Statistical Foundations: Replication Solutions

20 October 2021

Modern Research Methods

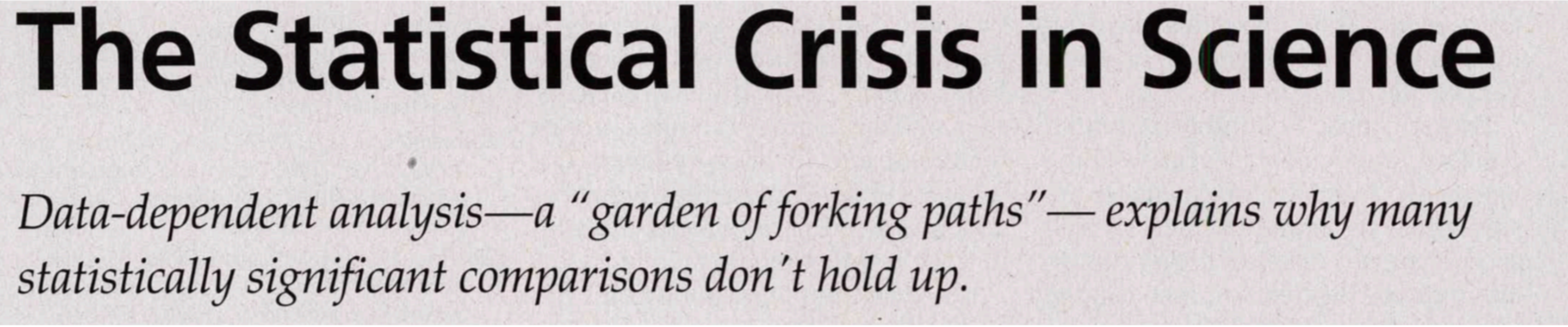
Logistics

- Due Thursday at noon (via Canvas)
- My office hours are virtual today – email/talk to me if you'd like to meet
- No lab Friday
- Quiz Monday, as usual

Last time: Potential reasons for replication failure

1. Fraud
2. Actual change in population effect
3. Error in reporting/analysis
4. Hidden moderator
5. Inadequate materials/description
6. Data-depending analysis ("p-hacking"/"HARKing")
7. File drawer problem ("publication bias")
8. Low study precision

Data-dependent analysis



The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Choosing your analysis based on seeing your data/the outcome of a test ("analytic flexibility")

"*p*-hacking"/"Questionable research practices" (QRP)

"…it is unacceptably easy to publish *statistically significant* evidence consistent with *any* hypothesis" -- Simmons, Nelson, & Simonsohn, 2011

Examples of p-hacking

Collect more data?

Should some observations be excluded? Which ones?

Which conditions should be combined with which ones?

Which measures should we analyze? Should we transform the measure?

Which control variables should we consider?

Should we include pilot data?

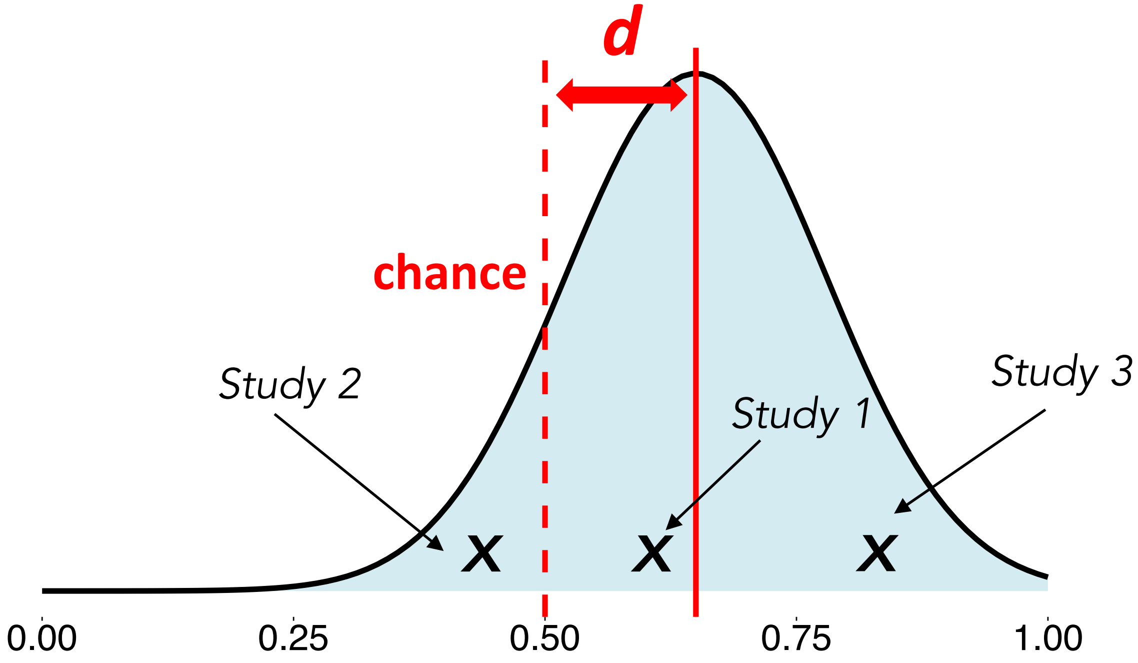
Try your hand at *p*-hacking!

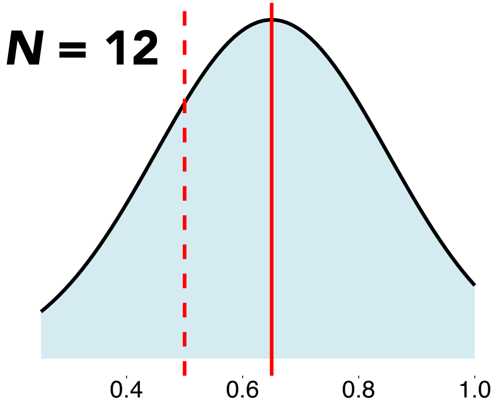
http://shinyapps.org/apps/p-hacker/

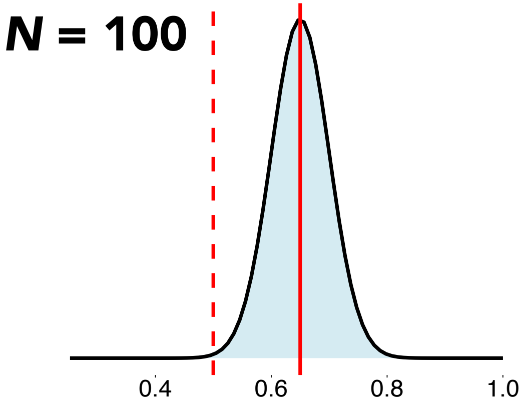
Goal: get a publishable finding (i.e. $p < .05$)

- How many significant p-values can you get?
- What things can you do to increase the likelihood of getting a significant *p*-value?

