Supplementary Information
for
A discipline-wide investigation of replicability in Psychology over the past 20 years

This document is structured as follows:

- **Supplementary Text 1** describes training and prediction samples.
  - **Tables S1, S2**
- **Supplementary Text 2** describes pre- and post-publication metrics of the papers.
  - **Figures S1**
- **Supplementary Text 3** presents the machine-learning model.
  - **Figures S2-S4**
- **Supplementary Text 4** presents additional results.
  - **Figures S5, S6**
- **Refences**

1. Data: the training and prediction samples

The data used in this paper were divided into a training and a prediction sample. The training sample consists of $N = 388$ studies from 12 manual replication projects in Psychology. Below we briefly describe these projects, and more details can be found on their websites. The training sample was used to calibrate and build a replicability prediction model. The model was then applied to the prediction sample of $N = 14{,}126$ papers, which is the focus of this project.

Table S1. Training sample by replication project / platform.

#	Project / Platform	Psychology Subfields	Number of Studies	Number of successful replications
1	RPP (1)	Cognitive, Social	96	37
2	RRR (2)	Cognitive, Social	8	1
3-6	ML1-4 (3-6)	Cognitive, Social, Personality, Organizational	42	22
7	JSP (7)	Social, Organizational	16	5
8	SSRP (8)	Cognitive, Social	18	10
9	LOOPR (9)	Personality	22	20
10	CORE (10)	Social, Organizational	39	30
11	Curate Science (11)	Cognitive, Social, Personality, Organizational	93	18
12	PFD (12)	Cognitive, Social, Organizational	54	25
			Total = 388	Overall success rate = 43.3%

1.1 RPP

The Reproducibility Project: Psychology by the Open Science Collaboration (1) completed 100 replicated studies sampled from three top Psychology journals published in 2008, using the same procedures as the original studies. Each study was replicated by a single lab. Three studies with null original results were excluded from the replication report. Two of the remaining 97 were replications of the same study and achieved highly similar results, thus combined into one record. The final sample consisted of 96 studies.

1.2 RRR

The Registered Replication Report is an initiative and a new article type in the journal Advances in Methods and Practices in Psychological Science (13). Each original study was replicated independently by multiple labs, which mimicked the original study protocol as closely as possible. In each replication report, the authors unequivocally stated whether they successfully replicated the original study. We note that the study by Srull and Wyer (1979) failed PDF conversion to text (see section 3.3 below for more explanation) and was therefore excluded from the sample.

1.3-1.6 ML1-4

The Many Lab Project is a large-scale replication project of five waves, classified as either Social or Cognitive Psychology studies, with each one informally referred to as "Many Labs" numbered 1 through 5. Thus far, Many Labs 1 (3), 2 (14), 3 (5) and 4 (6) have published their results. In each project, independent research teams first mimic the original study protocols as closely as possible, and then collectively produce a single replication report and state their overall conclusion regarding the replication outcome.

1.7 JSP

"Replications of Important Results in Social Psychology" is a special issue of the journal Social Psychology that includes 15 replication reports (7). All studies are on Social Psychology topics. Multiple labs, mimicking the original study protocol as closely as possible, replicated each original study independently. All were combined into a single replication report, declaring their overall conclusions regarding the replication outcome, as was done in the RRR.

1.8 SSRP

Social Sciences Replication Project was a project that replicated 21 existing Social science experiments using the same procedures as the original studies (8). We kept the 18 Psychology studies and excluded the three economic studies for our purpose. Each study was replicated by a single lab. We used the authors' interpretation from Figure 1c to decide replication outcomes.

1.9 LOOPR

The Life Outcomes of Personality Replication Project (9) was a project led by Professor Christopher J. Soto. He collected a large sample of survey responses to replicate literature on the relationship between personality and life outcomes. The replication results were grouped by the original study, with each study likely including multiple effects that were tested. The replication authors used the column "ReplicationSuccessByOutcome" to indicate the percentage of replication successes for all effects tested in the study. We treated studies with percentages larger than .75 as replication success, smaller than .25 as failure and in-betweens as mixed results.

1.10 CORE

Mass Replications & Extensions (10) is an ongoing project led by Professor Gilad Feldman, where his students attempted to conduct pre-registered replications in Psychology. The replication outcomes are labelled in a summary.

1.11 Individual efforts (Curate Science)

Besides the large-scale organized replication project, there are also published reports dedicated to replicating one effect at a time. Curate Science (15) is a website documenting and summarizing these individual efforts, as well as large-scale collective projects. We included replicated studies from Curate Science that were not part of any other projects described above.

1.12 Individual efforts (PFD)

PsychFileDrawer.org is an online tool designed to archive replication reports (12). Any user can upload such results. We included replicated studies on PFD that were not part of any other projects described above.

For all projects, we excluded studies with the following characteristics: 1) the original effect is null, 2) mixed, inconclusive results, 3) already included in another project, 4) text only available in PDF and failed PDF to text conversion and 5) published prior to 1970, because they were written in a format and style very different from recently published studies.

Table S2. Replicability prediction sample by journal.

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%**

For the prediction sample, we chose five journals considered top tier in a particular subfield, as well as *Psychological Science*, a top journal that publishes research in all Psychology subfields. The journals were chosen based on their subject SJR impact (16). All articles were published in 2000 - 2019, except those in *Psychological Science*, which include articles from 2003 to 2019. *Psychological science* published prior to 2003 were all only available in PDFs and the software used failed to convert them into

text (see section 3.1). We excluded papers that were 1) in the training sample, 2) retracted, 3) meta-analyses, reviews, or commentaries.

