The Language of (Non)replicable Social Science

Michal Herzenstein[*†]
Lerner College of Business and Economics
University of Delaware

Sanjana Rosario[*]
Columbia Business School
Columbia University

Shin Oblander[*]
Sauder School of Business
University of British Columbia

Oded Netzer[*]
Columbia Business School
Columbia University

Abstract

Using publicly available data from 299 pre-registered replications from the social sciences, we find that the language used to describe a study can predict its replicability above and beyond a large set of controls related to the paper characteristics, study design and results, author information, and replication effort. To understand why, we analyze the textual differences between replicable and nonreplicable studies. Our findings suggest that the language in replicable studies is transparent and confident, written in a detailed and complex manner, and generally exhibits markers of truthful communication, possibly demonstrating the researchers' confidence in the study. Nonreplicable studies, however, are vaguely written and have markers of persuasion techniques such as the use of positivity and clout. Thus, our findings allude to the possibility that authors of nonreplicable studies are more likely to make an effort, through their writing, to persuade readers of their (possibly weaker) results.

Keywords: Open Science, Replication Prediction, Text Analysis, Psychometric Properties of Language, Machine Learning Models, Computational Social Sciences.

[*] The authors contributed equally to this work.
[†] Michal Herzenstein is the corresponding author: michalh@udel.edu

Research Transparency Statement

General Disclosures

The authors declare no conflicts of interest. Funding: This research was supported by the authors' universities. Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research complies with ethical norms of secondary data collection and storage. Computational reproducibility: Code files and readme are uploaded to OSF, along with processed data files and results files.

Study Disclosures

Preregistration: This study was not preregistered. Materials: The list of original articles and their replication papers is publicly available at https://osf.io/qy8ev/. Data: The processed text and metadata are publicly available at https://osf.io/qy8ev/. Analysis scripts: The code is publicly available at https://osf.io/qy8ev/.

Statement of Relevance

The language used in academic studies in psychology and behavioral economics predicts whether their findings were successfully replicated by other researchers, which is important due to the growing concern of low replicability rates. To understand why, we examine the textual differences between replicable and nonreplicable studies. Replicable studies often have elaborated and confident narratives, which have been shown to be markers of truth-telling. Nonreplicable studies are often written vaguely and exhibit clout and positivity. Therefore, our results suggest that the way research is written likely reflects its authors' hunch about the veracity of their studies. Because these differences are mostly based on context-free language such as adjectives, quantifiers, and pronouns, we believe our results are relevant for the open science efforts in the social sciences and possibly other disciplines. However, given the relatively small sample of replication attempts, we advise repeating our analyses as more manual replications are published.

In a survey of over 1,500 scientists (Baker, 2016), 70% reported they tried and failed to replicate another scientist's experiment, and roughly 50% admitted they are sometimes unable to replicate their own work. When asked why, the answers alluded to "sloppy research" conduct—selective reporting, low statistical power and poor analysis, poor experimental design, and insufficient oversight. In this paper, we examine whether collectively these practices manifest in the way academic studies are written.

We hypothesize the answer is yes, because written words carry implications beyond their literal meanings. For example, word choices, whether conscious to the writer or not, have been associated with writers' state of mind (Ventrella, 2011) and intentions (Netzer et al., 2019). While it could seem obvious that writers reveal information about their mindset with their word choices in informal interpersonal communication, writers have also been shown to make such disclosures even in more formal and curated texts such as poems (Pennebaker, 2011), loan applications (Netzer et al., 2019), and presidential communications (Van Der Zee et al., 2021). Even when a text is edited by multiple authors, it carries valuable information. For example, the language in companies' 10-K filings has been associated with the company's stock return and volatility, trading volume, fraud, and unexpected earnings (Loughran & McDonald, 2011). Remarkably, the information embedded in word choice has been documented even after controlling for observed and verified information related to the text writer, such as credit scores when asking for a loan.

In this research, we explore the relations between the language used in academic studies and their replicability likelihood. Past research has established that metadata related to the paper, its authors, and the analyses' statistics can predict its replicability (Altmejd et al., 2019), and study text along with statistics related to its analysis is similarly predictive (Yang et al., 2020;

Youyou et al., 2023). In this paper we aim to understand whether the text is predictive of replicability even after controlling for a rich set of metadata variables related to the paper, study design, authors, and replication study. We find that the answer is yes, and this finding represents our first contribution. We then take the next step and aim to understand why the text has predictive abilities. Specifically, we explore how the language in replicable and nonreplicable papers differ, and whether understanding these differences can shed light on why language helps predict replication likelihood. We do so by complementing the machine learning textual features with linguistic style metrics. Indeed, using uninterpretable machine learning representations of the text in academic papers has been a limitation of past research, as Crockett et al. (2023) and Mottelson & Kontogiorgos (2023) point out.

Prior studies explored the relationship between language and the veracity of the research it describes in a variety of settings. For example, nonreplicable studies use more rare word combinations than replicable studies (Yang et al., 2020), and AI written fake research is more likely to include unusual language instead of common terms (e.g., “colossal information” instead of “big data”; Cabanac et al., 2021). In a similar, albeit more extreme vein, fraudulent research has been shown to include more words associated with deception, fraud, and obfuscation of information (Markowitz & Hancock, 2014; 2016). Our research extends these studies in several ways. First, we control for an extensive set of metadata variables (e.g., the paper’s keywords) to distinguish writing style from merely different research topics. Second, we explore a broad range of linguistic features such as writing style dictionaries. This allows us to better understand the role of the text in predicting replicability.

We contribute to the movement toward Open Science that aims to increase the openness, integrity, and reproducibility of scholarly research. As part of this movement, many researchers have attempted to replicate published papers with only a moderate rate of success. We assembled a dataset of 299 studies in psychology and behavioral economics whose replications were published and were not done by the original authors. Our data include information about the original paper, focal study (the one that other labs attempted to replicate), authors, and the text used in the abstract, focal study, and the entire paper, as well as information about the replication study.

Our first finding is that, controlling for a large set of metadata variables, the text in the focal study and in the entire paper improve predictions of replicability in holdout samples above and beyond a predictive model that uses only the metadata. We find this result consistently, in different slices of the dataset, with different methods of text analysis, and underlying models. Our large set of metadata variables allows us to alleviate concerns that the text reflects the authors' characteristics, the study's design and objective statistical power, how the paper was selected for replication (systematically or not, based on the replication project), the quality of the replication, the subfield of the paper, and even its general topic (e.g., goals, attitudes, or economic games).

To unpack the role of the text in predicting replicability, we utilize the Linguistic Inquiry and Word Count (LIWC 2015, www.liwc.app) dictionaries, which are a set of 92 nested and context-free psychometric dictionaries, along with measures of abstraction, obfuscation (Markowitz & Hancock, 2016), readability (Flesch, 1948), and narrative arcs that describe the structure of stories (Boyd et al., 2020). Controlling for the metadata, we find that the language in academic studies likely reflects their authors' intuition regarding their veracity, which may explain the language's predictive ability of replicability. Specifically, we find that replicable studies are often written in an elaborate and complex manner that expresses the writer's confidence in the research. Conversely, nonreplicable studies are usually written more vaguely but with clout and positivity, and exhibit an archetypical pattern of story arcs that poses a dilemma (or conflict) and then resolves it. These results suggest that authors of nonreplicable studies might make an effort through their word choices to influence their reviewers and readers to accept the paper's claims despite presenting possibly weaker evidence (Dahlstrom, 2014). While our findings are robust to multiple analysis methods, we note that the sample size of manual replication efforts is relatively small at present.

Compiling the Dataset

Using publicly available data on replication efforts of original studies in psychology and behavioral economics, we compiled a dataset of 299 studies: 96 studies were replicated by Open Science Collaboration (2015; RPP), 49 by Many Labs (Klein et al., 2014; 2018; 2022; Ebersole et al., 2016; ML), 18 by Camerer et al. (2016; EE), 22 by Camerer et al. (2018; SSRP), 8 by Zwaan et al. (2018), and a range of individual replications that were pre-registered, published in well-regarded journals, and were not performed by the original authors. Table S1 in the Supplemental Material lists the replication efforts in our dataset, and Table S2 presents the list of original papers included in our analyses and their published replication effort.

We collected four types of measures on the original and replication papers: (i) Following Altmejd et al. (2019), our focal dependent variable is a binary indicator of whether a study is replicated, based on the assessment of the replication team. Effectively in most replication studies, this means that the replication effort found a significant effect ($p \leq 0.05$) in the same direction as the original paper. Overall, 42% of the replication attempts in our dataset were successful. Our second dependent variable is the end price in prediction markets in which participants were experts from the field who bid on the likelihood a replication would be successful before it was carried out. It is a relevant dependent variable in our context because it helps assess whether after reading the paper, and possibly being influenced by its language, these experts could predict the paper's replicability. The experts received the paper, the hypothesis to be replicated and replication plan, and then traded contracts that pay $1 if the study was successfully replicated and $0 otherwise. Dreber et al. (2015) explain that this type of contract allows the end price to be interpreted as the predicted probability that the study would successfully replicate. We have end prices on 99 studies (Camerer et al., 2016; 2018; Dreber et al., 2015; Forsell et al., 2019); (ii) metadata from the original papers: the paper's discipline (social psychology, cognitive psychology, or economics); 45 keywords or JEL codes (see Supplemental Material for how we processed the keywords); publication year; information about the authors (number of authors, proportions of male authors, and of full professors); citation count collected from Google Scholar; the focal study's effect type (correlation, main effect, or interaction); number of participants; who they were (students, community, online, anyone); whether the study was done in the United States or elsewhere; and statistics reported in the text of the focal study (effect size converted to r, p-value, post-hoc power); (iii) metadata from the replication papers: whether the original author(s) advised the replication team, and indicators of the replication project—RPP, ML, EE, SSRP, or other; and (iv) the text of the original papers, broken into abstract, full text, and focal study. Some of our metadata come from Altmejd et al. (2019) and the rest was collected by us. Further details about our data collection effort (including how we handle missing data) and summary statistics are in the Supplemental Material.

We collected a secondary dataset of 2,420 papers from the same journal issues as the replicated papers, which contains papers from many domains, including the hard sciences, in order to train our text representation model (word embeddings). Training the model on the text in academic papers makes our representation learning model more relevant to our context than pretrained embedding models.

Processing the Text

We processed the text in each section in several ways. First, we created text embeddings using the Gensim library in Python to train a Word2Vec model (Mikolov et al., 2013) on a secondary dataset of 2,420 academic articles. We then used the trained embedding model to obtain 100-dimensional vector representations of the text in the original paper by averaging the word embeddings across all words in the relevant documents (abstract, focal study text, or full text). We use these averaged embeddings as features in our predictive analysis. Second, we rely on a well-researched and context-free dictionary, the Linguistic Inquiry and Word Count, that classifies words into 92 meaningful nested dictionaries, and calculated the frequency of each LIWC dictionary in the text. We filtered and cleaned up the dictionaries to ensure they are meaningful in our context (see Supplemental Material for more details); this was necessary because our starting point was all the dictionaries, in contrast to prior work that used a handful of dictionaries to test specific hypotheses related to academic publications (Markowitz & Hancock 2016; Wheeler et al., 2021). Third, for each section of the text we calculated the Flesch (1948) Reading Ease score, abstraction and obfuscation indexes (Markowitz & Hancock, 2016). Fourth, we passed the text files through the algorithm in arcofnarrative.com to obtain story arc scores that describe the flow of the narrative.

The Text in Academic Publications Alludes to Their Replicability Likelihood

Method

To predict the paper replicability using a host of metadata variables and textual features, we evaluate several machine learning models (ridge regression, elastic net, and XGboost), various ways to process the text (indicators for unique words, topic modeling, embeddings with different hyper parameters, and LIWC), different subsets of the metadata (excluding the variables capturing study results (effect size and p-value), or using only them following Yang et al. (2021)), and two sizes of train-test split (80%-20% and 70%-30%). We calibrate the model tuning parameters (e.g., ridge penalty) using 10-fold cross-validation within the training set, then estimate out-of-sample performance using the predicted values on the test set. Ridge models were best-performing, and the other variations were not meaningfully different, hence we present here results with ridge, 80%-20% split, and all controls. We show a meaningful subset of other models in Tables S7-8.

Results

Table 1 presents the replicability predictive ability of the text in each section of the paper separately—focal study, full text, and abstract as well as the metadata variables. Predictive accuracy is measured as the Area Under the Receiver Operating Characteristic Curve (AUC), and is compared across six models: metadata only, text embeddings only, embeddings and metadata, text features only (LIWC dictionaries, arc of narrative, and Flesch readability. We do not include obfuscation and abstractions in this analysis because they are nested within LIWC dictionaries), text features and metadata, and lastly all three sets combined. For the focal study text and full text, we find that the model that includes all text (i.e., embeddings and textual features) and metadata predicts replication better than the model that includes only the metadata ($\text{AUC}_{\text{metadata+all study text}}=0.725$ and $\text{AUC}_{\text{metadata+all full text}}=0.716$ versus $\text{AUC}_{\text{metadata}}=0.696$). The $\text{AUC}_{\text{metadata}}$ is the same for the focal study and full text because the dependent variable is at the study level. These AUCs are averaged over 10,000 random train-test splits, which allows us to assess the predictive improvement's reliability—the proportions of runs in which the model that includes the text and metadata predicts better than the model that includes only metadata are 70.30% for the study text and 64.08% for the full text. Interestingly, the text itself conveys substantial information about the paper's replicability likelihood as the models that predict replicability based on text features alone perform similarly to the ones that use only the metadata ($\text{AUC}_{\text{text embeddings}}=0.703$ and $\text{AUC}_{\text{text features}}=0.683$ for the focal study text, and $\text{AUC}_{\text{text embeddings}}=0.699$ and $\text{AUC}_{\text{text features}}=0.671$ for the full text versus $\text{AUC}_{\text{metadata}}=0.696$), suggesting that word choice captures roughly as much information about replicability as a comprehensive set of metadata variables. Looking at the interpretable text features, LIWC dictionaries, narrative arc, and readability, we find that they also improve predictions above and beyond the rich set of metadata ($\text{AUC}_{\text{metadata+study text features}}=0.713$ and $\text{AUC}_{\text{metadata+fulltext text features}}=0.702$ versus $\text{AUC}_{\text{metadata}}=0.696$). The text in the abstract alone is not as informative about the paper's replicability (Table 1, Panel C), which is not surprising given that many journals are quite prescriptive about how the abstract should be written (e.g., third person, present tense). Running the models with only papers from psychology (275 studies) led to similar results ($\text{AUC}_{\text{metadata+all study text}}=0.717$ and $\text{AUC}_{\text{metadata+all full text}}=0.709$ versus $\text{AUC}_{\text{metadata}}=0.686$).