Reason 8: Low study precision

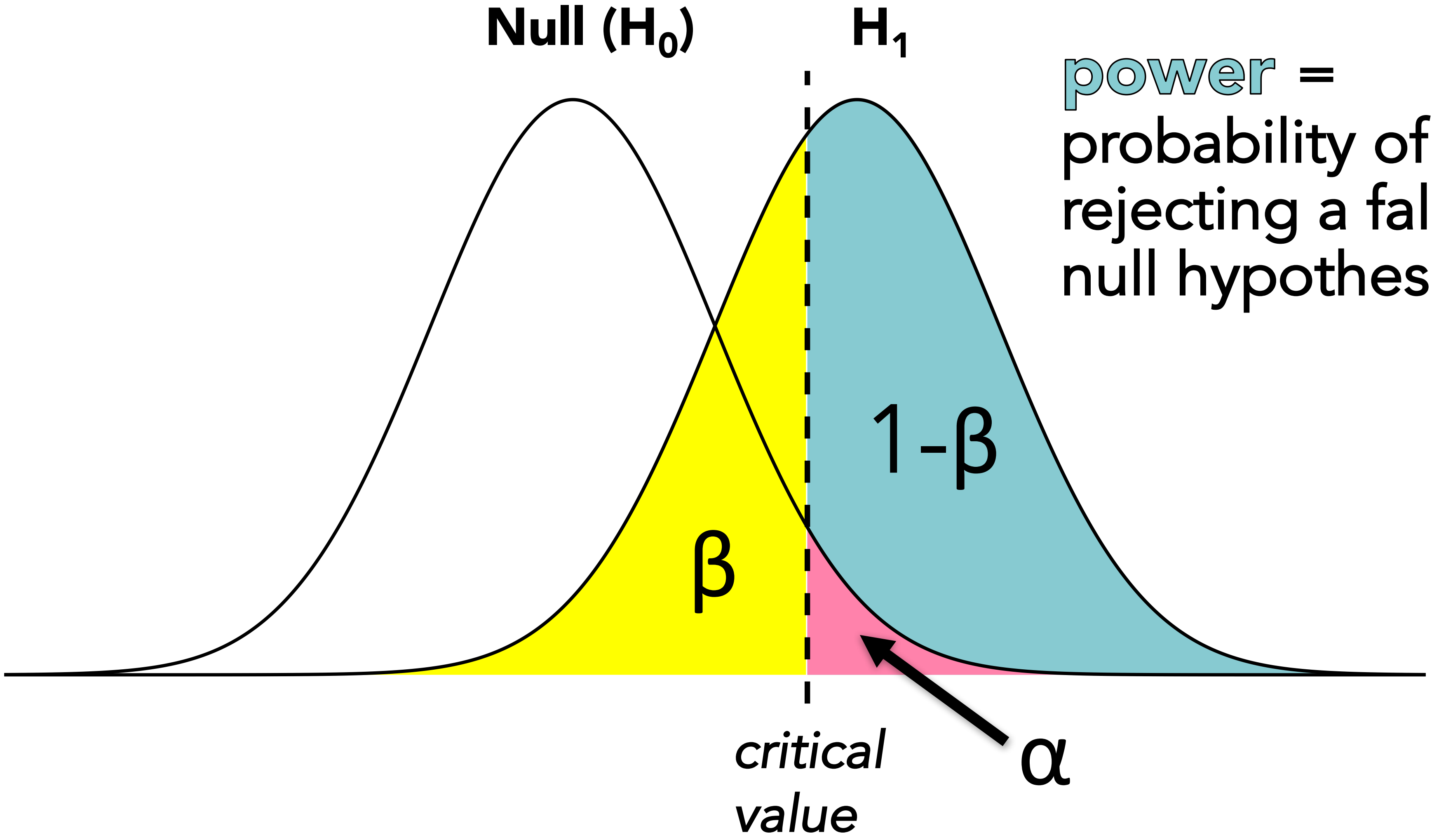






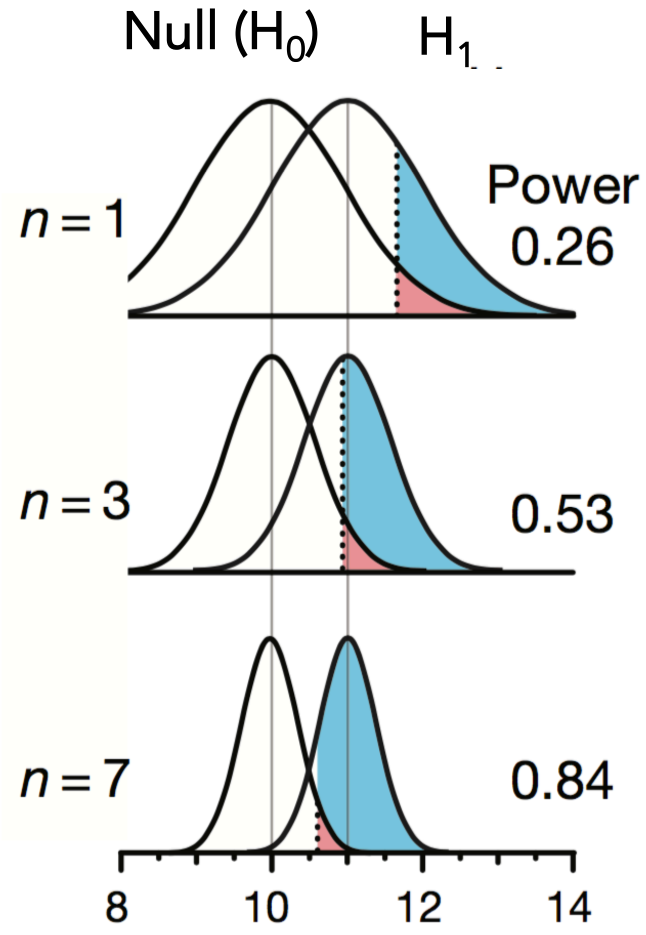
Some expected variability in effect size due to sample size; less variability with larger sample sizes.