2. Data: other pre- and post-publication metrics of the papers

We collected five metrics related to a paper and examined their relationship with the replicability of a paper. Three metrics occurred prior to the publication and two metrics occurred after the publication. All five metrics were collected for papers in both training and prediction samples.

Since the data spans six subfields and 20 years of publication, we must consider the differences in base rates for these research pre- and post-publication metrics (summarized below in Figure S1). For instance, Social Psychology receives more media attention than other subfields; the number of citations increases as a paper ages. We, therefore, normalized all metrics for each paper, dividing the raw score by the average in its subfield and publication year.

2.1 Authors' cumulative number of publications.

We collected the first and senior authors' cumulative number of publications respectively to measure the authors' research experience prior to the publication of the focal paper. A senior author is defined as the author on the team with the most cumulative number of citations when the focal paper was published. The data were retrieved from the *Dimensions* database (17). For each author, we counted the number of papers published by the author before the publication year of the focal paper.

2.2 Authors' citation impact.

We collected the first and senior authors' cumulative citation counts respectively to measure the authors' career impact prior to the publication of the focal paper. A senior author is defined as the author on the team with most cumulative number of citations when the focal paper was published. The data were retrieved from the *Dimensions* database (17). For each author, *Dimensions* tracked the number of citations received for all their publications in each year. We summed these yearly citation counts until the publication year of the focal paper.

2.3 Authors' institutional prestige.

We collected ranking information about the first and senior authors' institutions to measure the prestige of the research institution. First and senior authors' institutions were extracted from the *Dimensions* database (17) and matched with the rankings in 2021 *QS World University Rankings* (18). The *Rankings* listed 500 top universities around the world. We divided the institutions into tiers 1 to 5 based on their ranking (e.g., ranking 1-100 = tier 1, ranking 101-200 = tier 2 etc.). Categorizing the institutions into tiers allowed us to include institutions not in the world 500 top universities and assign them to be tier 6.

To further verify that ranking captures institutional prestige, we compared the categorized rankings against researchers' subjective impression of institution prestige. The RPP (1) once recruited research assistants to rate institution prestige. We correlated the subjective prestige ratings with our categorized rankings (tiers 1-6) for 173 institutions and observed that higher prestige ratings are strongly related to lower rankings (Spearman r = -0.77, Pearson r = -0.75, ps < 0.001).

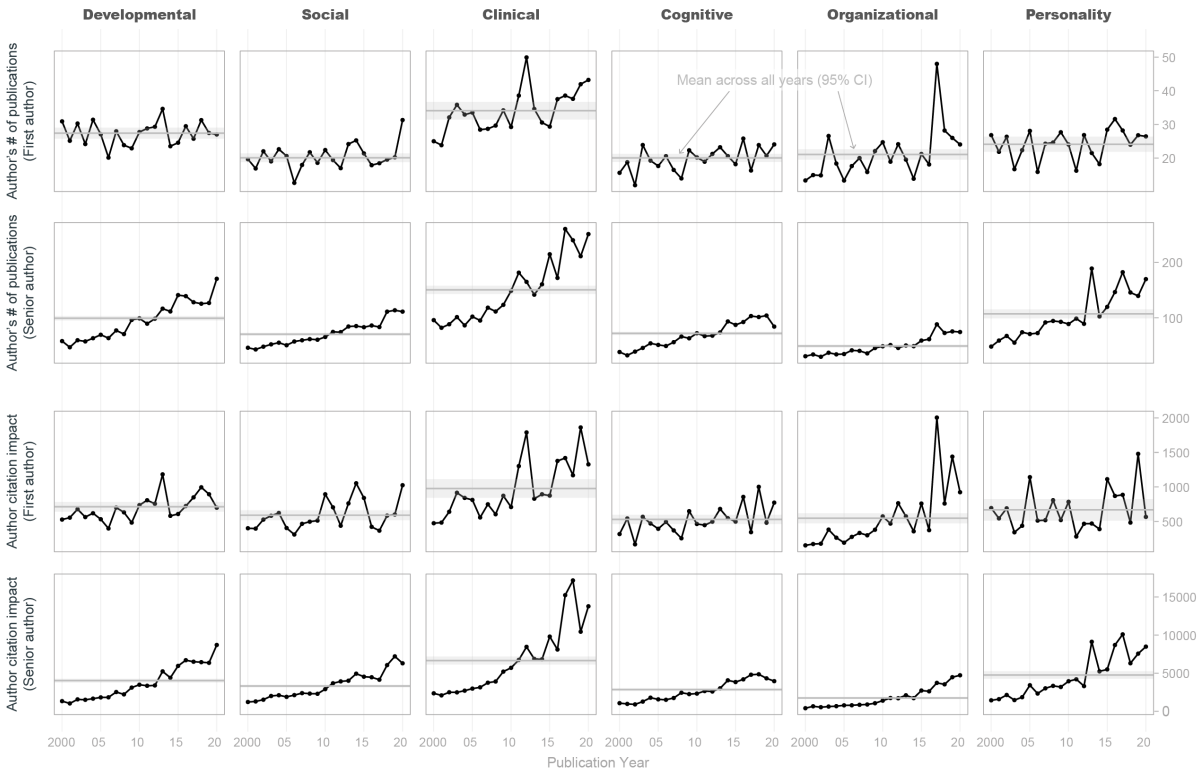
2.4 Paper's citation impact.

