Analyzing and Optimizing Replicability in the Behavioral, Social, and Cognitive Sciences

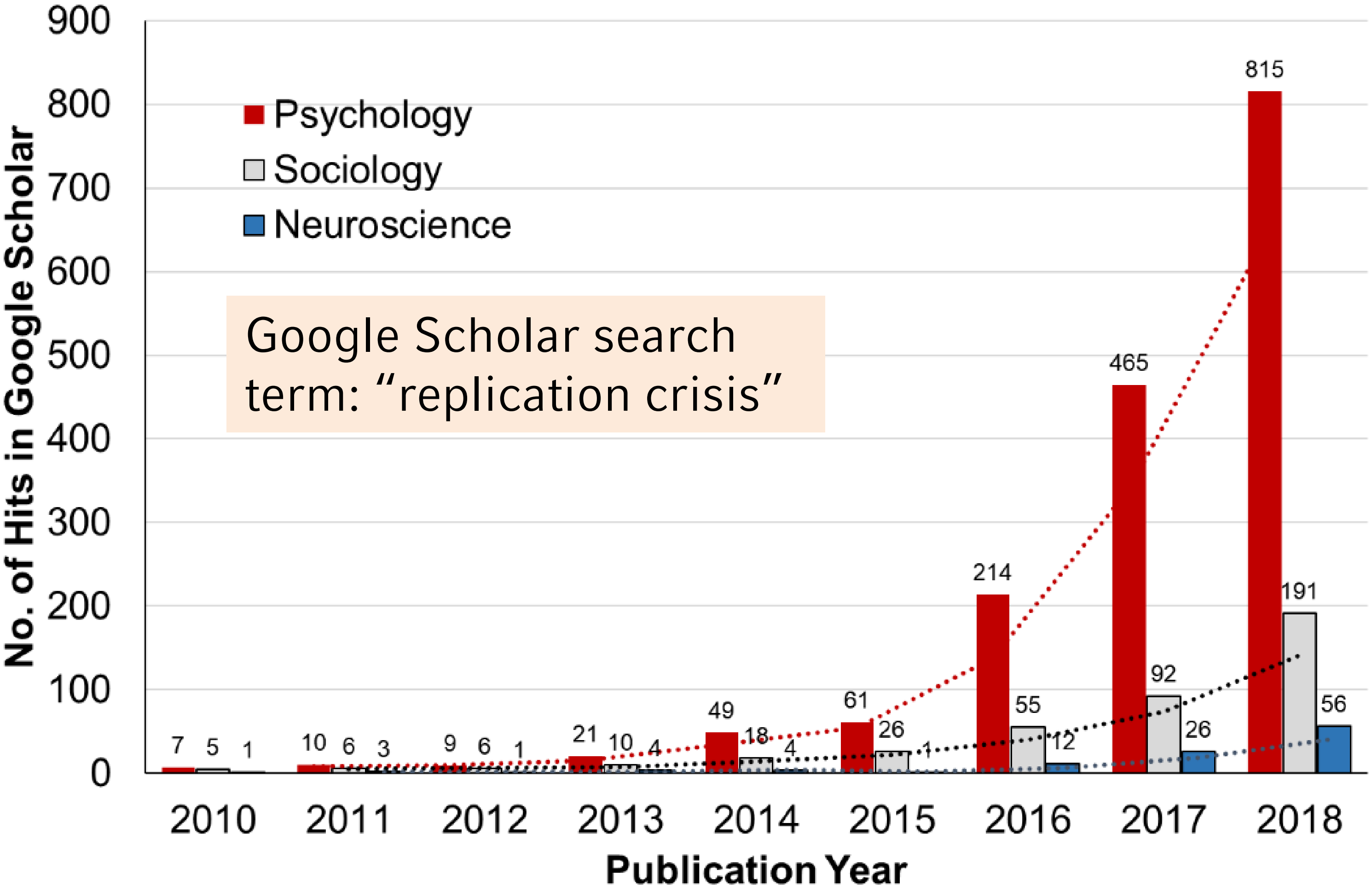
Mario Gollwitzer

The Symptoms

- Many empirical findings are apparently non-replicable:
  - RP:P Project (OSC, 2015): 100 selected findings (social/cognitive psych); one "direct replication" per finding; **replication success: 39%**
  - ManyLabs 1 (Klein et al., 2014; Social Psychology): 13 selected findings (social/econ); 36 samples each; **replication success: 77%**
  - ManyLabs 2 (Klein et al., 2018; AMPPS): 28 selected findings (social/cog/econ); >60 samples each; **replication success: 54%**
  - ManyLabs 3 (Ebersole et al., 2016; JESP): 10 selected findings (social psych); 20 samples each; **replication success: 30%**

  … (more ManyLabs/RRR projects on individual effects; even more underway)
- Replication rates lower in life sciences and neurosciences, higher in behavioral economics (e.g., Begley & Ellis, 2012; Camerer et al., 2016; Camerer et al., 2018; Prinz, Schlange, & Asadullah, 2011)

The Symptoms



Google Scholar search term: “replication crisis”

The Cure (I): Raising Methods Standards?

- Conduct sufficiently powered studies; justify sample size determination
- pre-register materials, design, hypotheses, and analyses
- correct for errors prior to submission (e.g., by using StatCheck; *PsychScience*)
- stricter significance levels (e.g., Ioannidis, 2018)
- report confidence interval estimates (e.g., *PSPB*)
- abandon NHST (and use Bayesian inference instead)
- ban the use of inferential statistics altogether (Trafimov & Marks, 2015; *BASP*; but see Fricker et al., 2019)

The Cure (II): Open Science?

