Evaluating Model Replication Efforts Through Prior Predictive Similarity Checking

Sonja D. Winter

(on behalf of my co-authors:
Wes Bonifay, Hana Skoblow, & Ashley Watts)

Statistics, Measurement, & Evaluation in Education
University of Missouri

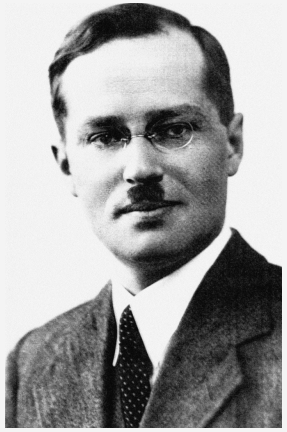
Comprehensive Statistical Model Evaluation

ncme 2024 ANNUAL MEETING · PHILADELPHIA ·

Model replication

Previous investigations of the replicability of studies and theories in psychological science have focused primarily on experimental effects

- But many subfields of psychology rely on statistical modeling rather than experimental designs
- Researchers in these areas, just like those in more experimental settings, must examine the degree to which their statistical models are replicable



"Models become plausible by repetition."

~ Jerzy Neyman

Model replication

Over time, many psychological researchers have assumed that model replication is indexed primarily by reproducing the goodness-of-fit (GOF) from a previous study, e.g.:

- "[the] improved model fit … replicate[d] the findings" of earlier research (Whiteman et al., 2022, p. 132)
- "the best fit … replicated previous findings" (Giuntoli et al., 2021, p. 1668)
- "substantially better fit … replicated the classical … approach" (Fernandez de la Cruz et al., 2018 p. 608)
- The "very good fit … **proved** the replicability of the overall structure" (Paruzel-Cazchura & Blukacz, 2021, p. 16)

Here, we argue **replicating the good fit of a model is NOT sufficient support for the original statistical model and its underlying theory**

Problems with good fit

Roberts & Pashler (2000) identified 3 aspects of GOF that preclude it from providing strong theoretical support.

Good fit…

1. Does not clarify what a model predicts
2. Does not clarify the variability of the data
3. Does not consider the a priori likelihood that the model will fit any plausible data

GOF only yields meaningful support for a theory when both data and theory are constrained, that is, when the data are not too variable and the theory is not too flexible

Good fit and replication

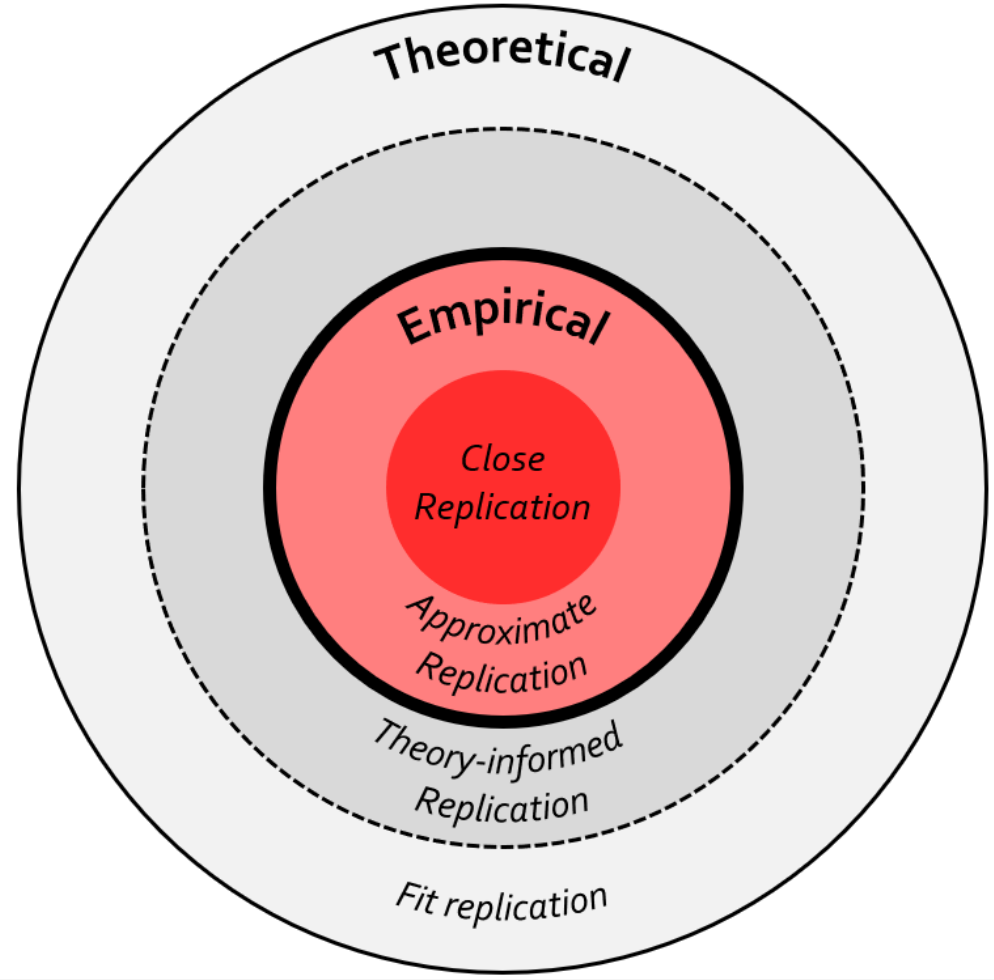
Extending Roberts & Pashler's points into the context of replication:

1. Regarding the original study that the researcher intends to replicate, GOF does not predict anything of substantive value about the replication outcomes
2. GOF reveals nothing about the similarity of the original and replication data
3. If a model has a high degree of *fitting propensity*, then good fit to the original and replication datasets will be unsurprising and of minimal scientific value

Target-setting

In psychology, perfect similarity between studies is impractical and likely unnecessary (McShane et al., 2019)

- Instead, researchers should focus their aim on the particular aspects of the original study that they wish to replicate



Prior predictive model checking

To formally investigate the similarity between the original and replicated replication and original data and parameter estimates, we will use **Bayesian prior predictive model checking**

- PrPMC consists of generating predictive samples for each observed variable in our model, based solely on the prior distributions placed on the model parameters
- These predictive samples represent hypothetical observed samples that are plausible under the expectations embedded in the prior distributions

Prior predictive model checking

A test statistic or quantity T in the observed sample y is compared to the same test statistic or quantity obtained from the simulated predictive samples y^{pred}

- A *prior predictive* p-*value* (*prpp*) is then computed to quantify the likelihood of T in the distribution of the prior predictive samples:

$$prpp = p\left(T(y) \geq T\left(y^{\text{pred}}\right)\right)$$

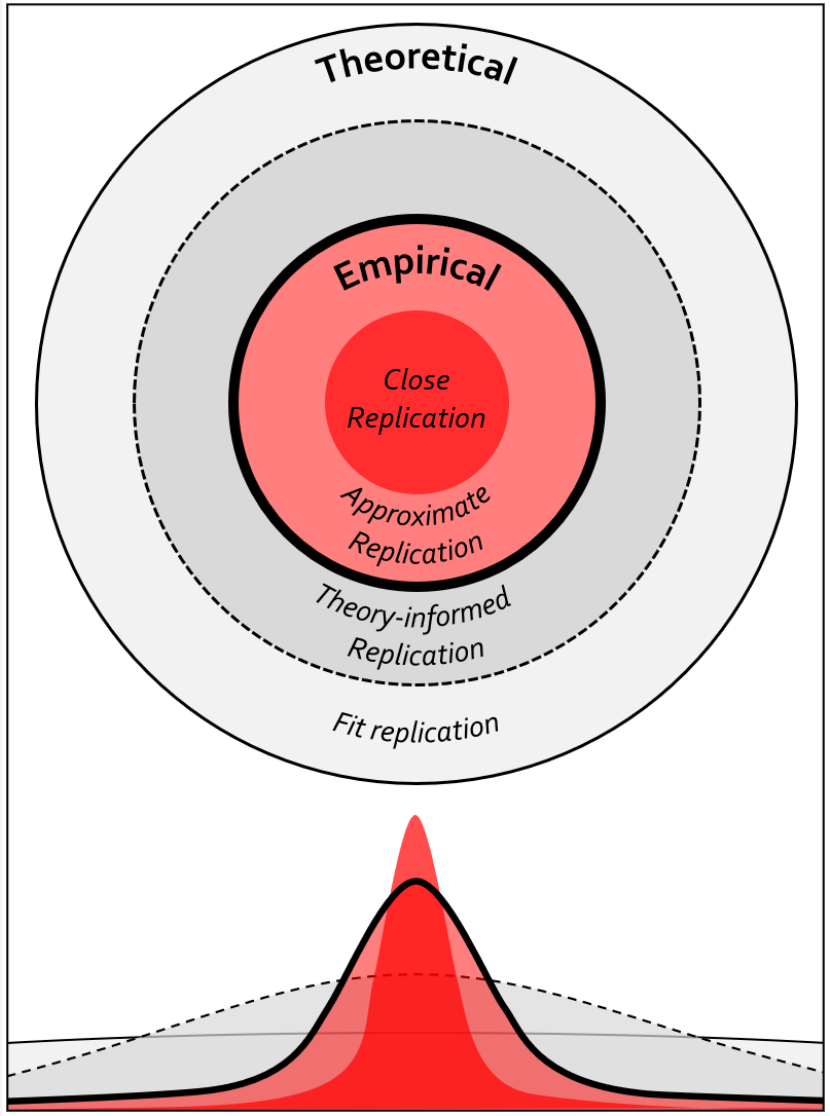
- $prpp \leq .05$ or $\geq .95$ indicates the presence of **systematic differences** between the observed sample and the prior predictive samples
- $prpp$ near .50 indicate that the observed T is **near the center** of the predictive distribution
  - The observed T aligns with the T we would expect to see, based on our priors

