PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

A discipline-wide investigation of the replicability of Psychology papers over the past two decades

Wu Youyou[a], Yang Yang[b], and Brian Uzzi[c,d,1]

Edited by Boleslaw K. Szymanski, Rensselaer Polytechnic Institute, Troy, NY; received May 23, 2022; accepted December 14, 2022 by Editorial Board Member Kenneth W. Wachter

Conjecture about the weak replicability in social sciences has made scholars eager to quantify the scale and scope of replication failure for a discipline. Yet small-scale manual replication methods alone are ill-suited to deal with this big data problem. Here, we conduct a discipline-wide replication census in science. Our sample (N = 14,126 papers) covers nearly all papers published in the six top-tier Psychology journals over the past 20 y. Using a validated machine learning model that estimates a paper's likelihood of replication, we found evidence that both supports and refutes speculations drawn from a relatively small sample of manual replications. First, we find that a single overall replication rate of Psychology poorly captures the varying degree of replicability among subfields. Second, we find that replication rates are strongly correlated with research methods in all subfields. Experiments replicate at a significantly lower rate than do non-experimental studies. Third, we find that authors' cumulative publication number and citation impact are positively related to the likelihood of replication, while other proxies of research quality and rigor, such as an author's university prestige and a paper's citations, are unrelated to replicability. Finally, contrary to the ideal that media attention should cover replicable research, we find that media attention is positively related to the likelihood of replication failure. Our assessments of the scale and scope of replicability are important next steps toward broadly resolving issues of replicability.

science of science | replication | machine learning | psychology

Replicability of research fortifies scientific predictions and strategies for improving living standards; it is also testimony for science being self-correcting. Carl Popper (1) concluded that replication in science ensures that "we are not dealing with a mere isolated "coincidence," but with events which, on account of their regularity and reproducibility, are in principle inter-subjectively testable." A turning point in testing for replication took place in 2011, when a controversial study on "time-reserved causality" (2) prompted a rare replication study. The replication failed (3, 4), leading to more replications and the discovery that replication failures are more than single incidents (5–10). Researchers were concerned about the implication of systematic replication failures, which include a weakened knowledge base, higher public distrust in science, and funding cuts (11–13). A poll of 1,500 scientists conducted by Nature in 2016 reported that 51% of respondents agreed that science is experiencing a replication crisis (14). This response compelled the United States Defense Advanced Research Projects Agency to create a program in 2018 for studying the scale and scope of replication failure in social science (15–17).

Despite growing concerns about replication failures, the sample of manual replication studies is small in number and a limited fraction of the total literature (18, 19). In Psychology—the scientific discipline that has conducted the most replication studies—the total number of direct, independent replications is less than 400. Moreover, the sample disproportionately represents classic papers by selected authors or specific subfields (20, 21). Most replications come from the subfields of Social Psychology and Cognitive Psychology, leading to speculation that Developmental Psychology, Clinical Psychology, and Education Psychology have similar rates of replication failure despite a lack of subfield-specific analyses (22, 23).

To expand and diversify replication data, researchers developed alternative methods to estimate a paper's likelihood of replication success (24). A prediction market has become a major approach to estimating a paper's replicability. It involves having experts wager whether a published paper will successfully replicate in a future manual replication test (25). The method's high accuracy has made prediction markets an effective solution for estimating a paper's replicability (16). Nonetheless, while prediction markets scale better than manual replications, they still require the recruitment of thousands of expert reviewers over many years to predict replicability for a large sample of papers (17).

Significance

The number of manually replicated studies falls well below the abundance of important studies that the scientific community would like to see replicated. We created a text-based machine learning model to estimate the replication likelihood for more than 14,000 published articles in six subfields of Psychology since 2000. Additionally, we investigated how replicability varies with respect to different research methods, authors 'productivity, citation impact, and institutional prestige, and a paper's citation growth and social media coverage. Our findings help establish large-scale empirical patterns on which to prioritize manual replications and advance replication research.

Author affiliations: [a]Department of Psychology and Human Development, Institute of Education, University College London, London WC1H 0AL; [b]Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556; [c]Kellogg School of Management, Northwestern University, Evanston, IL 60201; and [d]Northwestern Institute of Complex Systems, Northwestern University, Evanston, IL 60201

Author contributions: W.Y., Y.Y., and B.U. designed research; W.Y. and Y.Y. performed research; W.Y. and Y.Y. analyzed data; and W.Y., Y.Y., and B.U. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. B.K.S. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: uzzi@northwestern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2208863120/-/DCSupplemental.

Published January 30, 2023.

Table 1. Replicability prediction sample

Top-tier Psychology journal	Papers	Journal impact factor	Acceptance rate (%)
Journal of Abnormal Psychology	1,611	9.0	17
Journal of Experimental Psychology, Learning, Memory & Cognition	2,366	3.0	25
Child Development	2,677	5. 9	17
Journal of Applied Psychology	1,792	7.4	8
Journal of Personality and Social Psychology	2,611	7. 7	15
Psychological Science	3,069	7.0	6.3
	Total = 14,126	Weighted Mean = 6.53	Weighted Mean = 15%

The table lists the sample of papers used in the analysis to estimate the replicability of the Psychology literature. The sample of 14,126 papers are authored by a total of 26,349 distinct scholars with affiliations at 6,173 institutions. These papers have amassed 1,222,292 total citations and 27,447 total media mentions. The number of papers listed in each row constitutes the articles published in the corresponding journals over the time-period 2000 to 2019.

Machine learning methods have also been developed to predict replication outcomes. Machine learning models can predict replicability either from a study's narrative text (26) or from numerical characteristics, such as P values or sample sizes of the study (27, 28). Both types of models make accurate predictions that are on par with prediction markets (26). Text-based models quantify the narrative in a paper, including the description of a study's design and the interpretation of results (29), which are not captured in models based on solely on numerical characteristics. In addition, text quantification can be automated, thereby making the technique more scalable and reproducible than manually extracting numerical characteristics from manuscripts.

Here, we employ a text-based machine learning method to predict the likelihood of replication success for the Psychology literature. Our sample of Psychology papers covers nearly all of the papers published over a 20-y period in top-tier Psychology journals in six major subfields: Clinical Psychology, Cognitive Psychology, Developmental Psychology, Organizational Psychology, Personality Psychology, and Social Psychology. In total, the sample includes 14,126 papers by 26,349 distinct authors from 6,173 distinct institutions, with 1,222,292 total citations and 27,447 total media mentions.

Our analysis proceeds as follows: We first briefly describe our text-based machine learning model, which was previously validated and shown to accurately predict manual replication outcomes (26). We then apply the model to predict the replicability of the Psychology literature, with an eye to investigate how replicability varies across Psychology subfields, research methods, pre- and post-publication characteristics of the paper, as well as the expertise and experience of the authorship team.

