Estimating the reproducibility of psychological science

Open Science Collaboration*

INTRODUCTION: Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. Scientific claims should not gain credence because of the status or authority of their originator but by the replicability of their supporting evidence. Even research of exemplary quality may have irreproducible empirical findings because of random or systematic error.

RATIONALE: There is concern about the rate and predictors of reproducibility, but limited evidence. Potentially problematic practices include selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results. Direct replication is the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding and is the means of establishing reproducibility of a finding with new data. We conducted a large-scale, collaborative effort to obtain an initial estimate of the reproducibility of psychological science.

RESULTS: We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. There is no single standard for evaluating replication success. Here, we evaluated reproducibility using significance and P values, effect sizes, subjective assessments of replication teams, and meta-analysis of effect sizes. The mean effect size (r) of the replication effects (M_r = 0.197, SD = 0.257) was half the magnitude of the mean effect size of the original effects (M_r = 0.403, SD = 0.186), representing a substantial decline. Ninety-seven percent of original studies had significant results (P < .05). Thirty-six percent of replications had significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

CONCLUSION: 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 replications produced weaker evidence for the original findings despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original P value) was more predictive of replication success than variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but they did not appear to do so here.

Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication. 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. Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both. Replication can increase certainty when findings are reproduced and promote innovation when they are not. This project provides accumulating evidence for many findings in psychological research and suggests that there is still more work to do to verify whether we know what we think we know.

*Corresponding author. E-mail: nosek@virginia.edu

The list of author affiliations is available in the full article online.

Cite this article as Open Science Collaboration, Science 349, aac4716 (2015). DOI: 10.1126/science.aac4716

ON OUR WEB SITE
Read the full article at http://dx.doi.org/10.1126/science.aac4716

π

Original study effect size versus replication effect size (correlation coefficients). Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.
Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown. We conducted replications of 100 experimental and correlational studies published in three psychology journals using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline. Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original result; and if no bias in original results is assumed, combining original and replication results left 68% with statistically significant effects. Correlational tests suggest that replication success was better predicted by the strength of original evidence than by characteristics of the original and replication teams.

Researchers with their affiliations appear at the end of this paper.

Corresponding author. E-mail: nosek@virginia.edu
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 Frame-
work (OSF) (http://osf.io/ezcuj). In total, 84 of
the 100 completed replications (84%) were of
the last reported study in the article. On aver-
age, the to-be-replicated articles contained 2.99
studies (SD = 1.78) with the following distribu-
tion: 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 test, or correlation coefficient. The effect
was identified before data collection or analysis
and was presented to the original authors as part
of the design protocol for critique. Original au-
tors occasionally suggested that a different
effect be used, and by default, replication teams
turned to original authors' judgments. None-
theless, because the single effect came from a
single study, it is not necessarily the case that
the identified effect was central to the overall
aims of the article. In the individual replication
reports and subjective assessments of replica-
tion outcomes, more than a single result could
be examined, but only the result of the single
effect was considered in the aggregate analyses
[additional details of the general protocol and
individual study methods are provided in the
supplementary materials and (25)].

In total, there were 488 articles in the 2008
issues of the three journals. One hundred fifty-
eight of these (32%) became eligible for selec-
tion for replication during the project period,
between November 2011 and December 2014.
From those, 111 articles (70%) were selected by
a replication team, producing 113 replications.
Two articles had two replications each (supplemen-
tary materials). And 100 of those (58%) replications
were completed by the project deadline for in-
clusion in this aggregate report. After being
claimed, some studies were not completed be-
cause the replication teams ran out of time or
could not devote sufficient resources to com-
pleting the study. By journal, replications were
completed for 39 of 64 (61%) articles from
PSCI, 31 of 55 (56%) articles from JPSP, and 28
of 39 (72%) articles from JEP:LMC.

The most common reasons for failure to match
an article with a team were feasibility constraints
for conducting the research. Of the 47 articles
from the eligible pool that were not claimed, six
(13%) had been deemed infeasible to replicate
because of time, resources, instrumentation, de-
hpendence on historical events, or hard-to-access
samples. The remaining 41 (87%) were eligible
but not claimed. These often required specialized
samples (such as macaques or people with autism),
resources (such as eye tracking machines or func-
tional magnetic resonance imaging), or knowl-
edge making them difficult to match with teams.