Prior predictive *similarity* checking

In the context of model replication, we propose a novel application of PrPMC that we refer to as *prior predictive* ***similarity*** *checking*

This process will allow researchers to quantify the degree of similarity:

1. between the original and replication **model parameter estimates**, via test quantities
2. between the original and replication **data**, via test statistics



Empirical application

Two datasets:

- An “original” study: The National Comorbidity Survey (NCS; Kessler et al., 1994), *N* = 8,098
- And a “replication” study: The National Comorbidity Survey Replication (NCS-R; Kessler & Merikangas, 2004), *N* = 9,282

Model:

- Confirmatory factor model with 3 correlated factors:
  1. **Externalizing**; indicators: alcohol dependence, drug dependence, & conduct disorder
  2. **Distress**; indicators: major depression, dysthymia, & generalized anxiety disorder
  3. **Fear**; indicators: agoraphobia, panic disorder, social anxiety disorder, & specific phobia

Empirical application

Traditional approach:

- The model fit well to the original (NCS) data: $\chi^2(32) = 160.19$, $p < .001$, CFI = .982, TLI = .975, RMSEA = .022, 90% CI [.019, .026]
- The same model also fit well to the replication (NCS-R) data: $\chi^2(32) = 124.70$, $p < .001$, CFI = .991, TLI = .987, RMSEA = .018, 90% CI [.014, .021]

The replication study successfully reproduced the original fit!

…Hooray?