Data and Methods

Our analysis uses a diverse mix of bibliographic, author, and media coverage data sources. Data and code for generating the data have been deposited into Open Science Framework (30). Table 1 lists the large sample of journal publications used in the analysis, including five journals of specialized subfields and a general journal Psychological Science.

All papers were published between 2000 to 2019 and classified into a subfield based on the journal's subfield specialization: 1) Journal of Abnormal Psychology (Clinical Psychology), 2) Journal of Experimental Psychology, Learning, Memory & Cognition (Cognitive Psychology), 3) Child Development (Developmental Psychology), 4) Journal of Applied Psychology (Organizational Psychology) 5) Journal of Personality and Social Psychology (Social Psychology). There are two exceptions to the classification rule above. First, because personality research appears in all top-tier journals, we labeled articles as personality research if the word "personality" was in the title or abstract, irrespective of the journal in which they appear. Second, because Psychological Science publishes work from all subfields, we classified its papers based on which subfield specialty journal the authors of a paper mainly published in. The total sample includes 14,126 papers. All data were collected in accordance with publishers' terms of use and UK copyright law.

Machine Learning Model. Our machine learning model used an ensemble of random forest and logistic regression models to predict a paper's likelihood of replication based on the paper's text. The model was validated in a prior publication using stringent out-of-sample tests and was shown to have an accuracy that is on par with prediction markets (26). Specifically, the procedures for creating the model are as follows (see *SI Appendix*, Fig. S2 for a diagram illustrating the procedures): Step 1, converting individual English words into vectors. We trained a model using word2vec (31) on a corpus of 2 million social science publication abstracts published between 2000 and 2017 from the Microsoft Academic Graph (MAG) (32). The goal was to associate individual words with one another in the context of social science literature and to represent that association quantitatively in a 200-dimension vector. Step 2, converting publications into vectors. To do this, we multiplied the normalized frequency of each word in each paper in the training sample (Table 2) by its corresponding 200-dimension word vector, which produced a paper-level vector representing the textual content of the paper. Step 3, predicting each paper's replication outcome (pass/fail) from its paper-level vector using an ensemble of random forests and logistic regressions. To determine whether a study replicated or not, we used a common metric reported in all replication studies—the replication team's summary judgment of whether the study replicated or did not replicate ("yes" or "no"). Together, steps 1 to 3 created a machine learning model that uses a paper's text/narrative to predict its likelihood of replication, which we call the "replication score." *SI Appendix*, *Supplementary Text 3* provides details on all procedures.

Performance and Robustness Tests for the Machine Learning Model. We undertook a series of tests to evaluate the model's performance and robustness. First, a threefold cross-validation was employed to avoid over-fitting in the training set. The average Area Under the ROC Curve (AUC) of the threefold cross-validation was 0.74.

Second, we also evaluated the effect of imbalance in the training sample's composition of experimental vs. non-experimental research (81% and 19%, respectively). We manually coded each study's research method and calculated the model's performance

Table 2. Manual replication studies in Psychology used to train the replicability prediction model

#	Project/Platform	Psychology subfields	No. of studies	No. of successful replications
Pre-trained	MAG paper abstracts	Social Science	2 million	N/A
1	RPP (10)	Cognitive, Social	96	37
2	RRR (33)	Cognitive, Social	8	1
3 to 6	ML1-4 (8,34–36)	Cognitive, Social, Personality, Organizational	42	22
7	JSP (37)	Social, Organizational	16	5
8	SSRP (38)	Cognitive, Social	18	10
9	LOOPR (39)	Personality	22	20
10	CORE (40)	Social, Organizational	39	30
11	Curate Science (41)	Cognitive, Social, Personality, Organizational	93	18
12	PFD (42)	Cognitive, Social, Organizational	54	25
			Total = 388	Overall success rate = 43.3%

The table lists the manual replication studies used to train the machine learning model to predict a paper's estimated replicability based on the text in the manuscript. A total of 388 available manual replication studies in Psychology reported pass/fail replication outcomes. Column 2 lists the abbreviated names of coordinated replication projects that conducted the replications or platforms that curated the projects (see *SI Appendix, Supplementary Text* 1 for full descriptions); column 3 lists the subfields of Psychology covered in each project/platform; column 4 lists the number of studies in each project/platform; column 5 counts the number of successfully replicated studies. To determine whether a study was successfully replicated or not, we used a common metric reported in all replication studies—the replication team's summary judgement of whether the study replicated or did not replicate ("yes" or "no"). Note that for some replication projects, the number of studies included here might differ from the original number (e.g., RPP conducted 100 studies while only 96 are included here). See *SI Appendix, Supplementary Text* 1 for the exclusion criteria for each project. ML: Many Labs; RPP: The Reproducibility Project: Psychology; RRR: The Registered Replication Report; JSP: Journal Social Psychology; SSRP: The Social Science Reproduction Platform; LOOPR: The Life Outcomes of Personality Replication; CORE: Collaborative Open-science REsearch; PFD: Psych File Drawer.

separately for experimental (n = 314, AUC = 0.74) vs. non-experimental research (n = 72, AUC = 0.69). The smaller sample of non-experimental studies shows a performance difference, but the performance level is still acceptable for subsequent analyses in this paper.

Third, we assessed issues related to transfer learning. Transfer learning occurs when models are developed in one domain and applied to another (43). The practice arises in our study because the prediction sample contains papers from two subfields that are not present in the training sample, Clinical Psychology and Developmental Psychology. Manual replications from these two subfields are scarce, and it could take another decade to accumulate a sizeable sample (44). This raises a concern about whether the model can provide valid estimates for papers in Clinical Psychology and Developmental Psychology. To address this concern, we followed protocols and conducted three separate robustness tests (43, 45, 46).

(i) We used existing data on Social and Cognitive Psychology to simulate the transfer-learning process and estimate the performance of using a model trained on the manual replications of one psychology subfield to predict the replication failure of another psychology subfield, and to compare the model's prediction to actual manual replication data in the predicted subfield. Specifically, we examined how a model developed solely based on papers from Social Psychology (n = 256)—the main subfield in the training sample—would perform on papers from Cognitive Psychology (n = 90). We found that the performance of such transfer learning to Cognitive Psychology (AUC = 0.72) is comparable to when the model was applied to Social Psychology (benchmark AUC = 0.73). This provides support for transfer learning success between subfields in Psychology.

(ii) One might argue that the text model's successful transfer from Social to Cognitive Psychology does not guarantee its successful transfer to Clinical Psychology or Developmental Psychology. To answer this question, we compared the subfields for their topic and textual similarity. Prior machine learning, research has shown that transfer learning in text-based models is more successful when the textual features in the training and application domains are more alike (47). Therefore, if Social–Clinical and Social–Developmental similarities are comparable to or higher than Social-Cognitive similarity, we could then expect the model to be as valid in Clinical or Development as in Cognitive Psychology.