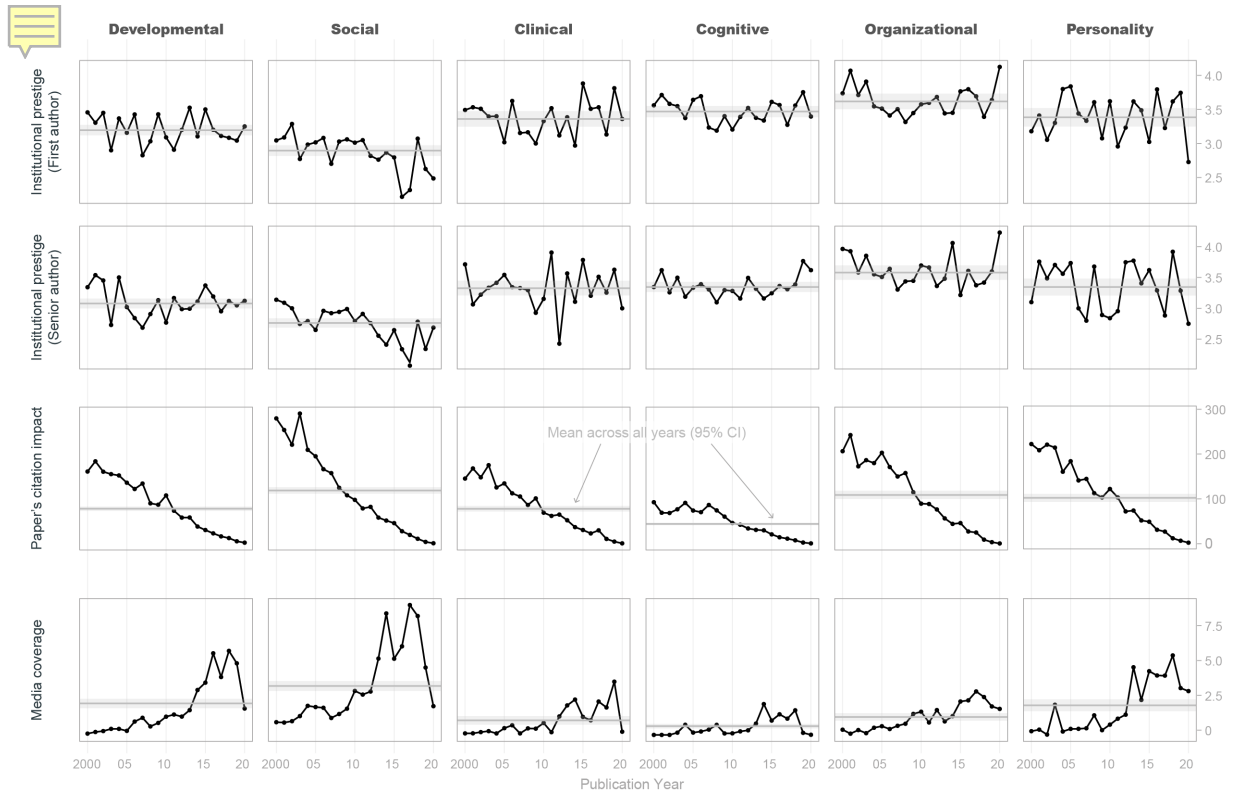
We collected all citations accrued by the focal paper from publication, as tracked in the *Dimensions* database (17).

2.5 Paper's media coverage.

We collected the number of media attentions accrued by the focal paper from publication until June 2020, as tracked by the *Altmetric* database (19). *Altmetric* tracks media coverage by mining a list of manually curated news sources. If a news article mentions a paper, the number of media attention increases by one for that paper. The media mentions used in our project do not include social media mentions like Facebook posts or Twitter posts.

Figure S1. Base rates of author and paper metrics for Psychology papers across six subfields. The gray horizontal lines in each panel represent the mean level for each subfield; the bands are 95% confidence intervals for the means. The black dots represent the means broken down by publication year. The base rates for each metric vary by subfield and by publication year. For instance, citation impact increases as a publication ages. Social Psychology on average receives more media coverage than other subfields. The averages were used to normalize the metrics for each paper according to its subfield and publication year in Figure 4 and Figure S5.





3. Methods

3.1 Converting published papers to text files.

For every downloaded article in HTML format, we identified the section heading for the block of text (e.g., title, abstract, authors, introduction, results, etc.) using HTML tags. We kept only the main text of each paper. Most papers have well-defined section boundaries, and the sections and lines that do not constitute main texts were excluded, including journal titles, page numbers, contributions, acknowledgments, abstracts and footnotes. Within the main text, we removed statistics, in-text citations, numbers, equations and other non-textual information like figures, tables and their captions.

A small proportion of the papers were only available in PDFs and we ran them through the published software GROBID that converts PDFs of scholarly articles into text and delineates sections based on machine learning (20). 66 papers in the training sample and 821 in the prediction sample were successfully converted into text files. This software is not perfect (F-Score = 0.78) (21). Indeed, we observed that a small percentage of the conversions produced disorganized and unreliable outputs. Because we seek to build a protocol that can be automatically applied to any paper, we did little manual follow-up processing and treated imperfect conversions as noises. We followed the same procedure of removing non-textual information as explained before. Some conversions even failed entirely, and those articles were excluded.