Precision related to probability of correctly rejecting null hypothesis



power =
probability of rejecting a false null hypothesis

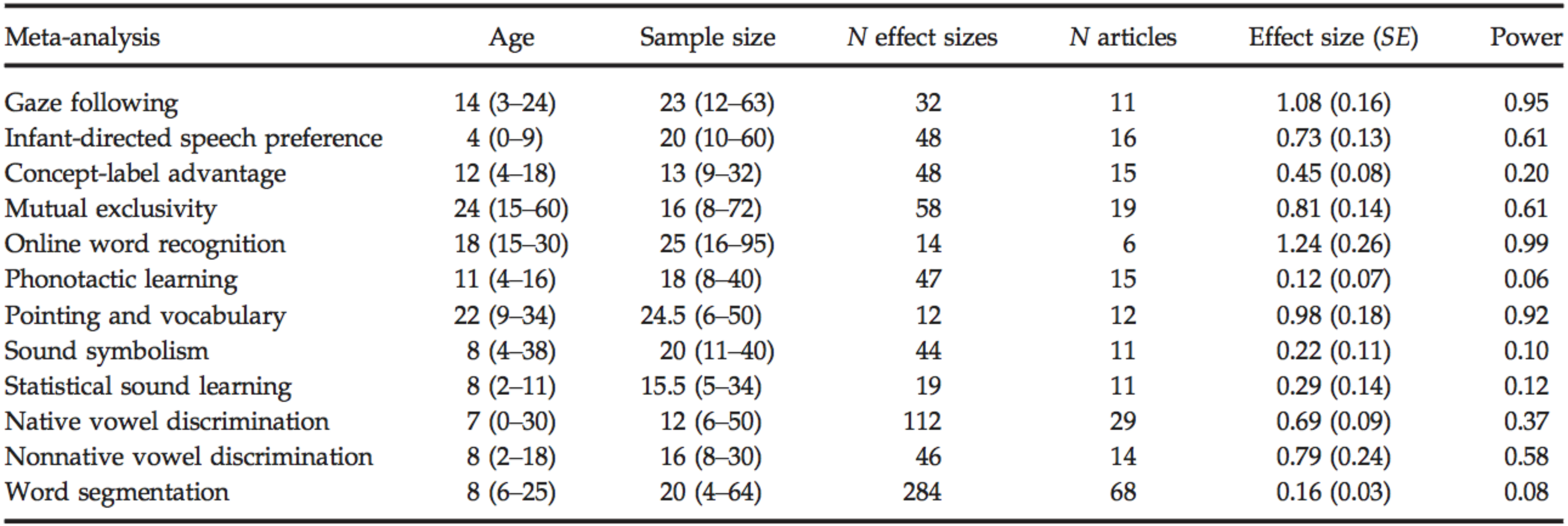
Power is related to sample size and effect size



Low precision -> low power

(Krzywinski & Altman, 2003)

Power is often low, leading overestimations of effect size



Meta-analysis	Age	Sample size	*N* effect sizes	*N* articles	Effect size (*SE*)	Power
Gaze following	14 (3–24)	23 (12–63)	32	11	1.08 (0.16)	0.95
Infant-directed speech preference	4 (0–9)	20 (10–60)	48	16	0.73 (0.13)	0.61
Concept-label advantage	12 (4–18)	13 (9–32)	48	15	0.45 (0.08)	0.20
Mutual exclusivity	24 (15–60)	16 (8–72)	58	19	0.81 (0.14)	0.61
Online word recognition	18 (15–30)	25 (16–95)	14	6	1.24 (0.26)	0.99
Phonotactic learning	11 (4–16)	18 (8–40)	47	15	0.12 (0.07)	0.06
Pointing and vocabulary	22 (9–34)	24.5 (6–50)	12	12	0.98 (0.18)	0.92
Sound symbolism	8 (4–38)	20 (11–40)	44	11	0.22 (0.11)	0.10
Statistical sound learning	8 (2–11)	15.5 (5–34)	19	11	0.29 (0.14)	0.12
Native vowel discrimination	7 (0–30)	12 (6–50)	112	29	0.69 (0.09)	0.37
Nonnative vowel discrimination	8 (2–18)	16 (8–30)	46	14	0.79 (0.24)	0.58
Word segmentation	8 (6–25)	20 (4–64)	284	68	0.16 (0.03)	0.08

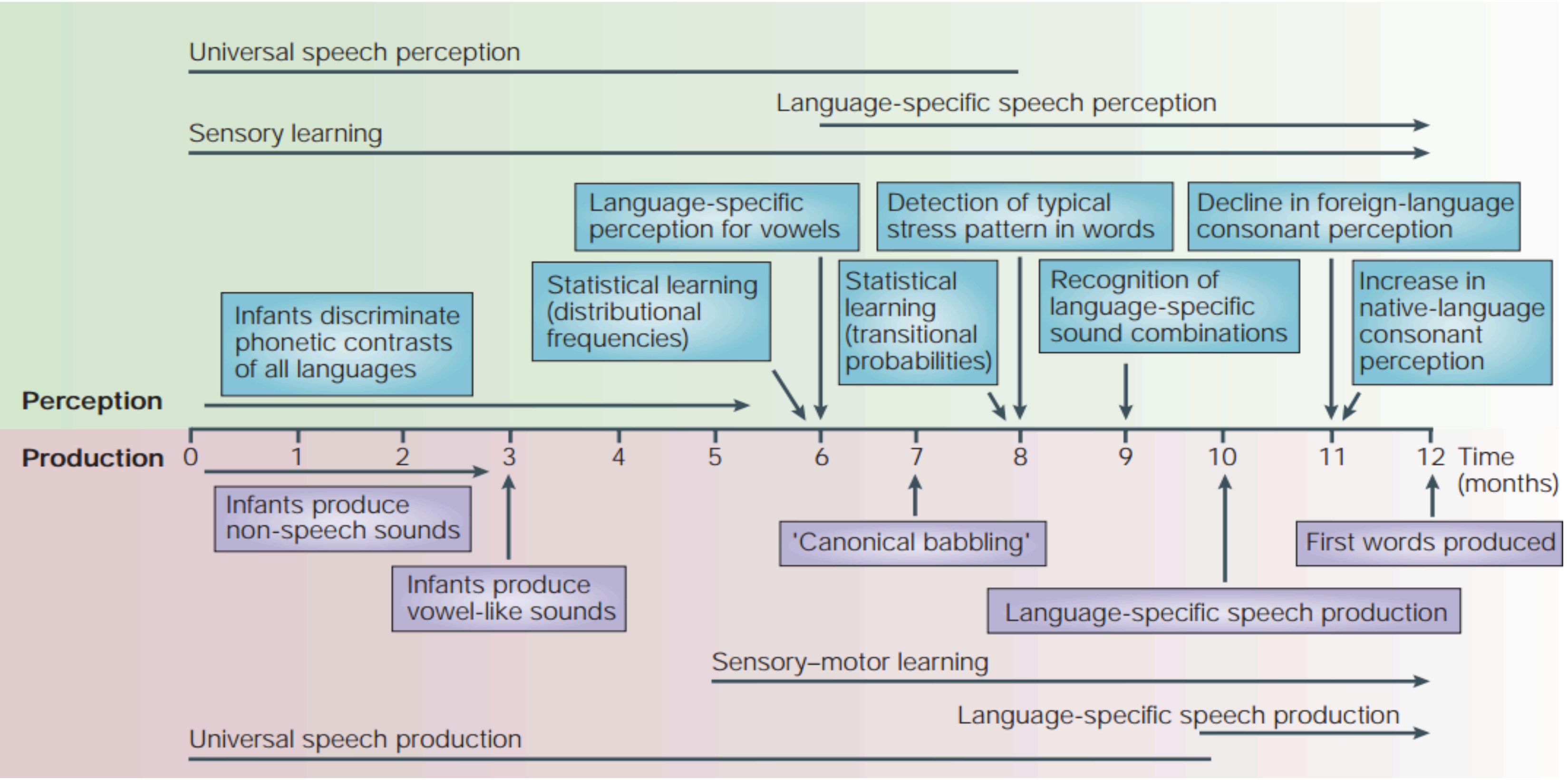
(Bergmann et al., 2018)