To measure the overlap in research topics between two subfields, we collected research topics for each paper in the testing sample from MAG database. To measure textual similarity between two subfields, we calculated cosine similarity and word mover's distance (WMD). *SI Appendix, Supplementary Text* 3.4.1 describes the methods in detail.

Results show that Clinical (57%) and Developmental papers (56%) overlap in higher percentages of topics with Social papers than Cognitive papers do (42%). In addition, all three subfields display equal levels of textual similarities with Social Psychology (cosine similarity = 0.90 to 0.91, WMD = 0.24 to 0.26). Because analysis (i) shows that a model built on Social Psychology is transferable to Cognitive Psychology, we can now expect the model to transfer to Clinical Psychology and Developmental Psychology where higher feature similarity with Social Psychology is observed.

(iii) We assessed how the predicted replication scores align with alternative indicators of replicability like sample sizes and P values in Clinical Psychology or Developmental Psychology papers. Both metrics are indicators of reliability because the risk of false positives decreases with larger sample sizes and lower P values (5, 48, 49). We stress that the prediction model contains no information about sample sizes and P values because papers in the training sample were stripped of all numbers or statistics. Thus, if the sample size and P value of a paper correlate with our model's replication predictions, it would provide independent support for the model's applicability in Clinical Psychology or Developmental Psychology.

Procedurally, we manually coded a random subset of studies in Clinical Psychology and Developmental Psychology from the prediction. To obtain sample sizes, we extracted the number of participants from the paper. If a paper has multiple studies, we took the average sample sizes of all studies in the paper. To obtain

P values, we located the first main claim of a paper from its abstract and extract the *P* value of the test associated with that main claim. The main claim is usually proceeded by phrases like “The results show that” or “Our analyses suggest that.” *SI Appendix*, *Supplementary Text* 3.4.2 provides further methodological details.

The results show that the predicted replication score correlates in rank order with both the original sample size $r(97) = 0.31$, $P = 0.002$, and the original *P* value $r(91) = -0.42$, $P < 0.001$. Since the prediction model contains no sample size and *P* value information, the results are therefore not tautological and add support for successful transfer learning to Clinical Psychology and Developmental Psychology.

Measures of Pre- and Post-publication Correlates of Replicability. To examine the link between replication likelihoods and a paper's other observable publication features, we constructed several key measures of observable features of a paper discussed in the replication literature. For example, replication outcomes have been hypothesized to be related to researchers' expertise (34) or a paper's media attention (50). We collected five measures that capture a paper's characteristics, three pre-publication measures that capture characteristics of the authorship team, and two post-publication measures that capture the dynamics of readers' reactions to the research. Pre-publication characteristics include the paper's first and senior authors' experience and competence, measured as their 1) cumulative number of publications, 2) citation impact prior to the publication of the focal paper, and 3) institutional prestige based on the rank of the first and senior authors' universities in 2021 QS World University Rankings (51). A senior author is defined as the author on the research team with the most cumulative citations when the focal paper was published. Post-publication characteristics include the focal paper's 4) citation count and 5) media mentions. Media mentions were computed by Altmetric (52). All other measures were taken from Dimensions (53), which approved our use of the data for this project. To control for publication age and subfield differences in these metrics (*SI Appendix*, Fig. S1), we normalized all metrics by dividing the observed score by the average in its subfield and publication year. *SI Appendix*, *Supplementary Text* 2 presents more details about the metrics and how they were normalized.

Results

Using the calibrated machine learning model described above, we predicted a replication score for each paper in the replicability prediction sample (n = 14,126). The score can be interpreted as the relative likelihood of replication success. In other words, a paper with a replication score of 0.80 is more likely to replicate than papers with lower replication scores and is twice as likely to replicate as a paper with a replication score of 0.40. Using the replication scores, we conducted three sets of analyses: First, we determined subfield differences in estimated replication rates, bridging the gaps in previous small-sample manual replications; second, we compared replication rates between experimental and non-experimental research designs; and third, we examined how replicability correlates with other pre- and post-publication characteristics of a paper.

Fig. 1 shows the predicted replication score distribution for all 14,126 Psychology papers (range = 0.10 to 0.86, mean = 0.42, median = 0.41, *SD* = 0.15, skewness = 0.31). Several findings are noteworthy. First, the distribution is broadly consistent with speculations from manual replications and the latest forecasts from prediction markets (15). Manual replications suggest that slightly more Psychology papers would fail rather than pass manual replication tests (43% overall success rate). The estimated distribution of the replication scores for the last 20 y of Psychology publications suggests a similar pattern. Second, it has been argued that the recent attention to replication failure in Psychology has improved replication rigor (54). When we plotted the mean replication scores over our 20-y period, we found that the replication scores are relatively stable. The average replication scores decreased by approximately 10% from 2000 and 2010 and then from 2010 and 2019 increased back to roughly the same level as the year 2000 (*SI Appendix*, Fig. S6)—a pattern that aligns with the observation that changes in research practice have potentially improved replication rates in Psychology (9, 21, 55). Third, we found that pooling replication scores across subfields of Psychology obscures important subfield differences. Below, we detail how replication rates differ across subfields in Psychology.

Comparative Replication Scores by Subfield. To address key questions of replicability variation among subfields (22, 56), we parsed the overall distribution by subfield. Previous manual replications provide some expectations, but only for three subfields (with $n \geq 30$): Personality Psychology (77% success rate, n = 30), Cognitive Psychology (50% success rate, n = 90), and Social Psychology (38% success rate, n = 256).