Aggregate data preparation

Each replication team conducted the study, ana-
alyzed their data, wrote their summary report,
and completed a checklist of requirements for
sharing the materials and data. Then, independ-
ent reviewers and analysts conducted a project-
wide audit of all individual projects, materials,
data, and reports. A description of this review
is available on the OSF (https://osf.io/ntme). More-
over, to maximize reproducibility and accuracy,
the analyses for every replication study were re-
produced by another analyst independent of
the replication team using the R statistical pro-
gramming language and a standardized analytic
format. A controller R script was created to re-
generate the entire analysis of every study
and recreate the master data file. This R script, avail-
able at https://osf.io/fkmwwg, can be executed to
produce the results of the individual studies. A
comprehensive description of this reanalysis
process is available publicly (https://osf.io/a2egy).

Measures and moderators

We assessed features of the original study and
replication as possible correlates of reproduc-
bility and conducted exploratory analyses to
inspire further investigation. These included
characteristics of the original study such as the
publishing journal; original effect size,
xperimentation condition, and other resources
such as eye tracking machines or functional
magnetic resonance imaging. The replications
were classified as statistically significant
(P ≤ 0.05), and were coded as representing
nonsignificant
(P > 0.05). The

Published journal and subdiscipline

Journals' different publishing practices may re-
sult in a selection bias that covaries with repro-
ducibility. Articles from three journals were
made available for selection: JPSP (n = 59 articles), JEP:
LMC (n = 40 articles), and PSCI (n = 68 articles). From
this pool of available studies, replications were
selected and completed from JPSP (n = 32 studies), JEP:
LMC (n = 28 studies), and PSCI (n = 40 studies) and were
coded as representing cognitive (n = 43 studies) or social-personality
(n = 57 studies) subdisciplines. Four studies that
would ordinarily be understood as "developmental psychology" be-
cause of studying children or infants were coded as having a cognitive or social
emphasis. Reproducibility may vary by subdiscipline
in psychology because of differing practices. For
example, within-subjects designs are more com-
mon in cognitive than social psychology, and
these designs often have greater power to detect
effects with the same number of participants.

Statistical analyses

There is no single standard for evaluating rep-
lication success (25). We evaluated reproduc-
ability using significance and P values, effect sizes,
subjective assessments of replication teams, and
meta-analyses of effect sizes. All five of these
indicators contribute information about the rel-
ations between the replication and original find-
ning and the cumulative evidence about the effect
and were positively correlated with one another
(r ranged from 0.22 to 0.96, median r = 0.57).
Results are summarized in Table 1, and full details
of analyses are in the supplementary materials.

Significance and P values

Assuming a two-tailed test and significance or
α level of 0.05, all test results of original and
replication studies were classified as statisti-
cally significant (P ≤ 0.05) and nonsignificant
(P > 0.05). However, original studies that inter-
preted nonsignificant P values as significant
were coded as significant (four cases, all with
P values < 0.06). Using only the nonsignificant
P values of the replication studies and applying
Fisher's method (26), we tested the hypothesis
that these studies had "no evidential value" (the
null hypothesis of zero-evidence holds for all these
studies). We tested the hypothesis that the propor-
tions of statistically significant results in the
original and replication studies are equal using
the McNemar test for paired nominal data and
calculated a CI of the reproducibility parame-
ter. Second, we compared the central tendency
of the distribution of P values of original and
replication studies using the Wilcoxon signed-rank test and the *t* test for dependent samples. For both tests, we only used study-pairs for which both *P* values were available.

**Effect sizes**

We transformed effect sizes into correlation coefficients whenever possible. Correlation coefficients have several advantages over other effect size measures, such as Cohen’s *d*. Correlation coefficients are bounded, well known, and therefore more readily interpretable. Most important for our purposes, analysis of correlation coefficients is straightforward because, after applying the Fisher transformation, their standard error is only a function of sample size. Formulas and code for converting test statistics *z*, *F*, *t*, and *χ*² into correlation coefficients are provided in the appendices at http://osf.io/ezum7. To be able to compare and analyze correlations across study-pairs, the original study’s effect size was coded as positive; the replication study’s effect size was coded as negative if the replication study’s effect was opposite to that of the original study.