In practice, when reading academic papers, readers often use their experience with prior papers to predict the replicability of newer papers. Accordingly, we test whether the text in older papers helps predict the replicability likelihood of newer papers. We split our dataset into older (published before 2012) and newer papers (published in 2012 or later), resulting in an approximately 80%-20% split for train and test samples. We find that textual information improves predictive ability even when split over time ($\text{AUC}_{\text{metadata+all study text}}=0.820$ and $\text{AUC}_{\text{metadata+all full text}}=0.795$ versus $\text{AUC}_{\text{metadata}}=0.673$), replicating our main result, and suggesting that the textual signals of replicability are persistent over time.

Finally, we tested whether the text captures similar information to the intuition of academic experts who bet *a priori* on the replicability likelihood of these papers. Since the prediction market outcomes were not used in training our models, we can treat them as another form of prediction test. Inspired by Camerer et al. (2016), we calculate the correlation between our models' predicted probability a study would replicate with the prediction market ending prices, and find that the correlation improves with the addition of the text ($r_{\text{metadata+all study text}}=0.615$ and $r_{\text{metadata+all full text}}=0.614$ versus $r_{\text{metadata}}=0.513$), highlighting that the text carries important replicability signals that participants in the prediction markets were able to detect. Put differently, the improvement in correlation that comes with the addition of the text suggests that prediction market participants' estimations of replicability made use of the paper's textual information (whether explicitly or implicitly).

In sum, expanding results documented by past research (Altmejd et al. 2019; Yang et al. 2020; and Youyou et al. 2023), we find that the language used in academic publications improves predictions of their replicability even after controlling for extensive metadata variables directly related to the probability of a successful replication, such as the subfield and keywords (e.g., some topics are easier to replicate), type of effect (e.g., main effects are more replicable than interactions; Altmejd et al., 2019), study statistics (Yang et al., 2020), and whether the original authors helped with the replication. Therefore, the improved predictive effect of the text features is likely driven by the writing style of the study rather than merely the topic of the paper or ease of replicability. We elaborate on these aspects next.

*** Table 1 here ***

The Language of (non)Replicability

Method

Why does the text in academic publications contain information regarding replicability beyond what is captured by the metadata? We hypothesize that authors' word choices likely reflect their intuition about their study's veracity. Because papers often include multiple studies and authors may be more confident about the replicability likelihood of some studies than others, it is expected that the language used to describe any specific study will be more predictive of its replicability than the language used in the entire paper. This premise is corroborated by our findings that the improvement in the replicability predictions of the text of the focal study is higher than that of the full text of the paper (the percent of runs in which the model including the metadata and text features predicted replicability better than the model with metadata was only is 64.5% for the study text compared with 55.4% in the full text, see Table 1). Therefore, in this section we focus on the language authors used in the focal study.

We ran multiple LASSO regressions on the text features, controlling for all the metadata variables (i.e., with no regularization on the metadata), which ensures that the linguistic features we identify do not merely reflect differences in disciplines' conventions (e.g., some disciplines write more parsimoniously than others) and subject matters (as captured by the paper's keywords. We note that in these analyses we used fewer keywords due to identifiability constraints. See Table S6 for that list), differences over time (e.g., older papers may document fewer results), or the result of more or less experienced original research teams or replication teams. To remedy for the problem of multicollinearity in LIWC dictionaries, we entered only the low-level dictionaries into LASSO. For example, the low-level dictionaries Sadness, Anxiety, and Anger are nested in Negative Emotions, which is nested in Affective Processes. Similarly, we excluded LIWC summary variables and the obfuscation and abstraction indexes from LASSO. However, there is still substantial collinearity among the LIWC low-level dictionaries because many words appear in multiple dictionaries (e.g., the word "were" appears in the dictionaries auxiliary verbs, common verbs, and past focus). Therefore, we also ran logistic regressions with one text feature at a time, controlling for all the metadata variables. The narrative arc measures were entered together and individually to logistic regressions with all metadata variables for each section of text (the abstract model includes 260 abstracts because we removed those with fewer than 100 words from this analysis). We present the coefficients for variables that were selected in the LASSO regression and the coefficient and statistical significance for the significant variables from the "one-at-a-time" analyses in Tables 2-3. For full results see Tables S9-11.

Results

The language of replicability has markers of complexity and truth-telling

Table 2 and Fig. 1 present the results for language markers of replicability, and provide the relevant statistics. Overall, the authors of replicable studies seem detailed, truthful, forthcoming, and trustworthy based on their word choices.

Replicable studies are characterized by informative, elaborated, and detailed language. They often include quantifying words (“more,” “each”, see more words and the LASSO and “one-at-a-time” logistic regression statistics in Table 2 and Fig. 1), and number words (“first,” “second”), which likely serve to elaborate on the results. Comparing the text in academic articles from predatory vs. real journals, Markowitz, Powel, & Hancock (2014) found that articles in real journals use more quantifiers and prepositions (which we discuss next) suggesting the text is more detailed and linguistically complex. Replicable studies also tend to have interrogative words (“which,” “when,” “whether,”), auxiliary verbs (“were,” “is”), and common verbs whose top words are identical to auxiliary verbs, providing readers with descriptive, specific, and concrete information (Pennebaker et al., 2014). Conversely, the abstraction index is associated with nonreplicability. This result is consistent with research in other domains that associated more informative text with truth-telling (Reboul, 2021), because readers perceive the writer as more committed to the ideas and positions in their text, and because concrete information reduces uncertainty which allows the reader to better evaluate the claims (Larrimore et al., 2011). Finally, the use of present tense verbs (“is,” “have”) is a marker of truth telling (Netzer et al., 2019) and is more common among replicable studies, while future tense verbs (“predict,” “expect”), which are often more speculative, were more common among nonreplicable studies. Taken together these results suggest that the authors of replicable studies tend to be more forthcoming and detailed.

Replicable studies are written with longer sentences, which is indicative of sophisticated and complex language (Markowitz et al., 2014). Additionally, replicable studies use more prepositions (“of,” “in,” and “to”) and space words (“in,” “on,” and “at”), which are often used when authors analyze and categorize complex ideas, thereby showcasing complex, analytical thinking and a formal language style (Pennebaker et al., 2014). Such words are also more evident in truthful versus false narratives (Ott et al., 2012). Another marker of complex text is the use of comparisons, and indeed words related to order (named “power” in LIWC; “higher,” “over”) and differentiation (“than,” “different”) are also more likely to appear in replicable studies. We interpret these dictionaries as providing context to the study by pointing out how it compares and contrasts with past research. Exclusions and negations are also markers of complex ideas (Conway et al., 2014) because they describe nouns that are either inside or outside a category. This corresponds well with many words in the comparative dictionaries weak (“health” in LIWC; “weak,” “weaker) and differentiation (“not,” “but”) which are associated with replicability.

Replicable studies exude confidence, with authors commonly using certainty words (“all,” “total”), while nonreplicable research is written more vaguely (which we discuss next). Past research associated certainty with truth-telling because liars lack conviction (Netzer et al., 2019). In our context, however, the fact that authors describe their nonreplicable studies with less confidence may highlight some truthfulness, reflecting their true confidence in the study’s replicability likelihood.

*** Table 2 and Fig.1 here ***

The language of nonreplicability has markers of deception and persuasion

Table 3 and Fig. 2 present the results and relevant statistics for linguistic markers of nonreplicability. Overall, the text in nonreplicable studies is vague, hyped-up, and has the archetypical structure of a story. These are different methods of persuasion, possibly employed to overcome weaker results.

Nonreplicable studies are vaguely written. Five text features support this assertion—abstraction index (Markowitz & Hancock, 2016), future tense verbs, impersonal pronouns, and the use of adjectives and articles (specifically the indefinite articles “a” and “an”). Higher abstraction index values suggest the text is vague, uncertain, and uncommitted (Larrimore et al., 2011), and is common in fraudulent research (Markowitz & Hancock, 2016), predatory journals (Markowitz et al., 2014), and deceptive financial reporting (Li, 2008). Text written in future tense is perceived as speculative and therefore less committed (Netzer et al., 2019). Adding to the vagueness of the language in nonreplicable studies are impersonal pronouns (“this,” “that”), also known as “vague pronouns,” and adjectives (“same,” “high”). Adjectives are considered ambiguous despite the illusion that a concrete claim was made (Warren, 1988), and are indeed more prevalent in fraudulent corporate reporting (Goel & Uzuner, 2016) and deceptive advertising (Burke et al., 1988). We caveat that while we reference research on deception, replicability likelihood does not necessarily imply lying, as researchers rarely explicitly lie.

The next set of results suggests that nonreplicable studies employ different persuasion tactics—relying on authors’ clout, positivity, and storytelling.

We find that nonreplicable studies often use third and first-person plural pronouns and the affiliation dictionary (“we,” “our”). Over 90% of the studies in our dataset were written by multiple authors, so the use of plural pronouns is not surprising, despite controlling for the number of authors; however, their prevalence in nonreplicable studies is noteworthy, and perhaps alludes to the authors’ need to lend credibility to the study (Hyland, 1996) and deflect responsibility (because “we” represents a large group, Pennebaker et al., 2014). Past research supports this interpretation, as the usage of first-person plural pronouns has been associated with clout (Jordan et al., 2019). Lastly, clout has been shown to be negatively associated with the LIWC dictionaries “certainty” and “differentiation” which were related to replicability, because these dictionaries convey finality and assertiveness, and therefore do not necessitate the use of the writer’s clout in delivering the ideas in the text (Moore, Yen, & Powers, 2020). As such, the finding that clout and certainty and differentiation land on opposite sides of the replicability divide mirrors past findings.

Authors of nonreplicable studies write more positively as evident by the following four dictionaries: reveal (see in LIWC; “see,” “revealed,” “showed”), positive emotions (“positive,” “strong,” “support”), achievement (“obtained,” “best”), and reward (“positive,” “obtained”). Overly positive writings have been associated with negative outcomes in other areas, such as firm under-performance (Kang, Park, & Han, 2018) and fake reviews (Li et al., 2014), because authors convey a level of optimism which is likely unrealistic. The dictionary “work” (“test,” “analysis”) is associated with nonreplicability and although these words are very common in academic studies, their prevalence specifically in nonreplicable studies alludes to the authors’ need to reiterate what they did. While this conclusion is based on one dictionary, it echoes findings from other areas. For example, borrowers who ended up defaulting on their loans felt the need to reiterate and explain their past when asking for the loan (Netzer et al., 2019).

Lastly, nonreplicable studies have the archetypical structure of many stories. Boyd et al. (2020) show that stories, regardless of their content, share a similar structure—first setting the stage and establishing the context (staging), then presenting the conflict the protagonists grapple with, and finally resolving it (cognitive tension). Academic articles that tell good stories help researchers persuade their readers of the thesis laid out in the article (Dahlstrom, 2014). Since the flow of the narrative plays a crucial role in the persuasiveness of the story (Nabi & Green, 2015), authors who attempt to persuade their readers in their thesis and results, are likely to follow a narrative structure that has more staging early on, followed by cognitive tension and resolution. Indeed, cognitive tension is negatively associated with replicability ($\beta = -0.358$, $p = 0.037$), and staging is marginally so ($\beta = -0.266$, $p = 0.095$). Moreover, cognitive tension is consistently associated with nonreplicability—in the full text of the paper ($\beta = -0.328$, $p = 0.043$), marginally so in the abstract ($\beta = -0.318$, $p = 0.082$), and in "one-at-a-time" analyses that include all the metadata ($\beta_{study} = -0.324$, $p = 0.053$; $\beta_{fulltext} = -0.327$, $p = 0.041$). See the full set of results in Table S11. These findings suggest that while good storytelling is a desirable trait of the narrative, it may make it easier for readers to believe nonreplicable results (Dahlstrom, 2014).

*** Table 3 and Fig. 2 here ***

Language reflects the authors' intuition about their study's veracity

While our analyses control for authors' characteristics, research topic, and other paper metadata, we cannot guarantee that other aspects, not controlled for in our analyses, may be correlated with the text of the paper. To attempt to hold almost "all-else-constant," we focus on six papers in our dataset that have at least one successfully replicated study and at least one unsuccessfully replicated study. This provides a clean comparison of the language using a "sibling" analysis design. While only a cursory analysis due to the small sample size (n=6) that does not permit formal statistical testing, it still provides important insights about the role of the text. We find that 26 text features out of the 62 we tested (all but high-level dictionaries) using paired differences yielded effect sizes of at least medium magnitude (Cohen's $d > 0.3$; Cohen, 1988). This result indicates that writing styles differ substantially between studies from the same paper.

Most of these text features (20/26=77%) are in the same direction as our main analysis, showing that text signals tend to be directionally consistent within and across papers (see Table S12). An analysis of the statistics from the sibling studies, shows that the p-values of replicable studies are lower (Cohen's *d* = 0.769, large) and effect sizes are higher (Cohen's *d* = 0.47, medium) than those of nonreplicable studies in the same paper (see Table S13). Taken together, these results allude to the mindset of the authors as they wrote up the focal studies holding constant the authors' and the papers' characteristics, reflecting the authors' intuition about their studies' veracity.

Discussion

Past research used machine learning models to predict replicability, using either metadata variables (Altmejd et al., 2019) or text features (Yang et al., 2020; Youyou et al., 2023). These efforts sparked a discussion about the benefits and caveats of such methods, particularly with respect to the nature of information captured by the textual features relative to the characteristics of the research itself (Crockett et al., 2023; Mottelson & Kontogiorgos, 2023). We attempt to shed light on some of this friction by combining the largest set of metadata variables on the study, research topic, author characteristics, and replication effort, with the most detailed set of text features, including writing style measures, in this type of research thus far. This allows us to explore not only whether the study text is predictive of replicability above and beyond a rich set of controls, but also specifically what type of language contains information about replicability, which furthers our understanding of why the study text improves predictions of replicability.

Exploring the text in replicable and nonreplicable studies suggests that, whether knowingly or not, authors express their study's replicability likelihood in the way they write it.

Indeed, the words that we find to be associated more with replicable studies are related to elaboration and concreteness, which may indicate how careful the authors were while designing the study and analyzing and interpreting the results. The presence of quantifying and interrogative words as well as numbers further suggest the authors provided objective statistics in the study result. Together, we take these results to mean that the authors were meticulous and transparent about the methods and results, leaving little room to cut corners. This echoes the survey results mentioned in the introduction (Baker, 2016). On the other hand, nonreplicable studies are vaguely written, perhaps purposefully so, and exhibit a variety of persuasion techniques. Bearing these results in mind, next we reflect on issues related to approaches to science, citations, and the review process.