Fig. 2*A* shows the distribution of replication scores grouped by Psychology's six major subfields. All distributions are normal (abs(skewness) < 0.50), except for Developmental Psychology, which is slightly right-skewed (skewness = 0.62). We found that the replicability rates reported by manual replications echoed the estimated

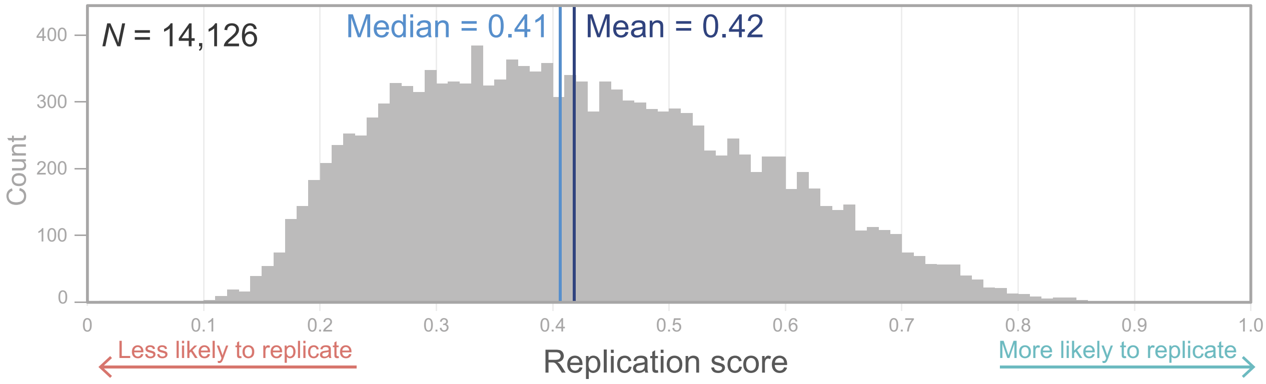

Fig. 1. Predicted Replication Scores of Psychology Literature, 2000 to 2019. Machine learning model prediction of replication likelihood for 14,126 papers published in the highest rated journals in the Psychology subfields of Developmental Psychology, Social Psychology, Clinical Psychology, Cognitive Psychology, Organizational Psychology, and Personality Psychology.

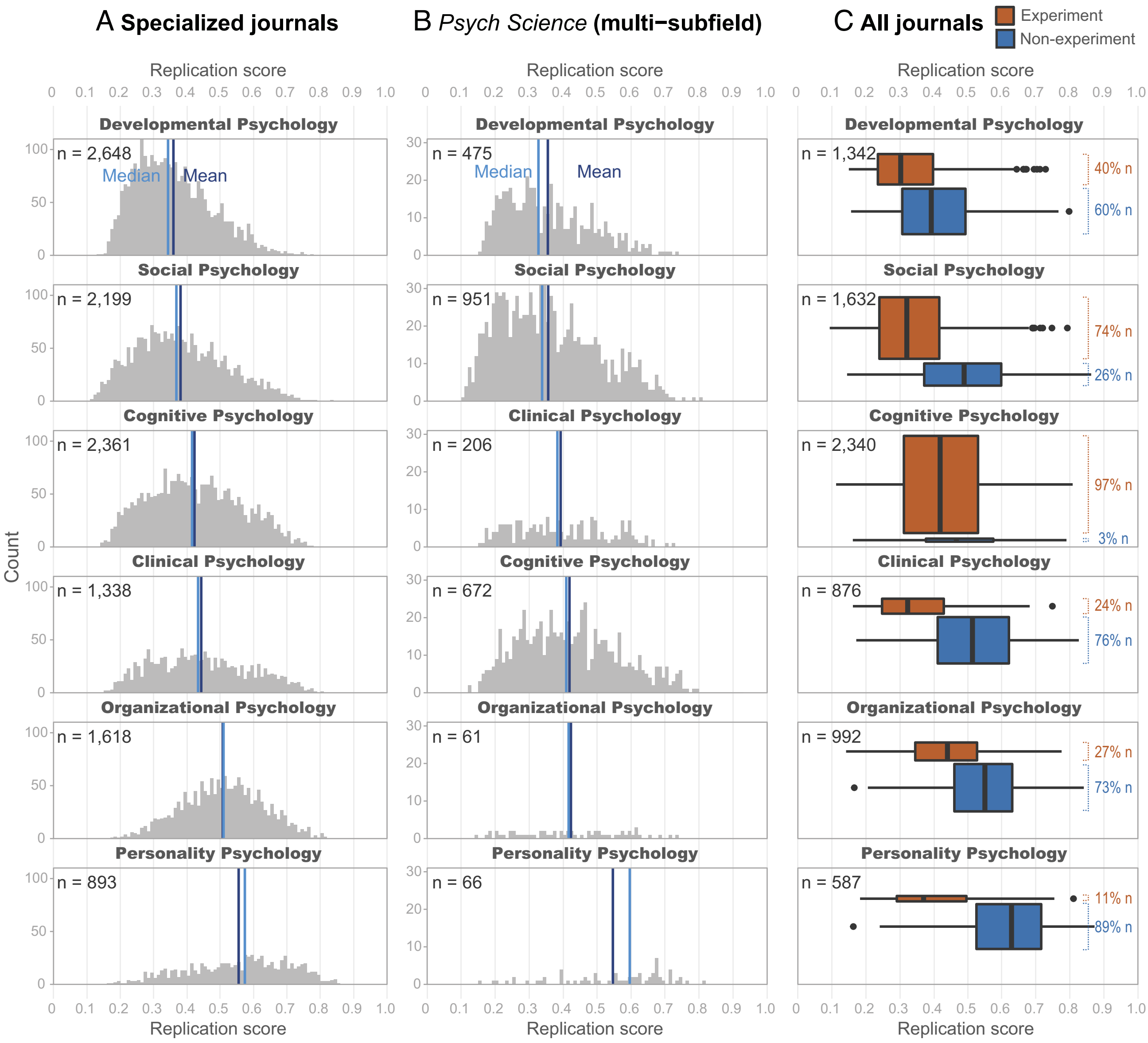

Fig. 2. Comparing Replicability for Six Psychology Subfields and Between Experimental and Non-experimental Research. Panel *A* shows the average replicability estimated for papers published in specialized journals, categorized into six subfields. The light blue vertical line represents the median for each subfield, and the dark blue line is the mean. Panel *B* also illustrates predicted replication scores, but the papers are all published in a single multi-subfield journal, Psychological Science. The subfield replicability rankings were largely consistent with the ones in specialized journals, except that the order of Cognitive and Clinical Psychology was reversed. To explain the subfield patterns, Panel *C* further breaks down the average replication scores by research methods for each subfield, comparing replicability between experimental (orange boxes) and non-experimental research (blue boxes). The proportion of experimental vs. non-experimental research in each subfield is marked as *k*% of total papers (e.g., 40% of Developmental Psychology papers are experimental). Experimental research on average has lower replication scores, and the proportion of experimental research partially explains the subfield differences in average replicability.

replication scores produced by our model. Personality Psychology had the highest estimated replication score (Mean = 0.55) followed by Organizational Psychology (Mean = 0.50). Cognitive Psychology (Mean = 0.42) had a higher score than Social Psychology (Mean = 0.37). The subfields of Development Psychology and Clinical Psychology, which have received relatively scant attention in manual replication studies, have means of 0.36 and 0.44, respectively.

To rule out the possibility that the above pattern reflects journal differences rather than subfield differences, we repeated the analysis using a single journal that publishes multiple subfields, Psychological Science. We assigned the Psychological Science papers into subfields according to which specialized journals the authors tend to publish in. For instance, if the author published mainly in Journal of Applied Psychology, the author's work in Psychological Science is then categorized as Organizational Psychology. If a paper has multiple authors, we choose the most common subfield across the authors. Using this approach, 2,431 papers were successfully classified into six subfields.