For the training sample, our unit of analysis is the unit of manual replication, which is usually a study or a set of studies in a paper that was replicated and has a single outcome. Hence, we only utilized text pertaining to the target study/studies to predict replication outcomes, rather than the full paper. For the prediction sample, the unit of analysis is the main text of the full paper.

3.2 Building a replicability prediction model.

The process of building a replicability prediction model largely follows our previous proof-of-concept paper (22). Figure S2 presents a graphic overview of the procedures. Each step is explained in detail below.

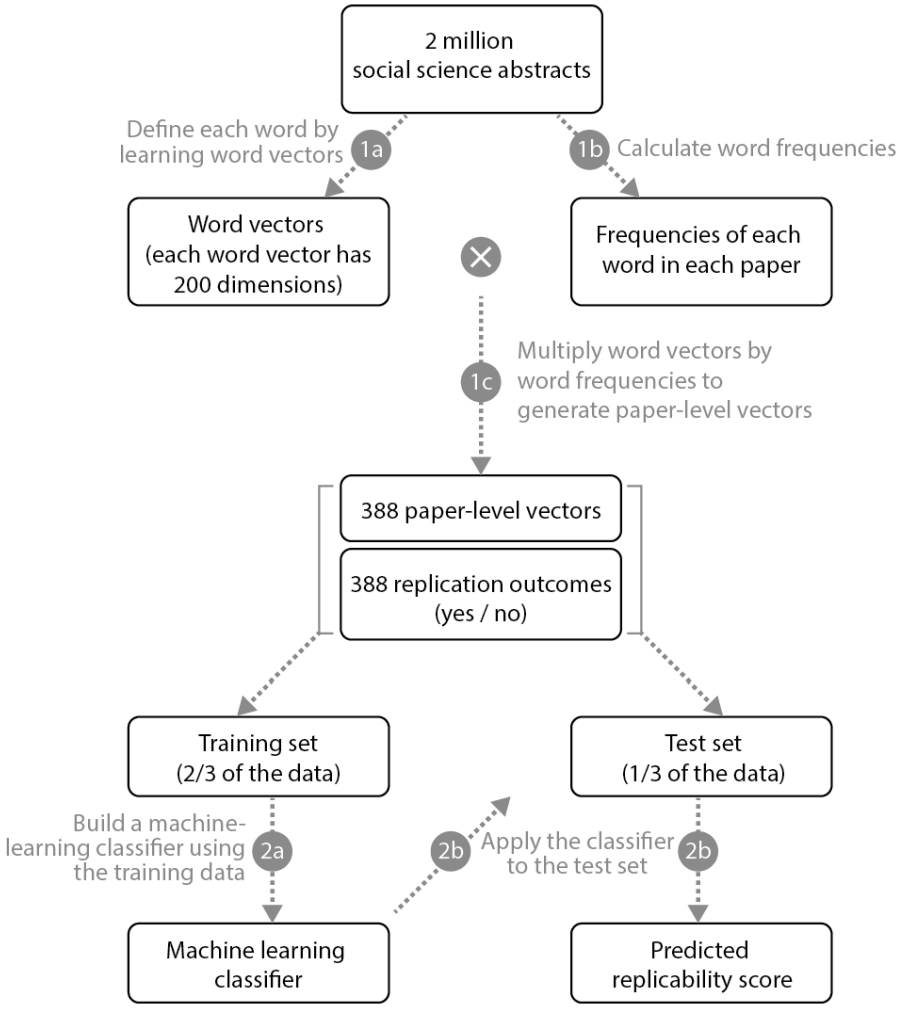


Figure S2: Overview of the replicability prediction model. Steps 1a – c convert text into vectors; Steps 2a and 2b perform machine learning and cross-validation.

3.2.1a. Converting words to word vectors.

A model does not take papers directly as input. Papers need to be first decomposed into individual words, which will be numerically defined (i.e., converted to word vectors). The word vectors will then

be reassembled to represent the content of the manuscript. In this subsection, we explain how to convert words into word vectors in a way that preserves their semantic meaning using word2vec (23).

The first step was to collect a rich, relevant corpus so we could teach the machine model to understand what each word means in the context of Psychology literature. Our choice is two million Social Science publication abstracts from the Microsoft Academic Graph (MAG) database (24). We extracted 18 million sentences from these abstracts, and the vocabulary size was roughly 200,000. These sentences provided "contexts" for the words we are about to define.

For each word in a sentence, we took the five nearby words as context words to pair with the target word. Figure S3 provides an example of this process iterating from the first word "there" in a sentence to the last word, "address," setting the context window size at five words. For example, for the target word "are," "there" is its nearby word to the left, and "many," "questions," "that," "our," and "analysis" are its four nearby words to the right. After going through all the sentences, we know how many times each word was a neighbor with all other words in the same five-word window. This relationship can be visualized as a co-occurrence matrix (Figure S4).

