RESEARCH ARTICLE

Estimating the reproducibility of psychological science

c *†

having only a small set of articles available at a time and matching studies with replication teams' interests, resources, and expertise.

By default, the last experiment reported in each article was the subject of replication. This decision established an objective standard for study selection within an article and was based on the intuition that the first study in a multiple-study article (the obvious alternative selection strategy) was more frequently a preliminary demonstration. Deviations from selecting the last experiment were made occasionally on the basis of feasibility or recommendations of the original authors. Justifications for deviations were reported in the replication reports, which were made available on the Open Science Framework (OSF) (http://osf.io/ezcuj). In total, 84 of the 100 completed replications (84%) were of the last reported study in the article. On average, the to-be-replicated articles contained 2.99 studies (SD = 1.78) with the following distribution: 24 single study, 24 two studies, 18 three studies, 13 four studies, 12 five studies, 9 six or more studies. All following summary statistics refer to the 100 completed replications.

For the purposes of aggregating results across studies to estimate reproducibility, a key result from the selected experiment was identified as the focus of replication. The key result had to be represented as a single statistical inference test or an effect size. In most cases, that test was a t test, F

out this test on the subset of study pairs in which both the correlation coefficient and its standard error could be computed [we refer to this data set as the meta-analytic (MA) subset]. Standard errors could only be computed if test statistics were r, t, or F(1,df

(M = 0.403, SD = 0.188) were reliably larger than replication effect sizes (M = 0.197, SD = 0.257), Wilcoxon's W = 7137, P < 0.001. Of the 99 studies for which an effect size in both the original and replication study could be calculated (30), 82 showed a stronger effect size in the original study (82.8%; P < 0.001, binomial test) (Fig. 1, right). Original and replication effect sizes were

in reproducibility across indicators. Replication success was more consistently related to the original strength of evidence (such as original P value, effect size, and effect tested) than to characteristics of the teams and implementation of the replication (such as expertise, quality, or challenge of conducting study) (tables S3 and S4).

No single indicator sufficiently describes replication success, and the five indicators examined here are not the only ways to evaluate reproducibility. Nonetheless, collectively, these results offer a clear conclusion: A large portion of repli-

challenges and may cross-fertilize strategies so as to improve reproducibility.

Because reproducibility is a hallmark of credible scientific evidence, it is tempting to think that maximum reproducibility of original results is important from the onset of a line of inquiry through its maturation. This is a mistake. If initial ideas were always correct, then there would hardly be a reason to conduct research in the first place. A healthy discipline will have many false starts as it confronts the limits of present understanding.

Innovation is the engine of discovery and is vital for a productive, effective scientific enterprise. However, innovative ideas become old news fast. Journal reviewers and editors may dismiss a new test of a published idea as unoriginal. The claim that "we already know this" belies the uncertainty of scientific evidence. Deciding the ideal balance of resourcing innovation versus verification is a question of research efficiency. How can we maximize the rate of research progress? Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both. The ideal balance is a topic for investigation itself. Scientific incentives—funding, publication, or awards—can be tuned to encourage an optimal balance in the collective effort of discovery (36, 37).

Progress occurs when existing expectations

The Open Science Collaboration