Academic writing could reflect the authors' approach to science—confirmatory versus exploratory. Research conducted with the confirmatory approach begins with clear hypotheses, grounded in theory, and then collects data that may or may not support the hypotheses (although due to the publication bias, published papers tend to report more supportive data). In comparison, research done with the exploratory approach aims to understand the data first, and then interprets the findings. This approach is common when theory is unable to advise predictions or when the researchers set out to find the unexpected. Arguments have been made both ways regarding replicability likelihood of either approach (Rubin & Donkin, 2022). If theory-based research has an archetypical story arc—more staging up front when the hypotheses are being set, and an inverted U shape for cognitive tension as the studies that confirm the hypotheses are presented, then our results could imply that this research may be less replicable.

Nonreplicable papers are cited more than replicable papers, possibly because they present more ostentatious findings (Serra-Garcia & Gneezy, 2021). Some papers create more excitement and buzz using exaggerated and inaccurate claims about their findings (Richie, 2020), consequently receiving more academic and popular media attention. These ideas correspond with our result that nonreplicable studies are likely to be presented more positively, even after controlling for the citation count.

Reviewers of new academic manuscripts can use our results to determine additional information to solicit from the authors. For example, if a manuscript is written rather vaguely or does not include interrogative words (e.g., what, when, why), the review team can ask the authors to elaborate some more. Reviewers can ask, for instance, when the results do not hold, why do they happen, and how do they relate to past results. Further, even when papers tell interesting stories, our results suggest that the review team should focus on the methods and results.

Similarly to other papers in this stream of science, our research has limitations, chief among them being the relatively small sample size. Manual replications are laborious, time consuming, and expensive, and thus relatively rare. Therefore, although the results we report are based on multiple pieces of evidence robust to a variety of methods, their underlying sample size should be borne in mind. The second limitation is related to generalization. While most of the results we report are based on context-free dictionaries such as adjectives, quantifiers, and pronouns and therefore could generalize to other fields, our sample nonetheless comes from one area, the social sciences. This potentially limits us from making general statements about the world of science (as advised by Crockett et al., 2023). Future research could expand our work to a larger sample size and other fields, as more papers are manually replicated. Finally, while older papers were able to predict the replicability of newer papers, this result might change in the future, perhaps due to the dissemination of our findings. Therefore, we recommend recalibrating our model as newer replications become available to identify possible temporal changes in the language of (non)replicable science.

References

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., ... & Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PloS one*, 14(12).

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604).

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science advances*, 6(32), eaba2196.

Burke, R. R., DeSarbo, W. S., Oliver, R. L., & Robertson, T. S. (1988). Deception by implication: An experimental investigation. *Journal of consumer Research*, *14*(4), 483-494.

Cabanac, G., Labbé, C., & Magazinov, A. (2021). Tortured phrases: A dubious writing style emerging in science. Evidence of critical issues affecting established journals. arXiv e-prints, arXiv-2107.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conway, L. G., Conway, K. R., Gornick, L. J., & Houck, S. C. (2014). Automated integrative complexity. *Political Psychology*, 35(5), 603-624.)

Crockett, M. J., Bai, X., Kapoor, S., Messeri, L., & Narayanan, A. (2023). The limitations of machine learning models for predicting scientific replicability. *Proceedings of the National Academy of Sciences*, 120(33), e2307596120.

Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences*, 111, 13614-13620.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50), 15343-15347.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., ... & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, 75, 102117.

Goel, S., & Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 215-239.

Hyland, K. (1996). Writing without conviction? Hedging in science research articles. *Applied linguistics*, 17(4), 433-454.

Jordan, K. N., Sterling, J., Pennebaker, J. W., & Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. Proceedings of the National Academy of Sciences, 116(9), 3476-3481.

Kang, T., Park, D. H., & Han, I. (2018). Beyond the numbers: The effect of 10-K tone on firms' performance predictions using text analytics. Telematics and Informatics, 35(2), 370-381.

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., ... & Ratliff, K. A. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, 8(1), 35271.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, 45(3), 142-152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1), 19-37.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45, 221-247.

Li, J., Ott, M., Cardie, C., & Hovy, E. (2014, June). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1566-1576).

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, 66(1), 35-65.

Markowitz, D. M., & Hancock, J. T. (2014). Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PloS one*, 9(8), e105937.

Markowitz, D. M., & Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, 35(4), 435-445.

Markowitz, D. M., Powell, J. H., & Hancock, J. T. (2014, June). The writing style of predatory publishers. In *2014 ASEE Annual Conference & Exposition* (paper ID#8614).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Moore, R. L., Yen, C. J., & Powers, F. E. (2021). Exploring the relationship between clout and cognitive processing in MOOC discussion forums. *British Journal of Educational Technology*, 52(1), 482-497.

Mottelson, A., & Kontogiorgos, D. (2023). Replicating replicability modeling of psychology papers. *Proceedings of the National Academy of Sciences*, 120(33), e2309496120.

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research*, 56(6), 960-980.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin*, 29, 665–675

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.

Ott, M., Cardie, C., & Hancock, J. (2012). Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web*. 201-210.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us.* Bloomsbury Press.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12), e115844.

Reboul, A. (2021). Truthfully Misleading: Truth, Informativity, and Manipulation in Linguistic Communication. *Frontiers in Communication*, 6, 62.

Richie, S. (2020). *Science fictions: How fraud, bias, negligence, and hype undermine the search for truth*. Metropolitan Books.

Rubin, M., & Donkin, C. (2022). Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philosophical Psychology*, 1-29.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. Science advances, 7(21), eabd1705.

Van Der Zee, S., Poppe, R., Havrileck, A., & Baillon, A. (2021). A personal model of Trumpery: linguistic deception detection in a real-world high-stakes setting. *Psychological science*, 33(1), 3-17.

Ventrella, J. (2011). *Virtual Body Language*. Pittsburgh: ETC Press.

Warren, B. (1988). Ambiguity and vagueness in adjectives. *Studia linguistica*, *42*(2), 122-172.

Wheeler, M. A., Vylomova, E., McGrath, M. J., & Haslam, N. (2021). More confident, less formal: stylistic changes in academic psychology writing from 1970 to 2016. *Scientometrics*, *126*, 9603-9612.

Yang, Y., Youyou, W., & Uzzi, B. (2020). Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(20), 10762-10768.

Youyou, W., Yang, Y., & Uzzi, B. (2023). A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proceedings of the National Academy of Sciences*, 120(6), e2208863120.

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, 25(5), 1968-1972.

Table 1. Predicting paper replicability in held-out samples by text section

Panel A: Study text

Train-test	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
239-60[1]	All papers in dataset	Average test AUC	0.6961	0.7028	0.7205	0.6827	0.7130	0.7245
		SD across splits	0.0607	0.0624	0.0610	0.0637	0.0612	0.0606
		% w/ improvement[2]		53.49%	68.32%	42.63%	64.51%	70.30%
220-55[3]	Psychology papers only	Average test AUC	0.6861	0.6913	0.7116	0.6749	0.7054	0.7169
		SD across splits	0.0664	0.0671	0.0658	0.0681	0.0663	0.0655
		% w/ improvement[2]		52.61%	66.69%	44.45%	64.04%	69.13%
236-63	Predicting new papers from old	Test AUC	0.6729	0.7816	0.8226	0.7384	0.7805	0.8204
299-99[4]	Market prediction	Correlation	0.5129	0.6054	0.6037	0.5794	0.6179	0.6147

Panel B: Full text[5]

Train-test	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
239-60[1]	All papers in dataset	Average test AUC	0.6961	0.6989	0.7140	0.6713	0.7022	0.7159
		SD across splits	0.0607	0.0647	0.0629	0.0664	0.0630	0.0630
		% w/ improvement[2]		51.60%	63.62%	35.82%	55.37%	64.08%
220-55[3]	Psychology papers only	Average test AUC	0.6861	0.6910	0.7053	0.6673	0.6909	0.7090
		SD across splits	0.0664	0.0688	0.0676	0.0708	0.0693	0.0676
		% w/ improvement[2]		52.78%	62.18%	40.34%	53.79%	63.96%
236-63	Predicting new papers from old	Test AUC	0.6729	0.7688	0.8171	0.7633	0.8016	0.7949
299-99[4]	Market prediction	Correlation	0.5129	0.5809	0.6046	0.5328	0.5844	0.6144

Panel C: Abstract text[6]

Train-test	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
208-52[1]	All papers in dataset that have an abstract of 100 words or more	Average test AUC	0.6752	0.6462	0.6753	0.5715	0.6651	0.6660
		SD across splits	0.0678	0.0716	0.0710	0.0722	0.0701	0.0721
		% w/ improvement[2]		35.82%	50.26%	12.52%	43.50%	44.21%
196-49[1]	Psychology papers only that have an abstract of 100 words or more	Average test AUC	0.6544	0.6377	0.6586	0.5990	0.6484	0.6497
		SD across splits	0.0728	0.0755	0.0757	0.0770	0.0764	0.0771
		% w/ improvement[2]		42.27%	52.60%	27.24%	46.77%	47.08%
206-54	Predicting new papers from old	Test AUC	0.7670	0.7415	0.8086	0.4730	0.6991	0.7840
260-84[7]	Market prediction	Correlation	0.4781	0.5078	0.5402	0.3589	0.5516	0.5622

Note: This table shows the results of logistic regressions with ridge regularization. The dependent variable is a binary indicator for replicability (1=replicable). There are three specifications: (1) Metadata model: includes only the metadata variables; (2) Text models: (a) Text embeddings: this model includes only the text represented by the embedding space (hyperparameters: continuous bag of words (CBOW), 3-word windows, 100 dimensions, 50 epochs, including stop words. Alternative specifications are presented in Table S7), (b) Text features: this model include LIWC, arc of narrative, and readability, (c) Text features + embeddings: this model includes (a) and (b); and (3) Metadata + Text: includes both textual and metadata variables. We present the average holdout predictions from 10,000 replications of a random 80% calibration-20% validation split of the papers in our sample. To evaluate the models' performance, we use the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The models are based on three slices of the paper text: Panel A = the text in the focal study, Panel B = the entire paper, or Panel C = the abstract; as well as three different slices of the data: all papers, only psychology papers, and over time (train on papers published before 2012 and predict papers published in 2012 or later). "Market predictions" models calculate the Pearson correlation between our models' predictions and end prices in prediction markets.

[1] Average across 10,000 splits. We note that because there are several papers with multiple studies being replicated, we split the sample by paper to make sure multiple studies of the same paper are always on the same side of the train-test split. Thus, the exact number of studies in the train and test sample can vary slightly by split.

[2] Proportion of splits (out of 10,000) with improved prediction over the model with metadata only. This measure is calculated for the main analysis and the papers in psychology.

[3] We dropped 24 studies whose field is economics to get 275 studies in psychology only.

[4] Trained on the entire dataset to predict replication outcome; predictions from the trained model were then correlated with the market predictions for the 99 studies/papers from the prediction markets.

[5] Analysis is still at the study level, though the text is for the full paper. 22 papers had more than one study replicated.

[6] 39 studies' corresponding papers did not have an abstract or had an abstract of under 100 words (the minimum required for the arc of narrative algorithm) and were removed from this analysis.

[7] 15 studies' corresponding papers did not have an abstract or had an abstract of under 100 words (the minimum required for the arc of narrative algorithm) and were removed from the prediction markets analysis.

Table 2. Text features associated with replicable research based on the focal study text

Linguistic signals	Evidence	LASSO coefficient	One-at-a-time coefficient [SE]	Top words in the study text corpus
1. Informative, elaborated, and detailed text	● Provision of information:			
	○ Numbers[1]	0.315	0.239 [0.169]	Two, one, first, three, second
	○ Quantifiers	0.310	0.467 [0.170]**	Each, more, all, both, average, any
	● Elaboration:			
	○ Interrogative[1]	0.350	0.262 [0.154]†	Which, when, who, whether, how
	○ Auxiliary verbs	0.297	0.311 [0.167]†	Were, was, is, are, be
	○ Common verbs	0.058	0.175 [0.163]	Were, was, is, are, be, one, would
2. Complex and analytical text	● Categorical language:			
	○ Prepositions[1]	0.301	0.363 [0.169]*	Of, in, to, for, with, as, on
	○ Space	0.252	0.201 [0.159]	In, on, at, both, high, low
	● Comparative language:			
	○ Order (Power in LIWC)[2]	0.134	0.032 [0.161]	High, low, higher, order, age, over
	○ Differentiation	0.119	0.034 [0.151]	Not, on, than, of, but, different
	○ Conjunctions[3]	0.047	0.142 [0.152]	And, as, when, if, but
	● Markers of complex text:			
	○ Longer sentences by word count[4]	0.201	0.083 [0.160]	
	○ Weak (Health in LIWC)[1,2]	0.026	0.049 [0.148]	Life, physical, weak, weaker, operation
3. Confident and truthful text	○ Present tense	0.344	0.200 [0.192]	Is, are, be, have, see
	○ Certainty	0.182	0.301 [0.149]*	All, positive, completed, total, accuracy
4. Other selected dictionaries	○ Leisure	0.333	0.212 [0.166]	Play, music, games, parties, family, novel
	○ Male references	0.117	0.141 [0.176]	Men, he, male, his, him, himself

Note: This table presents all LIWC low-level dictionaries (excluding punctuations) that were selected by LASSO and are associated with replicability.

[1] Although we do not interpret the full text due to its lower predictive ability of the outcome, we note that this dictionary is associated with replicability at a one-at-a-time analysis using the full text, but not in the LASSO analysis of the full text.

[2] We changed the name of the LIWC dictionary to be more meaningful to our content. The original name is in parenthesis.

[3] This dictionary is not associated with replicability in the full text.

[4] Words with 6 letters or more is also a marker of complex language, and its LASSO coefficient is positive (0.063) but its one-at-a-time coefficient is negative and not significant (β=-0.242, p=0.147). Therefore, we do not draw conclusion from that association.

† One-at-a-time coefficient is significant at $p < 0.1$; * $p < 0.05$; ** $p < 0.01$. See full results for the one-at-a-time regressions in Table S9.