We compared effect sizes using four tests. We compared the central tendency of the effect size distributions of original and replication studies using both a paired two-sample *t* test and the Wilcoxon signed-rank test. Third, we computed the proportion of study-pairs in which the effect of the original study was stronger than in the replication study and tested the hypothesis that this proportion is 0.5. For this test, we included findings for which effect size measures were available but no correlation coefficient could be computed (for example, if a regression coefficient was reported but not its test statistic). Fourth, we calculated “coverage,” or the proportion of study-pairs in which the effect of the original study was in the CI of the effect of the replication study, and compared this with the expected proportion using a goodness-of-fit *χ*² test. We carried

Table 1. Summary of reproducibility rates and effect sizes for original and replication studies overall and by journal/discipline. df/N refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size = 3 of correlation, and sample size for *z* and *χ*²). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are “replications *P* < 0.05” (3 original nulls excluded; *n* = 97 studies); “mean original and replication effect sizes” (3 excluded; *n* = 97 studies); “meta-analytic mean estimates” (27 excluded; *n* = 73 studies); “percent meta-analytic (*P* < 0.05)” (25 excluded; *n* = 75 studies); and, “percent original effect size within replication 95% CI” (5 excluded, *n* = 95 studies).

<table>
<thead>
<tr>
<th>Replications</th>
<th>Percent</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>Mean (SD)</th>
<th>Median (SD)</th>
<th>Average replication power</th>
<th>Meta-analytic mean (SD) estimate</th>
<th>Percent meta-analytic (<em>P</em> &lt; 0.05)</th>
<th>Percent original effect size within replication 95% CI</th>
<th>Percent subjective “yes” to “Did it replicate?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>35/97</td>
<td>36</td>
<td>0.403 (0.188)</td>
<td>54</td>
<td>0.197 (0.257)</td>
<td>68</td>
<td>0.92</td>
<td>0.309 (0.223)</td>
<td>68</td>
<td>47/39</td>
</tr>
<tr>
<td>JPSP, social</td>
<td>7/31</td>
<td>23</td>
<td>0.29 (0.10)</td>
<td>73</td>
<td>0.07 (0.11)</td>
<td>120</td>
<td>0.91</td>
<td>0.138 (0.087)</td>
<td>43</td>
<td>34/25</td>
</tr>
<tr>
<td>JEP/LMC, cognitive</td>
<td>13/27</td>
<td>48</td>
<td>0.47 (0.18)</td>
<td>36.5</td>
<td>0.27 (0.24)</td>
<td>43</td>
<td>0.93</td>
<td>0.393 (0.209)</td>
<td>86</td>
<td>62/54</td>
</tr>
<tr>
<td>PSCI, social</td>
<td>7/24</td>
<td>29</td>
<td>0.39 (0.20)</td>
<td>76</td>
<td>0.21 (0.30)</td>
<td>122</td>
<td>0.92</td>
<td>0.286 (0.228)</td>
<td>58</td>
<td>40/32</td>
</tr>
<tr>
<td>PSCI, cognitive</td>
<td>8/15</td>
<td>53</td>
<td>0.53 (0.2)</td>
<td>23</td>
<td>0.29 (0.35)</td>
<td>21</td>
<td>0.94</td>
<td>0.464 (0.221)</td>
<td>92</td>
<td>60/53</td>
</tr>
</tbody>
</table>

Table 2. Spearman’s rank-order correlations of reproducibility indicators with summary original and replication study characteristics. Effect size difference computed after converting *r* to Fisher’s *z*. df/N refers to the information on which the test of the effect was based (for example, *df* of *t* test, denominator *df* of *F* test, sample size = 3 of correlation, and sample size for *z* and *χ*²). Four original results had *P* values slightly higher than 0.05 but were considered positive results in the original article and are treated that way here. Exclusions (explanation provided in supplementary materials, A3) are “replications *P* < 0.05” (3 original nulls excluded; *n* = 97 studies); “effect size difference” (3 excluded; *n* = 97 studies); “meta-analytic mean estimates” (27 excluded; *n* = 73 studies); and, “percent original effect size within replication 95% CI” (5 excluded, *n* = 95 studies).

