



# Article Level Metrics and Many Labs Replication Outcomes



Erika Salomon (ecsalomon@gmail.com)  
Department of Psychology, University of Illinois at Urbana-Champaign

## Introduction

- Psychologists have started using metrics based on test statistics or p-values to infer the evidential value of published research findings.
  - At the same time, large-scale replication projects (e.g., Many Labs<sup>1,2</sup>, Reproducibility Project: Psychology<sup>3</sup>) have tested the large-scale replicability of psychological research.
  - Using data from Many Labs 1 and 3, this study examines whether paper-level metrics predict replication results.
- Research Question:** Is it possible to predict from a paper's statistics whether or how well its effects will replicate?

## Method

**Sample:** Original papers reporting effects included in Many Labs 1 and 3 were used because of the availability of meta-analytic estimates of replication success.

Only papers that reported sufficient information to calculate the appropriate metrics were included in each analysis.

**Predictors:** Using the p-checker app, six predictors were estimated for each original paper, based on tests of critical hypotheses, as recommended by Simonsohn, Nelson, & Simmons<sup>4</sup>.

- P-Curve: Evidential Value:** Test statistic ( $z$ ) for evidential value of a set of studies based on  $p$ -values<sup>5</sup>
- P-Curve: Lacks Evidential Value:** Test statistic ( $z$ ) for evidential value of a set of studies based on  $p$ -values<sup>5</sup>
- R-Index:** Based on the difference between the expected and actual number of significant results<sup>6</sup>
- Test of Insufficient Variance (TIVA):** The variance in the converted z-scores of test statistics<sup>7</sup>
- Correlation Between Effect Size and  $N$ :** Pearson correlation between the observed effect sizes and sample sizes in a paper
- N-Pact Factor:** Median sample size for included tests<sup>8</sup>

**Outcomes:** Replication outcomes were operationalized in two ways:

- Difference in Effect Size (continuous):** The difference between the original and replication effects, scaled to Cohen's  $d$
- Replication Success (dichotomous):** Whether the meta-analytic estimate of the effect size was significant at  $p < .05$ .

Fig 1. Predicting Differences Between Original and Replication Effect Sizes (d)