Fig. 2*B* visualizes the subfield differences in replicability for Psychological Science. The patterns largely mirror those observed previously in specialized journals (Fig. 2*A*). The only exception is that Cognitive Psychology's average replicability is slightly lower than Clinical Psychology's (t = −4.18, P < 0.001) in specialized journals, but higher in Psychological Science (t = 2.34, P = 0.02). These findings provide evidence for the claim that replication rates can vary widely by subfield within a particular discipline. Therefore, characterizations of replication rates should be made with respect to a subfield rather than to an entire discipline. This finding can also help identify possible determinants of replication failure and research improvement strategies.

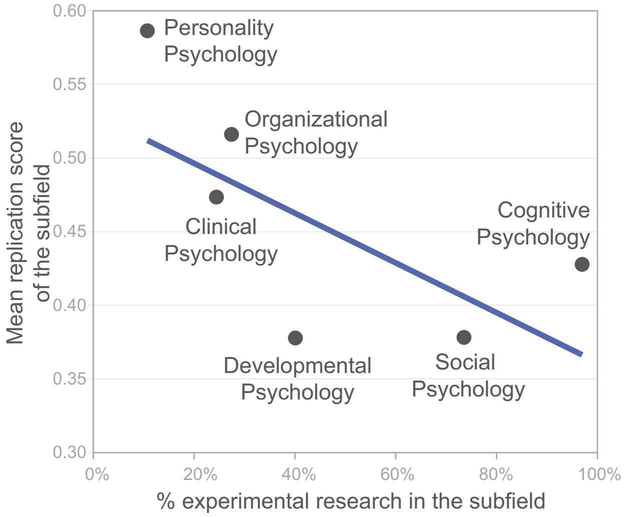

Fig. 3. Percentage of Experimental Research in Each Psychology Subfield and the Subfield's Mean Replication Score. Subfields with larger proportions of non-experiments (Personality Psychology, Organizational Psychology and Clinical Psychology) have a propensity for higher average replication scores. An exception is Developmental Psychology. It has the lowest average replicability and is mostly non-experimental. The discrepancy may be accounted for by the tendency of Developmental Psychology to study participants over their lifespans, from infancy to adulthood, which presents unique data collection challenges (57).

Comparative Replication Scores by Method. A key issue related to the possible sources of subfield variation in replication scores is the research methods used in the papers. We split the papers in the prediction sample into two groups: papers using experimental methods and papers using non-experimental methods. A paper is considered experimental if it has the word "experiment" in its titles, abstracts, or section headings. A paper is non-experimental if the word "experiment" is not present in any part of the paper. Using this method, we successfully categorized 8,159 papers in the prediction samples. We then computed average replication scores for each group.

Fig. 2*C* shows the estimated average replication scores for experimental vs. non-experimental papers by subfield. We report three results. First, experimental and non-experimental papers systematically differ in replicability. The mean replication score of non-experimental papers is significantly higher than the mean of experimental studies. Non-experimental papers have an overall mean replication score of 0.50, whereas experimental papers have a mean replication score of 0.39 (t = 34.22, P < 0.001, Cohen's d = 0.78). The difference is confirmed in the training sample as well, where the replication success rate is 69% for non-experimental and 37% for experimental studies respectively. Second, the difference in replicability between experimental vs. non-experimental studies generalizes across our six subfields. Within each subfield, non-experimental papers have a significantly greater average replication score than experimental studies in all six subfields (for Cognitive Psychology, t = 2.64, P = 0.01, Cohen's d = 0.32, all other P < 0.001 and Cohen's d = [0.65, 1.56]). Third, Fig. 3 shows that subfields with smaller proportions of experiments (Personality Psychology and Organizational Psychology) have a propensity for higher average replication scores. A notable exception is Developmental Psychology, which has the lowest average replicability though 60% of its research is non-experimental. One explanation for this pattern is that Developmental Psychology focuses on children and life courses, two areas in which researchers face unique difficulties in collecting large samples under controllable circumstances (57,58).

Pre- and Post-publication Correlates of Replicability. We examined the relationship between replicability and other characteristics of a paper. Three of these characteristics occur prior to the publication (the authors' cumulative publication number, citation impact, and institutional prestige) and two occur after the publication (the focal paper's citation impact and media coverage). Our analysis compares these metrics using the Mann-Whitney rank-sum test between 1) studies that passed vs. failed manual replication in the training sample, and 2) papers that are likely vs. unlikely to replicate in the prediction sample. Papers in the top 10% of replication scores were defined papers as likely to replication and papers in the bottom 10% of replication scores were defined as papers unlikely to replicate. We opted to focus on comparing the bottom vs. top 10% of the prediction data because the machine learning model is most accurate at these points in its distribution (*SI Appendix, Supplementary Text* 3.2.2b). This maximizes accuracy in the replicability estimates and reduces noise in the subsequent analyses with other publication metrics. Moreover, additional robustness checks presented in *SI Appendix, Supplementary Text* 4.1 and Fig. S5 confirm that the results hold over 5%, 15%, or 20% cutoffs of replication scores.

Our analysis begins with "researcher competence," which has been hypothesized to correlate with replication failures (14, 59, 60). Fig. 4 *A* and *B* indicate a statistically significant relationship between researcher competence and replication success in both training and prediction samples. In Fig. 4*A*, the training sample displays a significant and positive association between authors' cumulative number of publications and replication success for the first author (bi-serial correlation r = 0.17, P = 0.008), but not the senior author (P = 0.38). A senior author is defined as the author on the research team with most cumulative number of citations when the focal paper was published. In a similar vein, the training sample (Fig. 4*B*) shows that authors' cumulative citation impact is significantly and positively related to replication success for the first author (r = 0.19, P = 0.004), but not for the senior author (P = 0.77). The prediction sample results in Fig. 4 *A* and *B* indicate positive relationships between replication success and first and senior authors' number of publications, as well as between replication success and the authors' citation impact (all P < 0.001).

Second, we found no statistically significant evidence that replicability is linked to the prestige of the first or senior author's institution in the training sample or prediction sample, all P > 0.1 (Fig. 4*C*). This result suggests that the researchers' own records, rather than the institution's, is predictive of replicability.

Third, we found no significant difference in a paper's cumulative citation number between studies that passed vs. failed replications in our training sample (Fig. 4*D*, P = 0.51). This finding is consistent with results from prior research (26,61). By contrast, we found that in the prediction sample, the papers likely to replicate received significantly fewer citations than papers unlikely to replicate (Fig. 4*D*); however, the effect was negligible (biserial correlation r = −0.05, P = 0.04). Together, we conclude that citation number is weakly associated with replicability and is not diagnostic of a paper's replicability, despite citation impact being a widely accepted indicator that a paper's quality and importance (62–65).