Prior Specifications

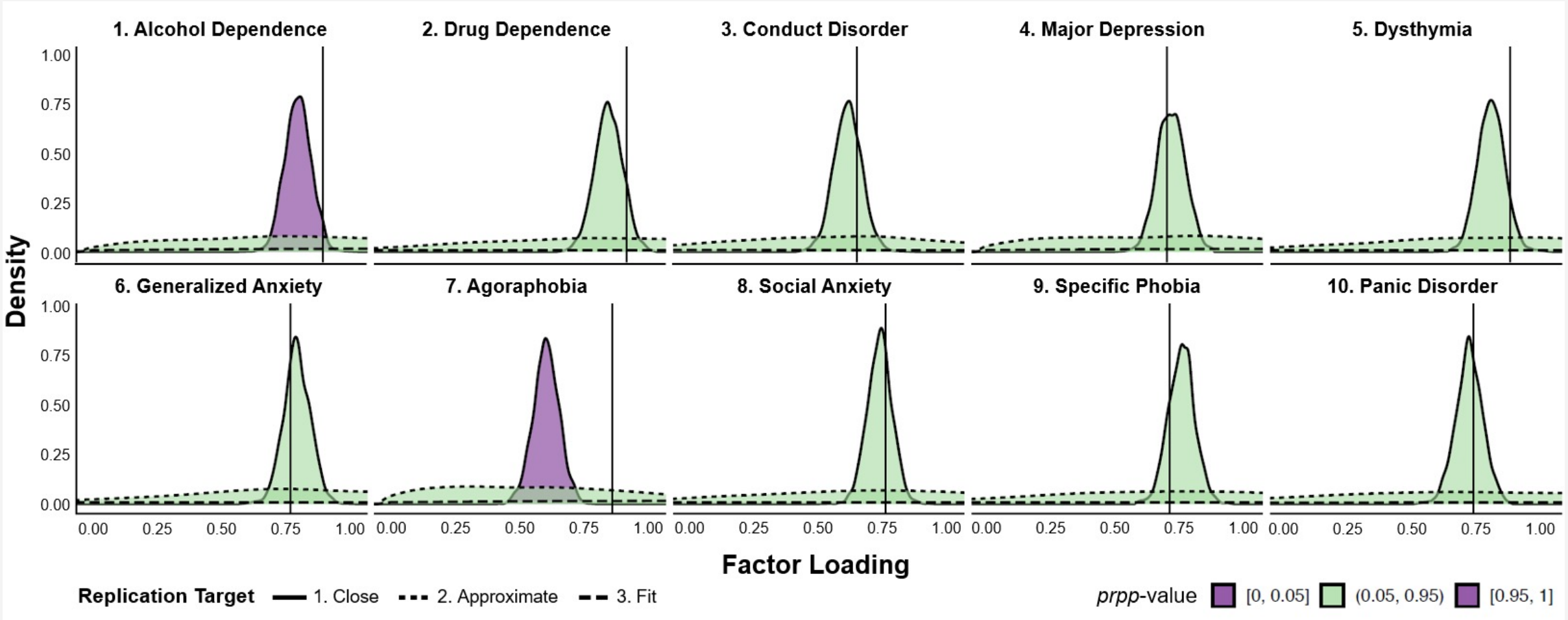
	Original Study			Replication Target Prior Distributions		
	Estimate	Std. Err.	β	Close	Approximate	Fit
Loadings						
Externalizing						
Alcohol dependence	1.00	–	.79	N(.79, .05)	N(.79, .50)	N(0, 5)
Drug dependence	1.18	.16	.84	N(.84, .05)	N(.84, .50)	N(0, 5)
Conduct disorder	.60	.06	.61	N(.61, .05)	N(.61, .50)	N(0, 5)
Distress						
Major depression	1.00	–	.70	N(.70, .05)	N(.70, .50)	N(0, 5)
Dysthymia	1.28	.14	.78	N(.78, .05)	N(.78, .50)	N(0, 5)
Generalized anxiety	1.28	.13	.79	N(.79, .05)	N(.79, .50)	N(0, 5)
Fear						
Agoraphobia	1.00	–	.60	N(.60, .05)	N(.60, .50)	N(0, 5)
Social anxiety	1.43	.14	.73	N(.73, .05)	N(.73, .50)	N(0, 5)
Specific phobia	1.48	.14	.74	N(.74, .05)	N(.74, .50)	N(0, 5)
Panic disorder	1.32	.16	.70	N(.70, .05)	N(.70, .50)	N(0, 5)

Prior Specifications

	Original Study			Replication Target Prior Distributions		
	Estimate	Std. Err.	β	Close	Approximate	Fit
Thresholds						
Alcohol dependence	1.69	.08	–	N(1.69, .05)	N(1.69, .50)	N(0, 3)
Drug dependence	2.58	.16	–	N(2.58, .05)	N(2.58, .50)	N(0, 3)
Conduct disorder	1.46	.04	–	N(1.46, .05)	N(1.46, .50)	N(0, 3)
Major depression	1.40	.05	–	N(1.40, .05)	N(1.40, .50)	N(0, 3)
Dysthymia	2.37	.10	–	N(2.37, .05)	N(2.37, .50)	N(0, 3)
Generalized anxiety	2.64	.12	–	N(2.64, .05)	N(2.64, .50)	N(0, 3)
Agoraphobia	2.09	.06	–	N(2.09, .05)	N(2.09, .50)	N(0, 3)
Social anxiety	1.64	.06	–	N(1.64, .05)	N(1.64, .50)	N(0, 3)
Specific phobia	1.84	.07	–	N(1.84, .05)	N(1.84, .50)	N(0, 3)
Panic disorder	2.57	.11	–	N(2.57, .05)	N(2.57, .50)	N(0, 3)
Correlations						
Externalizing–Distress	.50	.05	.40	Beta(97.7, 42.8)	Beta(16.5, 7.7)	LKJ(1)
Externalizing–Fear	.37	.05	.39	Beta(98.5, 43.9)	Beta(16.0, 7.6)	LKJ(1)
Distress–Fear	.46	.05	.63	Beta(96.3, 22.9)	Beta(24.5, 6.4)	LKJ(1)