<table>
<thead>
<tr>
<th>Replications</th>
<th>Percent</th>
<th>Effect size difference</th>
<th>Meta-analytic estimate</th>
<th>Original effect size within replication 95% CI</th>
<th>Subjective “yes” to “Did it replicate?”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original P value</td>
<td>−0.327</td>
<td>−0.057</td>
<td>−0.468</td>
<td>0.032</td>
<td>−0.260</td>
</tr>
<tr>
<td>Original effect size</td>
<td>0.304</td>
<td>0.279</td>
<td>0.793</td>
<td>0.121</td>
<td>0.277</td>
</tr>
<tr>
<td>Original df/N</td>
<td>−0.150</td>
<td>−0.194</td>
<td>−0.502</td>
<td>−0.221</td>
<td>−0.185</td>
</tr>
<tr>
<td>Importance of original result</td>
<td>−0.105</td>
<td>0.038</td>
<td>−0.205</td>
<td>−0.133</td>
<td>−0.074</td>
</tr>
<tr>
<td>Surprising original result</td>
<td>−0.244</td>
<td>0.102</td>
<td>−0.181</td>
<td>−0.113</td>
<td>−0.241</td>
</tr>
<tr>
<td>Experience and expertise of original team</td>
<td>−0.072</td>
<td>−0.033</td>
<td>−0.059</td>
<td>−0.103</td>
<td>−0.044</td>
</tr>
<tr>
<td>Replication characteristics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication P value</td>
<td>−0.828</td>
<td>0.621</td>
<td>−0.614</td>
<td>−0.562</td>
<td>−0.738</td>
</tr>
<tr>
<td>Replication effect size</td>
<td>0.731</td>
<td>0.586</td>
<td>0.850</td>
<td>0.611</td>
<td>0.710</td>
</tr>
<tr>
<td>Replication power</td>
<td>0.368</td>
<td>0.053</td>
<td>0.142</td>
<td>0.056</td>
<td>0.285</td>
</tr>
<tr>
<td>Replication df/N</td>
<td>−0.085</td>
<td>−0.224</td>
<td>−0.692</td>
<td>−0.257</td>
<td>−0.164</td>
</tr>
<tr>
<td>Challenge of conducting replication</td>
<td>−0.219</td>
<td>0.085</td>
<td>−0.301</td>
<td>−0.109</td>
<td>−0.151</td>
</tr>
<tr>
<td>Experience and expertise of replication team</td>
<td>−0.096</td>
<td>0.133</td>
<td>0.017</td>
<td>0.053</td>
<td>−0.068</td>
</tr>
<tr>
<td>Self-assessed quality of replication</td>
<td>−0.069</td>
<td>0.017</td>
<td>0.054</td>
<td>0.088</td>
<td>−0.055</td>
</tr>
</tbody>
</table>
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 $F(df_1, df_2)$. The expected proportion is the sum over expected probabilities across study-pairs. The test assumes the same population effect size for original and replication study in the same study-pair. For those studies that tested the effect with $F(df_1 > 1, df_2)$ or $\chi^2$, we verified coverage using other statistical procedures (computational details are provided in the supplementary materials).

**Meta-analysis combining original and replication effects**

We conducted fixed-effect meta-analyses using the R package metafor (27) on Fisher-transformed correlations for all study-pairs in subset MA and on study-pairs with the odds ratio as the dependent variable. The number of times the CI of all these meta-analyses contained 0 was calculated. For studies in the MA subset, estimated effect sizes were averaged and analyzed by discipline.

**Subjective assessment of “Did it replicate?”**

In addition to the quantitative assessments of replication and effect estimation, we collected subjective assessments of whether the replication provided evidence of replicating the original result. In some cases, the quantitative data anticipate a straightforward subjective assessment of replication. For more complex designs, such as multivariate interaction effects, the quantitative analysis may not provide a simple interpretation. For subjective assessment, replication teams answered “yes” or “no” to the question, “Did your results replicate the original effect?” Additional subjective variables are available for analysis in the full data set.