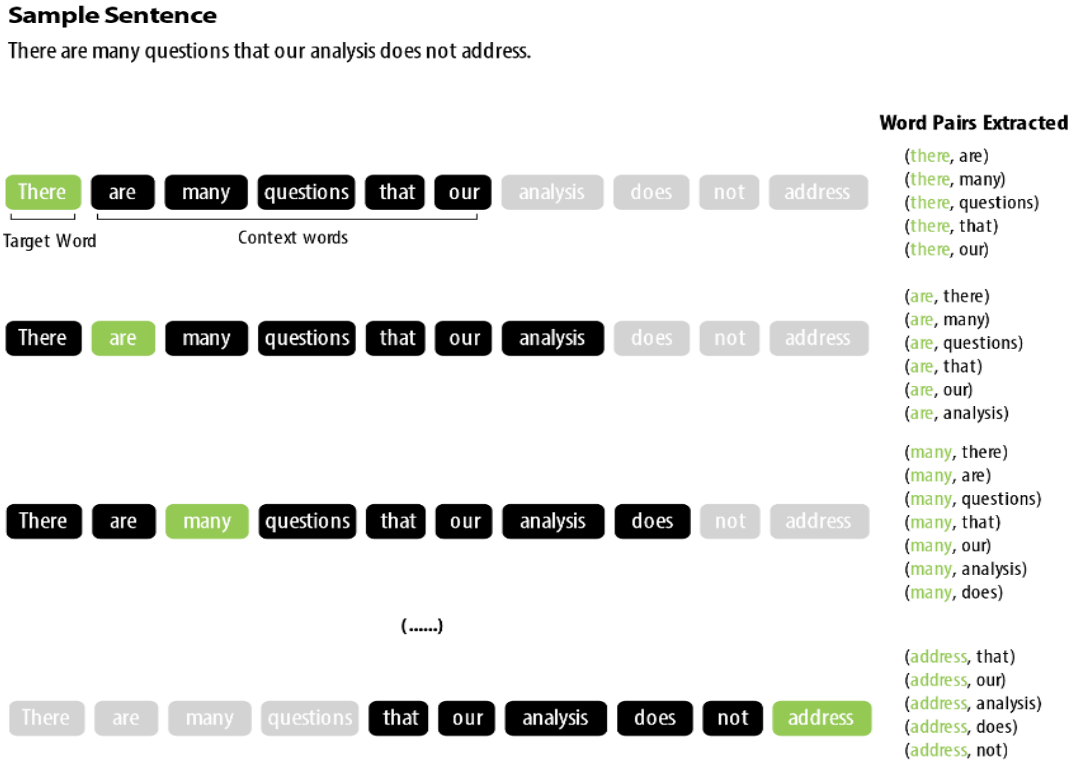


Figure S3: Extracting word pairs from papers using a context window of five.

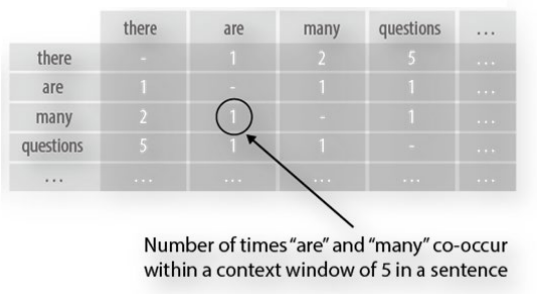


Figure S4: Word co-occurrence matrix for a context window of five.

The next step is to represent these co-occurrence relationships more efficiently. In theory, matrix factorization or principal component analysis can be performed to extract underlying "factors" to condense the matrix. However, considering the size and complexity of word pairs, we adopted word2vec as an alternative, conceptually equivalent technique to understand the co-occurrence relationships with more precision. Word2vec is a neural network-based model. As a result, each word could be effectively represented by a high-dimensional vector. We set the number of dimensions to be 200. Our previous work tested a different number of dimensions and the results are similar (22).

Conceptually, every word is defined by all other words that were a neighbor within the five-word window. For example, the word "cat" and "dog" often occur with the same set of words like "feed", "eat" or "sleep" in the same sentence. Because of that, "cat" and "dog" are considered semantically similar by the word2vec model and will be located close to each other in the 200-dimension vector space.

3.2.1b. Calculate word frequencies for each paper.

After obtaining quantitative definitions for individual words, the next step is to obtain quantitative representations of contents for individual papers. In a nutshell, this is achieved by multiplying the frequency of words in each paper with the word vectors derived in the previous steps. The process of calculating word frequencies is a non-trivial task. In addition to considering raw frequency, we also took into account the prevalence of a word in the corpus.

Term frequency (TF) measures the number of times a term (word) occurs in a document. Here, we defined the normalized term frequency of a term *t* in a document *d*:

$$TF_{t,d} = \frac{W_t}{W}$$

where W_t is the number of term *t* in a document *d* and W is the total number of terms in a document *d*. Raw term frequency, as noted above, suffers from a critical problem: all terms are considered equally important when it comes to assessing relevancy when, in fact, certain terms have little or no discriminating power (e.g., reports on the cellphone industry likely have the term "cellphone" in every document, but that frequency falsely inflates rather than accurately conveys the power of the word). To address this, we used an inverse document frequency (IDF) to attenuate the effect of terms that occur too

often in a collection; we scaled down the term weights of individual terms with high collection frequency across all documents using standard methods.

The inverse document frequency of a term *t* in a collection of documents is defined as:

$$IDF_t = \log \frac{N_t}{df_t}$$

where df_t is defined as the number of documents in the collection that contain a term *t*, and *N* is defined as the total number of documents in a collection. The definitions of term frequency (TF) and inverse document frequency (IDF) are combined to produce a composite weight for each term in each document (TF-IDF). Here, the TF-IDF weighting scheme assigned to term *t* a weight in document *d* calculated by

$$TF\text{-}IDF_t = TF_{t,d} \times IDF_t,$$

and we multiplied that term's normalized term frequency with its IDF in each document to calculate and characterize the prevalence of terms in each paper. TF and TF-IDF are essential to the accurate representation of paper contents.