Table 3. Text features associated with nonreplicable research based on the focal study text

Linguistic signals	Evidence	LASSO coefficient	One-at-a-time coefficient [SE]	Top words in the study text corpus
1. Vague and deceptive text	● Abstraction index[1]	N/A	-0.414 [0.189]*	
	● Vagueness:			
	○ Future tense	-0.306	-0.264 [0.157] †	Then, will, may, predicted, expected, might
	○ Impersonal pronouns	-0.068	-0.018 [0.166]	That, this, other, which, it, these
	○ Adjectives	-0.262	-0.085 [0.148]	As, then, after, same, high
	○ Articles	-0.335	-0.121 [0.169]	The, a, an
2. Text written with clout	○ Third person plural pronouns	-0.437	-0.287 [0.171]†	They, them, themselves
	○ Affiliation	-0.253	-0.161 [0.170]	We, our, interaction, groups, social
	○ First person plural pronouns[1,2]	-	-0.289 [0.173]†	We, our, us
3. Positivity	○ Reveal (See in LIWC)[3]	-0.278	-0.298 [0.160]†	See, revealed, showed, shows
	○ Positive emotions[2]	-0.239	-0.425 [0.180]*	Positive, value, greater, strong, support, important
	○ Achievement	-0.135	-0.290 [0.183]	First, obtained, best, better, efficiency
	○ Reward[4]	-0.106	-0.278 [0.198]	Positive, scores, obtained, good, best, better
4. Other selected dictionaries	○ Work[2]	-0.187	-0.279 [0.159]†	Test, analysis, performance, reported
	○ Anxiety	-0.341	-0.407 [0.184]*	Aversion, pressure, anxiety, fear, avoidance
	○ Feel	-0.278	-0.415 [0.182]*	Round, feelings, feel, hand, weight
	○ Risk	-0.071	-0.262 [0.159]†	Aversion, yielded, consequences, trust, problems
	○ Female references	-0.056	-0.139 [0.188]	Female, her, she, woman, mother, herself
5. Tells an interesting story	● Archetypical narrative of a story:			
	○ Cognitive tension arc[5]	β = -0.364, SE = 0.172, p = 0.035		
	○ Staging arc	β = -0.265, SE = 0.159, p = 0.096		

Note: This table presents all LIWC low-level dictionaries (excluding punctuations) that were selected by LASSO and are associated with nonreplicability.

[1] Missing LASSO coefficients mean that while the text feature was not selected by LASSO (likely due to collinearity with the other text features), it is associated with nonreplicability in the one-at-a-time analysis. N/A is for summary variables (such as abstraction) that were not part of LASSO regression.

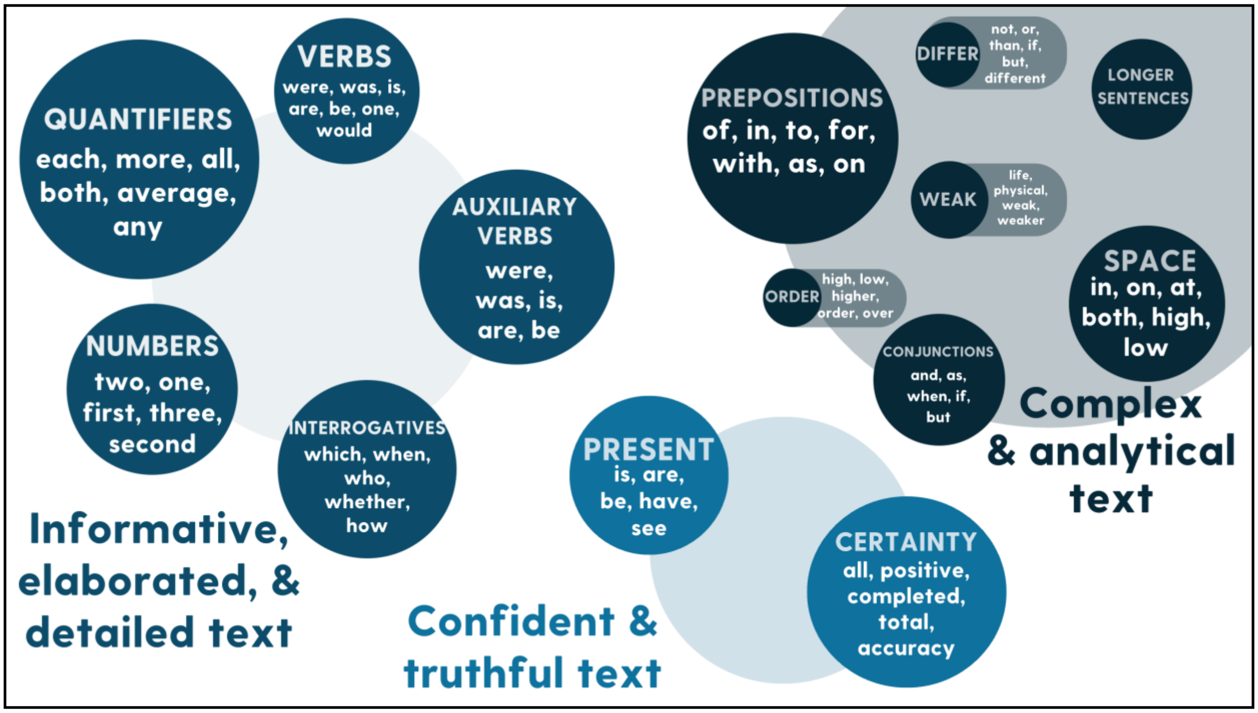
[2] Although we do not interpret the full text due to its lower predictive ability of the outcome, we note that this dictionary is associated with nonreplicability at a one-at-a-time analysis using the full text, but not in the LASSO analysis of the full text.

[3] We changed the name of the LIWC dictionary to be more meaningful to our content. The original name is in parentheses.

[4] This dictionary is not associated with replicability in the full text.

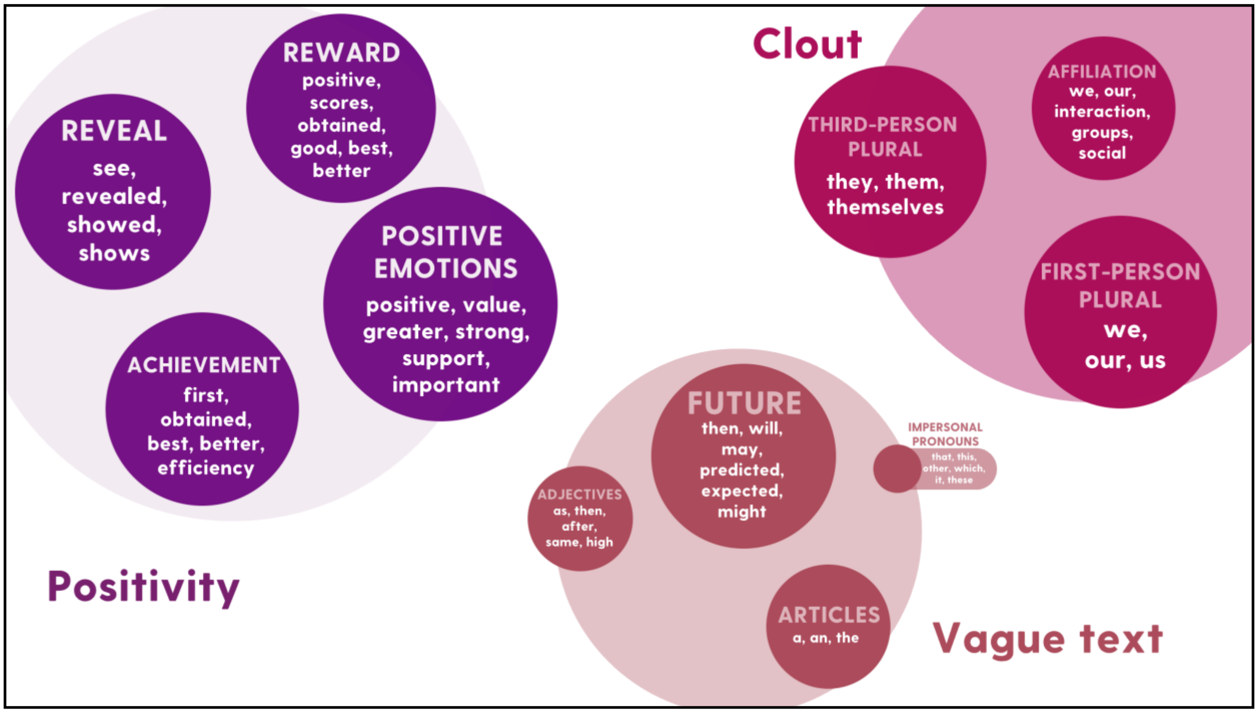
[5] Results of binary logit regressions (replicability=1) with all arc of narrative variables (staging, cognitive tension, and plot progression) and all the metadata, based on the focal study text. Cognitive tension arc is also significant at $p = 0.043$ in the full text (see all results in Table S11), and when ran alone with the metadata ($p = 0.053$ for study text, $p = 0.041$ for full text).

† One-at-a-time coefficient is significant at $p < 0.1$; * $p < 0.05$. See full results for the one-at-a-time regressions in Table S9.



VERBS
were, was, is, are, be, one, would
QUANTIFIERS
each, more, all, both, average, any
AUXILIARY VERBS
were, was, is, are, be
NUMBERS
two, one, first, three, second
INTERROGATIVES
which, when, who, whether, how
Informative, elaborated, & detailed text
PREPOSITIONS
of, in, to, for, with, as, on
DIFFER
not, or, than, if, but, different
LONGER SENTENCES
WEAK
life, physical, weak, weaker
ORDER
high, low, higher, order, over
SPACE
in, on, at, both, high, low
CONJUNCTIONS
and, as, when, if, but
Complex & analytical text
PRESENT
is, are, be, have, see
CERTAINTY
all, positive, completed, total, accuracy
Confident & truthful text

Fig. 1. Writing styles associated with replicable studies. The dictionary bubbles' sizes are based on the "one-at-a-time" coefficients from Table 2.



REWARD
positive, scores, obtained, good, best, better
REVEAL
see, revealed, showed, shows
POSITIVE EMOTIONS
positive, value, greater, strong, support, important
ACHIEVEMENT
first, obtained, best, better, efficiency
Positivity
Clout
THIRD-PERSON PLURAL
they, them, themselves
AFFILIATION
we, our, interaction, groups, social
FIRST-PERSON PLURAL
we, our, us
FUTURE
then, will, may, predicted, expected, might
IMPERSONAL PRONOUNS
that, this, other, which, it, these
ADJECTIVES
as, then, after, same, high
ARTICLES
a, an, the
Vague text

Fig. 2. Writing styles associated with nonreplicable studies. The dictionary bubbles' sizes are based on the "one-at-a-time" coefficients from Table 3.

Supplemental Materials for

The Language of (Non)replicable Social Science

Michal Herzenstein, Sanjana Rosario, Shin Oblander, Oded Netzer

Table of content

A. Data Sources and Procedures

Our dataset includes 272 papers and 299 different studies that were manually replicated (some papers had multiple studies manually replicated). Table S1 categorizes the original papers by their replication projects/papers.

Table S1: Replication efforts included in our data

Replication project	Cite	# of studies in our dataset
Reproducibility of psychological science project (RPP)	Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).	96

Many Labs (ML)	Klein Richard, et al. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, 45(3), 142-152. (**ML1**)
	Klein, R. A., et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490. (**ML2**)
	Ebersole, C. R., et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82. (**ML3**)
	Klein, R. A., et al. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, 8(1), 35271. (**ML4**)
	Notes:
	Two papers in ML2 have scenario-based experiments (Ross, Greene, and House 1977; Hauser, Cushman, Young, Kang-Xing Jin, and Mikhail 2007) and each scenario was replicated individually. Since the scenarios could not be easily separated in the study text, we treated these papers as if they were replicated only once. The text includes both scenarios and all replications were successful.
		ML1 = 13
		ML2 = 25
		ML3 = 10
		ML4 = 1

Social science replication project (SSRP)	Camerer, C. F., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644.	20
Experimental economics (EE)	Camerer, C. F., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.	18
Participant nonnaiveté (PN)	Zwaan, R. A., et al. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic bulletin & review*, 25(5), 1968-1972.	8

	Note: Table 1 in Zwaan et al. suggests that the motor task was reproduced from Forster and Davis (1984). But that paper does not have a motor activation task. Study 1 in Eimer and Schlaghecken (1999) is very similar to the study protocol in the replication effort and was mentioned in an earlier version of the replication paper. We set the latter as the paper that was replicated. We did not use the last replication presented in Table 1 because it is of a paper by Rolf Zwaan and we do not include replications done by the original author.	
Social psychology special issue (SPSI)	Social Psychology (2014) vol. 45, issue 3: special issue about replication.	22
Registered replication reports (RRR)	Individual replication reports from these journals: *Psychological Science, Perspectives on Psychological Science,* and *Advances in Methods and Practices in Psychological Science.* Note: Olson and Fazio (2001) was replicated by a group of researchers that then contacted the original authors and included the first author in the replication efforts, and made him a coauthor. While one of our criteria is that the replication was not done by the original author, this paper started by other researchers and therefore we include it.	22
Other direct replications (ODR)	Individual direct registered replications we collected based on CurateScience.org and Google Scholars search. Note: Inclusion criteria: (a) replication was published in a journal with impact factor of 2 or higher, (b) replication was not done by the original author, (c) replication exactly follows original study (i.e., direct rather than conceptual replication), (d) replication is of papers in cognitive psychology, social psychology, or economics.	64

Table S2 (uploaded here: https://osf.io/8ptwf) presents the full citation of each original paper and its replication paper, including an ID # that we use to connect that list with our two other datasets: the metadata on each paper, and the text files for each paper.

B. Variables

In what follows we describe the types of data/variables in our dataset:

1. Text files created from the original manuscripts
2. Metadata related to the paper, authors, and replication effort(s)
3. Independent variables and controls derived from the text
4. Prediction markets end prices
5. Secondary dataset of 2,420 papers' full text for training embedding models

1. Text files created from the original manuscript

We created three text files from the words in each paper:

a. **Abstract**: includes only the abstract of the paper. Text files for abstract are empty for the following paper IDs: 155, 157, 190, 196, 207, 208, 221, 231, 261, 263. These papers are dropped from the analysis. Following the publisher guidelines, the first paragraph was treated as the abstract in the paper IDs 156, 295.
b. **Full text**: includes the full text of each paper, from the first word of the first paragraph to the last word of the last paragraph. This excludes title, abstract, keywords, tables and figures and their titles and captions, proofs and equations, footnotes, authors' notes and acknowledgments, and reference list. This also excludes appendixes printed as part of the paper unless the appendix provided additional information about the experiment(s) that was/were replicated. Supplementary materials were also excluded from the full text files.
c. **Study**: includes all the text of the focal study, that one that was replicated (starting in methods and ending in discussion). If more than one study was replicated from the original paper, then a separate study file was created for each study that was replicated.

27 papers have studies that were replicated more than once or have multiple studies that were replicated. Table S3 details how we treated them.