**Analysis of moderators**

We correlated the five indicators evaluating reproducibility with six indicators of the original study (original $P$ value, original effect size, original sample size, importance of the effect, surprising effect, and experience and expertise of original team) and seven indicators of the replication study (replication $P$ value, replication effect size, replication power based on original effect size, replication sample size, challenge of conducting replication, experience and expertise of replication team, and self-assessed quality of replication) (Table 2). As follow-up, we did the same with the individual indicators comprising the moderator variables (tables S3 and S4).

**Results**

**Evaluating replication effect against null hypothesis of no effect**

A straightforward method for evaluating replication is to test whether the replication shows a statistically significant effect ($P < 0.05$) with the same direction as the original study. This dichotomous vote-counting method is intuitively appealing and consistent with common heuristics used to decide whether original studies “worked.” Ninety-seven of 100 (97%) effects from original studies were positive results (four had $P$ values falling a bit short of the 0.05 criterion—$P = 0.0508, 0.0514, 0.0516, 0.0567$—but all of these were interpreted as positive effects). On the basis of only the average replication power of the 97 original, significant effects ($M = 0.92$, median $(Mdn) = 0.95$), we would expect approximately 89 positive results in the replications if all original effects were true and accurately estimated; however, there were just 35 (36.1%; 95% CI = (26.6%, 46.2%)), a significant reduction [McNemar test, $\chi^2(1) = 59.1, P < 0.001$].

A key weakness of this method is that it treats the 0.05 threshold as a bright-line criterion between replication success and failure (28). It could be that many of the replications fell just short of the 0.05 criterion. The density plots of $P$ values for original studies (mean $P$ value = 0.028) and replications (mean $P$ value = 0.302) are shown in Fig. 1, left. The 64 nonsignificant $P$ values for replications were distributed widely. When there is no effect to detect, the null distribution of $P$ values is uniform. This distribution deviated slightly from uniform with positive skew, however, suggesting that at least one replication could be a false negative, $\chi^2(128) = 155.83, P = 0.048$. Nonetheless, the wide distribution of $P$ values suggests against insufficient power as the only explanation for failures to replicate. A scatterplot of original compared with replication study $P$ values is shown in Fig. 2.

**Evaluating replication effect against original effect size**

A complementary method for evaluating replication is to test whether the original effect size is within the 95% CI of the effect size estimate from the replication. For the subset of 73 studies in which the standard error of the correlation could be computed, 30 (41.1%) of the replication CIs contained the original effect size (significantly lower than the expected value of 78.5%, $P < 0.001$) (supplementary materials). For 22 studies using other test statistics ($F(df_1 > 1, df_2)$ and $\chi^2$), 68.2% of CIs contained the effect size of the original study. Overall, this analysis suggests a 47.4% replication success rate.

This method addresses the weakness of the first test that a replication in the same direction and a $P$ value of 0.06 may not be significantly different from the original result. However, the method will also indicate that a replication “fails” when the direction of the effect is the same but the replication effect size is significantly smaller than the original effect size (29). Also, the replication “succeeds” when the result is near zero but not estimated with sufficiently high precision to be distinguished from the original effect size.

**Comparing original and replication effect sizes**

Comparing the magnitude of the original and replication effect sizes avoids special emphasis on $P$ values. Overall, original study effect sizes

---

**Fig. 1. Density plots of original and replication $P$ values and effect sizes.** (A) $P$ values. (B) Effect sizes (correlation coefficients). Lowest quantiles for $P$ values are not visible because they are clustered near zero.
A scatterplot of the original study and replication P values for three psychology journals. Data points scaled by power of the replication based on original study effect size. Dotted red lines indicate P = 0.05 criterion. Subplot below shows P values from the range between the gray lines (P = 0 to 0.005) in the main plot above.

Fig. 2. Scatterplots of original study and replication P values for three psychology journals. Data points scaled by power of the replication based on original study effect size. Dotted red lines indicate P = 0.05 criterion. Subplot below shows P values from the range between the gray lines (P = 0 to 0.005) in the main plot above.