1. **Compliance with reporting standards**: report and justify analytical decisions in detail; report all basic and supplementary analyses in addition to main analyses in the paper or in the SOM
2. **Open materials**: provide all materials (e.g., stimuli, items) used in study; provide videos or protocols describing the experimental procedure
3. **Preregistration** of hypotheses, operationalizations, analysis plan/code, sampling procedure, power analyses; clear distinction between confirmatory and exploratory analyses
4. **Open data**: compliance with the FAIR principles (Wilkinson et al., 2016); compliance with data documentation standards ("meta-data")
5. **Reproducible analysis code**
6. **Sharing research output** and assessment; publication of pre-prints or green/golden open access post-prints; post-publication peer review

Three Perspectives

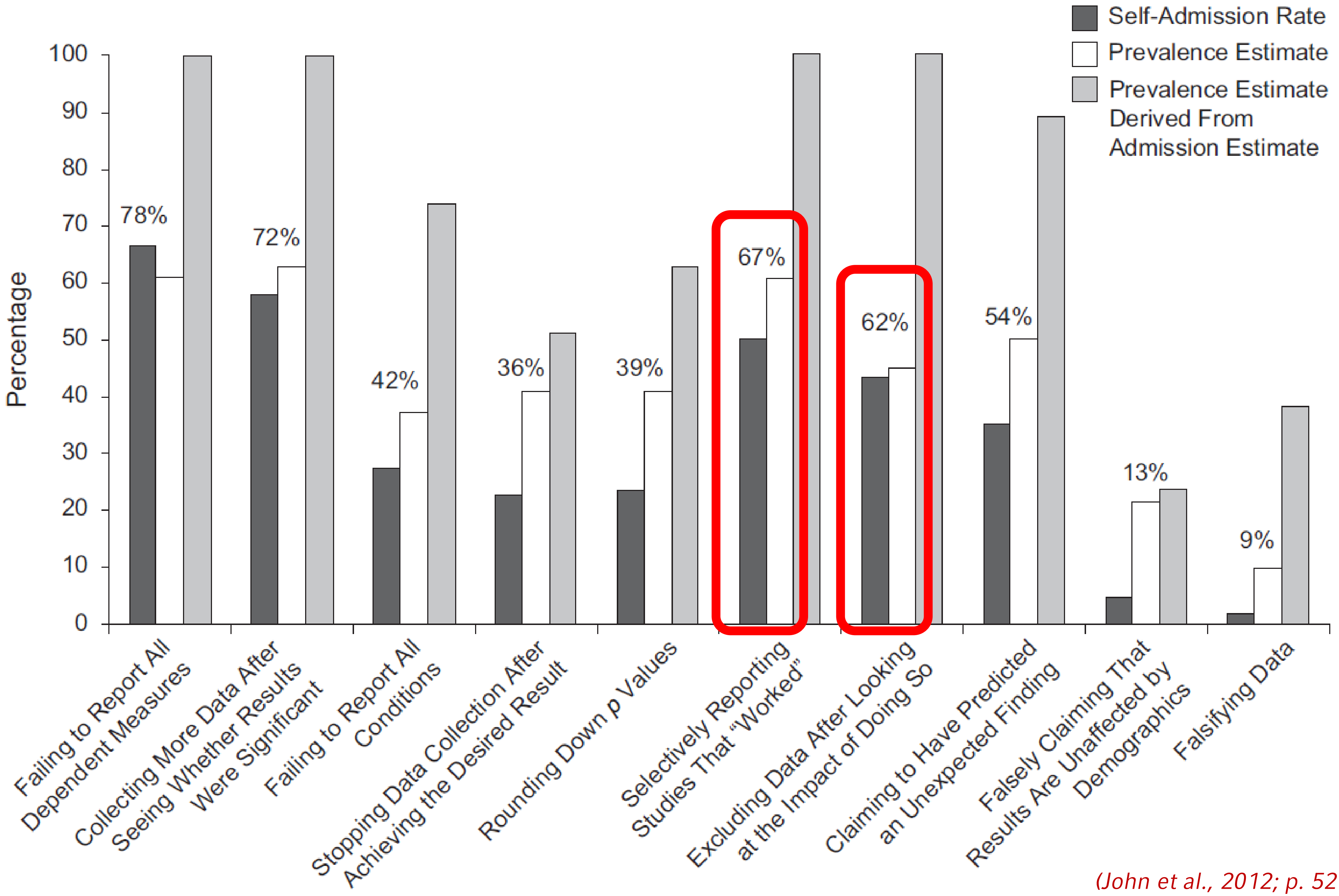
1. **False-Positive Perspective**
   - *“The literature is full of false positives that are obviously non-replicable.”*

2. **Context-Dependency Perspective**
   - *“Many effects are contingent on contextual conditions; if these are absent, the effect cannot be replicated.”*

3. **False-Negative Perspective**
   - *“The results from current replication projects underestimate true replicability rates.”*

