PNAS
www.pnas.org

Supplementary Information for

Estimating the deep replicability of scientific findings using human and artificial intelligence

Yang Yang, Youyou Wu, and Brian Uzzi (uzzi@northwestern.edu)

Correspondence are addressed to Brian Uzzi.
Email: uzzi@northwestern.edu

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1909046117

1. Replicability Literature Review

1.1. Definitions.

Our work investigates replicability, defined as “re-performing the experiment and collecting new data” (1). Reproducibility, a related term, is defined as “re-performing the same analysis with the same code” (1). Reproduction studies are small parts of our sample, and we specify them in the main text. To determine whether a study replicated or not, we used a common metric reported in each replication study—the replication team’s summary judgement of whether the study replicated or did not replicate (“yes” or “no”).

1.2. Statistics and Replicability.

Traditionally, statistical power has been regarded as the main indicator of replicability (2). Statistical power estimates the chance that a study will yield a significant effect. It is determined by the true effect size, a given sample size, and significance level. Because true effect size is unknown, power is usually estimated using reported effect size after a study is conducted. However, post hoc estimates of power suffer from selection bias because journals tend to publish only significant effects; thus, the use of post hoc power estimates in assessing replicability has not been very effective (2). For example, Schimmack and Laughton reported limited success predicting replication outcomes from post hoc power analysisof studies in the Reproducibility Project: Psychology (RPP) (3). Only 54 out of 94 studies (57%) were correctly classified. The RPP team also examined the relationship between each of the power-related measures and their replication outcome. The team found that these measures correlate in rank order with replication success at the level of r = -0.26 for p values, 0.28 for effect sizes, and -0.19 for sample sizes.

1.3. Subjective Perceptions and Replicability.

Subjective measures such as the surprisingness of the original finding and the contextual sensitivity of the study also correlate with replicability in the RPP at level of r = -0.24 and -0.23, respectively (4, 5). Both methods, however, are impractical to scale because they require manual coding. The RPP also tested measures such as the importance of the original finding and the experience and expertise of original to predict replication outcomes but found little association (r = -0.07 and -0.04, respectively).

1.4. Prediction Markets, Survey Belief, and Replicability.

Prediction markets and surveys attempt to use crowd wisdom to predict a paper’s replicability (6). Participants consist of researchers who are invited to “bet” on whether a study will or will not replicate before the replication is conducted. In addition to betting, participants also provide their subjective rating of the replication outcome. Specifically, participants receive information about the original study and the design of the replication study. Once the market opens, participants bet on the replication, with the final market price being equivalent to the probability of replication [0, 1]. Four sets of prediction markets and surveys took place to date (6-9). The results varied between different replication projects, but the general pattern indicates that both markets and surveys are effective in predicting replication outcome. However, one major limitation of prediction markets is that they require many participants betting over a relatively long period of time to be effective.

1.5. Other Procedures and Replicability.

Other procedures have been proposed to predict replicability in special circumstances, including the p-curve, Test of Insufficient Variance, meta-analysis, and pre-registration (10-13). However, these procedures are not applicable to the process of quantitatively predicting replicability of individual studies, the focus of our work.

2. Data.

The manually replicated ("ground truth") replication studies used to train and test our model came from eight projects in psychology and two projects in economics that were manually replicated. Below, we briefly described each project (see **Table S1** for a tabular summary). More details about these projects can be found on their websites listed below.

Table S1 Data Sources.

	Training vs Test	Project	# of studies	Discipline	# of journals	Original study year	Original study methodology	Replication methodology	Replication report publication status
1	Training set	RPP (4)	96	Cog Psych; Social/Personality Psych	3	2008	Experiments & correlational studies	Single-lab, same method	Published
2	Test set I	RRR (14)	8	Cog Psych; Social Psych	4	1988-2014	Experiments	Multi-lab, same method	Published
3-5		ML1 (15), ML2 (16), ML3 (17)	42	Cog Psych; Social Psych	22	1973-2013	Experiments	Multi-lab, same method	Published
6		JSP (18)	16	Social Psych	7	1999-2012	Experiments	Multi-lab, same method	Published
7		SSRP (8)	21	Cog Psych; Social Psych, Economics	2	2011-2015	Experiments	Single-lab, same method	Published
8		Individual Efforts (19)	33	Cog Psych; Social Psych	8	1972-2013	Experiments	Single-lab/Multi-lab, same method	Published
9	Test set II	PFD (20)	57	Cog Psych;	20	2001-2017	Experiments & correlational studies	Single-lab, same method	Unpublished, includes class projects
10	Test set III	EERP (7)	18	Economics	2	2011-2014	Experiments	Single-lab, same method	Published
11	Test set IV	ERW (21)	122	Economics	45	1973-2015	Experiments, correlational studies, & modeling	(i) same data, same code (ii) new data, same methods (iii) same data, new methods (iv) new methods, new data	Published
Total	Test set total = 317		413		80 unique	1972-2017			

2.1. RPP Data.

The Reproducibility Project: Psychology by the Open Science Collaboration (4) completed 100 replicated studies sampled from three top psychology journals published in 2008, using the same procedures as the original studies. All studies fell into either social/personality psychology or cognitive psychology. 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, and they were combined into one record. The final sample consisted of 96 studies. For each target study, we downloaded the original published journal article.

2.2. RRR.

The Registered Replication Report is an initiative and a new article type in the journal Advances in Methods and Practices in Psychological Science (22). Eight replication reports have been released during the time frame of the present project. All studies are classified as either social or cognitive psychology, and each original study was replicated independently by multiple labs, which mimicked the original study protocol as closely as possible. Collectively, these labs produced a single replication report synthesizing the results. One study with null original results was excluded. One report replicated two different studies from the same paper testing different effects, and we treated them as two separate replications. The final sample consisted of eight target studies from seven papers. For each study, we downloaded the original published journal article. In each replication report, the authors unequivocally stated whether they successfully replicated the original study.