Table S3: Papers with multiple replications

ID	Cite	Notes
7	Correll (2008)	The same study was replicated in RPP.63 and ODR.31. Both failed to replicate and therefore are treated as one observation.
17	Bressan and Stranieri (2008)	Was replicated in RPP.148 and RPP.149. Both failed to replicate and therefore are treated as one observation.
30	Schnall et al. (2008)	Study 1 was replicated in SPSI.11 and Study 2 was replicated in RPP.151. We have two observations for this paper.
34	Albarracín et al. (2008)	Study 5 was replicated in RPP.49 and Study 7 was replicated in RPP.50. We have two observations for this paper.
36	Risen and Gilovich (2008)	Study 6 was replicated in RPP.68 and Study 2 was replicated in ML2.22. We have two observations for this paper.
62	Vul and Pashler (2008)	Was replicated in RPP.116 and ODR.3. Both successfully replicated and therefore are treated as one observation.
121	Kidd and Castano	Was replicated in SSRP.13 and in ODR.44, 45, and 47. All

	(2013)	replications failed, so SSRP.13 and ODR44 are treated as one observation. ODR.45 and 47 replicate Studies 3 and 5 respectively so we have two separate observations.
151	Tversky and Kahneman (1981)	Problem 1 (Asian disease) was replicated in ML1.2 and Problem 10 (calculator on sale) was replicatedML2.20. We keep them as two different observations.
152	Jacowitz and Kahneman (1995)	Was replicated in ML1.3, ML1.4, ML1.5, and ML1.6. The four different replications of the anchoring effect documented in the original paper used four out of the original 15 stimuli. Given that the original paper does not provide separate statistics for each stimulus, and that all replication attempts were successful, we pooled the replication data together by taking a weighted average of the effect sizes and use this as one observation.
161	Nosek, Banaji, and Greenwald (2002)	Was replicated in ML1.15 and ML1.16. Since one replication is of Study 1 and the other is based on data from Studies 1 and 2, we treat them as two different observations. The text file for study161a includes the text of Study 1, and the text file for study161b includes the text from Studies 1 and 2.
206	Rand et al. (2012)	Was replicated in SSRP.4 and RRR.1. Since two different studies were replicated (Study 7 and Study 8 respectively), we have two observations for this paper.
207	Asch (1946)	Study 7 was replicated in SSRP.4 and Study 8 was replicated in RRR.1. We have two observations for this paper.
217	Eyal et al. (2008)	Studies 2-4 were replicated in SPSI.14-16. We have three observations for this paper.
219	Sachdeva et al. (2009)	Studies 1, 3, and 4 were replicated in SPSI.1-3. We have three observations for this paper.
221	Banerjee et al. (2012)	Studies 2-4 were replicated in SPSI.14-16. We have three observations for this paper.
227	Schooler and Engstler-Schooler (1990)	Studies 1 and 3 were replicated in SPSI.18-19. We have two observations for this paper.
228	Bargh, Chen, Burrows (1996)	Studies 1 and 2 were replicated in SPSI.21-22. We have two observations for this paper.
248	Tykocinski et al. (1995)	Studies 1 and 4 were replicated in RRR.7 and RRR.11. We have two observations for this paper.
255	Heintzelman et al. (2013)	Studies 2a and 2b were replicated in ODR.1. We have two observations for this paper.
259	Kahneman and Miller (1986)	Studies 1, 2, 4, and 6 were replicated in ODR.19-21. We have four observations for this paper.
275	Bargh et al. (2001)	Studies 2 and 4 were replicated in ODR.27-8. We have two observations for this paper.
277	Bem (2011)	Studies 8 and 9 were replicated in ODR.55-6. We have two observations for this paper.
297	Schmeichel (2007)	Study 1 was independently replicated by three papers. The first one successfully replicated the original findings with n=38 and

		the other two failed to replicate with n=138, 200. By majority rule we treat this as a failed replication.
305	Fernbach et al. (2013)	Studies 2 and 3 were replicated in RRR.18-19. We have two observations for this paper.
306	Gebauer et al. (2018)	Studies 1 and 2 were replicated in RRR.20-21. We have two observations for this paper.
309	Clark et al. (2014)	Studies 1 and 2 were replicated in ODR.69-70. We have two observations for this paper.
313	Hsee and Zhang (2004)	Studies 1 and 2 were replicated in ODR.74-75. We have two observations for this paper.

Whenever we have more than one observation for a given paper, we made sure they always appear in the same sample (training or holdout) for prediction.

2. Metadata variables related to the original paper, authors, and replication effort(s)

We received data from Altmejd et al. (2019) that included all replicated papers from replication projects RPP, SSRP, EE, ML1-3 (although Altmejd and his colleagues did not include papers from ML2 replication project in their analysis, they collected partial data on those papers).
We verified all the data that come from the citation: author names, number of authors, male authors, title, journal, volume, issue, year, and discipline. We verified the following for all ML2 papers and a sample of the other papers: type of effect (main vs. interaction), p-value, information about the lab: US or elsewhere, lab vs. online, subjects (students, community, or anyone), and compensation.

We complement the variables in Altmejd et al. (2019) with additional variables that we hand-collected for all papers: rank of each author (assistant/associate/full professor, PhD student, research fellow etc.) at the time the paper was published based on their resumes posted online, we then calculated % of full professors, number of tables, figures, references, and studies in the paper, whether the paper has online supplementary materials, and citation count based on Google Scholar in January 2022. For some papers, we were not able to find the resumes of all authors online, and thus were not able to determine the proportion of full professors on the author team (paper IDs 227, 283, 288, and 318); for these observations, we treat the full professor ratio variable as missing.

Several replications in our paper did not appear in Altmejd et al. (2019). These include SPSI, RRR, PN, and ODR. For these paper we collected author names, number of authors, male authors (based on names and pictures we found online), and full professors, title, journal, volume, issue, year, discipline, type of effect (main effect, interaction, correlation), p-value, effect size, information about the lab: US or elsewhere, lab vs. online, subjects (students, community, or anyone), and type of compensation (cash or course credit).
We were unable to collect all variables in all papers. Table S4 provides a list of special cases and how we resolved them. The cases that could not be resolved were left as missing.

Table S4: Notes regarding the statistics in the original and replication papers

ID	Cite	Notes
155	Hyman & Sheatsley (1950)	The original paper does not provide enough results to calculate the statistics in the study. We therefore treat the p-value, effect size, and sample size as missing.
156	Rugg (1941)	Same as paper 155.
157	Lorge & Curtiss (1936)	Same as paper 155, but sample size is not missing.
170	Inbar et al. (2009)	The original paper uses Cohen's q as a measure of effect size as it compares correlations between two constructs under 2 different conditions. The Cohen's q can't be converted to correlation coefficient (which is the scale all other effect sizes have been converted to), so we kept the effect size as q.
186	Schwarz et al. (1991)	Same as paper 170.
207	Asch (1946)	The paper does not provide any statistics or enough results to calculate the statistics of the comparisons that were made in the replication paper. Hence, we treat the p-value and effect size as missing. The replication paper was unable to reproduce the studies in the original paper, but also claim that the results in the original paper are not unequivocal.
208	Schachter (1951)	The ANOVA comparing the number of communications directed at the mode, slider, and deviate was replicated. We calculated the t-test comparing the slider to the deviate and use the p-value and r for that (from Tables 1-2 in the original paper and from the data uploaded by the replication paper). The sample size (n) we used for the original paper is # of groups not participants because data is presented in groups.
209	Driscoll, Davis, and Lipetz (1972)	The correlation we used in our analysis is the difference in love between time 1 and time 2 for unmarried couples (based on Table 2 in the original paper).
214	Shackelford et al. (2004)	We used the comparison between older men and women in the original and study 4 in the replication (this is the only study that was done in English in both papers).
246	Oosterhof and Todorov (2008)	The finding being replicated is that in doing PCA on people's judgments of faces, the first principal component correlates with trustworthiness but not dominance, and the opposite for the second component. We treat the "effect size" of interest as the correlation between the first principal component and the trustworthiness rating. We categorize this as a correlation rather than a main effect. Defining a p-value is difficult since the principal component is by definition a function of the trustworthiness rating, but we heuristically calculate a p-value using a z-test on the Fisher transformed correlation. This yields a p-value of effectively zero. The replication paper reports several results for subject pools from different countries, but since the subject pool in the original study is American, we take the USA and Canada

		correlation to be the replication effect size.
249	Schwarz and Clore (1983)	Based on the supplemental material of the replication paper (https://osf.io/s8apm/) we believe there is a typo in the df of the main F-test (should be 1 not 2) and recalculated its p-value accordingly. The effect size for the replication is based on the first experiment (d=0.1). The second experiment found a similar result (d=0.11).
266	Murray et al. (2002)	To calculate the effect size of the replication, we used the data and code provided by the replication authors (https://osf.io/4rkx9/). We added lines to the code to compute the direct and partial correlation between the interaction term and the partner enhancement DV.
275	Bargh et al. (2001)	We were unable to calculate the effect size in study 3 and left it as missing.
301	Olson and Fazio (2001)	The authors use a paired t-test since the effect is within-subjects. We converted it to correlation assuming that Cohen's d for within-subjects converts to correlation in the same way as between subjects.
313	Hsee and Zhang (2004)	While the main result is that there is no difference between single and joint evaluations, this is not formally tested. We used the effect size of the test for the single evaluation of 0 vs. 80 books.
315	Przybylski and Weinstein (2013)	The original paper says the condition coded as -1/1 while the replicating paper has it coded as 0/1. The replicating paper reports both effect sizes in the same table apparently without correcting for this difference, so we assumed the original paper misreported.
317	Snyder et al. (2015)	There is no effect size for this observation because it is a validation of the first factor in a factor analysis study. We therefore treat the effect size as missing.

Five metadata variables have missing values: p-value (2.7% missingness), effect size (3% missingness), sample size (0.7% missingness), proportion of authors who were full professors (1.7% missingness), and post-hoc power (3% missingness; derived from effect size and sample size, so missing when either are missing).To ensure that we adequately preserve the relationships between variables when controlling for these variables in our analyses, we apply iterative random forest imputation (as implemented by the R package “missForest”), using all metadata and keyword variables to perform imputation. These imputed values are used for all analyses that include metadata variables.

Additionally, we collected the following variables:

From the replication papers (and the individual reports from the large replication projects) we collected information regarding whether the original authors were involved in the replication. If the original authors shared their original material or commented on the data collection plan and/or analysis, then we marked them as involved.

We collected keywords from the original papers, wherever they were available. A psychology major undergraduate research assistant filled out keywords for the psychology papers that did not include them (60 cognitive psychology papers and 61 social psychology papers). Here are their ID numbers:

- Cognitive psychology papers: 1, 8, 9, 18, 24, 38, 39, 57, 62, 64, 77, 87, 89, 97, 110, 112, 115, 116, 120, 122, 124, 126, 129, 130, 154, 173, 187, 189, 190, 195, 196, 210, 213, 225, 226, 230, 232, 233, 234, 241, 244, 249, 250, 254, 260, 261, 263, 265, 266, 273, 285, 286, 287, 288, 289, 293, 295, 296, 300, 316.
- Social psychology papers: 3, 5, 11, 16, 17, 19, 26, 28, 30, 37, 45, 58, 73, 74, 78, 79, 81, 83, 88, 94, 95, 96, 98, 103, 111, 113, 114, 123, 127, 131, 152, 155, 156, 157, 163, 167, 169, 172, 175, 185, 186, 188, 191, 192, 193, 194, 198, 208, 209, 211, 212, 216, 222, 243, 251, 274, 283, 301, 307, 314, 318.

The RA used lists of keywords that appeared in at least two papers, in total 21 keywords in social psychology papers and 46 keywords in cognitive psychology papers. The RA assigned them to the papers missing keywords based on their reading of the abstracts. They could assign as many keywords as they saw fit, with the understanding that the first keyword should be the most representative one for the paper.

The RA then went over all the papers that had keywords and wherever possible chose a more general first keyword. For example, the first keyword for paper #55 is “additive effects” and the RA changed it to math/stat. But when the first keyword was general, it was left as is, for example paper #13 whose first keyword is “attention”.

After these changes we are left with 12 papers/28 studies without keywords (6 papers/14 studies each in social and cognitive psychology) which get filled in as “other social psych” and “other cognitive psych” respectively.

For the economics papers, we collected the JEL codes specified in the papers. There are six economics papers with no keywords or JEL codes: IDs 117-119, 125, 137, 148. Two authors went over their abstracts and assigned them JEL codes. After reassigning specific JEL codes to more general ones, the economics papers in our dataset are assigned to one of four keywords: individual behavior, group behavior, non-cooperative games, and consumer economics.

In total there are 45 keywords (43 + the 2 “other” categories). To avoid collinearity with the intercept and the discipline indicator for economics, we further dropped the dummy variables of “other cognitive psych” and “non-cooperative games” keywords. Hence, we use 43 keywords as controls in the prediction models. In the interpretation models, we collapsed them to 19 keywords so that each keyword is assigned to at least 4 studies (2 with successful replication and 2 with failed replication) to ensure identifiability of the model coefficients (as the keywords are unregularized in the interpretation model). Specifically, keywords with fewer than 2 successful and 2 unsuccessful replications were collapsed into one of the 2 “other” categories. For economics papers, “group behavior” fell below this threshold and was collapsed into “non-cooperative games.” Lastly, for the interpretation analysis of abstracts using narrative arc variables, dropping abstracts of less than 100 words leads to only a single paper (Bartling et al., 2012) with the “consumer economics” keyword; for this analysis only, to ensure identifiability of metadata coefficients, we further collapse “consumer economics” into the “non-cooperative games” keyword.