3.2.1c. Generate paper-level vectors.

We multiplied the TF and TF-IDF in each paper with word vectors respective to generate paper-level vectors. Now, each paper is represented by a 200-dimension vector. These paper-level vectors will be used to predict replication outcomes in the next step.

3.2.2a. Build a machine-learning classifier.

We trained an ensemble algorithm of bagging with a random forest model (25) and bagging with simple logistic regression to predict a binary replication outcome using paper-level vectors as features. Moreover, we considered all TF vectors (times word vectors) as one feature vector, and TF-IDF (times word vectors) as a second feature vector. The final predicted score of each paper was an average of predictions trained on these two vectors.

To alleviate the small sample issue, we kept the machine-learning algorithms simple and used the ensemble strategy. The depth of trees in our random forest model was kept to a shallow maximum depth of three, with the minimum number of instances per leaf set to five. In addition, we used logistic regressions and conducted several robustness tests, where we found that a maximum depth ranging from two to eight gave us almost identical results (22).

3.2.2b. Three-fold cross-validation and accuracy.

We employed repeated three-fold cross-validation in building the classifier to avoid overfitting. Specifically, we randomly split the data into three subsets, training the classifier on two-thirds of the data and applying the classifier to the rest to predict reliability. This ensured that the predictions represented new data that the model had not already seen. By rotating the training versus test set among

the three subsets, we could predict replication scores for the entire sample. The predicted replication score was a continuous variable with range [0, 1.0].

To assess accuracy, we compared continuous replication score and the actual replication outcomes (yes/no). The area under the ROC curve (AUC) is 0.74. When we use a replication score of 0.5 as the cut-off, labelling papers with a score > 0.5 as "success" and the rest as "failure", the model was correct for about 68% of the papers. We also discovered that the model is most accurate at two ends. If we label papers with top 10% replication scores as "success" and the bottom 10% as "failure", the model is then accurate about 82% of the papers.

3.3 Procedural differences with previous proof-of-concept paper.

Although the procedures taken to build the machine-learning model are largely the same as our previous proof-of-concept paper (22), we highlight a few distinctions of the current methods from the previous ones. We included two more replication projects; the current training sample has expanded to $n = 388$ studies. The present machine-learning model is built using cross-validation on all studies. Previously, we built the model on $n = 96$ RPP studies (1) only and conducted four out-of-sample tests on $n = 221$ other studies.

3.4 Performance and robustness tests for the machine-learning model

3.4.1. Topic and textual similarity analysis.

To measure the overlap in research topics between two subfields, we collected research topics for each paper in the testing sample from Microsoft Academic Graph (MAG) database (24). Topics were pooled by subfield. We then calculated the topic similarity between Social Psychology and Cognitive Psychology using two metrics: (1) Jaccard index: the number of common topics between Social and Cognitive, divided by the number of topics in the two subfields combined; (2) number of common topics divided by the number of total topics in the one subfield with smaller number of topics. We repeated the same process to measure topic similarity between Social and Clinical, and Social and Developmental respectively. For Jaccard index, the topic similarity is 42% between Social and Cognitive, and 57% between Social and Clinical, and 56% between Social and Developmental; For metric (2), the topic similarity is 24% between Social and Cognitive, and 26% between Social and Clinical, and 37% between Social and Developmental. Both metrics show that Social-Cognitive have a lower level of textual similarity than Social-Clinical or Social-Development.

To measure textual similarity in research topics between two subfields, each paper was converted into a vector using the techniques described in Section 3.2.1a-c. For each paper in Social Psychology, we computed its cosine similarity with each paper in Cognitive Psychology and took an average (mean = 0.90). We repeated the process for Clinical Psychology (mean = 0.91) and Developmental Psychology (mean = 0.91). We also measured textual similarity using word mover's distance (WMD). The WMD distance measures the dissimilarity between two papers as the minimum amount of distance that the embedded words of one paper need to "travel" to reach the embedded words of another vectors (26).

The average WMD is 0.26 from Social to Cognitive, and 0.24 from Social to Clinical, and 0.25 from Social to Developmental. Both metrics show that Social-Cognitive have the same level of textual similarity as Social-Clinical or Social-Development.

3.4.2. Correlate replication score with original sample size and *p*-value.

We manually coded a random set of 100 papers from Clinical Psychology (n = 50) and Developmental Psychology (n = 50). 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. One paper from the Developmental Psychology paper is theorical and therefore has no participant. 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." In Clinical Psychology, one paper with an original null effect, and two with only descriptive statistics were excluded. Two more papers with no clear statistics and therefore no *p*-value were also excluded.

4. Supplementary results

4.1 Pre- and post-publication correlates of replicability

In the main text, we analyzed correlates of replicability by comparing papers that are likely vs unlikely to replicate on five pre- and post-publication metrics. We defined likely vs 'unlikely to replicate' as papers with a top vs bottom 10% predicted replication score. We opted to focus on comparing the bottom vs top 10%, because the machine learning model was more accurate at two ends, as demonstrated in section 3.3.2b. This will reduce noise in the replication score and make subsequent analyses with other metrics more meaningful. Here, we further expand the analysis to alter the arbitrary cutoff of 10%. Figure S5 shows that the results and conclusions remain unchanged when using 5%, 15% or 20% cut-offs.