2.3. - 2.5. ML1, ML2, ML3.

The Many Lab Project is a large-scale replication project of five waves that are classified as either social or cognitive psychology studies, and each one is informally referred to as "Many Labs" numbered 1 through 5. Thus far, Many Labs 1 (15), 2 (16), and 3 (17) have published their results, first mimicking the original study protocol as closely as possible, and then collectively producing a single replication report synthesizing the results. One study with null original result was excluded. We further eliminated four studies first published prior to 1970 because they were written in a format and style very different from most of the more recently published studies. Two of the remaining studies were replications of the same original study and obtained highly similar results, we therefore combined them into a single record. Additionally, two papers that failed PDF conversion were eliminated. The final sample consisted of 42 target studies. For each study, we downloaded the original published journal article. The replication teams also stated their overall conclusion regarding the replication outcome, like in the RRR. In Many Lab 3, the authors also summarized in Table S3 whether each study is "highly replicable" or "unknown replicability" (i.e., failure to replicate in their attempts). We used the table as indication of replication outcome.

2.6. JSP.

"Replications of Important Results in Social Psychology" is a special issue of the journal Social Psychology that includes 15 replication reports (18). All studies are on social psychology topics. Each original study was replicated independently by multiple labs, mimicking the original study protocol as closely as possible, and all were combined into a single replication report synthesizing the results. One study with null original result was excluded. Report that were already part of the RPP or Many Labs 1 replication studies were excluded to avoid duplications. We further eliminated two studies first published prior to 1970 for reasons explained above in Section 2.3. One study was replicated twice by two independent teams with each team producing conflicting results and was therefore eliminated. Two reports replicated multiple studies from the same paper testing different effects, which we treated as separate target studies. The final sample consisted of 16 studies from 11 papers. For each study, we downloaded the original published journal article. The replication teams also declared their overall conclusions regarding the replication outcome, as was done in the RRR.

2.7. SSRP.

Social Sciences Replication Project was a project that replicated 21 existing social science experiments using the same procedures as the original studies (8). Among them, 18 were psychology studies and three were economics studies. Each study was replicated by a single lab. For each target study, we downloaded the original published journal article.

2.8. Individual efforts.

Besides the large-scale organized replication project, there are also published reports dedicated to replicating one effect at a time. Curate Science (23) is a website documenting and summarizing these individual efforts as well as large-scale collective projects. We collected replication outcomes on 33 psychology studies from 25 papers compiled by Curate Science. For each study, we downloaded the original published journal article.

2.9. PFD.

PsychFileDrawer.org is an online tool designed to archive replication reports (20). Any user can upload such results. Because reports on the website have not been peer-reviewed and published, we treated this dataset with caution and only used it for validation as a test set separate from the rest of the test sample. On the site, there were 88 full, direct replication reports that have not already been included in the other projects at the time of access. All studies are classified as either social or cognitive psychology, and each was replicated by a single lab. Eleven original studies from 27 reports were replicated two or more times. Most of these multiple replications achieved consistent results and were combined into a signal record. Three original studies, however, were reported to have different results across replications. Six replication reports pertaining to these studies were excluded, as were nine conceptual replications (i.e., non-direct replication). The final sample consisted of 57 studies. For each study, we downloaded the original published journal article. The authors also subjectively reported their replication as either a “success” or “failure”.

2.10. EERP.

Experimental Economics Replication Project was a project that replicated 18 existing economic experiments using the same procedures as the original studies (7). Each study was replicated by a single lab. For each target study, we downloaded the original published journal article. All were available in PDF format.

2.11. ERW.

Economics Replication Wiki is a website compiling and storing published replications of empirical studies in economics (21). At the time of access, there were 153 full replication reports on the website clearly defined as “successful” or “failed” replications, and they cover a wide range of economics subdisciplines. Each study was replicated by a single lab. Among these reports, eight original studies from 21 reports were replicated at least twice. These multiple replications achieved either consistent success/failure results or had at least one success/failure result with others partially successful. We combined each set of results into a signal record, further eliminated three studies first published prior to 1970, and removed eight book chapters and seven papers that failed to be converted from PDFs . The final sample was 122 papers. For each study, we downloaded the original published journal article. All papers were available in PDF format. Different from replications in psychology, economics replications are not necessarily direct replications of experiments. For instance, sometimes economists replicate a forecasting model by simply re-analyzing the same data using the same code provided by the original authors while other times they use the same data but verify the same hypothesis with a different analysis strategy. The replication reports were categorized by the website into four types accordingly: (i) same data, same code (N = 26), (ii) new data, same methods (N = 31), (iii) same data, new methods (N = 32), (iv) new methods, new data (N = 28), and N = 5 unknown. We included studies from all four categories in our analysis.

3. Methods.

3.1. Method for Converting Published Articles 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.) related to the manually replicated study 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. For papers with more than one study, we identified texts describing the studies targeted for replication; for single-study papers, we identified methods, results, and discussion sections. Texts pertaining to a target study as well as the discussion sections were the focus of the present research, rather than the full paper itself. Some papers contained multiple studies replicated by separate research teams. In such scenarios, we considered each study a separate record with its own input features (texts) and dependent variable (replication outcome).

