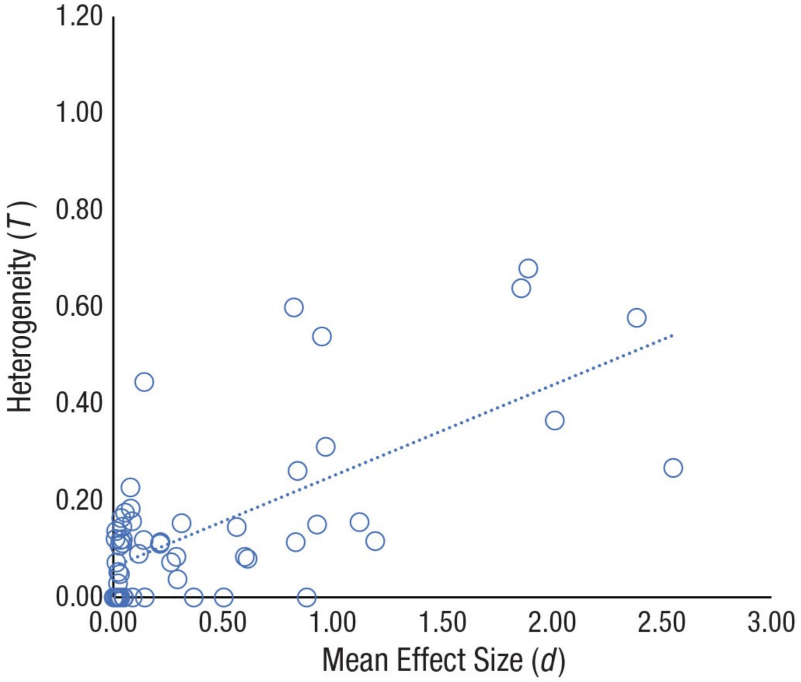
- Estimated heterogeneity of $\hat{\tau} = 0.09$ (95% CI: 0.06, 0.16) for the difference between original study and independent replications.

Heterogeneity in multi-lab replication studies

- In a systematic review of 57 pre-registered, multi-lab replication studies, Linden and Honekopp (2021) found median cross-lab heterogeneity of $\hat{\tau} = 0.09$ (see also Olsson et al., 2020)



Source: Linden and Honekopp (2021), Figure 4



Source: Linden and Honekopp (2021), Figure 5

What are we replicating?

- Cohen and colleagues' study focused on replication of **average causal effects** of coaching supports in mixed reality simulation.

		Effect Size		Adjusted Effect Size	
Study	**Source of Variation**	**Est.**	**(SE)**	**Est.**	**(SE)**
Study 1	Timing of study	1.69	(0.22)	1.45	(0.33)
Study 2	Teaching task	1.41	(0.21)	1.42	(0.21)
Study 3	Reference	1.41	(0.21)	1.41	(0.21)
Study 4	Participant background	0.39	(0.23)	0.67	(0.26)
Study 5	Delivery mode	1.62	(0.19)	1.37	(0.23)

Source: Cohen, Erickson, Krishnamachari, & Wong (2023), Table 2

- Should we focus only on average impact estimates, or should we also think about replication of ***other features of the data-generating process***?
  - Moderation by participant characteristics?
  - Outcome variance / variance ratios?
  - Coach-specific variance in outcomes?

What's the role of *independent* replication?

- Some research traditions put special emphasis on independent replications, conducted by ***researchers other than original investigators***.
- Knowledge claims and methods are conveyed through language and require interpretation.
    - From Cohen et al.:

      > To support candidates' practice and learning in the simulation sessions, we employ a directive, 4-step coaching model where coaches provide targeted feedback on a specific set of instructional skills. The coach first observes the candidate's simulated practice and diagnoses the instructional needs along a skill progression. Second, the coach gauges the candidate's perception of their performance (e.g., "How are you feeling about the simulation?") before identifying strengths and improvement targets. Third, the coach provides detailed information about the features of high-quality enactment of the targeted skill, how and why it supports positive student outcomes, and specific strategies the candidate can utilize in subsequent simulations. Finally, the coach engages in a role-playing exercise with the candidate, providing opportunities to rehearse a targeted skill.

    - Independent replication tests the ***sufficiency of description*** and can help to ***surface implicit assumptions*** embedded in a claim or method.
- Independent replication also ***distributes knowledge and expertise*** across multiple investigators and requires focusing on questions that are of ***collective interest***.

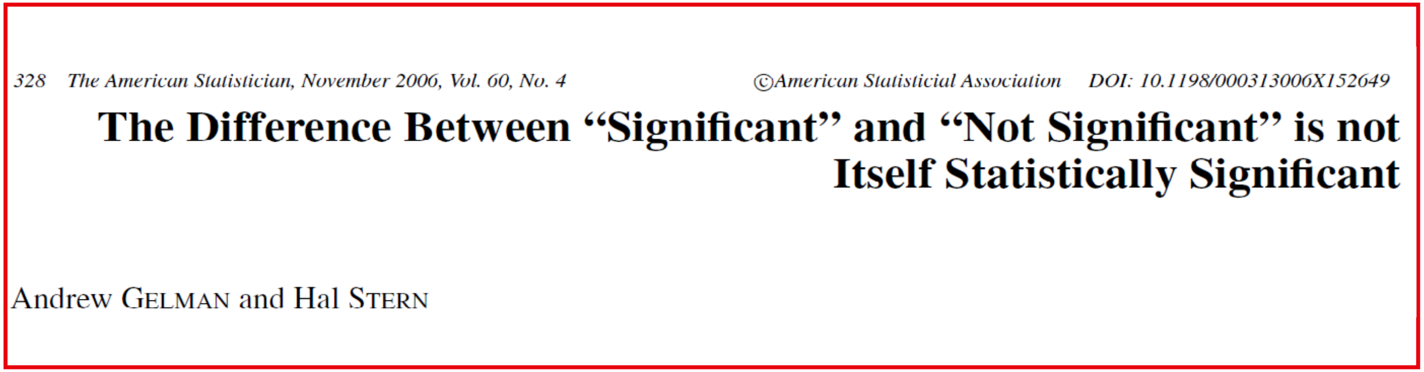
Extra

Anticipating synthesis

- Replication designs described by Wong, Anglin, and Steiner (2022) will contribute ***complex*** and ***multi-faceted*** evidence for research syntheses.
    - Replicators will need to think carefully about how their results will fit into a future synthesis.

Correspondence in significance is dumb

- Correspondence in significance patterns is **a bad measure** of replication.
  Can we please stop using it?



328 The American Statistician, November 2006, Vol. 60, No. 4 ©American Statisticial Association DOI: 10.1198/000313006X152649

The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant

Andrew GELMAN and Hal STERN