The final publication metric we examined with replicability is media coverage. Ideally, media should cover credible and rigorous research. Yet in reality, the mainstream media tends to highlight research that finds surprising, counterintuitive results (66). A small sample of replications has shown that the more surprising a study's finding, the less likely it is to replicate (10). Our analysis more directly tested the association between media coverage and replicability and found similar results. Both training and prediction samples indicate that media attention and replication success are

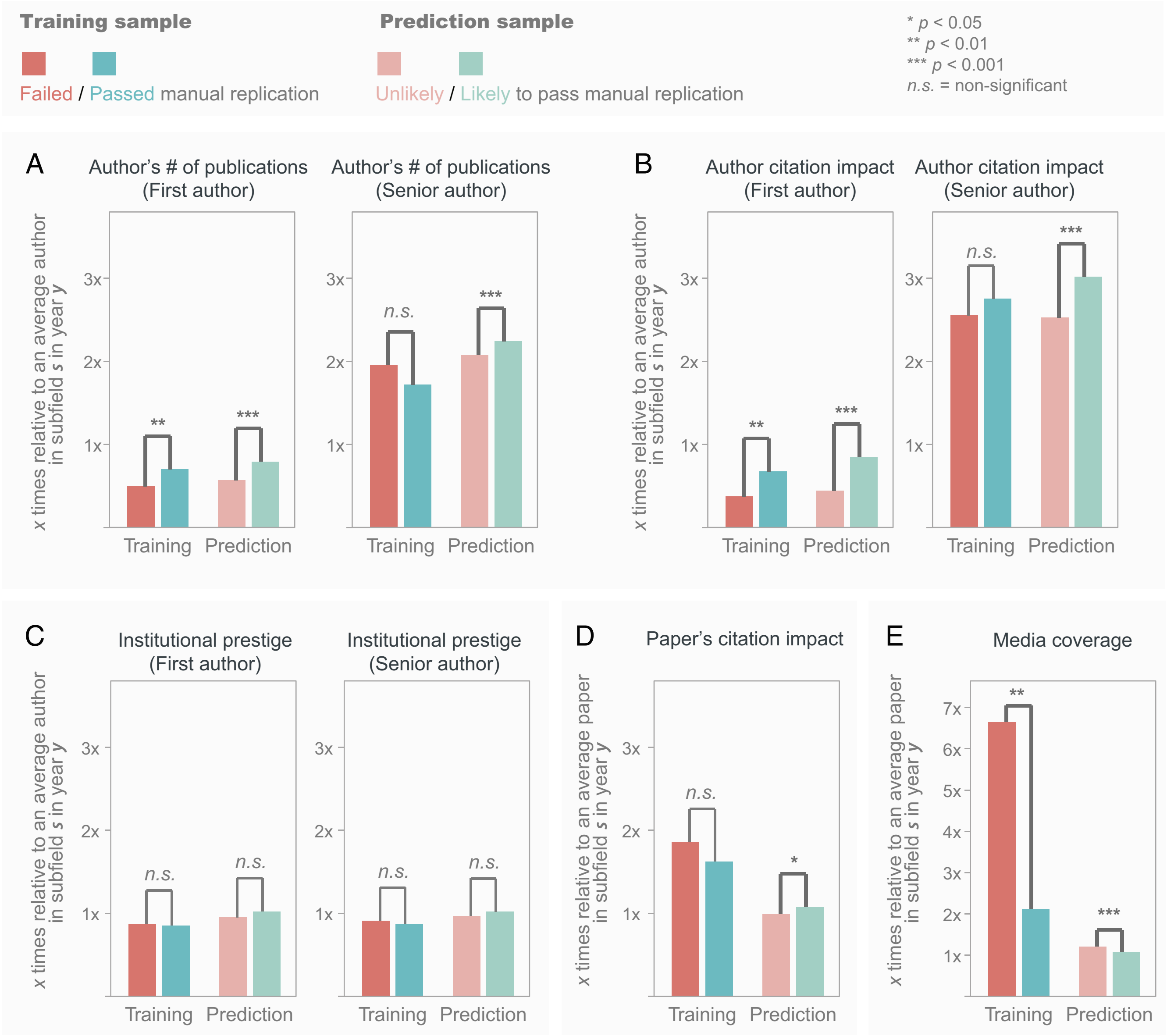

Fig. 4. The Relationship Between Replicability and Other Metrics of a Paper. Replicability is shown to have a small positive association with the first author's previous number of publications and citation impact, no association with institutional prestige of authors, a negligible association with the paper's citation impact and a negative association with media coverage. All analyses compare the target metric between 1) studies that passed vs. failed manual replication in the training sample and 2) papers likely vs. unlikely to replicate in the prediction sample. The comparisons are done using Mann–Whitney rank-sum tests. n.s. = not significant. All metrics were normalized for each paper using the averages of its subfield and publication year (illustrated in *SI Appendix*, Fig. S1). The "1x" on the y-axis in each panel represents the baseline of that metric (i.e., an average paper's level, or a multiple of one). For instance, Panel *E* shows that the mean media coverage for failed replications in the training sample is about 6.5 times an average paper's media coverage. The other panels compares replicable versus non-replicable papers on their A) authors' publication number, B) authors' citation impact, C) institutional prestige, and D) citation impact respectively.

negatively correlated (Fig. 4*E*). Biserial correlations are $r = -0.21$, $P = 0.001$ in the training sample, and $r = -0.13$, $P < 0.001$ in the prediction sample.

Discussion

This research uses a machine learning model that quantifies the text in a scientific manuscript to predict its replication likelihood. The model enables us to conduct the first replication census of nearly all of the papers published in Psychology's top six subfield journals over a 20-y period. The analysis focused on estimating replicability for an entire discipline with an interest in how replication rates vary by subfield, experimental and non-experimental methods, the other characteristics of research papers. To remain grounded in the human expertise, we verified the results with available manual replication data whenever possible. Together, the results further provide insights that can advance replication theories and practices.

A central advantage of our approach is its scale and scope. Prior speculations about the extent of replication failure are based on relatively small, selective samples of manual replications (21). Analyzing more than 14,000 papers in multiple subfields, we showed that replication success rates differ widely by subfields. Hence, not one replication failure rate estimated from a single replication project is likely to characterize all branches of a diverse discipline like Psychology. Furthermore, our results showed that subfield rates of replication success are associated with research methods. We found that experimental work replicates at significantly lower rates than non-experimental methods for all subfields, and subfields with less experimental work replicate relatively better. This finding is worrisome, given that Psychology's strong scientific reputation is built, in part, on its proficiency with experiments.