A small number of psychology papers were only available in PDFs. After downloading these files, we ran them through the published software GROBID that converts PDFs of scholarly articles into texts and delineates sections based on machine learning (24). This software is not perfect. Sometimes it mislabels sections (F-Score = 0.78) (25), and we have observed that a small percentage of the conversions indeed produced disorganized and unreliable outputs. Because we seek to build a protocol that can be automatically applied to any paper in the future, we did little manual follow-up processing and treated imperfect conversions as noises. As with methods explained above, we followed the same procedure of removing non-textual information, and we kept only the main text of each paper and texts related to the target studies.

On the Economics Replication Wiki site, 116 out of 122 papers were only available as PDFs. Because the rest came from five journals with different HTML formats, it was impractical to write codes to process them individually. We thus applied GROBID to convert all papers into PDFs. Most papers in economics do not have clear boundaries between studies, and sometimes descriptions of key effects targeted for replication are scattered in various parts of a paper. It was difficult to separate studies targeted for replication from the rest. Consequently, we sampled texts from the whole paper rather than only texts pertaining to the replicated study like in psychology papers.

3.2. Model Details.

We built our model in stages to best observe differences in predictive power according to the information used. The first model is based on narrative (text) only; the second is based on a list of original study statistics; the third combines the two. The models were trained and calibrated using the RPP data. Fig. S1 presents a graphic overview of how the narrative-only model is developed. The statistics-only model and the combined models use the same design. Each step is explained in detail below.

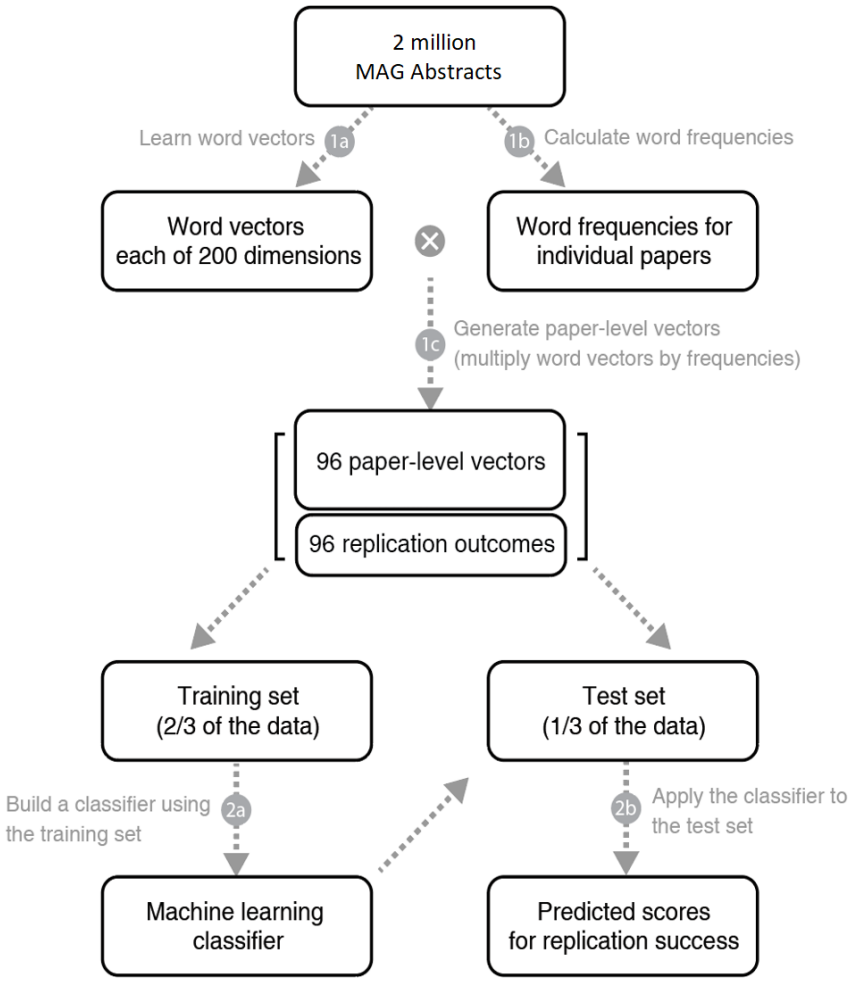

Figure S1 Overview of the Narrative-only Model. Steps 1a – c perform text feature extraction; Steps 2a and 2b perform machine learning and cross-validation.

3.2.1a. Converting Words to Numerical Representations of Words (i.e.,, word vectors). The narrative-only model does not take texts directly as inputs. In this section, we explain how to numerically represent individual words in papers in a way that preserves their semantic meaning. Using two million psychology and economics abstracts from MAG (26) as corpus, we used an algorithm called word2vec (27) to represent each word as a vector using an algorithm so that semantically related words would have similar vector representation.

We first extracted a total of 18 million sentences from two million abstracts. The token size was roughly 0.2 million. For each word in a sentence, we took the five nearby words (windows size) as context words to pair with the target word. Fig. S2 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 nearby words to the right.

It is helpful to visualize the outputs as a sparse co-occurrence matrix. In Fig. S3, the illustrative co-occurrence matrix lists example frequencies of two words being one another's nearby words in the corpus. The co-occurrence matrix allows us to know for a specific word (e.g., "are") how likely another word (e.g., "many") is a nearby word. In theory, matrix factorization or principle analysis can be performed to extract underlying "factors" to efficiently represent the co-occurrence

relationships among all words in the vocabulary (28). Given the size and complexity of word pairs, however, we adopted word2vec as an alternative, conceptually equivalent technique to understand the co-occurrence relationships with more precision. The word2vec model trained on two million MAG abstracts is called mag-200d.