One qualification about this result is the possibility that the original studies have inflated effect sizes due to publication, selection, reporting, or other biases (9,12–23). In a discipline with low-powered research designs and an emphasis on positive results for publication, effect sizes will be systematically overestimated in the published literature. There is no publication bias in the replication studies because all results are reported. Also, there are no selection or reporting biases because all were confirmatory tests based on pre-analysis plans. This maximizes the interpretability of the replication P values and effect estimates. If publication, selection, and reporting biases completely explain the effect differences, then the replication estimates would be a better estimate of the effect size than would the meta-analytic and original results. However, to the extent that there are other influences, such as moderation by sample, setting, or quality of replication, the relative bias influencing original and replication effect size estimation is unknown.

Subjective assessment of “Did it replicate?”

In addition to the quantitative assessments of replication and effect estimation, replication teams provided a subjective assessment of replication success of the study they conducted. Subjective assessments of replication success were very similar to significance testing results (39 of 100 successful replications), including evaluating “success” for two null replications when the original study reported a null result and “failure” for a P < 0.05 replication when the original result was a null.

Correlates of reproducibility

The overall replication evidence is summarized in Table 1 across the criteria described above and then separately by journal/discipline. Considering significance testing, reproducibility was stronger in studies and journals representing cognitive psychology than social psychology topics. For example, combining across journals, 14 of 55 (25%) of social psychology effects replicated by the P < 0.05 criterion, whereas 21 of 42 (50%) of cognitive psychology effects did so. Simultaneously, all journals and disciplines showed substantial and similar χ²(3) = 2.45, P = 0.48 declines in effect size in the replications compared with the original studies. The difference in significance testing results between fields appears to be partly a function of weaker original effects in social psychology studies, particularly in JPSP, and perhaps of the greater frequency of high-powered within-subjects manipulations and repeated measurement designs in cognitive psychology as suggested by high power despite relatively small participant samples. Further, the type of test was associated with replication success. Among original, significant effects, 23 of the 49 (47%) that tested main or simple effects replicated at P < 0.05, but just 8 of the 37 (22%) that tested interaction effects did.

Correlations between reproducibility indicators and characteristics of replication and original studies are provided in Table 2. A negative correlation of replication success with the original study P value indicates that the initial strength of evidence is predictive of reproducibility. For example, 26 of 63 (41%) original studies with P < 0.02 achieved P < 0.05 in the replication, whereas 6 of 23 (26%) that had a P value between 0.02 < P < 0.04 and 2 of 11 (18%) that had a P value > 0.04 did so (Fig. 2). Almost two thirds (20 of 32, 63%) of original studies with P < 0.001 had a significant P value in the replication.

Larger original effect sizes were associated with greater likelihood of achieving P < 0.05 (r = 0.304) and a greater effect size difference between original and replication (r = 0.279). Moreover, replication power was related to replication success via significance testing (r = 0.368) but not with the effect size difference between original and replication (r = 0.053). Comparing effect sizes across indicators, surprisingsness of the original effect, and the challenge of conducting the replication were related to replication success for some indicators. Surprising effects were less reproducible, as were effects for which it was more challenging to conduct the replication. Last, there was little evidence that perceived importance of the effect, expertise of the original or replication teams, or self-assessed quality of the replication accounted for meaningful variation.
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).

**Discussion**

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 replications produced weaker evidence for the original findings (31) despite using materials provided by the original authors, review in advance for methodological fidelity, and high statistical power to detect the original effect sizes. Moreover, correlational evidence is consistent with the conclusion that variation in the strength of initial evidence (such as original $P$ value) was more predictive of replication success than was variation in the characteristics of the teams conducting the research (such as experience and expertise). The latter factors certainly can influence replication success, but the evidence is that they did not systematically do so here. Other investigators may develop alternative indicators to explore further the role of expertise and quality in reproducibility on this open data set.

**Insights on reproducibility**

It is too easy to conclude that successful replication means that the theoretical understanding of the original finding is correct. Direct replication mainly provides evidence for the reliability of a result. If there are alternative explanations for the original finding, those alternatives could likewise account for the replication. Understanding is achieved through multiple, diverse investigations that provide converging support for a theoretical interpretation and rule out alternative explanations.

It is also too easy to conclude that a failure to replicate a result means that the original evidence was a false positive. Replications can fail if the replication methodology differs from the original in ways that interfere with observing the effect. We conducted replications designed to minimize a priori reasons to expect a different result by using original materials, engaging original authors for review of the designs, and conducting internal reviews. Nonetheless, unanticipated factors in the sample, setting, or procedure could still have altered the observed effect magnitudes (32).