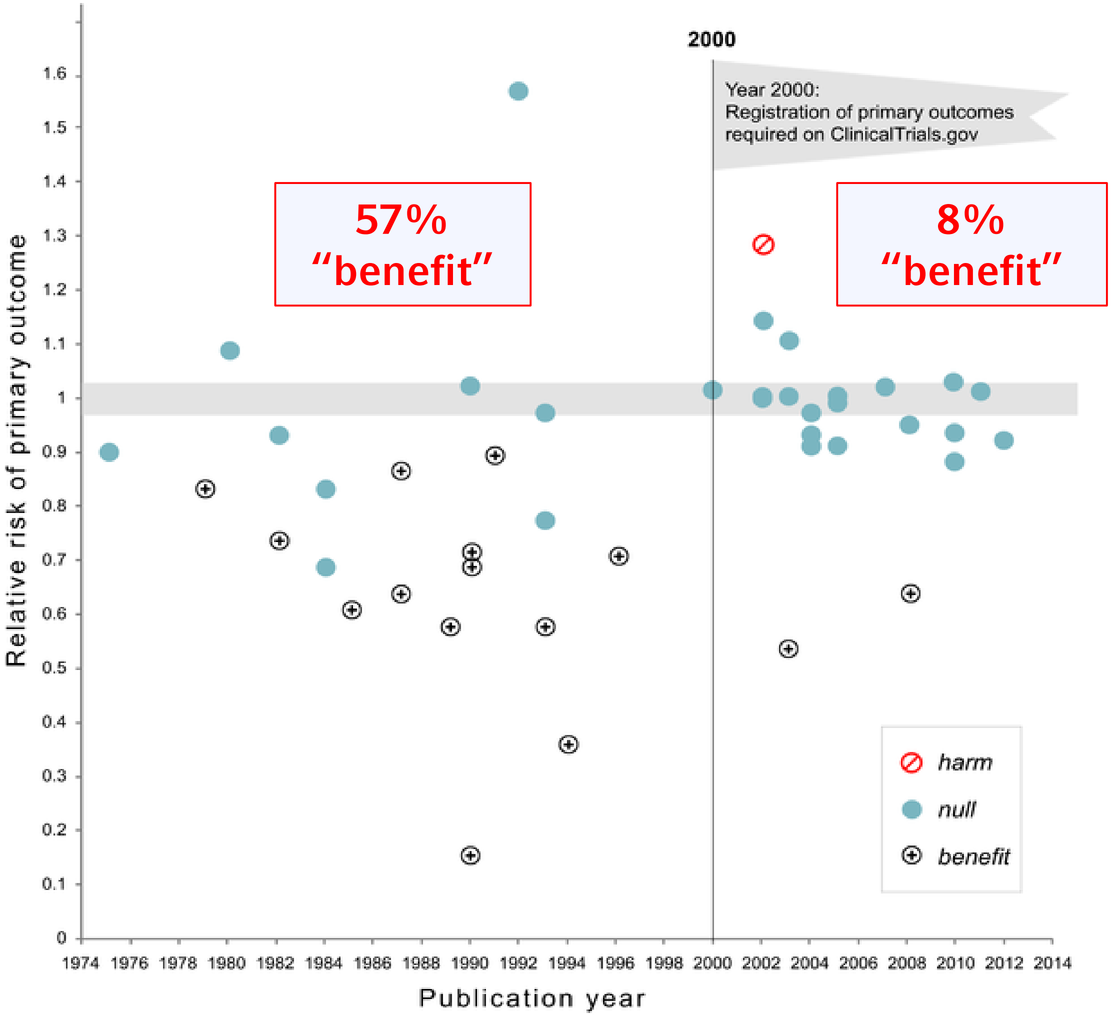
1. False-Positive Perspective

- What’s the evidence?
  - Relatively high prevalence of „questionable research practices“ (QRPs; Bakker et al., 2012; John et al., 2012; Simmons et al., 2011; Fiedler & Schwarz, 2015)
  - QRPs can inflate false-positive rates (e.g., Francis, 2012)
  - “Closed science” culture (Wicherts et al., 2011); many errors (incl. honest mistakes) remain undetected
  - Current incentive structure rewards quantity over quality; speed over accuracy; hypothesis-confirming over disconfirming findings (e.g., Ioannidis, 2012; Nosek et al., 2012; Smaldino & McElreath, 2016); publication bias



(John et al., 2012; p. 527)

1. False-Positive Perspective



Year 2000:
Registration of primary outcomes
required on ClinicalTrials.gov

57%
"benefit"

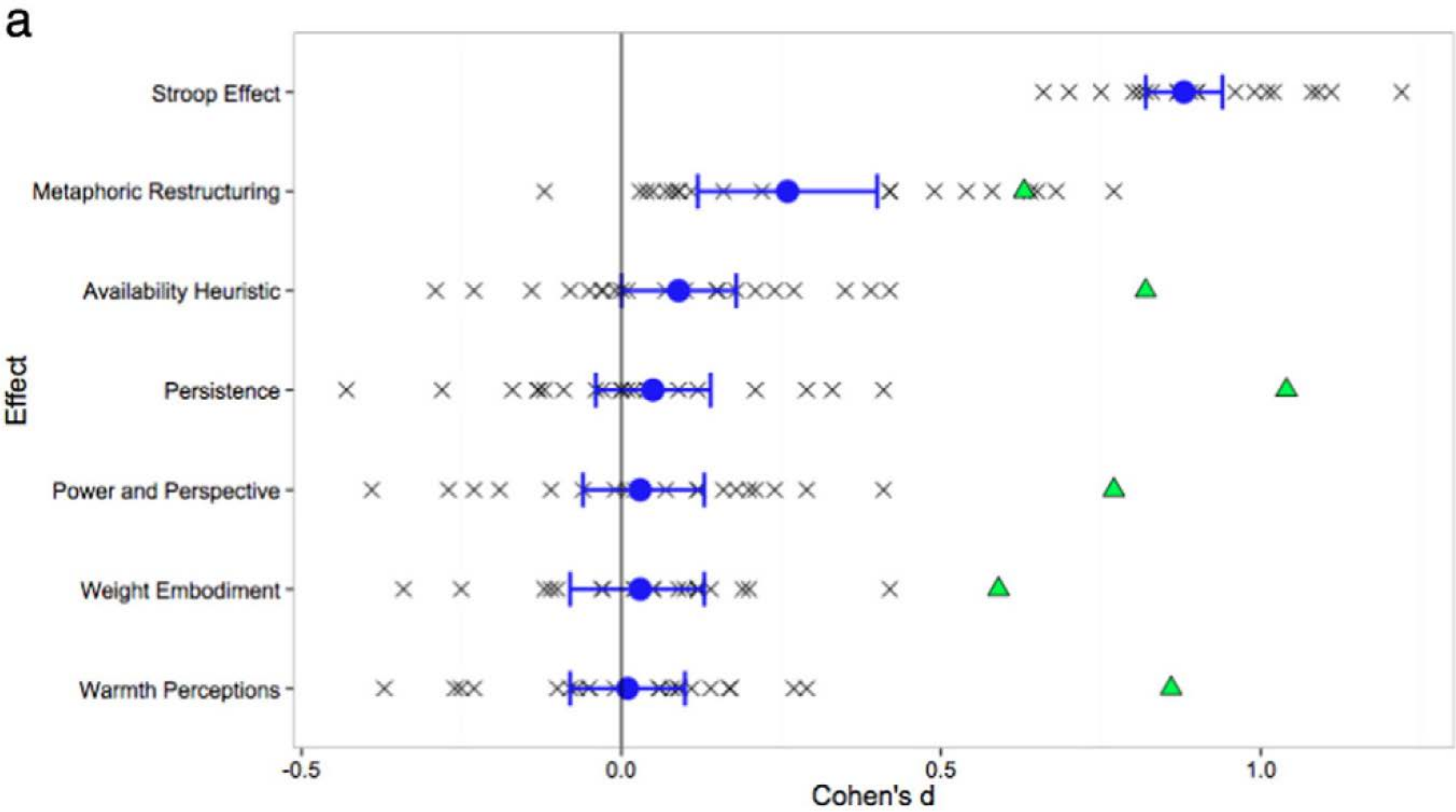
8%
"benefit"

After making pre-registration mandatory for pharmacological trials (in 2000), the frequency of statistically significant effects has dramatically decreased...