Besides calibrating the word2vec model on the two million abstracts, we also verified the performance of other word2vec models. The first word2vec model was trained on Wikipedia 2017, the UMBC web base corpus, and the statmt.org news dataset (29) and is denoted as wikinews-300d. The second one was trained on English wiki corpus (30) and is denoted as enwiki-300d. The third one was trained on two billion tweets (31) and is named twitter-200d. The fourth model was based on Google news (32) and is named googlenews-300d. In **S3.2.2b** below, we further discuss the performance of these word2vec models.

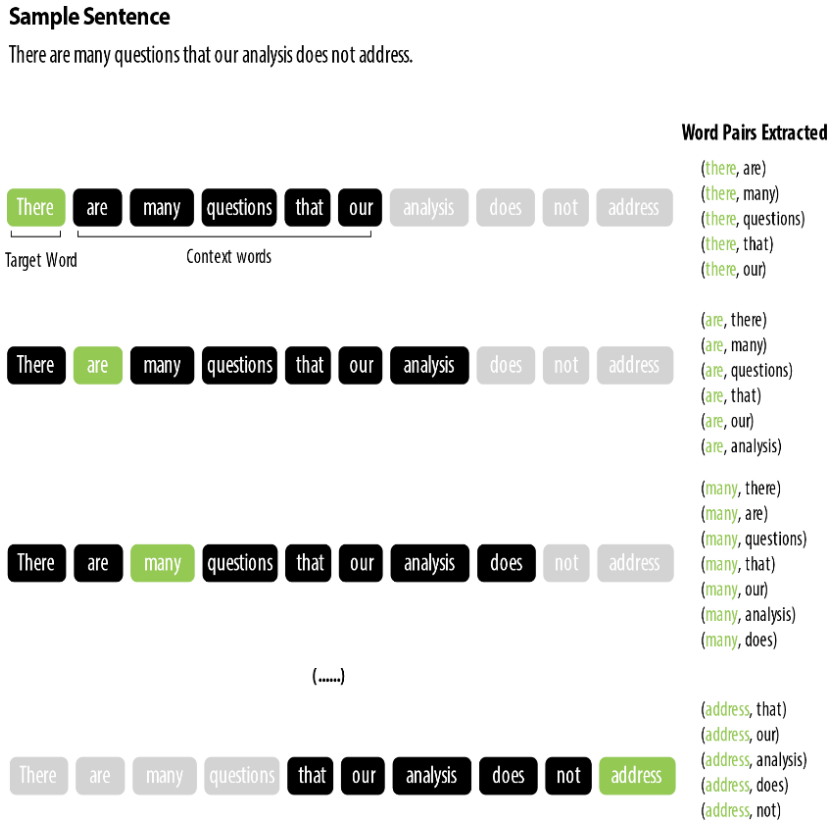

Figure S2 Extracting Word Pairs from Papers Using a Context Window of Five.

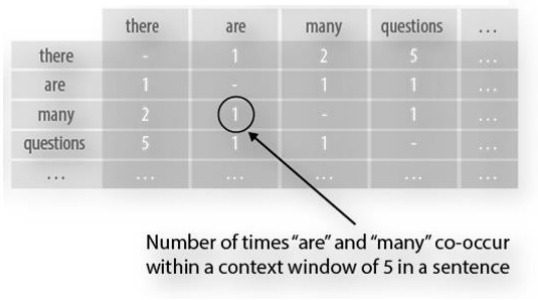

Figure S3 Word Co-occurrence Matrix for a Context Window of Five.

Word2vec is a neural network model with an input layer, a hidden layer, and an output layer. Fig. S4 lays out how this neural network is trained to understand word co-occurrences. The neural network takes one target word as input each time, represented by a vector of 0.2M components (one for every word in our vocabulary). The position corresponding to the word is set to "1" and the positions for all other words in the vector are set to "0". In the example in Fig. S4, the position for the target word "are" is highlighted in green. From the input layer to the hidden layer, there is a weight matrix of 0.2M rows and 200 columns. Each row in this input-to-hidden layer weight matrix is supposed to be the vector representing the corresponding target word. The matrix is initialized randomly and updated periodically. Therefore, multiplying the input vector with this matrix selects the vector for the target word "are." From the hidden layer to the output layer, there is another weight matrix of 200 rows and 0.2M columns. Each column in this hidden-to-output layer weight matrix represents a context word. We multiplied the word vector for "are" with the matrix and used a softmax regression to predict the probability of each word being a nearby word for that target word "are." The predictions were compared with the actual word pair frequencies, and updates were made to the weight matrices retrospectively to better approximate the ground truth. We repeated this procedure multiple times so that every word in the vocabulary was treated as the target word. The final trained model minimized the errors (i.e., maximizes the agreement) between the estimated co-occurrence probabilities and the ground-truth information.