Analyzing replicability alongside other metrics of a paper, we found that while replicability is positively correlated with researchers' experience and competence, other proxies of research quality, such as an author's university prestige and the paper's citations, showed no association with replicability in Psychology. The findings highlight the need for both academics and the public to be cautious when evaluating research and scholars using pre- and post-publication metrics as proxies for research quality.

We also correlated media attention with a paper's replicability. The media plays a significant role in creating the public's image of science and democratizing knowledge, but it is often incentivized to report on counterintuitive and eye-catching results. Ideally, the media would have a positive relationship (or a null relationship) with replication success rates in Psychology. Contrary to this ideal, however, we found a negative association between media coverage of a paper and the paper's likelihood of replication success. Therefore, deciding a paper's merit based on its media coverage is unwise. It would be valuable for the media to remind the audience that new and novel scientific results are only food for thought before future replication confirms their robustness.

We envision two possible applications of our approach. First, the machine learning model could be used to estimate replicability for studies that are difficult or impossible to manually replicate, such as longitudinal investigations and special or difficult-to-access populations. Second, predicted replication scores could begin to help prioritize manual replications of certain studies over others in the face of limited resources. Every year, individual scholars and organizations like Psychological Science Accelerator (67) and Collaborative Replication and Education Project (68) encounter the problem of choosing from an abundance of Psychology studies which ones to replicate. Isager and colleagues (69) proposed that to maximize gain in replication, the community should prioritize replicating studies that are valuable and uncertain in their outcomes. The value of studies could be readily approximated by citation impact or media attention, but the uncertainty part is yet to be adequately measured for a large literature base. We suggest that our machine learning model could provide a quantitative measure of replication uncertainty.

We note that our findings were limited in several ways. First, all papers we made predictions about came from top-tier journal publications. Future research could examine papers from lower-rank journals and how their replicability associate with pre- and post-publication metrics (70). Second, the estimates of replicability are only approximate. At the subfield-level, five out of six subfields in our analysis were represented by only one top journal. A single journal does not capture the scope of the entire subfield. Future research could expand the coverage to multiple journals for one subfield or cross-check the subfield pattern derived using other methods (e.g., prediction markets). Third, the training sample used to develop the model used nearly all the manual replication data available, yet still lacked direct manual replication for certain psychology subfields. While we conducted a series of transfer learning analyses to ensure the model's applicability beyond the scope of the training sample, implementation of the model in the subfields of Clinical Psychology and Developmental Psychology, where actual manual replication studies are scarce should be done judiciously. For example, when estimating a paper's replicability, we advise users to review a paper's other indicators of replicability, like original study statistics, aggregated expert forecast, or prediction market. Nevertheless, our model can continue to be improved as more manual replication results become available.

Future research could go in several directions: 1) our replication scores could be combined with other methods like prediction markets (16) or non-text-based machine learning models (27, 28) to further refine estimates for Psychology studies; 2) the design of the study could be repeated to conduct replication censuses in other disciplines; and 3) the replication scores could be further correlated with other metrics of interest.

The replicability of science, which is particularly constrained in social science by variability, is ultimately a collective enterprise improved by an ensemble of methods. In his book *The Logic of Scientific Discovery,* Popper argued that "we do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them" (1). However, as true as Popper's insight about repetition and repeatability is, it must be recognized that tests come with a cost of exploration. Machine learning methods paired with human acumen present an effective approach for developing a better understanding of replicability. The combination balances the costs of testing with the rewards of exploration in scientific discovery.

Data, Materials, and Software Availability. Data, and code for generating the data have been deposited in [Open Science Framework] (https://osf.io/f5sxn/).

1. K. Popper, *The Logic of Scientific Discovery* (Routledge, 2005).
2. D. J. Bem, Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *J. Personality Soc. Psychol.* **100**, 407 (2011).
3. J. Galak, R. A. LeBoeuf, L. D. Nelson, J. P. Simmons, Correcting the past: Failures to replicate psi. *J. Personality Soc. Psychol.* **103**, 933 (2012).
4. S. J. Ritchie, R. Wiseman, C. C. French, Failing the future: Three unsuccessful attempts to replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PloS One* **7**, e33423 (2012).
5. K. S. Button *et al.*, Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365 (2013).
6. J. B. Asendorpf *et al.*, Recommendations for increasing replicability in psychology. *Eur. J. Personality* **27**, 108–119 (2013).
7. M. Boss, J. Kleinert, Explaining social contagion in sport applying Heider's balance theory: First experimental results. *Psychol. Sport Exercise* **16**, 160–169 (2015).
8. R. A. Klein *et al.*, Investigating variation in replicability: A "many labs" replication project. *Social Psychol.* **45**, 142–152 (2014).
9. B. A. Nosek, T. M. Errington, Making sense of replications. *eLife* **6**, 4–7 (2017).
10. OSC, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
11. J. Freese, D. Peterson, Replication in social science. *Annu. Rev. Sociol.* **43**, 147–165 (2017).
12. E. Yong, Nobel laureate challenges psychologists to clean up their act. *Nat. News* (2012).
13. L. P. Freedman, I. M. Cockburn, T. S. Simcoe, The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, e1002165 (2015).
14. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nat. News* **533**, 452 (2016).
15. M. Gordon *et al.*, Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *R. Soc. Open Sci.* **7**, 200566 (2020).
16. M. Gordon, D. Viganola, A. Dreber, M. Johannesson, T. Pfeiffer, Predicting replicability–Analysis of survey and prediction market data from large-scale forecasting projects. *PLoS One* **16**, e0248780 (2021).
17. N. Alipourfard *et al.*, Systematizing confidence in open research and evidence (SCORE). *SocArXiv [Preprint]* (2021). https://osf.io/preprints/socarxiv/46mnb/.
18. H. Pashler, C. R. Harris, Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* **7**, 531–536 (2012).
19. T. E. Hardwicke *et al.*, Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspect. Psychol. Sci.* **17**, 239–251 (2022).
20. E. Forsell *et al.*, Predicting replication outcomes in the many labs 2 study. *J. Eco. Psychol.* **75**, 102117 (2018), 10.1016/j.joep.2018.10.009.
21. B. A. Nosek *et al.*, Replicability, robustness, and reproducibility in psychological science. *Annual review of psychology* **73**, 719–748 (2022).
22. J. L. Tackett, C. M. Brandes, K. M. King, K. E. Markon, Psychology's replication crisis and clinical psychological science. *Annu. Rev. Clin. Psychol.* **15**, 579–604 (2019).
23. B. Owens, Replication failures in psychology not due to differences in study populations. *Nature* **19**, 19 (2018).
24. A. Koul, C. Becchio, A. Cavallo, Cross-validation approaches for replicability in psychology. *Front. Psychol.* **9**, 1117 (2018).
25. A. Dreber *et al.*, Using prediction markets to estimate the reproducibility of scientific research. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 15343–15347 (2015).
26. Y. Yang, W. Youyou, B. Uzzi, Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10762–10768 (2020).
27. A. Altmejd *et al.*, Predicting the replicability of social science lab experiments. *PloS One* **14**, e0225826 (2019).