Kaplan & Irvin (2015)
https://doi.org/10.1371/journal.pone.0132382

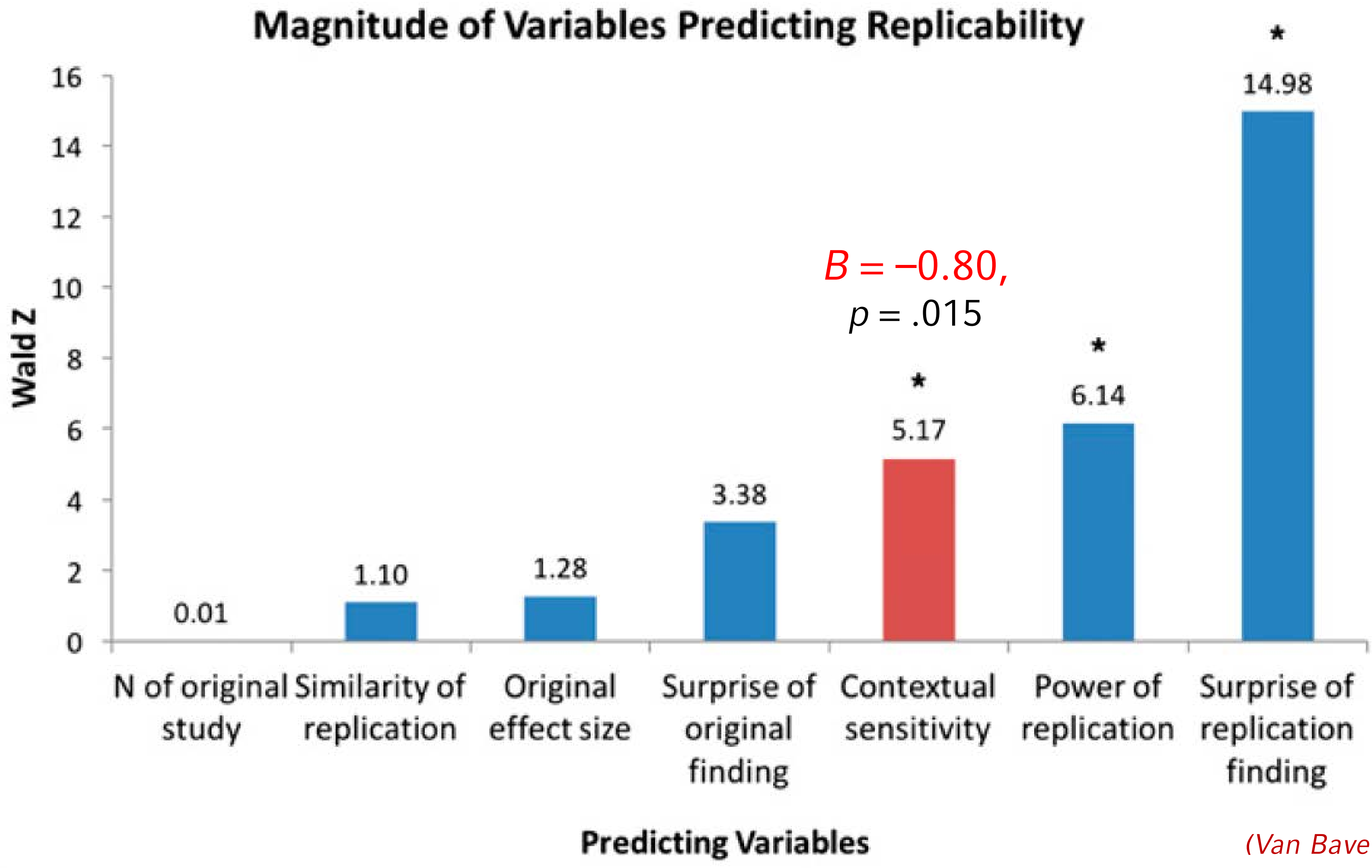
2. Context-Dependency Perspective

- What's the evidence?
  - Substantial heterogeneity of effect sizes across study sites in the „ManyLabs" projects
  - Evidence for contextual effects (in multilevel terms) in some well-known social psych findings (e.g., intergroup contact; Pettigrew, 2018)
  - Context-dependency predicts replicability (Van Bavel et al., 2016; but see Inbar, 2016)



(Ebersole et al., 2016)