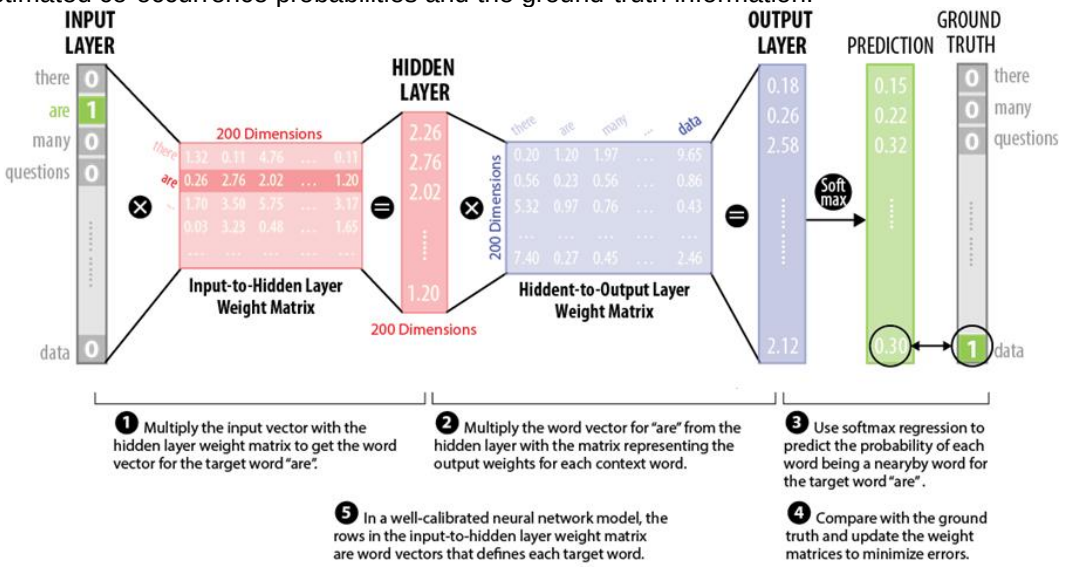

Figure S4 Training Word2vec Three-layer Neural Network Using Word Co-occurrence Information

In a well-calibrated neural network, the rows in the input-to-hidden layer weight matrix are 200-dimensional word vectors that define each word in our vocabulary. The 200 dimensions can be understood as the "factors" that capture the semantic meanings of a full matrix of word pair co-occurrences but in a more economical way. Broadly speaking, this step is akin to factor analysis or principle components analysis in that the neural network reduces a high dimensional variable space to a few factors that capture the same information. However, the neural network procedures are not strictly linear in the way factor analysis or principle component procedures are. These word vectors formed the building blocks of our analysis.

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. This is achieved by multiplying the frequency of words in each paper with word vectors. 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, 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.

Formally, 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.

When we multiplied the TF in a paper with word vectors (derived from step 1a in Fig. S1), we determined that paper's TF vector. Similarly, we multiplied the TF-IDF in a paper with word vectors (step 1c) and derived that paper's TF-IDF vectors. These TF and TF-IDF vectors quantitively represented the content of individual papers. More specifically, we generated the TF and TF-IDF vectors for each paper using the paper's replicated study section and the general discussion section (Fig. S4, red panels and blue panels, respectively). We also calculated the similarity between the paper's replicated section and discussion section to measure whether the findings and conclusions were consistent.

3.2.2a. Build Machine-learning Classifier.

We trained an ensemble algorithm of bagging with a random forest model (33) 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 the logistic regression and conducted several robustness tests, where we found that a maximum depth ranging from two to eight gave us almost identical results.

3.2.2b. Repeated 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 between the three subsets, we could predict replicability scores for the entire sample. The predicted score was a continuous variable with range [0, 1]. To assess accuracy, we examined the

relationship between the continuous replicability score and our binary outcome variable of replicability, and we calculated the accuracy and top k precision. To gauge performance variability due to sampling, we further iterated the above procedure 100 times based on a new random three-subset partition of the data that we applied every time. Accuracy and top-k precision were calculated after each iteration to get a distribution of 100 scores per paper.

In 67 out of 100 cross-validation rounds, the accuracy of the narrative-only model was better than the reviewer metrics. And in 73 out of 100 cross-validation rounds, the top-k precision of the narrative-only model was better than the reviewer metrics. Based on binomial testing, the p-values were less than 0.01.

In 65 out of 100 cross-validation rounds, the accuracy of the combined model (narrative + reviewer metrics) was better than that of the narrative-only model, and in 64 out of 100 cross-validation rounds, the top-k precision of the combined model was better than that of the narrative-only model. Based on binomial test, the p-values were less than 0.01 in terms of accuracy and top-k precision.

The average accuracy of the narrative-only model was 0.69 (SE = 0.032), and the top-k precision of the narrative-only model was 0.75 (SE = 0.026). As we mentioned in **S3.2.1a**, we tested the performance of our method using four different word2vec models. When using wikinews-300d, the accuracy of the narrative-only model was about 0.66 (SE = 0.030), and the top-k precision was 0.72 (SE = 0.025). When we replaced the mag-200d with enwiki-300d, the accuracy and top-k precision of the narrative-only model were 0.66 (SE = 0.028) and 0.72 (SE = 0.023), respectively. Similarly, the accuracy and top-k precision of the narrative-only model when using googlenews-300d were 0.63 (SE = 0.034) and 0.70 (SE = 0.028) The accuracy and top-k precision of the narrative-only model when using twitter-200d were 0.61 (SE = 0.038) and 0.68 (SE = 0.031). Training word2vec on academic paper abstracts (MAG dataset) clearly produced the best performances, with the runner-up performances produced by the models trained on Wiki corpus and on news. The model trained on tweets had the worst performance.

3.3. Comparing the Narrative-only, Statistics-only, and Combined Models.

To benchmark the model's performance against existing measures of replicability, we built a prediction model with a collection of replicability-relevant metrics introduced previously in section 1 (Replicability Literature Review), including original study p-values, original sample sizes, original effect sizes, and original statistical powers. For the 96 RPP papers, we collected four variables computed from original studies by the RPP team, and we referred to them as the "reviewer metrics." Table S2 includes full descriptions of these variables and their data sources. We used the same random forest algorithm used above in a narrative-only model (step 3.2.2a-b) but used the four reviewer metrics as input features rather than narrative vectors. We also trained a classifier using both the reviewer metrics and narrative vectors. Both procedures involved the repeated three-fold cross-validations, and hence each produced a distribution of 100 accuracy/top-k precision scores. We also measured the differences in performance distributions between these models and the narrative-only model to assess whether the narrative-only model improved upon the reviewer metrics model and whether the two could be combined to achieve higher accuracy.