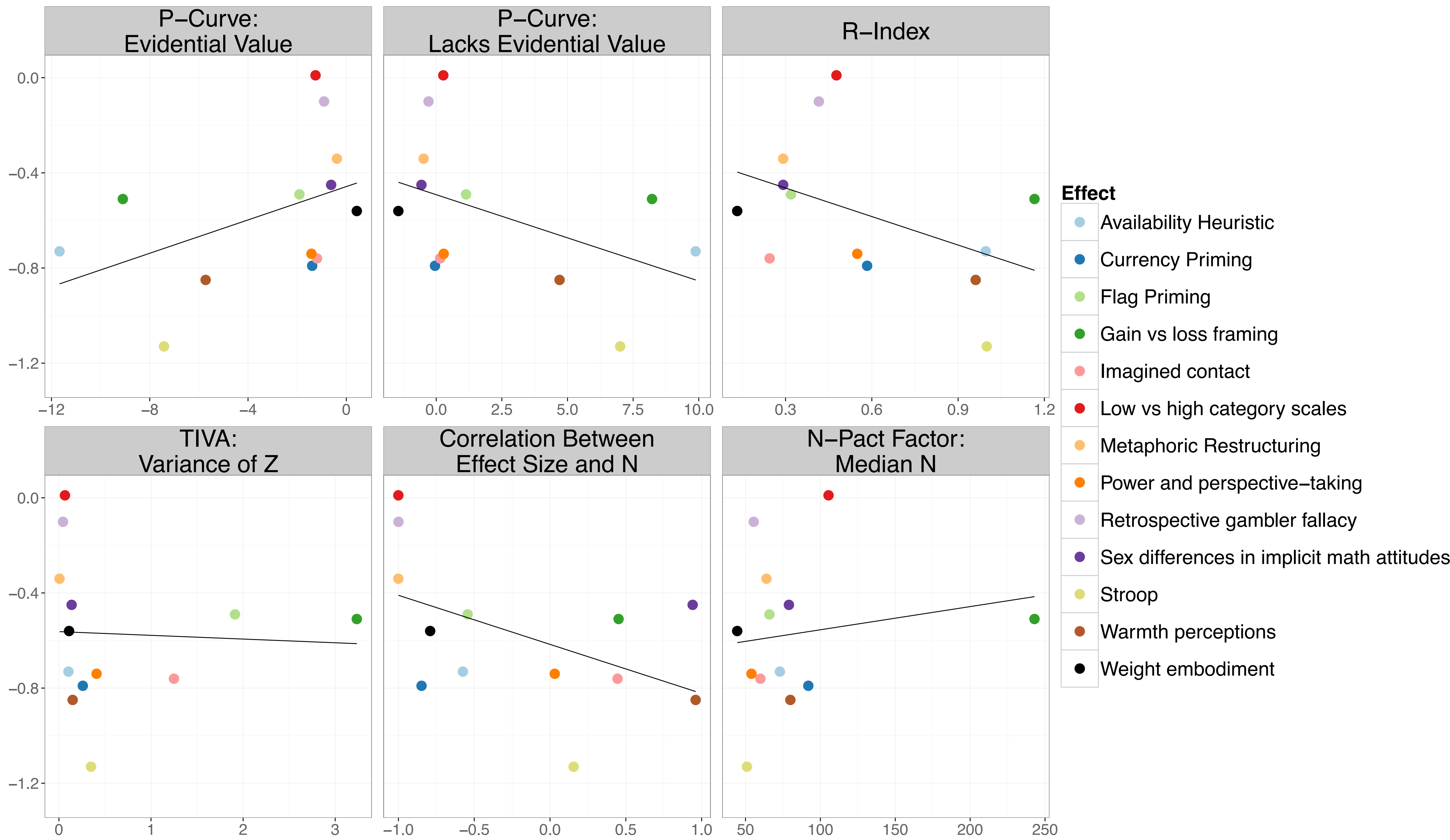
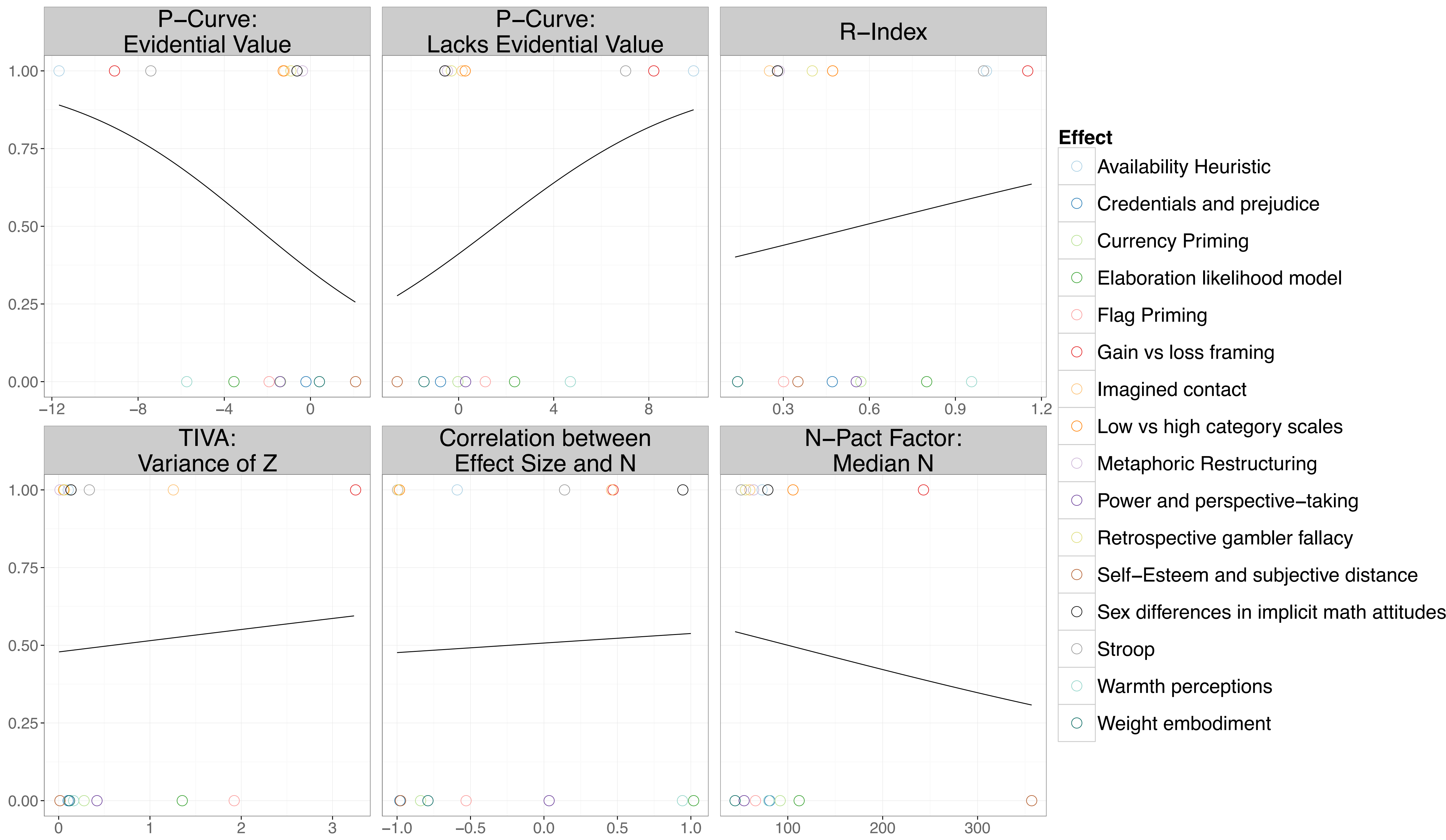