The metadata file is uploaded here: https://osf.io/38wyr

3. Processing the text in the papers

We took the following steps in processing the text and creating textual variables:

1. **Text embeddings:** We use the gensim Python package to train a word2vec model on a secondary dataset of academic articles (described below). We train word2vec using 100-dimensional embeddings with a window size of ±3 words, trained for 50 epochs, including stopwords, and using continuous bag of words (CBOW). All other settings are set to their gensim defaults. In Table S7 we present robustness checks results using alternate settings, namely: 50- or 150-dimensional embeddings (instead of 100), a window size of 5 (instead of 3), training for 100 epochs (instead of 50), excluding stopwords (instead of including them), or using skipgram (instead of CBOW), to demonstrate that our results are not sensitive to the specific settings of the word2vec model. We use the word embeddings, averaged over all words in a given paper (or section of a paper), as text features in our predictive analysis.
2. **LIWC**: We use the LIWC category scores obtained by passing the plain text files through the LIWC 2015 software. Before feeding text files into the LIWC software, we looked at the top 50 most common tokens in each dictionary and removed spurious words whose meanings in the academic context did not match the intended meaning in LIWC: namely, author names (e.g., LIWC erroneously counts the name "Nosek" as a "biology" word due to beginning with "nose"), abbreviations (e.g., "k" was erroneously counted as an abbreviation for the word "okay"), and the word "dummy" (which was erroneously counted as an insult). LIWC scores were then calculated after removing these spurious words. Out of the 92 dictionaries LIWC returns (including summary variables like Clout, and punctuation count), we removed those that appeared in less than 50% of documents so the results are not based on niche/uncommon language. Because abstracts are short and tend to use a more limited range of language than the text of the paper itself, that means more dictionaries get filtered out in the pre-processing step. Specifically:
    a. Full text: 7 removed (family, sexual, death, swear, filler, exclamation marks, nonfluencies).
    b. Study: 18 removed (first person singular, second person, third person singular, family, friend, hear, body, sexual, ingestion, home, religion, death, swear, assent, nonfluencies, filler, question marks, exclamation marks)
    c. Abstract: 36 removed (first person singular, first person plural, second person, third person singular, anxiety, anger, sadness, family, friend, female, male references, hear, feel, biological processes, body, health, sexual, ingestion, leisure, home, money, religion, death, informal, swear, netspeak, assent, nonfluencies, filler, Colon, Semi Colons, Question marks, exclamation marks, quotation marks, apostrophes, other punctionations)
    d. Additionally, we removed from all analyses the summary variable Informal and the dictionary Netspeak which includes abbreviations and therefore is irrelevant to our context.
3. **Word counts**: The LIWC software also returns document word counts and the words per sentence.
4. **Readability measure:** We use the koRpus package in R to calculate the Flesch Reading Ease score. A higher score indicates that the text is easier to read.
5. **Abstraction index:** We calculated the abstraction index based on Markowitz and

Hancock (2016) as the (negative) sum of the LIWC dictionaries prepositions, quantifiers, and articles after standardizing each dictionary to unit variance.

6. **Linguistic obfuscation:** We calculated the obfuscation index based on Markowitz and Hancock (2016) as: % of causal terms + abstraction index + jargon (% of words not in any LIWC dictionary) - positive emotion words - Flesch Reading Ease score. All terms in this sum are standardized before combining.
7. **Arc of narrative:** we passed the text files through the algorithm in www.arcofnarrative.com to obtain a score for each of the story stages (staging, plot progression, and cognitive tension) for each text section. Scores range from -100 to 100 where 100 means that the stage in the text is archetypical (as expected by Boyd, Blackburn, & Pennebaker, 2020): staging—more in the beginning of the text and less later, plot progression—increasing as the text progresses, cognitive tension—increases until about the middle of the text, and then decreases. A score of -100 means that the text has the opposite structure (i.e., less staging up front and more later). A score of 0 means that the story stage arc does not appear in the text. The algorithm requires that the text will have a minimum of 100 words in order to determine its arc of narrative score. The following abstracts have no abstract or abstracts of fewer than 100 words and therefore were excluded from analyses of abstracts involving narrative arc variables: 1, 5, 28, 62, 74, 116, 123, 131, 132, 135, 136, 137, 139, 141, 143, 144, 145, 150, 151, 155, 156, 157, 189, 190, 196, 197, 207, 208, 210, 221, 229, 261, 263, 289, and 295.

The processed text used in our analyses, as well as the code, are posted here: https://osf.io/qy8ev/

4. Prediction markets

In total we have prediction market end prices for 99 papers: 38 from Dreber et al. 2015 (predicted RPP replications), 18 from Camerer et al 2016, 21 from Camerer et al. 2018, and 22 from Forsell et al 2019 (predicted ML2).

Replication prediction markets in these studies work as follows: Participants were scientists who are experts in the field of the original study (and were not involved in the replication of the study). They received the original study focal hypothesis that was to be replicated and the replication plan, and then they traded contracts that pay $1 if the study was successfully replicated and $0 otherwise. Dreber et al. (2015) explain that this type of contract allows the price to be interpreted as the predicted probability that the study would successfully replicate.
Notes:

1. Dreber et al. (2015) ran prediction markets on 44 studies from RPP. We follow exclusions made by Altmejd et al. (2020) who used the results of prediction markets on 39 studies. In addition, we excluded one more study, Bressan & Stranieri (2008), because we pooled the results of the two replications of this study (hence our data reflects the weighted average of two replications whereas the prediction market is only for one of them).
2. Forsell et al. (2019) ran prediction markets on 24 studies, two of them are papers with experiments with multiple scenarios (Ross, Greene, and House, 1977; Hauser, Cushman, Young, Kang-Xing Jin, and Mikhail, 2007). We treat these papers as if they were

replicated only once because the different scenarios cannot be cleanly separated. Therefore, our study text files include all scenarios tested.

5. Compiling the secondary dataset of paper text for training embedding models

We created our own training set for estimating a word embedding model. There are publicly available pre-trained word2vec models that have been trained on Google News (Mikolov et al., 2013), Wikipedia (Pennington et al., 2014), or on academic abstracts (Yang et al., 2020) but we believe that the style of writing in all these sources is different than in the main text of an academic paper, motivating training our own embeddings on a newly collected dataset.

This secondary dataset includes the full text of 2,420 papers following these criteria:

1. We harvested the text from all papers in the issue of a paper being replicated (except for that replicated papers themselves). If there were three or more papers being replicated from the same issue, we harvested papers from the following issue in the same journal.
2. We did not include papers that are in picture format (non-editable PDFs) or old papers (earlier than 1985).

We harvested the raw text in each paper, which includes the full text of the paper, and does not include the title, abstract, acknowledgements, footnotes, references, and appendices. We also did not include any figures, tables, and their captions. We also removed equations, except for those embedded in-line.

C. Basic statistics and correlations

Table S5: Summary statistics

Panel A: Dichotomous variables (frequencies)

Variable	Not Replicated	Replicated	Total
Supplementary materials	0.243	0.262	0.251
Study done in US	0.734	0.714	0.726
Effect type: Main effect	0.642	0.810	0.712
Effect type: Correlation	0.075	0.063	0.070
Effect type: Interaction	0.283	0.127	0.217
Discipline: Economics	0.052	0.119	0.080
Discipline: Social psychology	0.509	0.373	0.452
Discipline: Cognitive psychology	0.439	0.508	0.468
Participants: Anyone	0.133	0.135	0.134
Participants: Community	0.064	0.095	0.077
Participants: Online	0.046	0.048	0.047
Participants: Students	0.757	0.722	0.742
Replication project: EE	0.040	0.087	0.060
Replication project: ML	0.133	0.206	0.164
Replication project: RPP	0.353	0.278	0.321
Replication project: SSRP	0.046	0.095	0.067
Replication project: Other	0.428	0.333	0.388
Original author(s) endorsement	0.740	0.651	0.702
Keywords : Cognitive other	0.040	0.056	0.047
Keywords : Social other	0.052	0.040	0.047
Keywords : Anchoring	0.017	0.016	0.017
Keywords : Attention	0.029	0.024	0.027
Keywords : Attitude	0.058	0.032	0.047
Keywords : Automaticity	0.000	0.016	0.007
Keywords : Beliefs	0.006	0.008	0.007
Keywords : Bias	0.035	0.024	0.030
Keywords : Choice	0.023	0.008	0.017
Keywords : Cognitive processes	0.058	0.040	0.050
Keywords : Conflict	0.006	0.008	0.007
Keywords : Construal Level	0.017	0.016	0.017
Keywords : Consumer economics	0.012	0.016	0.013
Keywords : Culture	0.000	0.016	0.007
Keywords : Death	0.012	0.008	0.010
Keywords : Embodiment	0.058	0.008	0.037
Keywords : Emotion	0.023	0.024	0.023
Keywords : Fluency	0.006	0.008	0.007
Keywords : Goals	0.029	0.008	0.020
Keywords : Group behavior	0.006	0.024	0.013
Keywords : Happiness	0.006	0.008	0.007
Keywords : Individual behavior	0.017	0.032	0.023
Keywords : Information processing	0.000	0.024	0.010
Keywords : Language	0.035	0.008	0.023

Keywords : Learning	0.017	0.040	0.027
Keywords : Math/Stat	0.000	0.048	0.020
Keywords : Memory	0.081	0.095	0.087
Keywords : Metaphor	0.006	0.008	0.007
Keywords : Money	0.012	0.016	0.013
Keywords : Moral	0.064	0.056	0.060
Keywords : Motivation	0.012	0.024	0.017
Keywords : Non-cooperative games	0.017	0.056	0.033
Keywords : Perception	0.040	0.032	0.037
Keywords : Personality	0.012	0.008	0.010
Keywords : Persuasion	0.006	0.008	0.007
Keywords : Prejudice	0.017	0.024	0.020
Keywords : Priming	0.017	0.008	0.013
Keywords : Recognition	0.006	0.008	0.007
Keywords : Relationship	0.058	0.008	0.037
Keywords : Similarity	0.023	0.016	0.020
Keywords : Social class	0.012	0.008	0.010
Keywords : Stereotype	0.006	0.024	0.013
Keywords : The self	0.040	0.024	0.033
Keywords : Time	0.000	0.016	0.007
Keywords : Visual perceptions	0.012	0.008	0.010

Panel B: Continuous variables (averages)

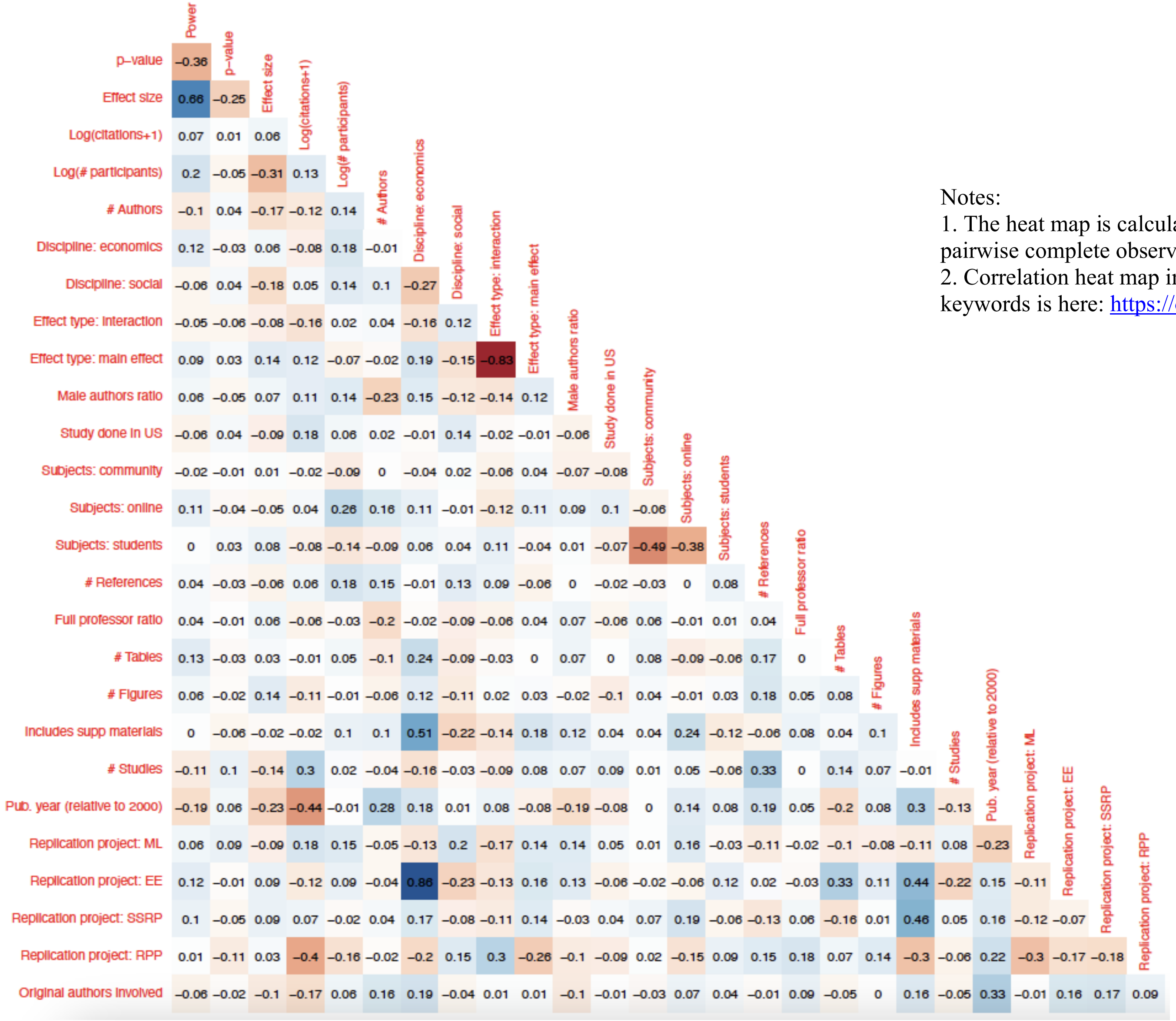
Variable	Not Replicated	Replicated	Total
Prediction market end price	0.521	0.723	0.619
Citation count[1,2]	672.532	1367.754	965.502
Publication year (relative to 2000)	6.364	0.119	3.732
Effect size (original paper)	0.344	0.472	0.398
P-value (original paper)	0.027	0.012	0.021
Post-hoc power (original paper)	0.674	0.812	0.731
Number of participants[2]	103.121	1997.427	894.010
Number of figures	2.139	2.738	2.391
Number of references	41.480	36.008	39.174
Number of studies	3.283	2.984	3.157
Number of tables	2.017	2.341	2.154
Number of authors	3.121	2.635	2.916
Rate of full professor authors	0.500	0.532	0.513
Rate of male authors	0.668	0.780	0.715
Word count abstract	144.169	147.400	145.524
Word count study	1432.780	1800.587	1587.776
Word count paper	6761.358	6870.833	6807.492

[1] Using only RPP, SSRP, EE replications, citation counts are 314.2 and 281.4, respectively, which replicates Serra-Garcia & Gneezy (2021).

[2] Citation count and Number of participants have right skewed distributions. The median citation counts are 299 (not replicated)/380.5 (replicated)/345 (total) and the median number of participants are 67 (not replicated)/78.5 (replicated)/72 (total).

The summary statistics for effect size, p-value, number of participants, post-hoc power, and rate of full professor authors are calculated exclusive of missing values.