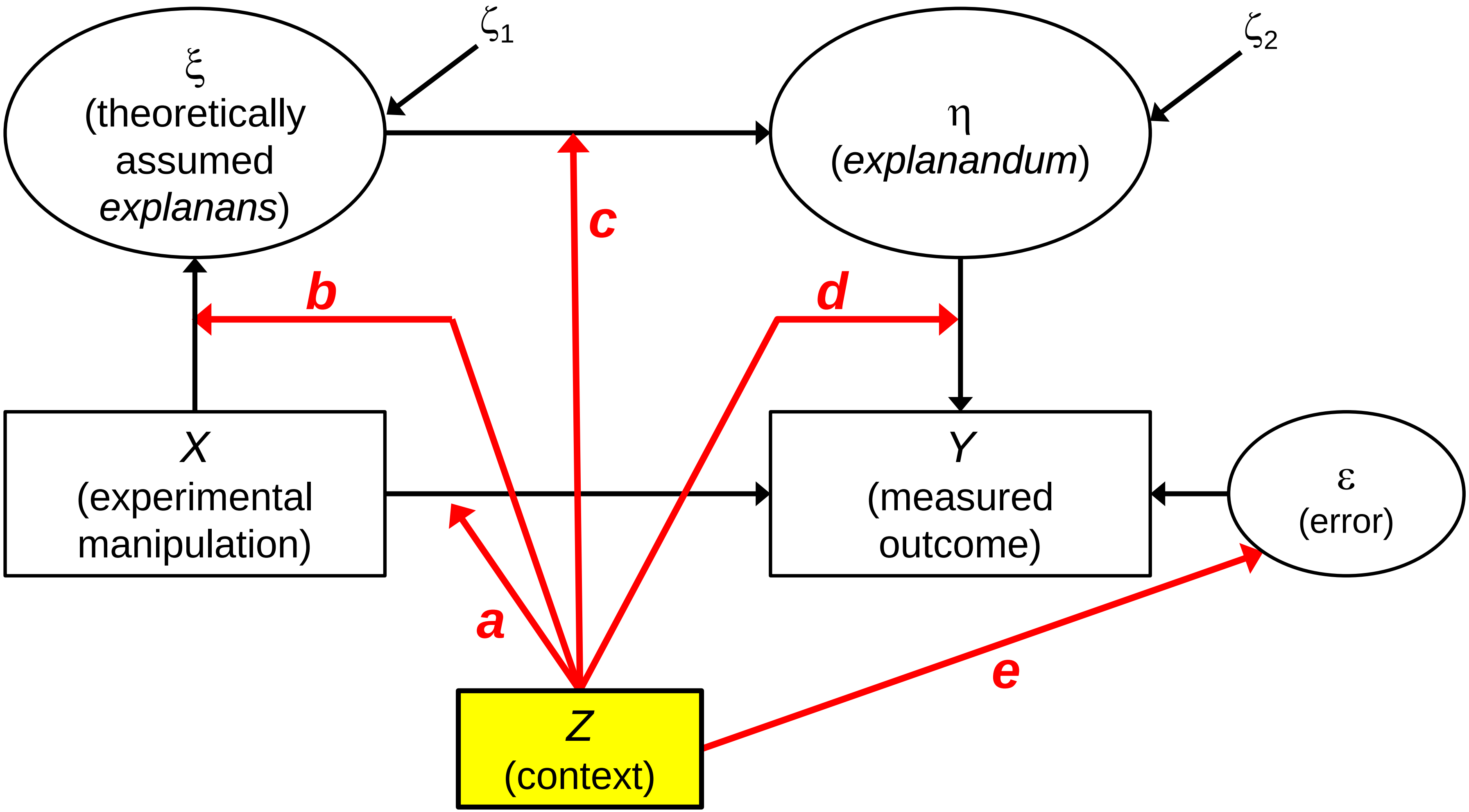
2. Context-Dependency Perspective



$B = -0.80$,
$p = .015$

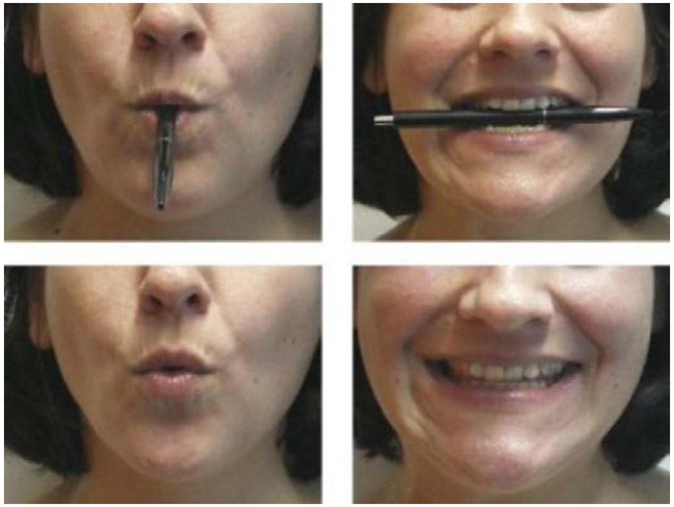
(Van Bavel et al., 2016)