We also attempted to conduct the pre- and post-publication correlates analysis by subfields but did not manage to do the analysis due to small sample sizes. First, when the training sample was parsed into subfields, only Social Psychology had $n > 30$ studies with all metrics available in both conditions (passed vs failed replication). Second, recall that we took top and bottom 10% of the studies as papers likely vs 'unlikely to replicate' and compared them in the prediction sample. However, each subfield's number of papers in the top/bottom 10% replication scores are uneven. Only Developmental, Social, Cognitive and Clinical Psychology had $n > 30$ studies with all metrics available in both conditions (likely vs unlikely to replicate).

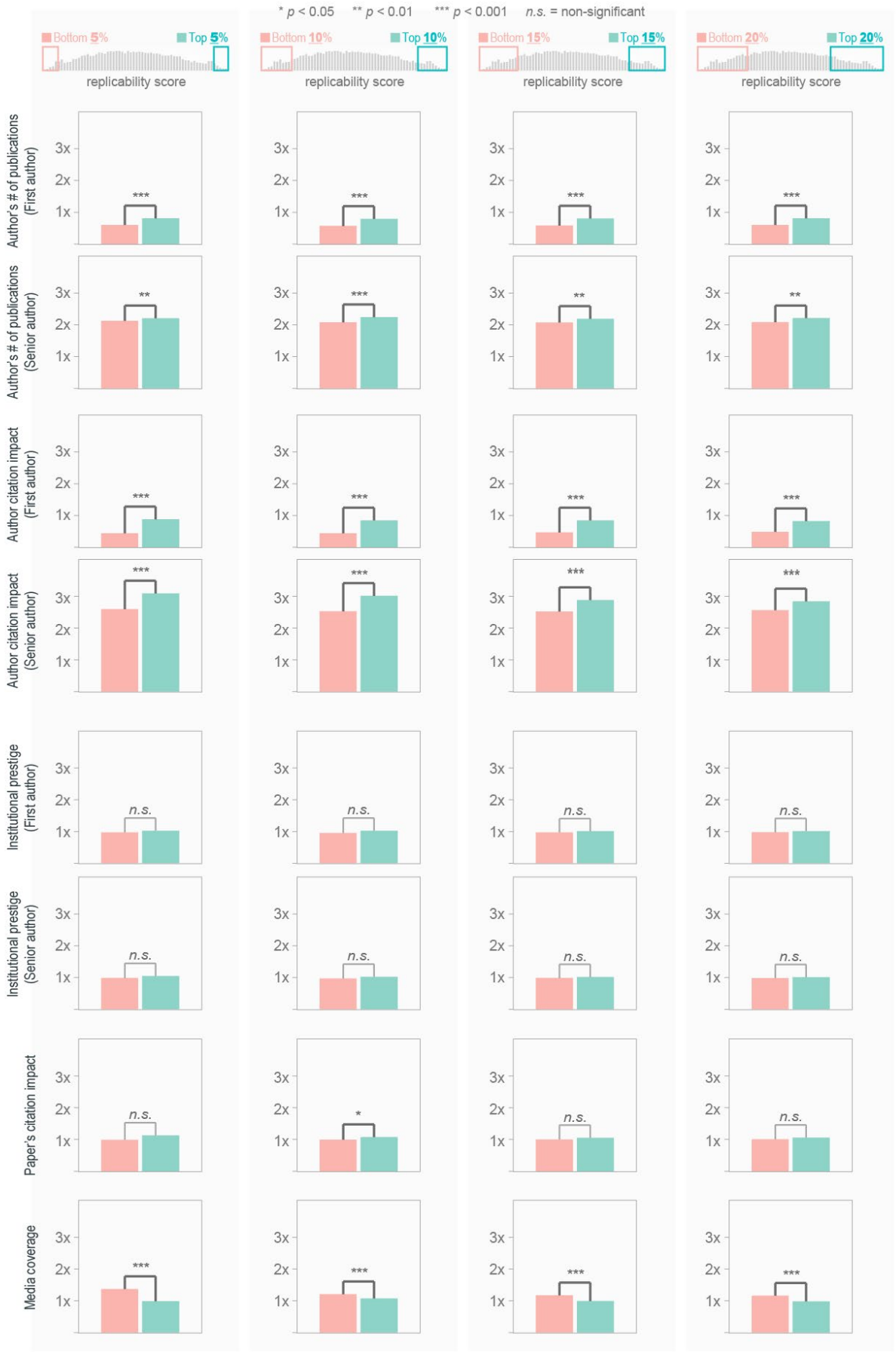

* p < 0.05 ** p < 0.01 *** p < 0.001 n.s. = non-significant
Bottom 5% Top 5%
Bottom 10% Top 10%
Bottom 15% Top 15%
Bottom 20% Top 20%
replicability score
Author's # of publications (First author)
Author's # of publications (Senior author)
Author citation impact (First author)
Author citation impact (Senior author)
Institutional prestige (First author)
Institutional prestige (Senior author)
Paper's citation impact
Media coverage
3x
2x
1x
***
**
*
n.s.

Figure S5: The relationship between replicability and other metrics of a paper. This figure expands upon Figure 4, where five pre- and post-publication metrics are compared between papers predicted to be likely vs unlikely to replicate. The comparisons are done using Mann-Whitney rank-sum tests. In Figure 4, likely vs unlikely to replicate is defined as papers with top vs bottom 10% predicted replication score. The arbitrary 10% cutoff is changed to 5%, 15%, and 20% respectively here as a robustness test. For all metrics, the difference between the least and the most replicable studies were consistent when varying the cutoffs from 5 to 20%. For instance, the differences in citation impact are all significant between bottom and top k% of papers when k = 5, 10, 15 or 20. The "1x" (1 time) on the y-axis in each panel represents the baseline of that metric—an average paper's level.