Panel C: Correlation table



	Power	p–value	Effect size	Log(citations+1)	Log(# participants)	# Authors	Discipline: economics	Discipline: social	Effect type: interaction	Effect type: main effect	Male authors ratio	Study done in US	Subjects: community	Subjects: online	Subjects: students	# References	Full professor ratio	# Tables	# Figures	Includes supp materials	# Studies	Pub. year (relative to 2000)	Replication project: ML	Replication project: EE	Replication project: SSRP	Replication project: RPP
p–value	–0.36																									
Effect size	0.66	–0.25																								
Log(citations+1)	0.07	0.01	0.06																							
Log(# participants)	0.2	–0.05	–0.31	0.13																						
# Authors	–0.1	0.04	–0.17	–0.12	0.14																					
Discipline: economics	0.12	–0.03	0.06	–0.08	0.18	–0.01																				
Discipline: social	–0.06	0.04	–0.18	0.05	0.14	0.1	–0.27																			
Effect type: interaction	–0.05	–0.06	–0.08	–0.16	0.02	0.04	–0.16	0.12																		
Effect type: main effect	0.09	0.03	0.14	0.12	–0.07	–0.02	0.19	–0.15	–0.83																	
Male authors ratio	0.06	–0.05	0.07	0.11	0.14	–0.23	0.15	–0.12	–0.14	0.12																
Study done in US	–0.06	0.04	–0.09	0.18	0.06	0.02	–0.01	0.14	–0.02	–0.01	–0.06															
Subjects: community	–0.02	–0.01	0.01	–0.02	–0.09	0	–0.04	0.02	–0.06	0.04	–0.07	–0.08														
Subjects: online	0.11	–0.04	–0.05	0.04	0.26	0.16	0.11	–0.01	–0.12	0.11	0.09	0.1	–0.06													
Subjects: students	0	0.03	0.08	–0.08	–0.14	–0.09	0.06	0.04	0.11	–0.04	0.01	–0.07	–0.49	–0.38												
# References	0.04	–0.03	–0.06	0.06	0.18	0.15	–0.01	0.13	0.09	–0.06	0	–0.02	–0.03	0	0.08											
Full professor ratio	0.04	–0.01	0.06	–0.06	–0.03	–0.2	–0.02	–0.09	–0.06	0.04	0.07	–0.06	0.06	–0.01	0.01	0.04										
# Tables	0.13	–0.03	0.03	–0.01	0.05	–0.1	0.24	–0.09	–0.03	0	0.07	0	0.08	–0.09	–0.06	0.17	0									
# Figures	0.06	–0.02	0.14	–0.11	–0.01	–0.06	0.12	–0.11	0.02	0.03	–0.02	–0.1	0.04	–0.01	0.03	0.18	0.05	0.08								
Includes supp materials	0	–0.06	–0.02	–0.02	0.1	0.1	0.51	–0.22	–0.14	0.18	0.12	0.04	0.04	0.24	–0.12	–0.06	0.08	0.04	0.1							
# Studies	–0.11	0.1	–0.14	0.3	0.02	–0.04	–0.16	–0.03	–0.09	0.08	0.07	0.09	0.01	0.05	–0.06	0.33	0	0.14	0.07	–0.01						
Pub. year (relative to 2000)	–0.19	0.06	–0.23	–0.44	–0.01	0.28	0.18	0.01	0.08	–0.08	–0.19	–0.08	0	0.14	0.08	0.19	0.05	–0.2	0.08	0.3	–0.13					
Replication project: ML	0.06	0.09	–0.09	0.18	0.15	–0.05	–0.13	0.2	–0.17	0.14	0.14	0.05	0.01	0.16	–0.03	–0.11	–0.02	–0.1	–0.08	–0.11	0.08	–0.23				
Replication project: EE	0.12	–0.01	0.09	–0.12	0.09	–0.04	0.86	–0.23	–0.13	0.16	0.13	–0.06	–0.02	–0.06	0.12	0.02	–0.03	0.33	0.11	0.44	–0.22	0.15	–0.11			
Replication project: SSRP	0.1	–0.05	0.09	0.07	–0.02	0.04	0.17	–0.08	–0.11	0.14	–0.03	0.04	0.07	0.19	–0.06	–0.13	0.06	–0.16	0.01	0.46	0.05	0.16	–0.12	–0.07		
Replication project: RPP	0.01	–0.11	0.03	–0.4	–0.16	–0.02	–0.2	0.15	0.3	–0.26	–0.1	–0.09	0.02	–0.15	0.09	0.15	0.18	0.07	0.14	–0.3	–0.06	0.22	–0.3	–0.17	–0.18	
Original authors involved	–0.06	–0.02	–0.1	–0.17	0.06	0.16	0.19	–0.04	0.01	0.01	–0.1	–0.01	–0.03	0.07	0.04	–0.01	0.09	–0.05	0	0.16	–0.05	0.33	–0.01	0.16	0.17	0.09

Notes:
1. The heat map is calculated based on pairwise complete observations.
2. Correlation heat map including all keywords is here: https://osf.io/w456x

D. Additional Results

Table S6: Logistic regression coefficients for metadata variables (1 = replicated)

Variable group	Variable name	Full Text / Study	Abstract
Study controls	Power	-0.341	-0.298
	p-value	-0.616	-0.639
	Effect size	0.812**	0.742*
	Log(# participants)	0.634*	0.623*
	Study done in US	0.009	0.025
	Subjects: community	0.189	0.053
	Subjects: online	-0.009	-0.140
	Subjects: students	-0.069	-0.093
	Effect type: interaction	-0.313	-0.262
	Effect type: main effect	0.121	0.189
Paper controls	Discipline: economics	0.019	-0.110
	Discipline: social psychology	0.139	0.170
	Log(citations+1)	0.027	0.018
	# References	-0.154	-0.357†
	# Tables	-0.013	0.164
	# Figures	0.291†	0.434*
	# Studies	-0.248	-0.071
	Includes supp materials	-0.152	-0.189
	Pub. year (relative to 2000)	-0.430†	-0.485*
Authors controls	# Authors	-0.061	-0.065
	Full professor ratio	0.026	-0.027
	Male authors ratio	0.183	0.221
Replication controls	Replication project: EE	0.115	0.111
	Replication project: ML	0.211	0.092
	Replication project: RPP	0.144	0.076
	Replication project: SSRP	0.314	0.329
	Original authors involved	-0.167	-0.162
Keywords	Keyword: anchoring	-0.073	-0.078
	Keyword: attention	-0.129	-0.152
	Keyword: attitude	-0.139	-0.194
	Keyword: bias	-0.057	-0.060
	Keyword: cognitive processes	-0.096	-0.163
	Keyword: construal level	-0.058	-0.081
	Keyword: consumer economics	-0.088	-0.022
	Keyword: emotion	-0.133	-0.135

	Keyword: individual behavior	-0.032	0.027
	Keyword: learning	0.017	0.013
	Keyword: memory	-0.170	-0.209
	Keyword: money	0.064	0.053
	Keyword: moral	-0.176	-0.205
	Keyword: motivation	0.019	0.018
	Keyword: perceptions	-0.297	-0.337†
	Keyword: prejudice	-0.041	-0.071
	Keyword: similarity	-0.356*	-0.321
	Keyword: the self	-0.211	-0.191
	Keyword: social psych other	-0.559†	-0.645*

Note. This table presents coefficients for the metadata we control for in all our analyses, organized by type of text in the analysis. The dependent variable is replication (binary). The models are based on binary logit model with standardized variables. Study and Full text models are the same because the models include only the metadata (no text variables) and the dependent variable (successfully replicated =1) is at the study level; the abstract model differs slightly due to excluding papers with no abstract (n = 286).
** p < 0.01, * p < 0.05, † p < 0.1

Table S7: Predictions with different hyperparameters (Ridge)

Hyperparameters: continuous bag of words (CBOW), 3-word window, 100 dimensions, 50 epochs, without stop words

Text section	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
Study	All papers in dataset	Average test AUC	0.6961	0.6881	0.7110	0.6827	0.7130	0.7210
		SD across splits	0.0607	0.0648	0.0622	0.0637	0.0612	0.0612
		% w/ improvement		45.52%	61.18%	42.63%	64.51%	67.44%
	Psychology papers only	Average test AUC	0.6861	0.6761	0.7020	0.6749	0.7054	0.7162
		SD across splits	0.0664	0.0687	0.0665	0.0681	0.0663	0.0654
		% w/ improvement		44.84%	60.78%	44.45%	64.04%	68.27%
	Predicting new papers from old	Test AUC	0.6729	0.7506	0.7794	0.7384	0.7805	0.7894
	Market prediction	Correlation	0.5129	0.4996	0.5471	0.5794	0.6179	0.5933
Full text	All papers in dataset	Average test AUC	0.6961	0.6846	0.7078	0.6713	0.7022	0.7143
		SD across splits	0.0607	0.0653	0.0624	0.0664	0.0630	0.0627
		% w/ improvement		43.58%	59.54%	35.82%	55.37%	62.95%
	Psychology papers only	Average test AUC	0.6861	0.6776	0.7009	0.6673	0.6909	0.7089
		SD across splits	0.0664	0.0688	0.0666	0.0708	0.0693	0.0670
		% w/ improvement		45.69%	59.83%	40.34%	53.79%	63.89%
	Predicting new papers from old	Test AUC	0.6729	0.7555	0.7749	0.7633	0.8016	0.7583
	Market prediction	Correlation	0.5129	0.5668	0.5895	0.5328	0.5844	0.6103
Abstract	All papers in dataset	Average test AUC	0.6752	0.6329	0.6721	0.5715	0.6651	0.6645
		SD across splits	0.0678	0.0726	0.0702	0.0722	0.0701	0.0721
		% w/ improvement		30.36%	48.33%	12.52%	43.50%	43.75%
	Psychology papers only	Average test AUC	0.6544	0.6255	0.6581	0.5990	0.6484	0.6472
		SD across splits	0.0728	0.0771	0.0757	0.0770	0.0764	0.0778
		% w/ improvement		37.15%	53.08%	27.24%	46.77%	46.34%
	Predicting new papers from old	Test AUC	0.7670	0.7184	0.7963	0.4730	0.6991	0.7654
	Market prediction	Correlation	0.4781	0.4941	0.5178	0.3589	0.5516	0.5575

Hyperparameters: continuous bag of words (CBOW), 3-word window, 100 dimensions, 100 epochs, with stop words

Text section	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
Study	All papers in dataset	Average test AUC	0.6961	0.7020	0.7199	0.6827	0.7130	0.7240
		SD across splits	0.0607	0.0626	0.0612	0.0637	0.0612	0.0608
		% w/ improvement		52.82%	67.90%	42.63%	64.51%	70.00%
	Psychology papers only	Average test AUC	0.6861	0.6907	0.7110	0.6749	0.7054	0.7165
		SD across splits	0.0664	0.0674	0.0660	0.0681	0.0663	0.0658
		% w/ improvement		51.91%	65.99%	44.45%	64.04%	68.89%
	Predicting new papers from old	Test AUC	0.6729	0.7749	0.8193	0.7384	0.7805	0.8204
	Market prediction	Correlation	0.5129	0.5988	0.6017	0.5794	0.6179	0.6228
Full text	All papers in dataset	Average test AUC	0.6961	0.6985	0.7136	0.6713	0.7022	0.7158
		SD across splits	0.0607	0.0648	0.0630	0.0664	0.0630	0.0632
		% w/ improvement		51.24%	63.59%	35.82%	55.37%	64.20%
	Psychology papers only	Average test AUC	0.6861	0.6916	0.7053	0.6673	0.6909	0.7098
		SD across splits	0.0664	0.0689	0.0677	0.0708	0.0693	0.0677
		% w/ improvement		53.15%	62.88%	40.34%	53.79%	64.52%
	Predicting new papers from old	Test AUC	0.6729	0.7622	0.8049	0.7633	0.8016	0.8149
	Market prediction	Correlation	0.5129	0.5780	0.6014	0.5328	0.5844	0.6116
Abstract	All papers in dataset	Average test AUC	0.6752	0.6457	0.6742	0.5715	0.6651	0.6641
		SD across splits	0.0678	0.0717	0.0713	0.0722	0.0701	0.0721
		% w/ improvement		36.08%	49.55%	12.52%	43.50%	43.02%
	Psychology papers only	Average test AUC	0.6544	0.6382	0.6585	0.5990	0.6484	0.6492
		SD across splits	0.0728	0.0760	0.0760	0.0770	0.0764	0.0768
		% w/ improvement		42.64%	52.83%	27.24%	46.77%	46.92%
	Predicting new papers from old	Test AUC	0.7670	0.7230	0.8071	0.4730	0.6991	0.7716
	Market prediction	Correlation	0.4781	0.5168	0.5707	0.3589	0.5516	0.5646

Hyperparameters: continuous bag of words (CBOW), 3-word window, 150 dimensions, 50 epochs, with stop words

Text section	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
Study	All papers in dataset	Average test AUC	0.6961	0.7107	0.7248	0.6827	0.7130	0.7265
		SD across splits	0.0607	0.0626	0.0607	0.0637	0.0612	0.0606
		% w/ improvement		58.32%	69.72%	42.63%	64.51%	70.69%
	Psychology papers only	Average test AUC	0.6861	0.7027	0.7179	0.6749	0.7054	0.7199
		SD across splits	0.0664	0.0666	0.0651	0.0681	0.0663	0.0652
		% w/ improvement		58.17%	69.01%	44.45%	64.04%	69.87%
	Predicting new papers from old	Test AUC	0.6729	0.8115	0.8237	0.7384	0.7805	0.8248
	Market prediction	Correlation	0.5129	0.5619	0.6229	0.5794	0.6179	0.6213
Full text	All papers in dataset	Average test AUC	0.6961	0.6974	0.7118	0.6713	0.7022	0.7140
		SD across splits	0.0607	0.0646	0.0632	0.0664	0.0630	0.0632
		% w/ improvement		50.50%	61.67%	35.82%	55.37%	62.37%
	Psychology papers only	Average test AUC	0.6861	0.6886	0.7028	0.6673	0.6909	0.7057
		SD across splits	0.0664	0.0688	0.0678	0.0708	0.0693	0.0681
		% w/ improvement		50.93%	60.43%	40.34%	53.79%	61.69%
	Predicting new papers from old	Test AUC	0.6729	0.7833	0.8016	0.7633	0.8016	0.8027
	Market prediction	Correlation	0.5129	0.6106	0.6288	0.5328	0.5844	0.6342
Abstract	All papers in dataset	Average test AUC	0.6752	0.6332	0.6685	0.5715	0.6651	0.6614
		SD across splits	0.0678	0.0720	0.0716	0.0722	0.0701	0.0723
		% w/ improvement		30.70%	46.16%	12.52%	43.50%	41.69%
	Psychology papers only	Average test AUC	0.6544	0.6301	0.6525	0.5990	0.6484	0.6467
		SD across splits	0.0728	0.0751	0.0757	0.0770	0.0764	0.0761
		% w/ improvement		38.50%	48.21%	27.24%	46.77%	45.12%
	Predicting new papers from old	Test AUC	0.7670	0.7168	0.8025	0.4730	0.6991	0.7485
	Market prediction	Correlation	0.4781	0.5365	0.5695	0.3589	0.5516	0.5762