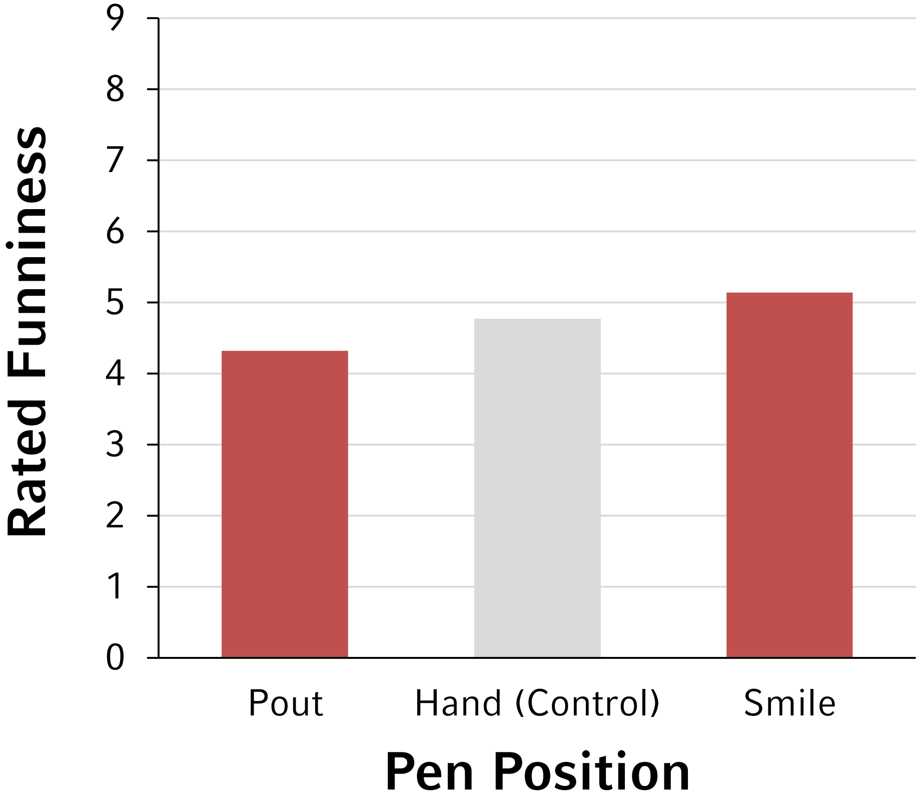
2. Context-Dependency Perspective



2. Context-Dependency Perspective

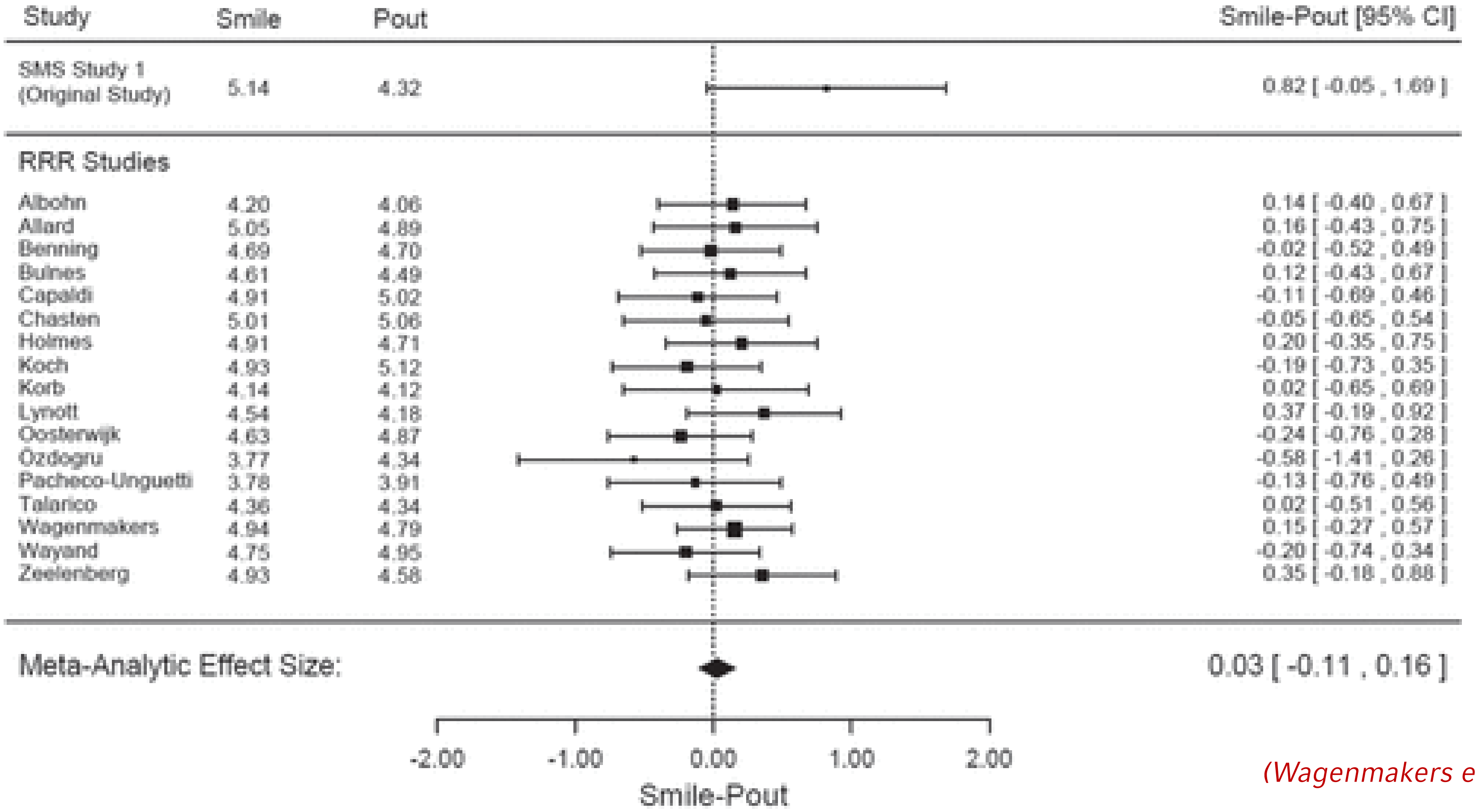
Example: Facial Feedback Effect (Strack, Martin, & Stepper, 1988)





2. Context-Dependency Perspective

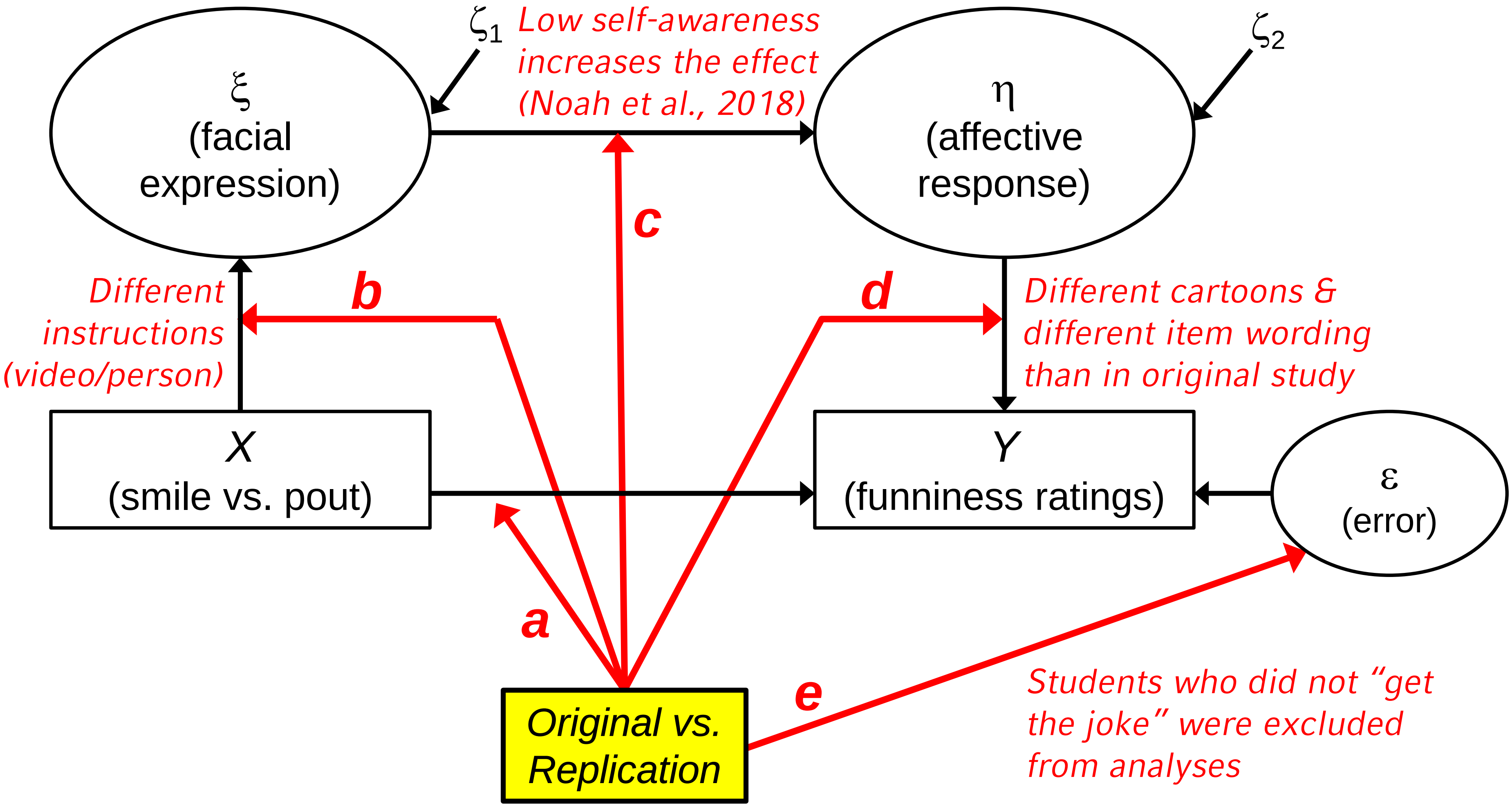
Example: Facial Feedback Effect (Strack, Martin, & Stepper, 1988)