3.4. Out-of-sample Tests.

We conducted out-of-sample tests to examine the generalizability of the model, testing 317 papers from 80 different journals in diverse disciplines that had undergone manual replications. The results in the main text showed that the model does well in predicting the replicability of papers it has never seen. We grouped the eleven additional replication projects introduced in Data into four samples: (i) 117 high-quality, multi-lab psychology publications from RRR, ML1-ML3, JSP, SSRP, and individual efforts compiled by Curate Science (three Economics papers were removed from SSRP); (ii) 57 unpublished psychology publications from PFD; (iii) 18 published economics experiments from EERP; and (iv) 122 economics publications from the ERW (citations to all datasets appear in Table 1 in main text). We applied the narrative model calibrated on the RPP to the out-of-sample

studies to compute predicted replicability scores. Paper-level narrative vectors were generated in the same way as described in section 3.2.1b-c. Not all four reviewer metrics were available for these out-of-sample studies. For sample (i), (ii), and (iii), we included original p values, effect sizes, sample sizes, and post-hoc power estimates. Most of these metrics were summarized by the replication team, and in cases where they were not reported, we extracted them from the original study. For sample (iv), none of the reviewer metrics were reported on the website. Furthermore, because the original study type varies in sample (iv), metrics such as P values, effect sizes, and sample sizes were either not applicable or inconsistent with the measures used in the other samples. We therefore did not include reviewer metrics and applied the narrative-only model to sample (iv). The machine learning algorithm used in out-of-sample tests was a simple combination of random forest and logistic regression. Full results are described in detail in the main text.

3.5. Comparison with Prediction Markets and Survey Beliefs.

The machine's performance (out-of-sample) was further compared to that of the prediction markets and surveys, two prevailing approaches. Both prediction market and survey data were available for a subset of N = 100 studies from four replication projects: RPP (N = 41), ML2 (N = 20), SSRP (N = 21), and EERP (N = 18). For EERP, we used pre-market survey ratings; for the other three projects, we used weighted survey ratings.

3.6. Four Statistical Tests to Compare Narrative and Reviewer Metrics.

We performed four statistical tests to confirm the significant differences between the narrative based model and the reviewer metrics based model. These four statistical tests were the KS test, the Anderson Darling test, the Camerer-von Mises test, and the Wilcox Rank Sum test. All confirmed that the narrative based model had a better record of performance than the reviewer metrics based model ($p < 0.001$) in terms of both accuracy and top-k precision.

3.7. Cross-Validation Tests on Full Data.

To further investigate the prediction power of our method, we conducted cross-validation tests on the combination of all psychology papers from RPP, RRR, ML1, ML2, ML3, JSP, SSRP, Individual Efforts, and PFD. The accuracy and top-k precision of the narrative-only model were 0.65 (SE=0.021) and 0.72 (SE=0.017), which are very similar to those in the RPP dataset. The accuracy and top-k precision of the narrative-only model as used on the combination of all economics papers (the majority came from ERW) were 0.61 (SE=0.032) and 0.67 (SE=0.027).

The average accuracy and top-k precision measurements of the narrative-only model applied to all papers in our study were 0.63 (SE=0.017) and 0.69 (SE=0.014). As mentioned in section S3.1, the low text quality of ERW may have contributed to the performance decrease, and if we were to remove the ERW from the data, the accuracy and top-k precision of the narrative-only model would increase to 0.66 (SE=0.019) and 0.72 (SE=0.016).

4. Machine Model Bias Inheritability Analysis.

We conducted a series of analyses exploring the mechanism of the narrative-only model, as described in the main text and summarized below.

4.1. Prestige of Authors and Institutions.

Adding author or affiliation prestige information (author's citation and institution rank) to the training RRP training data did not produce an accuracy significantly greater than the model without author or affiliation information (p=0.36).

4.2. Sex of Authors.

Adding the sex of authors to the RRP training data did not produce an accuracy statistically greater than the model without sex of author information (p=0.49).

4.3. Disciplinary Quality Perceptions.

Adding an indicator of social/cognitive assessment to the RRP training data did not produce an accuracy statistically greater than the model without the information (p=0.29). Cognitive psychology had a replication base rate of 53% whereas social/personality psychology had a replication base rate of 28%.

4.4. Journal Prestige.

Papers based in both the cognitive and social subdisciplines of psychology were sampled from different journals, and no significant differences (p=0.16) in the reproducibility rate were found.

4.5. Logistic Regression Analysis.

We also ran logistic regression on RPP datasets in which we included both the first and senior authors' citations, institutional ranks, and genders, all indicators of social and cognitive studies with journal fixed effect. We found that none were significantly indicative of replicability. The p-values of their coefficients' significances were 0.86, 0.93, 0.08, 0.63, 0.42, 0.87, and 0.38.

4.6. Content Words and Function Words.

We used content words and function words to further explore whether the AI model distinguished between replicated and non-replicated papers. We generated paper-level vectors for all studies according to the procedures included in Fig. S1, but with all the nouns (representing content words) and stop words (representing function words) removed in step 1b. The procedures remained the same for step 2. We obtained two distributions of cross-validated accuracies and top-k precisions with functions and content words removed and conducted KS tests comparing them with the main results that employed all words.

4.7. Word Frequency.

We conducted a simple word frequency analysis to observe if our AI model could be simplified to achieve any linear effect between word choices and replication outcomes. Specifically, we extracted individual words from 273 papers included in the eight replication projects, keeping only words used in at least 1% of the papers. All word and phrase counts were normalized by each paper's total number of words used. We correlated the normalized frequency with the replication outcome. Since we were exploring many variables at once, we treated a test as significant if it had less than a Bonferroni-corrected two-tailed p value of 0.05. Only one correlation—the word "experimenter" ($r = -0.26$, $p < 0.001$)—withstood the test. Two-hundred-seventy papers provided a limited sample for such a test, thus this analysis, while informative, should be continually expanded upon as more data becomes available.