Similarity of Parameter Estimates

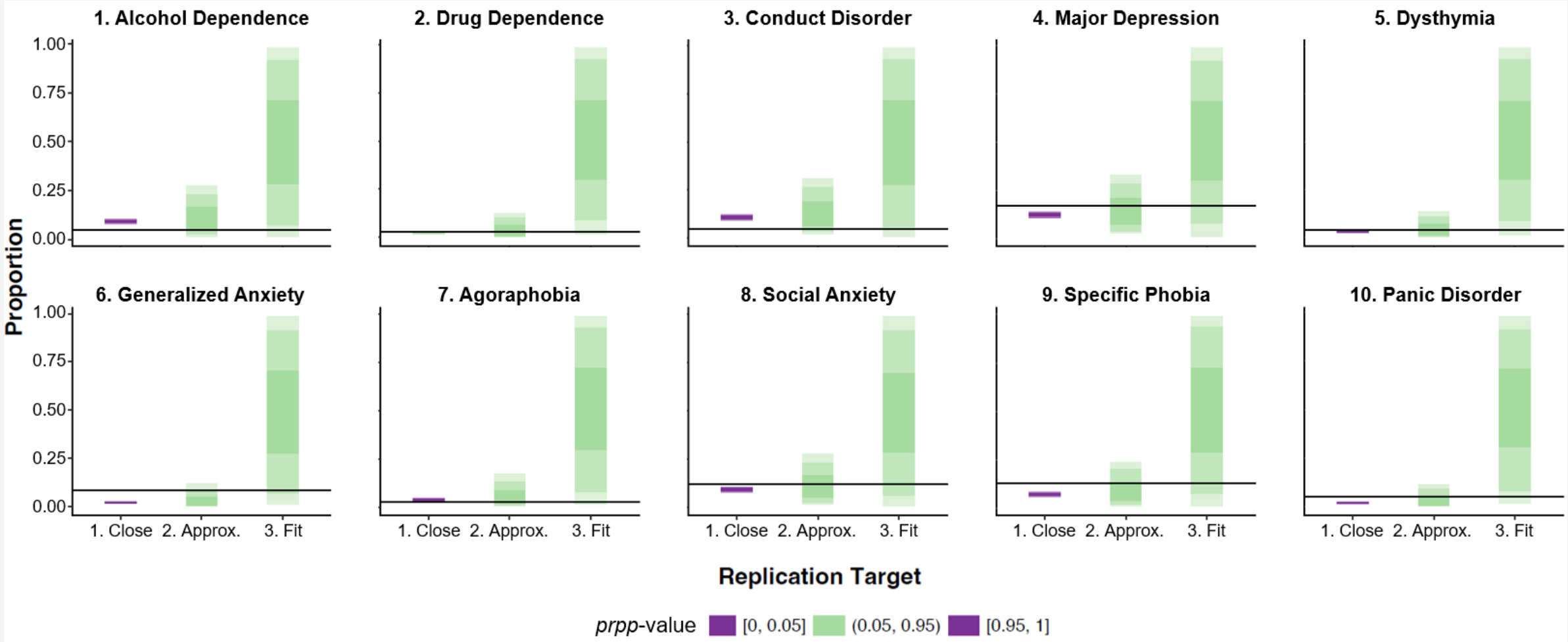
Factor Loadings



Green = hit; purple = miss.