Potential reasons for replication failure

1. Fraud
2. Actual change in population effect
3. Error in reporting/analysis
4. Hidden moderator
5. Inadequate materials/description
6. Data-depending analysis (“p-hacking”/”HARKing”)
7. File drawer problem (“publication bias”)
8. Low study precision

Why does replicability matter?

Our goal as scientists is to build predictive theories


Universal speech perception
Language-specific speech perception
Sensory learning
Language-specific perception for vowels
Detection of typical stress pattern in words
Decline in foreign-language consonant perception
Infants discriminate phonetic contrasts of all languages
Statistical learning (distributional frequencies)
Statistical learning (transitional probabilities)
Recognition of language-specific sound combinations
Increase in native-language consonant perception
Perception
Production
0
1
2
3
4
5
6
7
8
9
10
11
12
Time (months)
Infants produce non-speech sounds
Infants produce vowel-like sounds
'Canonical babbling'
First words produced
Language-specific speech production
Sensory–motor learning
Language-specific speech production
Universal speech production

Kuhl (2004)

But, if the experiments our theories are built on aren't real, then our theories are bad too.

How can we increase replicability?

Solution	
	1. Fraud
	2. Actual change in population effect
Reproducibility practices	3. Error in reporting/analysis
	4. Hidden moderator
Reproducibility practices	5. Inadequate materials/description
Strategies for reducing rates for failed replications due to false positives	6. Data-depending analysis ("p-hacking"/"HARKing")
	7. File drawer problem ("publication bias")
	8. Low study precision

Solutions

- Designed to reduce "questionable research practices", like *p*-hacking.
- And, therefore **increase replicability**

1. Pre-registration – register your hypothesis and analysis plan publically before you collecting your data
2. Registered reports – write the paper and have it reviewed before you collect your data.

Daniel Simons on replication solutions:

https://www.youtube.com/watch?v=LYfZr5poCkQ

Pre-registration

- OSF great tool for registration (essentially, time-stamping)
- All levels of specificity
    - AsPredicted template (“prereg-lite”) - https://aspredicted.org/
    - All the way through full analytic script capture

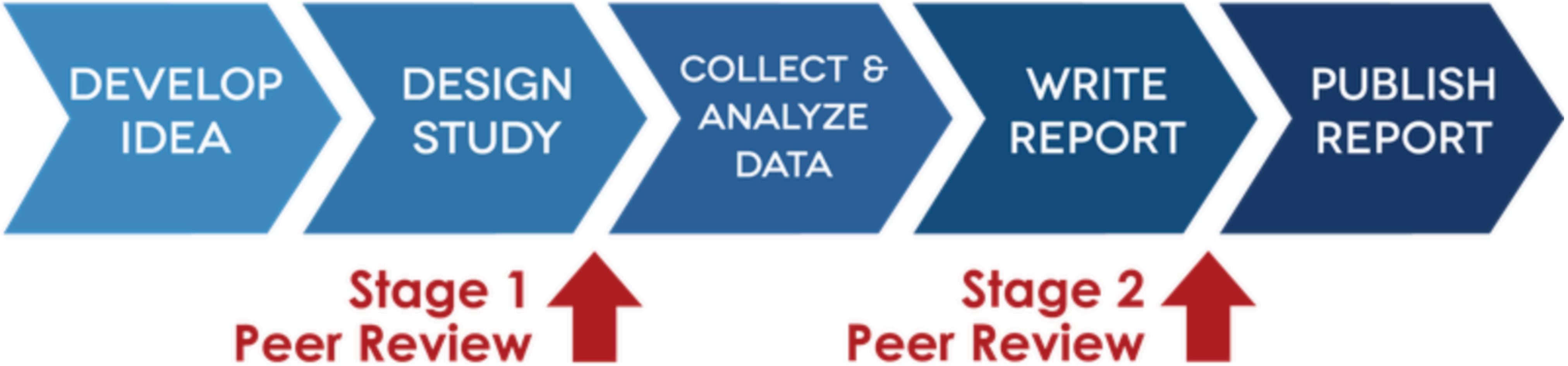
- Example preregistration from Lewis & Frank, 2018: https://osf.io/bxpke

Preregistration – in sum:

- It costs nothing and makes you feel good.
- If you’re running a study, just try it.
- It'll make you feel like a scientist.

Slide adopted from
Mike Frank

Registered Report



Are the hypotheses well founded?

Are the methods and proposed analyses feasible and sufficiently detailed?

Have the authors included sufficient positive controls to confirm that the study will provide a fair test?

"provisionally accepted"

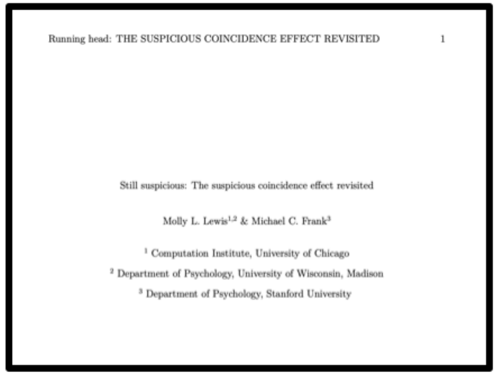
Did the authors follow the approved protocol?