28. J. Wu *et al.*, Predicting the reproducibility of social and behavioral science papers using supervised learning models. arXiv preprint arXiv:2104.04580 (2021). https://doi.org/10.48550/arXiv.2104.04580.
29. I. Boutron, P. Ravaud, Misrepresentation and distortion of research in biomedical literature. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2613–2619 (2018).
30. W. Youyou, Y. Yang. A discipline-wide investigation of replicability in Psychology over the past 20 years. *Open Science Framework*. Retrieved from osf.io/f5sxn.
31. T. Mikolov *et al.*, Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Proc. Syst.* **26** (2013).
32. K. Wang *et al.*, A review of microsoft academic services for science of science studies. *Front. Big Data* **2**, 45 (2019).
33. D. J. Simons, A. O. Holcombe, Registered replication reports. *APS Observer* **27** (2014).
34. R. A. Klein *et al.*, Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychol.* **8**, 35271 (2019).
35. R. A. Klein *et al.*, Many Labs 2: Investigating variation in replicability across sample and setting. *Adv. Methods Pract. Psychol. Sci.* **1**, 443–490 (2019).
36. C. R. Ebersole *et al.*, Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *J. Exp. Soc. Psychol.* **67**, 68–82 (2016).
37. B. A. Nosek, D. Lakens, Replications of ImportantResults in Social Psychology [Special Issue]. *Soc. Psychol.* **45**, 3 (2014).
38. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637 (2018).
39. C. J. Soto, How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychol. Sci.* **30**, 711–727 (2019).
40. Collaborative Open-science REsearch, Replications and extensions of classic findings in Social Psychology and judgment and decision making (2022), 10.17605/osf.io/5z4a8.
41. A. A. Aarts, E. P. LeBel, Curate science: A platform to gauge the replicability of psychological science (2016) https://curatescience.org. Accessed 1 June 2019.
42. H. Pashler, B. Spellman, S. Kang, A. Holcombe, Archive of replication attempts in experimental psychology (2019). http://psychfiledrawer.org/view_article_list.php. Accessed 1 June 2019. http://psychfiledrawer.org/.
43. O. Day, T. M. Khoshgoftaar, A survey on heterogeneous transfer learning. *J. Big Data* **4**, 1–42 (2017).
44. G. J. Duncan, M. Engel, A. Claessens, C. J. Dowsett, Replication and robustness in developmental research. *Dev. Psychol.* **50**, 2417 (2014).
45. M. Rohrbach, S. Ebert, B. Schiele, Transfer learning in a transductive setting. *Adv. Neural Inf. Process. Syst.* **26** (2013).
46. S. J. Pan, Q. Yang, A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2009).
47. X. Dai, S. Karimi, B. Hachey, C. Paris, *Using Similarity Measures to Select Pretraining Data for NER* (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), pp. 1460–1470.
48. Z. Maniadis, F. Tufano, J. A. List, One swallow doesn't make a summer: New evidence on anchoring effects. *Am. Eco. Rev.* **104**, 277–290 (2014).
49. R. C. Fraley, S. Vazire, The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One* **9**, e109019 (2014).
50. L. Malich, M. R. Munafò, Introduction: Replication of crises: interdisciplinary reflections on the phenomenon of the replication crisis in psychology. *Rev. General Psychol.* **26**, 10892680221077996 (2022).
51. Symonds, Quacquarelli. "QS World University Rankings 2020." (2021). Accessed 1 June 2019.
52. N. S. Trueger *et al.*, The altmetric score: A new measure for article-level dissemination and impact. *Ann. Emergency Med.* **66**, 549–553 (2015).
53. C. Bode, C. Herzog, D. Hook, R. McGrath, A guide to the Dimensions data approach. *Dimensions Report*. Cambridge, MA: Digital Science (2018).
54. B. A. Spellman, D. Kahneman, What the replication reformation wrought. *Behav. Brain Sci.* **41**, e149 (2018).
55. M. Motyl *et al.*, The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *J. Personality Social Psychol.* **113**, 34 (2017).
56. M. C. Makel, J. A. Plucker, Facts are more important than novelty: Replication in the education sciences. *Educ. Res.* **43**, 304–316 (2014).
57. H. Jeličić, E. Phelps, R. M. Lerner, Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Dev. Psychol.* **45**, 1195 (2009).
58. D. Peterson, The baby factory: Difficult research objects, disciplinary standards, and the production of statistical significance. *Socius* **2**, 2378023115625071 (2016).
59. J. M. Starck, *"Why Peer Review?" in Scientific Peer Review* (Springer, 2017), pp. 11–14.
60. L. E. Burman, W. R. Reed, J. Alm, A call for replication studies. *Public Finance Rev.* **38**, 787–793 (2010).
61. M. Serra-Garcia, U. Gneezy, Nonreplicable publications are cited more than replicable ones. *Sci. Adv.* **7**, eabd1705 (2021).
62. S. Wuchty, B. F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036–1039 (2007).
63. B. F. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008).
64. S. Mukherjee, D. M. Romero, B. Jones, B. Uzzi, The nearly universal link between the age of past knowledge and tomorrows breakthroughs in science and technology: The hotspot. *Sci. Adv.* **3**, 2017 (2017).
65. B. Uzzi, S. Mukherjee, M. Stringer, B. Jones, Atypical combinations and scientific impact. *Science* **342**, 468–472 (2013).
66. M. G. Pellechia, Trends in science coverage: A content analysis of three US newspapers. *Public Understanding Sci.* **6**, 49 (1997).
67. H. Moshontz *et al.*, The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Adv. Methods Practices Psychol. Sci.* **1**, 501–515 (2018).
68. J. R. Wagge *et al.*, A demonstration of the Collaborative Replication and Education Project: Replication attempts of the red-romance effect. *Collabra: Psychol.* **5**, 5 (2019).
69. P. M. Isager *et al.*, Deciding what to replicate: A decision model for replication study selection under resource and knowledge constraints. *Psychol. Methods* (2021) https://doi.org/10.1037/met0000438.
70. M. R. Dougherty, Z. Horne, Citation counts and journal impact factors do not capture some indicators of research quality in the behavioural and brain sciences. *Royal Society Open Science* **9**, 220334 (2022).