4.8. Words of Estimative Probability and Persuasion.

We examined whether research findings in replicated and non-replicated papers are described with different degrees of certainty and confidence. The U.S. intelligence agency have compiled a list of words of estimative probability (34) used in analytical reports. These words and phrases describe the certainty of a statement being made or the likelihood of an event happening (35). Terms like "almost certainly" or "highly likely" convey high confidence, while phrases like "little chance" or "highly unlikely" suggest high uncertainty. We counted the occurrences of these words of estimative probability in 273 papers from the eight psychology replication projects. All word and phrase counts were normalized by each paper's total number word count. We then weighed the frequencies by the subjective probability assigned them and summed across each paper to get an overall estimative probability score. Fig. S5 compares the distribution of estimative probability scores between non-replicated papers and replicated papers and shows no significant differences ($t = 0.20$, $p = 0.84$), suggesting a similar level of confidence in the language.

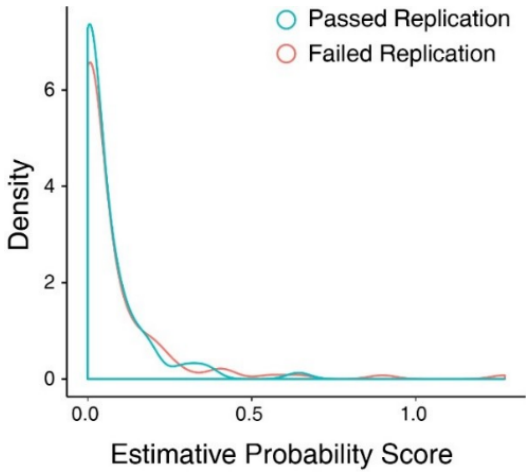

Figure S5 Replicated and Non-replicated Papers Use Similar Persuasion Words in Similar Frequencies. The estimative probability score was calculated for each paper by counting occurrences of terms included in a list of probabilistic words (e.g., “highly unlikely” or “almost certainly”) compiled by U.S. intelligence agencies and quantified with subjective likelihood.

4.9. LIME Explanation of AI model.

Local Interpretable Model-agnostic Explanations (LIME) (36) is an explanation technique that clarifies what data was used in what predictions by a classifier in an interpretable manner

LIME can identify the most important words that make a classifier believe a paper is replicated or not. Given a calibrated classifier C (our AI model trained on the training set), a interpretable data representation W (bag of words), and a single paper p (i.e., a paper in the test set), LIME is able to tell the important top-k words in W that make the classifier C think the paper p is replicated or not.

To identify top-k important words, LIME first performs a perturbed sampling around a single paper p, sampling incarnations of the paper p by drawing words uniformly at random from the original set of words in p. Each incarnation can be represented as a vector where a single element indicates whether a word W is included or not. The calibrated classifier C estimates new probabilities of replication based on these perturbed samples. Then, by using these perturbed vectors (regressors) and newly predicted probabilities (dependent variable), we can run a linear regression and get an estimation of the words (based on coefficients in the linear regression) most important to making the classifier think the paper p is replicated or not. This procedure is done locally, so the explanation for each paper p may vary from the others.

We applied LIME on our AI model and found that there are no common words in different papers, which makes machine consider it as replicated or non-replicated (see Fig. S6). In the whole corpus, we did not find any significant signal of single words that lead to machine’s judgement of replicability. This again suggests we should look into the high-dimension information embedded in narratives.

In our whole corpus, we do not find any significant signal of single words that lead to machine’s judgement of replicability. For example, the reason that machine considered paper 36 as non-replicated because it has words, such as “learning” and “experiment”. In contrast, the reason that machine considered paper 161 as replicated because it has words, “relevant”, “stimuli”, and “experiment”. This suggests us to look into the high-dimension information embedded in narratives.

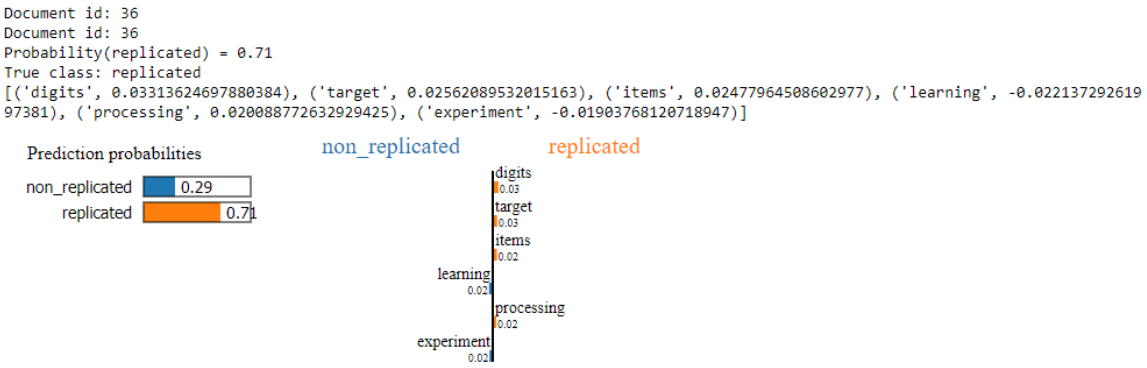

Figure S6 A Sample Output of LIME. The output of LIME provides the top six words that are considered crucial for the machine's judgement. Here, machine considered paper 36's probability to be replicated at 0.71. Machine considered it replicated because it included words such as "digits", "target", "items," and etc. and considered it non-replicated because it included words like "learning" and "experiment." In contrast, machine found paper 161 to be replicated because it included the words "relevant," "stimuli," and "experiment." In our whole corpus, we did not find any significant signal of single words that lead to machine's judgement of replicability. This again suggests we investigate the high-dimension information embedded in narratives.