Did positive controls succeed?

Are the conclusions justified by the data?

Slide adopted from
Chris Chambers

c.f. Typical peer-review process



Running head: THE SUSPICIOUS COINCIDENCE EFFECT REVISITED 1

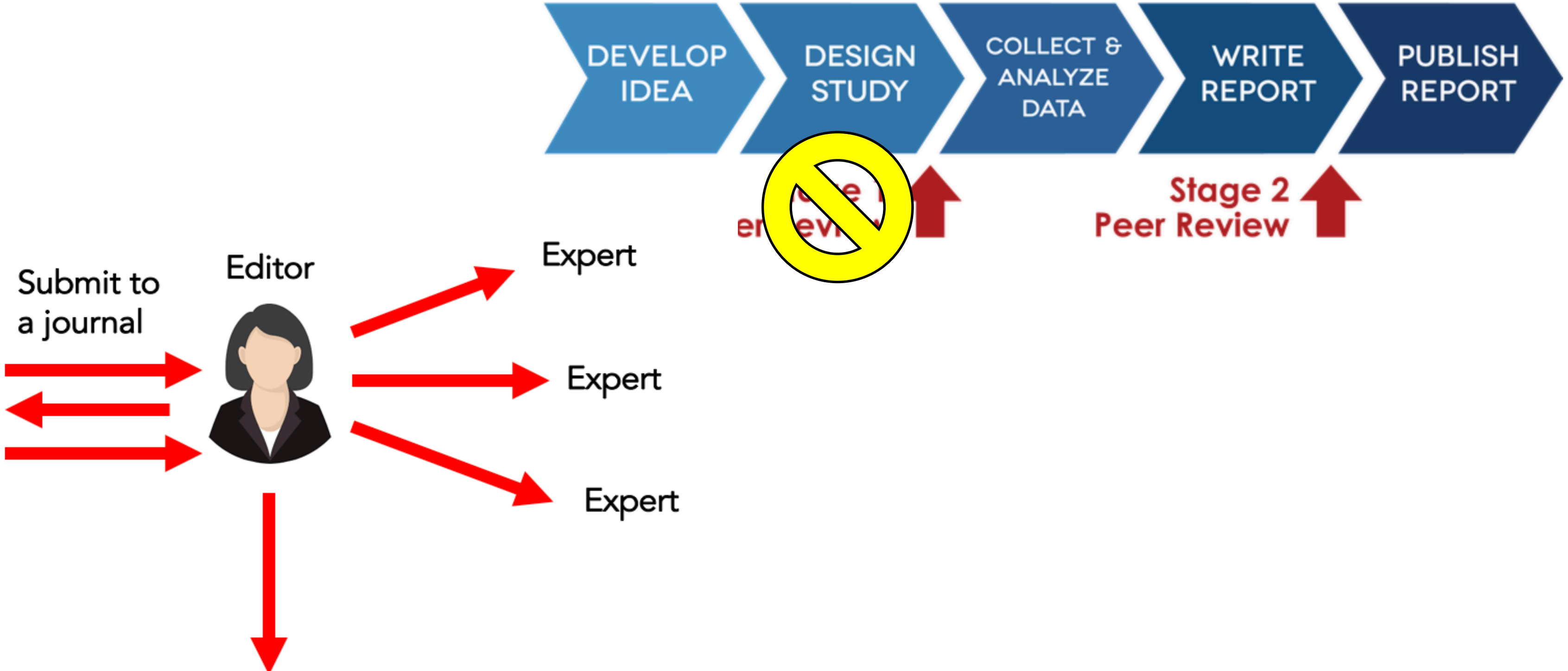
Still suspicious: The suspicious coincidence effect revisited

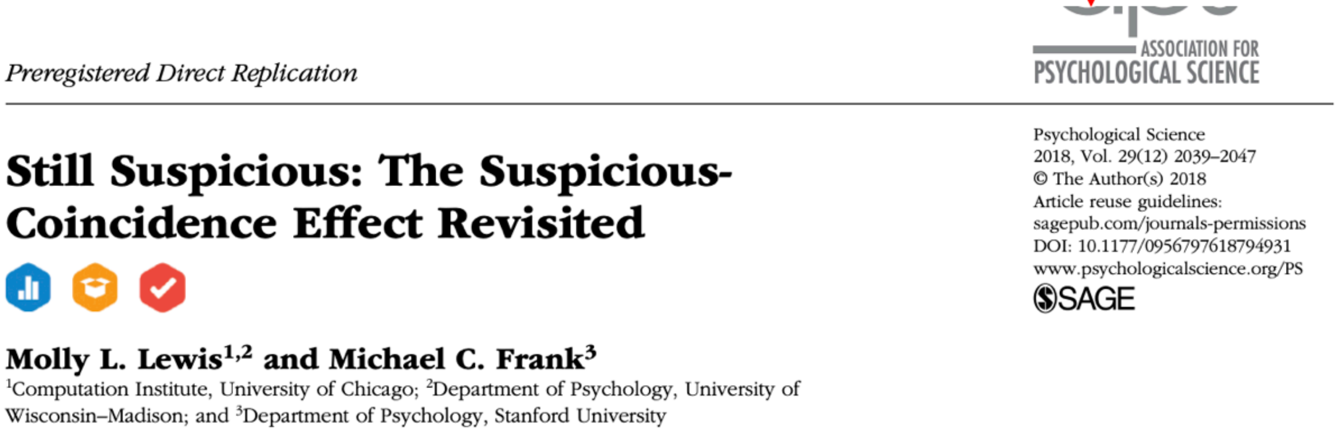
Molly L. Lewis[1,2] & Michael C. Frank[3]

[1] Computation Institute, University of Chicago

[2] Department of Psychology, University of Wisconsin, Madison

[3] Department of Psychology, Stanford University





ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Preregistered Direct Replication

Still Suspicious: The Suspicious-Coincidence Effect Revisited

Psychological Science
2018, Vol. 29(12) 2039–2047
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797618794931
www.psychologicalscience.org/PS
SAGE

Molly L. Lewis[1,2] and Michael C. Frank[3]
[1]Computation Institute, University of Chicago; [2]Department of Psychology, University of Wisconsin–Madison; and [3]Department of Psychology, Stanford University