4.2 Replication scores by year

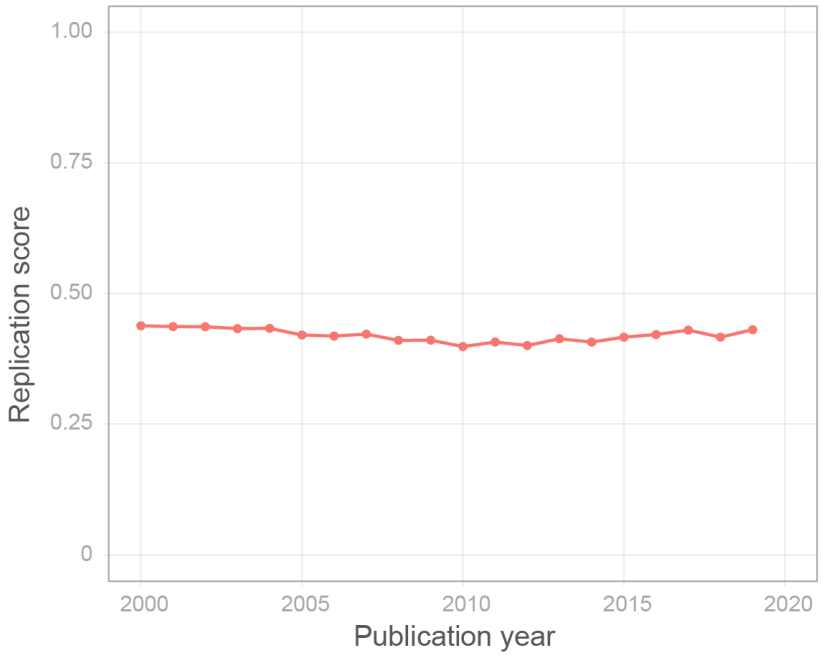


Figure S6: Mean predicated replication scores for Psychology papers by year. The average replication scores decreased between 2000 and 2010 by approximately 10% and then increased between 2010 and 2019 to roughly the same level as 2000. This indicates a pattern that aligns with the observation that changes in research practice have potentially improved replication rates in Psychology (27-30).

References

1. Open-Science-Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
2. Registered Replication Reports (https://www.psychologicalscience.org/publications/replication.
3. R. A. Klein *et al.*, Investigating variation in replicability: A "many labs" replication project. *Social Psychology* **45**, 142-152 (2014).
4. R. A. Klein *et al.*, Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science* (2019).
5. C. R. Ebersole *et al.*, Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* **67**, 68-82 (2016).
6. R. A. Klein *et al.*, Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. (2019).
7. B. A. Nosek, D. Lakens, Replications of ImportantResults in Social Psychology [Special Issue]. *Social Psychology* **45** (2014).
8. C. F. Camerer *et al.*, Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour* **2**, 637 (2018).
9. C. J. Soto, How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science* **30**, 711-727 (2019).
10. C. O.-s. REsearch, Replications and extensions of classic findings in Social Psychology and Judgment and Decision Making. 10.17605/OSF.IO/5Z4A8 (2022).
11. A. A. Aarts, E. P. LeBel, Curate science: A platform to gauge the replicability of psychological science. (2016).
12. H. Pashler, B. Spellman, S. Kang, A. Holcombe, PsychFileDrawer: archive of replication attempts in experimental Psychology. *Online< http://psychfiledrawer. org/view_ article_list. php*.
13. R. R. Reports (Registered Replication Reports. (https://www.psychologicalscience.org/publications/replication).
14. R. A. Klein *et al.*, Many Labs 2: Investigating variation in replicability across sample and setting. (2018).
15. C. Science (Curate Science. (http://curatescience.org/#about).
16. B. González-Pereira, V. P. Guerrero-Bote, F. Moya-Anegón, A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of informetrics* **4**, 379-391 (2010).
17. Digital-Science (2018-) Dimensions [Software] available from https://app.dimensions.ai.
18. Anonymous (2020) QS World University Rankings.
19. Anonymous (Altmetric.
20. P. Lopez, GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. 10.1007/978-3-642-04346-8_62.
21. M. Singh *et al.*, OCR ++ : A Robust Framework For Information Extraction from Scholarly Articles.
22. Y. Yang, W. Youyou, B. Uzzi, Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences* **117**, 10762-10768 (2020).
23. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space. 10.1162/153244303322533223, 1-12 (2013).
24. K. Wang *et al.*, A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data* **2**, 45 (2019).
25. L. Breiman, Random forests. *Machine Learning*, 5-32 (2001).
26. M. Kusner, Y. Sun, N. Kolkin, K. Weinberger (2015) From word embeddings to document distances. in *International conference on machine learning* (PMLR), pp 957-966.

27. B. A. Nosek *et al.*, Replicability, robustness, and reproducibility in psychological science. (2021).
28. B. A. Nosek, T. M. Errington, Making sense of replications. *eLife* **6**, 4-7 (2017).
29. B. A. Nosek, D. Lakens, Replications of Important Results in Social Psychology [Special Issue]. *Social Psychology* **45** (2014).
30. M. Motyl *et al.*, The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of personality and social psychology* **113**, 34 (2017).