More generally, there are indications of cultural practices in scientific communication that may be responsible for the observed results. Low-power research designs combined with publication bias favoring positive results together produce a literature with upwardly biased effect sizes (14, 16, 33, 34). This anticipates that replication effect sizes would be smaller than original studies on a routine basis—not because of differences in implementation but because the original study effect sizes are affected by publication and reporting bias, and the replications are not. Consistent with this expectation, most replication effects were smaller than original results, and reproducibility success was correlated with indicators of the strength of initial evidence, such as lower original $P$ values and larger effect sizes. This suggests publication, selection, and reporting biases as plausible explanations for the difference between original and replication effects. The replication studies significantly reduced these biases because replication preregistration and pre-analysis plans ensured confirmatory tests and reporting of all results.

The observed variation in replication and original results may reduce certainty about the statistical inferences from the original studies but also provides an opportunity for theoretical innovation to explain differing outcomes, and then new research to test those hypothesized explanations. The correlational evidence, for example, suggests that procedures that are more challenging to execute may result in less reproducible results, and that more surprising original effects may be less reproducible than less surprising original effects. Further, systematic, repeated replication efforts that fail to identify conditions under which the original finding can be observed reliably may reduce confidence in the original finding.

**Implications and limitations**

The present study provides the first open, systematic evidence of reproducibility from a sample of studies in psychology. We sought to maximize generalizability of the results with a structured process for selecting studies for replication. However, it is unknown the extent to which these findings extend to the rest of psychology or other disciplines. In the sampling frame itself, not all articles were replicated; in each article, only one study was replicated; and in each study, only one statistical result was subject to replication. More resource-intensive studies were less likely to be included than were less resource-intensive studies. Although study selection bias was reduced by the sampling frame and selection strategy, the impact of selection bias is unknown.

We investigated the reproducibility rate of psychology not because there is something special about psychology, but because it is our discipline. Concerns about reproducibility are widespread across disciplines (9–21). Reproducibility is not well understood because the incentives for individual scientists prioritize novelty over replication (20). If nothing else, this project demonstrates that it is possible to conduct a large-scale examination of reproducibility despite the incentive barriers. Here, we conducted single-replication attempts of many effects obtaining broad-and-shallow evidence. These data provide information about reproducibility in general but little precision about individual effects in particular. A complementary narrow-and-deep approach is characterized by the Many Labs replication projects (32). In those, many replications of single effects allow precise estimates of effect size but result in generalizability that is circumscribed to those individual effects. Pursuing both strategies across disciplines, such as the ongoing effort in cancer biology (35), would yield insight about common and distinct

---

**Fig. 3. Original study effect size versus replication effect size (correlation coefficients).** Diagonal line represents replication effect size equal to original effect size. Dotted line represents replication effect size of 0. Points below the dotted line were effects in the opposite direction of the original. Density plots are separated by significant (blue) and nonsignificant (red) effects.

<table>
<thead>
<tr>
<th>$p$-value</th>
<th>Not Significant</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
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 progress accumulates empirical evidence. In some cases, additional, confirmatory evidence. In some cases, single study almost never provides definitive resolution. As much as we might wish it to be otherwise, a crisis of confidence? Innovation points out paths that are likely; progress relies on both. The ideal balance is a topic for investigation itself. Scientific progress accumulates empirical evidence. In some cases, additional, confirmatory evidence. In some cases, single study almost never provides definitive resolution. As much as we might wish it to be otherwise, a crisis of confidence?

We conducted this project because we care deeply about the health of our discipline and believe in its promise for accumulating knowledge about human behavior that can advance the quality of the human condition. Reproducibility is central to that aim. Accumulating evidence is the scientific community’s method of self-correction and is the best available option for achieving that ultimate goal: truth.

REFERENCES AND NOTES


9. J. P. Ioannidis, What have we learned from published research findings are false. PLOS Med. 2, e124 (2005). doi:10.1371/journal.pmed.0020124; pmid:1606772


ACKNOWLEDGMENTS