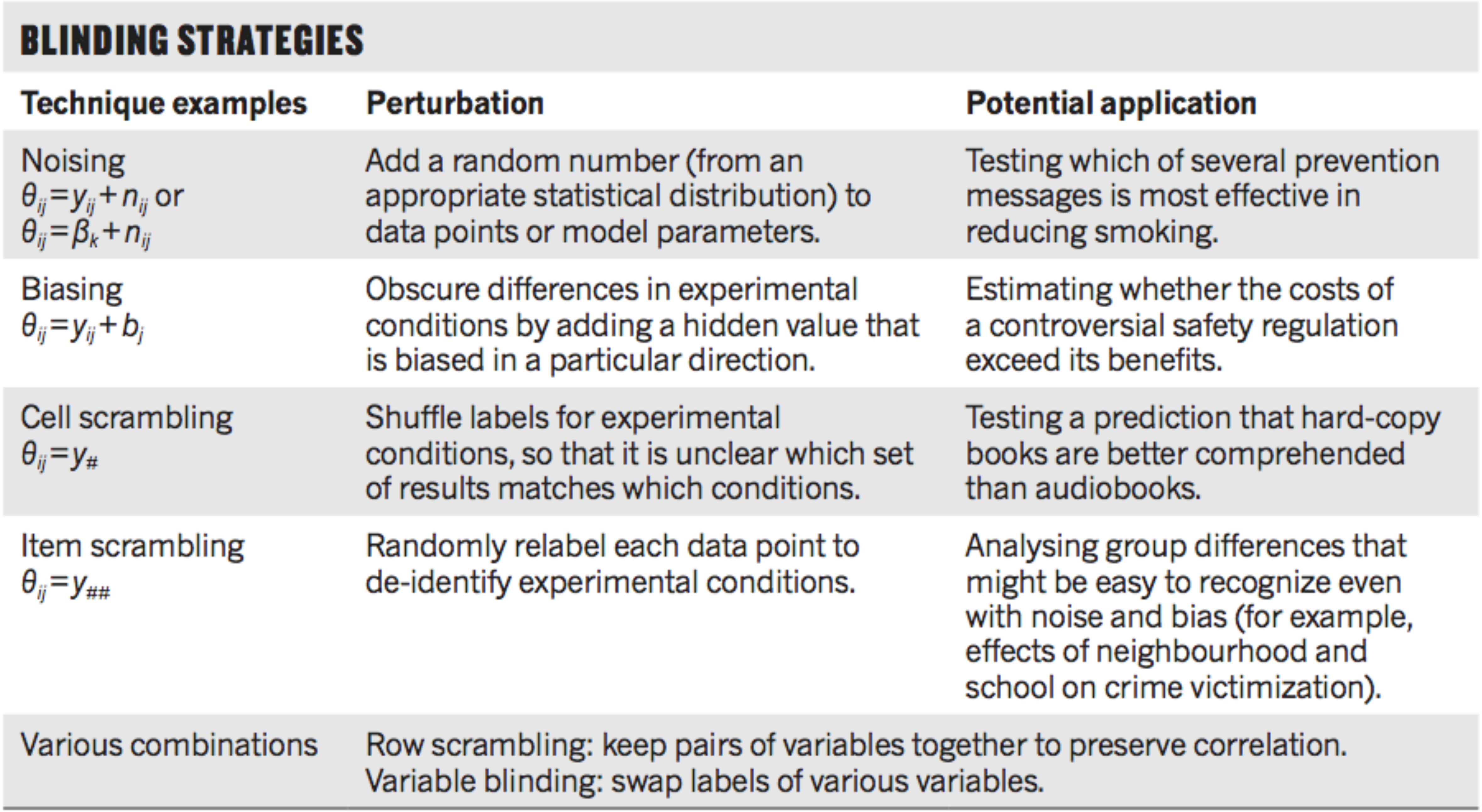
Solutions

- Designed to reduce “questionable research practices”, like *p*-hacking.
- And, therefore **increase replicability**

1. Pre-registration – register your hypothesis and analysis plan publically before you collecting your data
2. Registered reports – write the paper and have it reviewed before you collect your data.

Alternative: Blind analysis (MacCoun & Perlmutter, 2015)

- Practice in particle physics
- Define analysis plan after conducted study, but not on "real data"
- Have a third party perturb the data in some way (data values, labels, or both), then conduct analysis "in the dark"
- Once agreed upon analysis plan, unblind the data and reveal results



BLINDING STRATEGIES

Technique examples	Perturbation	Potential application
Noising $\theta_{ij}=y_{ij}+n_{ij}$ or $\theta_{ij}=\beta_k+n_{ij}$	Add a random number (from an appropriate statistical distribution) to data points or model parameters.	Testing which of several prevention messages is most effective in reducing smoking.
Biasing $\theta_{ij}=y_{ij}+b_j$	Obscure differences in experimental conditions by adding a hidden value that is biased in a particular direction.	Estimating whether the costs of a controversial safety regulation exceed its benefits.
Cell scrambling $\theta_{ij}=y_{\#}$	Shuffle labels for experimental conditions, so that it is unclear which set of results matches which conditions.	Testing a prediction that hard-copy books are better comprehended than audiobooks.
Item scrambling $\theta_{ij}=y_{\#\#}$	Randomly relabel each data point to de-identify experimental conditions.	Analysing group differences that might be easy to recognize even with noise and bias (for example, effects of neighbourhood and school on crime victimization).
Various combinations	Row scrambling: keep pairs of variables together to preserve correlation. Variable blinding: swap labels of various variables.	

MacCoun & Perlmutter, 2015

Alternative: Multiverse analysis (Steegen, et al, 2016)

- Do all the paths!
- “performing the analysis of interest across the whole set of data sets that arise from different reasonable choices for data processing”
- Then examine how sensitive p-value is to different choices

Multiverse analysis example

- Durante, Rae, and Griskevicius, 2013
- Being fertile led single women to become more liberal, less religious, and more likely to vote for Barack Obama.
- In contrast, being fertile led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney

Multiverse analysis example

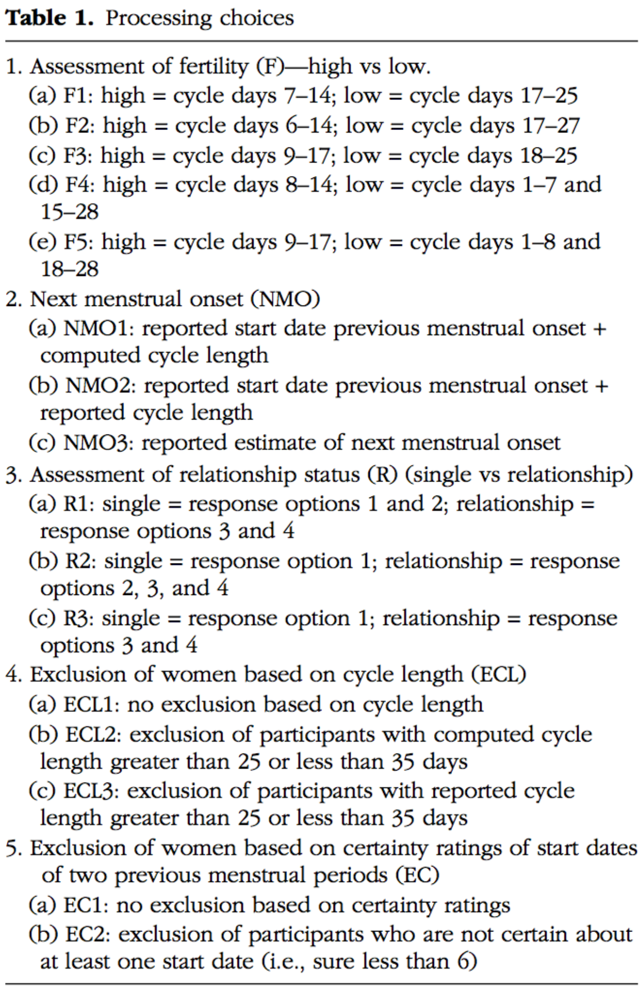
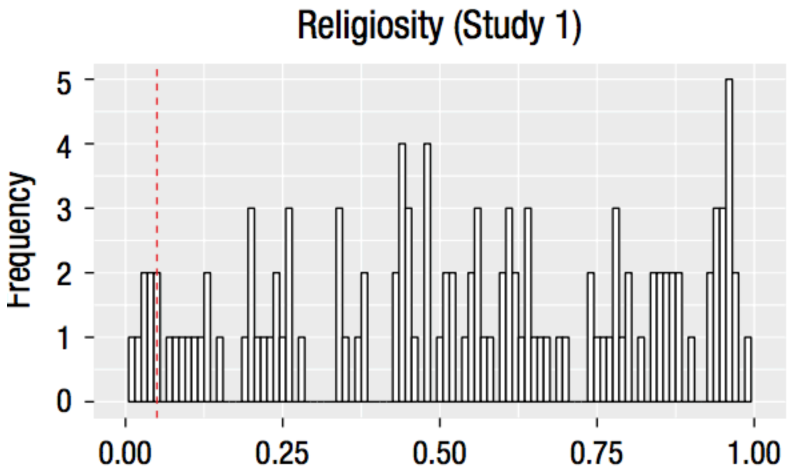
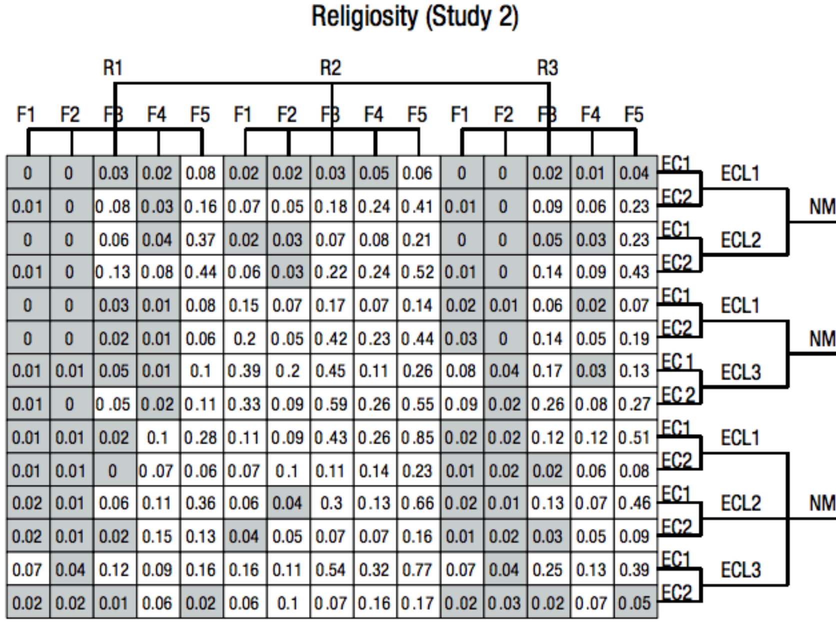


Table 1. Processing choices

1. Assessment of fertility (F)—high vs low.
   - (a) F1: high = cycle days 7–14; low = cycle days 17–25
   - (b) F2: high = cycle days 6–14; low = cycle days 17–27
   - (c) F3: high = cycle days 9–17; low = cycle days 18–25
   - (d) F4: high = cycle days 8–14; low = cycle days 1–7 and 15–28
   - (e) F5: high = cycle days 9–17; low = cycle days 1–8 and 18–28
2. Next menstrual onset (NMO)
   - (a) NMO1: reported start date previous menstrual onset + computed cycle length
   - (b) NMO2: reported start date previous menstrual onset + reported cycle length
   - (c) NMO3: reported estimate of next menstrual onset
3. Assessment of relationship status (R) (single vs relationship)
   - (a) R1: single = response options 1 and 2; relationship = response options 3 and 4
   - (b) R2: single = response option 1; relationship = response options 2, 3, and 4
   - (c) R3: single = response option 1; relationship = response options 3 and 4
4. Exclusion of women based on cycle length (ECL)
   - (a) ECL1: no exclusion based on cycle length
   - (b) ECL2: exclusion of participants with computed cycle length greater than 25 or less than 35 days
   - (c) ECL3: exclusion of participants with reported cycle length greater than 25 or less than 35 days
5. Exclusion of women based on certainty ratings of start dates of two previous menstrual periods (EC)
   - (a) EC1: no exclusion based on certainty ratings
   - (b) EC2: exclusion of participants who are not certain about at least one start date (i.e., sure less than 6)