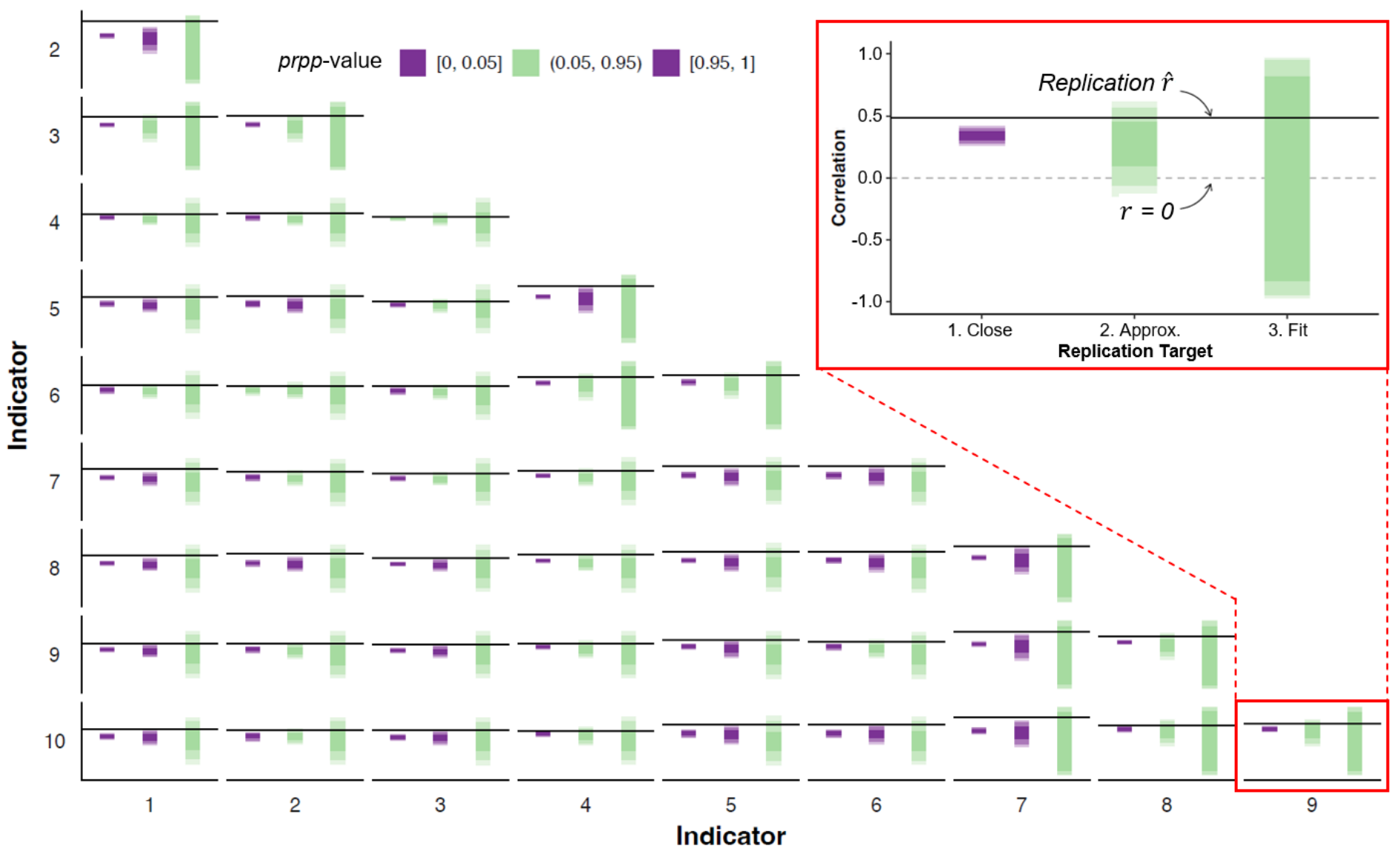
Similarity of Data

Response Proportions



Similarity of Data

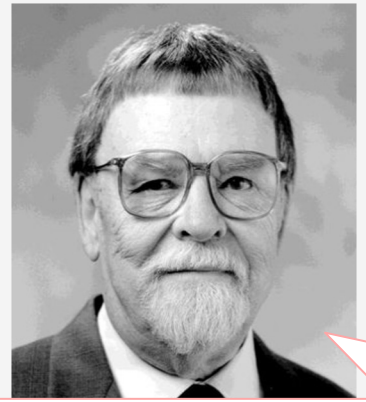
Interitem correlations



Summary

Nosek & Errington (2020): “Replication is a study for which any outcome would be considered diagnostic evidence about a claim from prior research” (p. 2)

- Prior predictive similarity checking offers **a confrontation of theory** rather than a confirmation of theory
  - The findings of any single (“original”) study are characterized by uncertainty
  - Failing the check: *Adds* to our uncertainty about original findings
  - Passing the check: *Reduces* our uncertainty about original findings
    - Riskier targets, if hit, offer a greater reduction of that uncertainty



“All we can do is subject theories … to grave danger of refutation. … A theory is corroborated to the extent that we have subjected it to such risky tests; **the more dangerous tests it has survived, the better corroborated it is.**”

~ Paul Meehl, 1978

Recommendations

Researchers should specify exact priors, informed by the original data, aimed at the preferred replication target

- Include these priors in a preregistration/registered report
- Accept failure and be transparent about it!

Gold medalist Doreen Wilber on failing to hit the target: “Even when I shoot a bad arrow, I don't get angry. I'm a very cool person. Nothing upsets me.” (*Des Moines Register*, 1981)

Scan this for access to:

- A preprint
- Thoroughly annotated R code
- A curated list of readings on Bayesian inference and prior specification

Thank you!

sdwinter@missouri.edu

@winterstat