Study	Smile	Pout	Smile-Pout [95% CI]
SMS Study 1 (Original Study)	5.14	4.32	0.82 [-0.05 , 1.69]
RRR Studies			
Albohn	4.20	4.06	0.14 [-0.40 , 0.67]
Allard	5.05	4.89	0.16 [-0.43 , 0.75]
Benning	4.69	4.70	-0.02 [-0.52 , 0.49]
Bulnes	4.61	4.49	0.12 [-0.43 , 0.67]
Capaldi	4.91	5.02	-0.11 [-0.69 , 0.46]
Chasten	5.01	5.06	-0.05 [-0.65 , 0.54]
Holmes	4.91	4.71	0.20 [-0.35 , 0.75]
Koch	4.93	5.12	-0.19 [-0.73 , 0.35]
Korb	4.14	4.12	0.02 [-0.65 , 0.69]
Lynott	4.54	4.18	0.37 [-0.19 , 0.92]
Oosterwijk	4.63	4.87	-0.24 [-0.76 , 0.28]
Ozdogru	3.77	4.34	-0.58 [-1.41 , 0.26]
Pacheco-Unguetti	3.78	3.91	-0.13 [-0.76 , 0.49]
Talarico	4.36	4.34	0.02 [-0.51 , 0.56]
Wagenmakers	4.94	4.79	0.15 [-0.27 , 0.57]
Wayand	4.75	4.95	-0.20 [-0.74 , 0.34]
Zeelenberg	4.93	4.58	0.35 [-0.18 , 0.88]
Meta-Analytic Effect Size:			0.03 [-0.11 , 0.16]

(Wagenmakers et al., 2016)

2. Context-Dependency Perspective



Low self-awareness increases the effect (Noah et al., 2018)

Different instructions (video/person)

Different cartoons & different item wording than in original study

Students who did not "get the joke" were excluded from analyses

3. False-Negative Perspective

- Selection bias in replication projects: Effects selected for RP:P and ManyLabs had a low replication chance *a priori* (Gilbert et al., 2016)
- It is unclear what the "replicandum" should be: significance? Effect size similarity? Confidence interval? Conditional causal effect? ... (Fiedler, 2018; Wong & Steiner, 2018)
- Just "counting asterisks" (no. of significant effects) is an inappropriate estimate of replicability (Patil et al., 2016)
- Replication rates need to be compared against a proper base rate of "true" effects (Bird, 2018; Miller, 2009)
- "Failed" replications may be a regression artifact (Fiedler & Prager, 2018)
- Even more highly-powered replication projects may still not have enough power to find the assumed effect (Erdfelder & Ulrich, 2018); especially if there is publication bias...