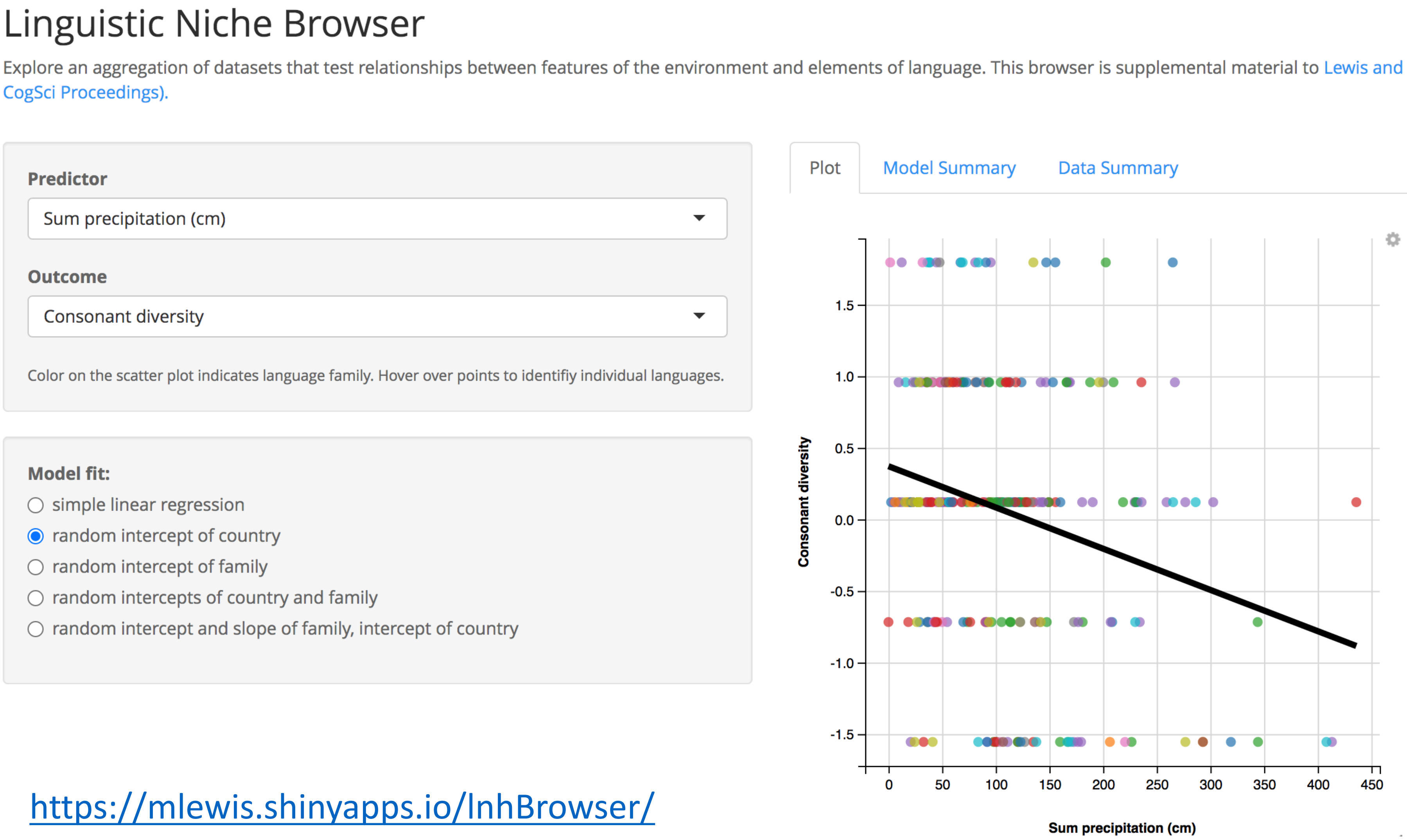


Religiosity (Study 2)

R1					R2					R3							
F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5			
0	0	0.03	0.02	0.08	0.02	0.02	0.03	0.05	0.06	0	0	0.02	0.01	0.04	EC1	ECL1	NM
0.01	0	0 .08	0.03	0.16	0.07	0.05	0.18	0.24	0.41	0.01	0	0.09	0.06	0.23	EC2		
0	0	0.06	0.04	0.37	0.02	0.03	0.07	0.08	0.21	0	0	0.05	0.03	0.23	EC1	ECL2	
0.01	0	0 .13	0.08	0.44	0.06	0.03	0.22	0.24	0.52	0.01	0	0.14	0.09	0.43	EC2		
0	0	0.03	0.01	0.08	0.15	0.07	0.17	0.07	0.14	0.02	0.01	0.06	0.02	0.07	EC1	ECL1	NM
0	0	0.02	0.01	0.06	0.2	0.05	0.42	0.23	0.44	0.03	0	0.14	0.05	0.19	EC2		
0.01	0.01	0.05	0.01	0.1	0.39	0.2	0.45	0.11	0.26	0.08	0.04	0.17	0.03	0.13	EC1	ECL3	
0.01	0	0 .05	0.02	0.11	0.33	0.09	0.59	0.26	0.55	0.09	0.02	0.26	0.08	0.27	EC2		
0.01	0.01	0.02	0.1	0.28	0.11	0.09	0.43	0.26	0.85	0.02	0.02	0.12	0.12	0.51	EC1	ECL1	NM
0.01	0.01	0	0 .07	0.06	0.07	0.1	0.11	0.14	0.23	0.01	0.02	0.02	0.06	0.08	EC2		
0.02	0.01	0.06	0.11	0.36	0.06	0.04	0.3	0.13	0.66	0.02	0.01	0.13	0.07	0.46	EC1	ECL2	
0.02	0.01	0.02	0.15	0.13	0.04	0.05	0.07	0.07	0.16	0.01	0.02	0.03	0.05	0.09	EC2		
0.07	0.04	0.12	0.09	0.16	0.16	0.11	0.54	0.32	0.77	0.07	0.04	0.25	0.13	0.39	EC1	ECL3	
0.02	0.02	0.01	0.06	0.02	0.06	0.1	0.07	0.16	0.17	0.02	0.03	0.02	0.07	0.05	EC2		

Explore effect of decisions yourself:
https://explorablemultiverse.github.io/examples/dataverse/

Interactive apps to explore multiverse



Linguistic Niche Browser

Explore an aggregation of datasets that test relationships between features of the environment and elements of language. This browser is supplemental material to Lewis and Frank (2016, CogSci Proceedings).

Predictor

Sum precipitation (cm)

Outcome

Consonant diversity

Color on the scatter plot indicates language family. Hover over points to identifiy individual languages.

Model fit:

○ simple linear regression
● random intercept of country
○ random intercept of family
○ random intercepts of country and family
○ random intercept and slope of family, intercept of country

https://mlewis.shinyapps.io/lnhBrowser/

Solutions

- Designed to reduce “questionable research practices”, like *p*-hacking.
- And, therefore **increase replicability**

1. Pre-registration – register your hypothesis and analysis plan publically before you collecting your data
2. Registered reports – write the paper and have it reviewed before you collect your data.
3. Blind analysis – conduct analysis blind to meaning in data
4. Multiverse analysis – do all ”forking paths” in analysis

Next up: Meta-analysis!

- How do you combine effect sizes from lots of different replications?
- “Quantitative literature review”