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 and Registered Replication Reports at Perspectives on Psychological Science) 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. Is it possible to infer from a paper’s statistics whether (or how well) its effects will replicate? Across the effects in Many Labs 1 and 3, I find that p-curve, replication index, test of insufficient variance, and average sample size do not predict replication outcomes. This is true whether replication results are operationalized as significant effects in the same direction or as differences in effect size from the original findings. These results suggest caution in using paper-level metrics to infer the evidential value of individual effects.

Across the effects in Many Labs 1 and 3, I find that p-curve, replication index, test of insufficient variance, and average sample size do not predict replication outcomes. These results suggest caution in using paper-level metrics to infer the evidential value of individual effects.