3. False-Negative Perspective

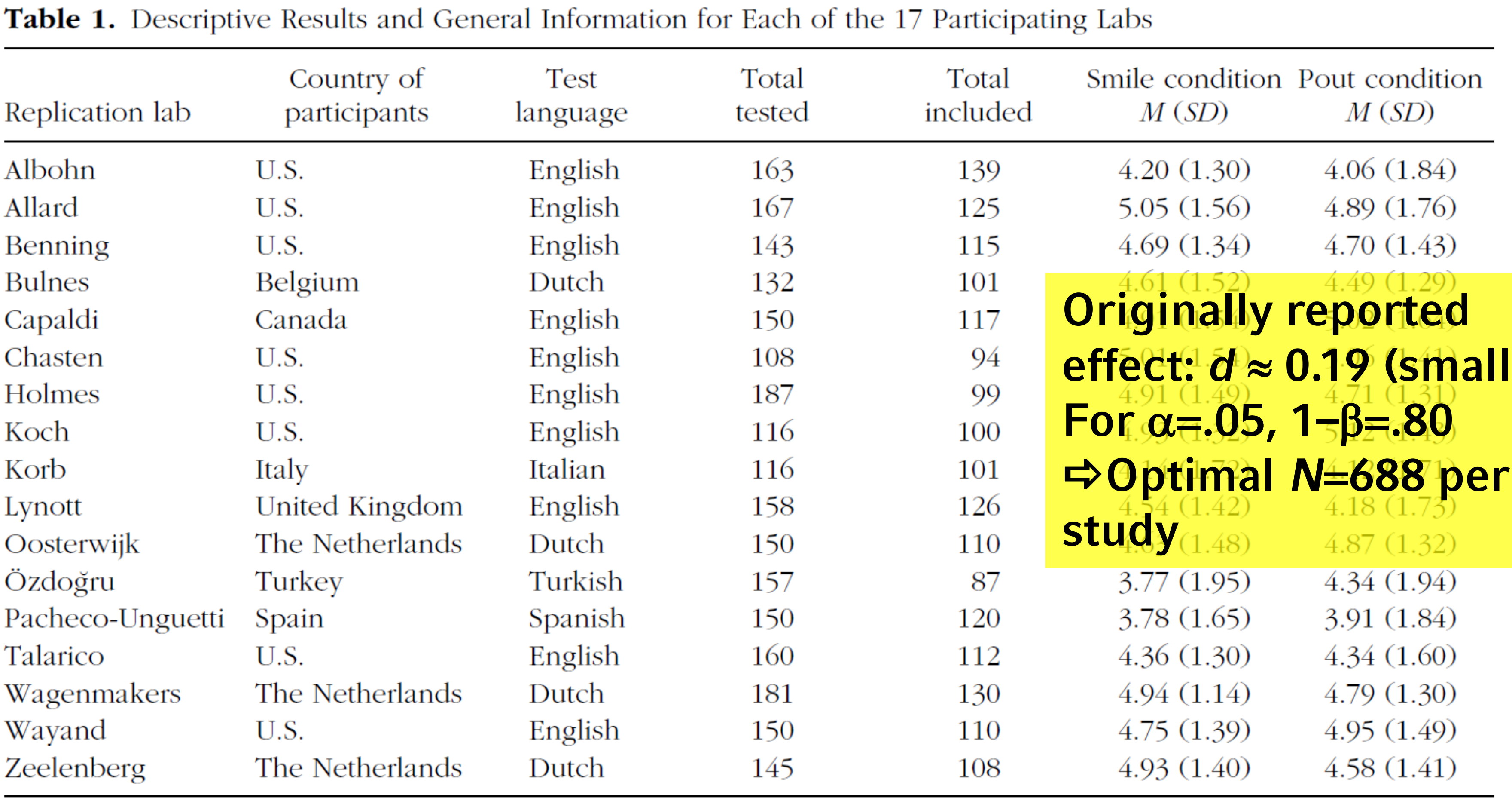


Table 1. Descriptive Results and General Information for Each of the 17 Participating Labs

Replication lab	Country of participants	Test language	Total tested	Total included	Smile condition *M* (*SD*)	Pout condition *M* (*SD*)
Albohn	U.S.	English	163	139	4.20 (1.30)	4.06 (1.84)
Allard	U.S.	English	167	125	5.05 (1.56)	4.89 (1.76)
Benning	U.S.	English	143	115	4.69 (1.34)	4.70 (1.43)
Bulnes	Belgium	Dutch	132	101	4.61 (1.52)	4.49 (1.29)
Capaldi	Canada	English	150	117	[illegible]	[illegible]
Chasten	U.S.	English	108	94	[illegible]	[illegible]
Holmes	U.S.	English	187	99	4.91 (1.49)	4.71 (1.31)
Koch	U.S.	English	116	100	[illegible]	[illegible]
Korb	Italy	Italian	116	101	[illegible]	[illegible]
Lynott	United Kingdom	English	158	126	4.54 (1.42)	4.18 (1.73)
Oosterwijk	The Netherlands	Dutch	150	110	[illegible]	4.87 (1.32)
Özdoğru	Turkey	Turkish	157	87	3.77 (1.95)	4.34 (1.94)
Pacheco-Unguetti	Spain	Spanish	150	120	3.78 (1.65)	3.91 (1.84)
Talarico	U.S.	English	160	112	4.36 (1.30)	4.34 (1.60)
Wagenmakers	The Netherlands	Dutch	181	130	4.94 (1.14)	4.79 (1.30)
Wayand	U.S.	English	150	110	4.75 (1.39)	4.95 (1.49)
Zeelenberg	The Netherlands	Dutch	145	108	4.93 (1.40)	4.58 (1.41)

Originally reported effect: *d* ≈ 0.19 (small)
For α=.05, 1–β=.80
⇨Optimal *N*=688 per study

(Wagenmakers et al., 2016)

Summary

- We are just beginning to understand when and why so many empirical effects are (non)replicable.
- It is unclear to what extent non-replicability is due to (1) high false-positive rates in the literature, (2) context dependency, and/or (3) false negatives in replication studies.
- The phenomenon of (non)replicability should be treated with as much scientific rigor as possible (“replication science”).
- Open Science may be a cure, but only if the “false positive” diagnosis was correct. But: Open Science is laudable *per se*!
- Context dependency is a concept that requires more theoretical and empirical elaboration. It should never be used to immunize an effect a posteriori (Meehl, 1990).

SPP 2317 "META-REP"

meta rep

DFG-Schwerpunktprogramm ("Priority Program") "META-REP: A Meta-scientific Program to Analyze and Optimize Replicability in the Behavioral, Social, and Cognitive Sciences"

https://leibniz-psychology.org/metarep

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

UKE HAMBURG

MAX-PLANCK-GESELLSCHAFT

DFG Deutsche Forschungsgemeinschaft

SPP 2317 “META-REP”

- Collaborative project platform (20-30 individual projects)
- Duration: 2 × 3 years (2021-2024 and 2024-2027)
- Overarching aim: To empirically investigate …

1. **WHAT** “replicability” means (and when a replication can be regarded as successful vs. failed) in different behavioral, social, and cognitive sciences,
2. **WHY** replication rates are (sometimes) lower than expected; i.e., which factors predict/explain the replicability of effects in different behavioral, social, and cognitive sciences (e.g., QRPs, contextual influences, etc.)
3. **HOW** an acceptable level of replicability can be achieved and maintained in different behavioral, social, and cognitive sciences.

Data Management Recommendations

Deutsche Gesellschaft für Psychologie

- originally issues 2016; now revised (2020):

  https://psyarxiv.com/hcxtm/ (German)

  https://psyarxiv.com/24ncs/ (English)

- Topics covered in the text:
  - Definitions ("raw data," "primary data," "secondary data");
  - Legal aspects (e.g., data protection, copyright, licenses);
  - Requirements for eligible repositories;
  - Two types of data sharing (data sharing to reproduce published findings; data sharing in the context of funded research projects);
  - Access restrictions ("open data," "conditional access," "restricted access," "secure data") and scientific use files;
  - Structural challenges and incentives; conflicts of interests; disputes

Thank You!