Fig 2. Predicted Probability of Replication Success



## Results

### Difference in Effect Size

**Table 1** presents correlations between the predictor variables and the difference in effect size between original and replication results. **None of the relationships reached significance at  $p < .05$ .** However, **Figure 1** shows restricted range among many predictors.

### Replication Success

**Table 2** presents logistic regression coefficients from models predicting success from each of the predictors. **Figure 2** shows the predicted probability of successful replication from each model.

Table 1. Correlations Among Predictors and Continuous Outcome

	1	2	3	4	5	6
1. Difference in Effect Size						
2. P-Curve: Evidential Value	<b>.44</b>					
3. P-Curve: Lacks Evidential Value	<b>-.45</b>	-.99				
4. R-Index	<b>-.44</b>	-.89	.89			
5. TIVA: Variance in Z	<b>-.05</b>	-.37	.39	.35		
6. Correlation Between Effect Size and N	<b>-.49</b>	-.37	.37	.42	.40	
7. N-Pact Factor	<b>.16</b>	.06	-.03	.15	.27	-.06

Table 2. Six Different Logistic Regression Models Predicting Replication Success

	$b$	$SE$	$z$	$p$	$OR$
1. P-Curve: Evidential Value	-0.23	0.18	-1.31	.19	0.79
2. P-Curve: Lacks Evidential Value	0.23	0.18	1.33	.18	1.26
3. R-Index	0.93	1.64	0.56	.57	2.53
4. TIVA: Variance in Z	0.14	0.58	0.25	.80	1.16
5. Correlation Between Effect Size and N	0.12	0.66	0.19	.85	1.13
6. N-Pact Factor	-0.00	0.01	-0.48	.63	1.00

## Conclusions

- Reporting practices should be improved** so that papers consistently report information needed for meta-analytic and meta-scientific research (e.g., cell sizes, effect sizes, full model details).
- Researchers should exercise caution in making inferences from metrics based on single papers or small sets of studies.**
- P-curve may show promise** in predicting replication outcomes but is far from definitive when used on single papers.

## References

<sup>1</sup>Ebersole et al., 2015 <sup>2</sup>Klein et al., 2014 <sup>3</sup>Open Science Collaboration, 2015 <sup>4</sup>Simonsohn, Simmons, & Nelson, 2015 <sup>5</sup>Simonsohn, Nelson, & Simmons, 2015, Quantifying Statistical Research Integrity: The Replicability-Index <sup>6</sup>Schimmack, 2014, The Test of Insufficient Variance (TIVA): A New Tool for the Detection of Questionable Research Practices <sup>7</sup>Fraley & Vazire, 2014

## PDF, Data, & Code

<https://github.com/ecsalomon/TSR---Test-Stats-Replication>