4.10. N-grams Analysis

We can count the number of times a specific n-gram appears in the entirety of the corpus and compare it to its frequency of occurrence by chance. We can generate the simulated chance data for each paper by preserving the number of sentences and number of words but shuffling the order of words within each individual sentence. We ran this simulation 10,000 times for each paper. Then, for each n-gram, we found 10,000 corresponding simulated occurrences. Trivially, each n-gram now had a corresponding z-score:

$$Z(n-gram) = ((real - avg(sim)))/std(sim)$$

In this way, each n-gram in a paper had a corresponding z-score now, and each paper had a distribution of z-scores. We found that with the increase of N (dimension), replicated and non-replicated papers alike showed significant differences in terms of their frequency of use of unusual and common n-grams. Non-replicating papers used more unusual n-grams and fewer common n-grams than did replicating papers (KS tests, p <0.02).

Table S2 Association Between the Reviewer Metrics and Replicability Reported in Prior Research.

Variable	Description	Data Source	Association between the reviewer metric and replicability reported in prior research
Original P value	P-values of the original study grouped into four intervals recommended by the RPP	RPP(4)	*r* = -0.26
Original effect size	Standardized effect size of the original study	RPP	*r* = 0.28
Original df or *N*	Sample size of the original study	RPP	*r* = -0.29
Post hoc power	Computed by converting the test statistic into a z-score and then determining the probability of obtaining that z-score if there is a true effect in the population	Calculated according to Schimmack's method (3)	57% classification accuracy

SI References

1. P. Patil, R. D. Peng, J. Leek, A statistical definition for reproducibility and replicability. *bioRxiv*, 066803 (2016).
2. J. M. Hoenig, D. M. Heisey, The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* **55**, 19-24 (2001).
3. U. Schimmack, A. Laughton (2015) Predictions about Replication Success in OSF-Reproducibility Project.
4. Open-Science-Collaboration, Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
5. J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, D. A. Reinero, Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences* **113**, 6454-6459 (2016).
6. A. Dreber *et al.*, Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences* **112**, 15343-15347 (2015).
7. C. F. Camerer *et al.*, Evaluating replicability of laboratory experiments in economics. *Science* **351**, 1433-1436 (2016).
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. E. Forsell *et al.*, Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology* (2018).
10. U. Simonsohn, L. D. Nelson, J. P. Simmons, P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General* **143**, 534 (2014).
11. U. Simonsohn, L. D. Nelson, J. P. Simmons, p -Curve and Effect Size : Correcting for Publication Bias Using Only Significant Results. 10.1177/1745691614553988 (2014).
12. U. Simonsohn, J. P. Simmons, L. D. Nelson, Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General* **144**, 1146-1152 (2015).
13. B. A. Nosek, D. S. Lindsay, Preregistration becoming the norm in psychological science. *APS Observer* **31** (2018).
14. Registered Replication Reports (https://www.psychologicalscience.org/publications/replication).
15. R. A. Klein *et al.*, Investigating variation in replicability: A "many labs" replication project. *Social Psychology* **45**, 142-152 (2014).
16. R. A. Klein *et al.*, Many Labs 2: Investigating variation in replicability across sample and setting. (2018).

17. 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).
18. B. A. Nosek, D. Lakens, Replications of Important Results in Social Psychology [Special Issue]. *Social Psychology* **45** (2014).
19. A. A. Aarts, E. P. LeBel, Curate science: A platform to gauge the replicability of psychological science. (2016).
20. H. Pashler, B. Spellman, S. Kang, A. Holcombe, PsychFileDrawer: archive of replication attempts in experimental Psychology. *Online*< *http://psychfiledrawer. org/view_ article_list. php*.
21. J. H. Höffler, ReplicationWiki: Improving Transparency in Social Sciences Research. *D-Lib Magazine* **23**, 1 (2017).
22. R. R. Reports (Registered Replication Reports). (https://www.psychologicalscience.org/publications/replication).
23. C. Science (Curate Science. (http://curatescience.org/#about).
24. P. Lopez, GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. 10.1007/978-3-642-04346-8_62.
25. M. Singh *et al.*, OCR ++ : A Robust Framework For Information Extraction from Scholarly Articles.
26. K. Wang *et al.*, A Review of Microsoft Academic Services for Science of Science Studies. *Frontiers in Big Data* **2**, 45 (2019).
27. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space. 10.1162/153244303322533223, 1-12 (2013).
28. O. Levy, Y. Goldberg, Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, pp 2177-2185 (2014).
29. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405* (2017).
30. I. Yamada, A. Asai, H. Shindo, H. Takeda, Y. Takefuji, Wikipedia2Vec: An Optimized Tool for Learning Embeddings of Words and Entities from Wikipedia. *CoRR* (2018).
31. J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp 1532-1543 (2014).
32. GoogleNews-vectors-negative300.bin.gz. (2013).
33. L. Breiman, Random Forests. *Machine Learning*, 5-32 (2001).
34. S. Kent, Words of estimative probability. (1964).

35. J. A. Friedman, R. Zeckhauser, Handling and mishandling estimative probability: likelihood, confidence, and the search for Bin Laden. *Intelligence and National Security* **30**, 77-99 (2015).
36. M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM), pp 1135-1144 (2016).