Hyperparameters: continuous bag of words (CBOW), 3-word window, 50 dimensions, 50 epochs, with stop words

					Metadata +		Metadata	Metadata +
Text				Text	text	Text	+ text	text features
section	Prediction	Stat	Metadata	embeddings	embeddings	features	features	+ embeddings
Study	All papers in dataset	Average test AUC	0.6961	0.6989	0.7228	0.6827	0.7130	0.7233
		SD across splits	0.0607	0.0637	0.0612	0.0637	0.0612	0.0610
		% w/ improvement		51.01%	71.13%	42.63%	64.51%	69.91%
	Psychology papers only	Average test AUC	0.6861	0.6898	0.7165	0.6749	0.7054	0.7178
		SD across splits	0.0664	0.0683	0.0660	0.0681	0.0663	0.0658
		% w/ improvement		51.61%	71.07%	44.45%	64.04%	70.01%
	Predicting new papers from old	Test AUC	0.6729	0.7650	0.8149	0.7384	0.7805	0.8182
	Market prediction	Correlation	0.5129	0.5480	0.5704	0.5794	0.6179	0.6305
Full text	All papers in dataset	Average test AUC	0.6961	0.6934	0.7131	0.6713	0.7022	0.7134
		SD across splits	0.0607	0.0650	0.0626	0.0664	0.0630	0.0632
		% w/ improvement		48.36%	64.09%	35.82%	55.37%	63.06%
	Psychology papers only	Average test AUC	0.6861	0.6863	0.7064	0.6673	0.6909	0.7066
		SD across splits	0.0664	0.0695	0.0676	0.0708	0.0693	0.0684
		% w/ improvement		49.77%	64.02%	40.34%	53.79%	63.26%
	Predicting new papers from old	Test AUC	0.6729	0.7744	0.8060	0.7633	0.8016	0.8060
	Market prediction	Correlation	0.5129	0.5603	0.6119	0.5328	0.5844	0.6161
Abstract	All papers in dataset	Average test AUC	0.6752	0.6387	0.6812	0.5715	0.6651	0.6699
		SD across splits	0.0678	0.0740	0.0709	0.0722	0.0701	0.0718
		% w/ improvement		32.75%	53.99%	12.52%	43.50%	46.58%
	Psychology papers only	Average test AUC	0.6544	0.6302	0.6647	0.5990	0.6484	0.6543
		SD across splits	0.0728	0.0775	0.0756	0.0770	0.0764	0.0776
		% w/ improvement		39.01%	56.88%	27.24%	46.77%	50.02%
	Predicting new papers from old	Test AUC	0.7670	0.7369	0.8241	0.4730	0.6991	0.7901
	Market prediction	Correlation	0.4781	0.4947	0.5419	0.3589	0.5516	0.5691

Hyperparameters: continuous bag of words (CBOW), 5-word window, 100 dimensions, 50 epochs, with stop words

Text section	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
Study	All papers in dataset	Average test AUC	0.6961	0.6936	0.7158	0.6827	0.7130	0.7208
		SD across splits	0.0607	0.0635	0.0615	0.0637	0.0612	0.0611
		% w/ improvement		47.84%	65.17%	42.63%	64.51%	67.91%
	Psychology papers only	Average test AUC	0.6861	0.6809	0.7061	0.6749	0.7054	0.7139
		SD across splits	0.0664	0.0687	0.0668	0.0681	0.0663	0.0660
		% w/ improvement		47.06%	62.95%	44.45%	64.04%	67.88%
	Predicting new papers from old	Test AUC	0.6729	0.7949	0.8115	0.7384	0.7805	0.8104
	Market prediction	Correlation	0.5129	0.5893	0.6052	0.5794	0.6179	0.6149
Full text	All papers in dataset	Average test AUC	0.6961	0.6920	0.7090	0.6713	0.7022	0.7127
		SD across splits	0.0607	0.0653	0.0636	0.0664	0.0630	0.0635
		% w/ improvement		47.46%	60.51%	35.82%	55.37%	61.93%
	Psychology papers only	Average test AUC	0.6861	0.6837	0.7005	0.6673	0.6909	0.7066
		SD across splits	0.0664	0.0693	0.0683	0.0708	0.0693	0.0680
		% w/ improvement		49.05%	59.34%	40.34%	53.79%	62.71%
	Predicting new papers from old	Test AUC	0.6729	0.7644	0.7949	0.7633	0.8016	0.7971
	Market prediction	Correlation	0.5129	0.5870	0.6072	0.5328	0.5844	0.6156
Abstract	All papers in dataset	Average test AUC	0.6752	0.6483	0.6801	0.5715	0.6651	0.6702
		SD across splits	0.0678	0.0713	0.0703	0.0722	0.0701	0.0712
		% w/ improvement		37.26%	53.33%	12.52%	43.50%	46.93%
	Psychology papers only	Average test AUC	0.6544	0.6330	0.6597	0.5990	0.6484	0.6513
		SD across splits	0.0728	0.0746	0.0749	0.0770	0.0764	0.0760
		% w/ improvement		39.95%	53.34%	27.24%	46.77%	48.07%
	Predicting new papers from old	Test AUC	0.7670	0.7539	0.8071	0.4730	0.6991	0.7824
	Market prediction	Correlation	0.4781	0.5128	0.5650	0.3589	0.5516	0.5689

Hyperparameters: skipgram, 3-word window, 100 dimensions, 50 epochs, with stop words

Text section	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
Study	All papers in dataset	Average test AUC	0.6951	0.6961	0.7124	0.7273	0.6827	0.7130
		SD across splits	0.0608	0.0607	0.0610	0.0603	0.0637	0.0612
		% w/ improvement			59.06%	73.44%	42.63%	64.51%
	Psychology papers only	Average test AUC	0.6848	0.6861	0.7021	0.7180	0.6749	0.7054
		SD across splits	0.0665	0.0664	0.0655	0.0653	0.0681	0.0663
		% w/ improvement			57.68%	70.79%	44.45%	64.04%
	Predicting new papers from old	Test AUC	0.6729	0.6729	0.8381	0.8370	0.7384	0.7805
	Market prediction	Correlation	0.5134	0.5129	0.6264	0.6180	0.5794	0.6179
Full text	All papers in dataset	Average test AUC	0.6951	0.6961	0.6892	0.7094	0.6713	0.7022
		SD across splits	0.0608	0.0607	0.0659	0.0641	0.0664	0.0630
		% w/ improvement			45.78%	60.43%	35.82%	55.37%
	Psychology papers only	Average test AUC	0.6848	0.6861	0.6835	0.7007	0.6673	0.6909
		SD across splits	0.0665	0.0664	0.0698	0.0684	0.0708	0.0693
		% w/ improvement			48.66%	59.48%	40.34%	53.79%
	Predicting new papers from old	Test AUC	0.6729	0.6729	0.7733	0.8049	0.7633	0.8016
	Market prediction	Correlation	0.5134	0.5129	0.5883	0.6107	0.5328	0.5844
Abstract	All papers in dataset	Average test AUC	0.6856	0.6752	0.6355	0.6723	0.5715	0.6651
		SD across splits	0.0645	0.0678	0.0737	0.0717	0.0722	0.0701
		% w/ improvement			30.76%	48.06%	12.52%	43.50%
	Psychology papers only	Average test AUC	0.6689	0.6544	0.6265	0.6569	0.5990	0.6484
		SD across splits	0.0700	0.0728	0.0783	0.0758	0.0770	0.0764
		% w/ improvement			37.34%	50.91%	27.24%	46.77%
	Predicting new papers from old	Test AUC	0.6411	0.7670	0.7369	0.8040	0.4730	0.6991
	Market prediction	Correlation	0.5124	0.4781	0.5180	0.5541	0.3589	0.5516

Note: This table is similar to Table 1 in the manuscript, just with different hyperparameters, showing results are consistent and robust to these parameters.

Table S8: Predictions with different models (mean AUC over 10,000 runs)

		Elastic net		XGBoost	
Text	Specification	Ave. AUC	% runs with improvement[1]	Ave. AUC	% runs with improvement[1]
Study	Metadata	0.6936		0.6815	
	Text embeddings	0.6737	39.61%	0.6244	25.10%
	Metadata + embeddings	0.7131	65.47%	0.6741	46.76%
	Text features	0.6529	30.15%	0.6440	32.45%
	Metadata + text features	0.6818	43.72%	0.6885	54.61%
	Metadata + text features + embeddings	0.7014	57.38%	0.6769	48.09%
Full text	Metadata	0.6936		0.6815	
	Text embeddings	0.6845	44.65%	0.6665	42.81%
	Metadata + embeddings	0.6935	52.77%	0.6856	53.01%
	Text features	0.6703	38.42%	0.6118	20.76%
	Metadata + text features	0.6882	48.86%	0.6803	49.93%
	Metadata + text features + embeddings	0.6984	55.66%	0.6887	54.51%
Abstract	Metadata	0.0706		0.6611	
	Text embeddings	0.0755	22.68%	0.5735	19.42%
	Metadata + embeddings	0.0690	18.63%	0.6203	30.65%
	Text features	0.0754	12.45%	0.5440	12.22%
	Metadata + text features	0.0710	30.88%	0.6225	30.15%
	Metadata + text features + embeddings	0.0685	15.29%	0.6220	31.54%

Note: We estimate the XGBoost model using the xgboost package in R, while we estimate elastic net and ridge using the glmnet package in R. As with the ridge model used in our main analyses, all AUC figures are calculated by averaging over repeated train/test splits where, at each split, 80% of observations are used to train the model and 20% are used to evaluate the test AUC. Within each training set, we first tune the model hyperparameter(s) using 10-fold cross-validation, then use the chosen hyperparameter(s) to train the model on the full training data and then use the predictions from this full model to make predictions on the test set. For XGBoost, we tune over the learning rate, number of rounds and maximum depth. We leave other hyperparameters to be unconstrained or set at their default value. For elastic net, we place an equal 50%-50% weight on the ℓ_1 and ℓ_2 penalties and calibrate the size of the total penalty. The embeddings hyperparameters are CBOW, 3-window, 100 dimensions, 50 epochs, with stop words (the same as in Table 1 in the paper).

[1] The % of runs out of 10,000 in which the model is predicting replication better than the model with metadata only.

Table S9: Text features standardized LASSO and one-at-a-time coefficients

Panel A: Study text

Text features associated with nonreplicability	LASSO coefficient	One-at-a-time coefficient [SE]	Text features associated with replicability	LASSO coefficient	One-at-a-time coefficient [SE]
Third person plural	-0.437	-0.287 [0.171]†	Periods	0.432	0.164 [0.172]
Semicolons	-0.409	-0.479 [0.178]**	Interrogatives	0.350	0.262 [0.154]†
Anxiety	-0.341	-0.407 [0.184]*	Present focus	0.344	0.200 [0.192]
Articles	-0.335	-0.121 [0.169]	Leisure	0.333	0.212 [0.166]
Future focus	-0.306	-0.264 [0.157]†	Number	0.315	0.239 [0.169]
Reveal (See in LIWC)	-0.278	-0.298 [0.160]†	Quantifiers	0.310	0.467 [0.170]**
Feel	-0.278	-0.415 [0.182]*	Prepositions	0.301	0.363 [0.169]*
Adjectives	-0.262	-0.085 [0.148]	Auxiliary verbs	0.297	0.311 [0.167]†
Affiliation	-0.253	-0.161 [0.170]	Space	0.252	0.201 [0.159]
Positive emotion	-0.239	-0.425 [0.180]*	Words per sentence	0.201	0.083 [0.160]
Work	-0.187	-0.279 [0.159]†	Certainty	0.182	0.301 [0.149]*
Colons	-0.172	-0.224 [0.171]	Order (Power in LIWC)	0.134	0.032 [0.161]
Achievement	-0.135	-0.290 [0.183]	Differentiation	0.119	0.034 [0.151]
Reward	-0.106	-0.278 [0.198]	Male references	0.117	0.141 [0.176]
Risk	-0.071	-0.262 [0.159]†	Words > 6 letters[1]	0.063	-0.242 [0.167]
Impersonal pronouns	-0.068	-0.018 [0.166]	Common verbs	0.058	0.175 [0.163]
Female references	-0.056	-0.139 [0.188]	Conjunctions	0.047	0.142 [0.152]
Other punctuations	-0.019	-0.115 [0.173]	Weak (Health in LIWC)	0.026	0.049 [0.148]
Abstraction	N/A	-0.414 [0.189]*	Parentheses	0.011	0.034 [0.183]
First person plural	-	-0.289 [0.173]†	Word count	-	0.179 [0.222]
Clout	N/A	-0.259 [0.187]	Past focus	-	0.125 [0.173]
Social processes	N/A	-0.227 [0.191]	Readability	N/A	0.103 [0.159]
Analytical thinking	N/A	-0.177 [0.157]	Discrepancy	-	0.075 [0.165]
Insight	-	-0.128 [0.157]	Authenticity	N/A	0.069 [0.156]
Dictionary words	N/A	-0.126 [0.179]	Adverbs	-	0.057 [0.157]
Emotional tone	N/A	-0.086 [0.162]	Motion	-	0.049 [0.155]
Dashes	-	-0.075 [0.159]	Sadness	-	0.047 [0.159]
Quatation marks	-	-0.071 [0.165]	Commas	-	0.045 [0.168]
Money	-	-0.070 [0.181]	Obfuscation	N/A	0.028 [0.175]
Tentative	-	-0.057 [0.147]	Comparisons	-	0.024 [0.153]
Anger	-	-0.053 [0.162]	Apostrophes	-	0.008 [0.156]
Time	-	-0.013 [0.156]	Negation	-	0.005 [0.149]
Causation	-	-0.002 [0.155]			

Note for all panels: these tables present the coefficients for LIWC low-level dictionaries and summary variables, readability, abstraction, and obfuscation. Missing in LASSO means the dictionary was not selected by LASSO and N/A means the variable was not entered into LASSO due to redundancy with other dictionaries.

[1] The only text feature that is associated with replicability (LASSO) and nonreplicability (one-at-a-time).

** $p < 0.01$, * $p < 0.05$, † $p < 0.1$

Panel B: Full text

Text features associated with nonreplicability	LASSO coefficient	One-at-a-time coefficient [SE]	Text features associated with replicability	LASSO coefficient	One-at-a-time coefficient [SE]
Anxiety	-2.458	-0.897 [0.283]**	Auxiliary verbs	0.005	0.220 [0.167]
Other punctuations	-2.328	-0.130 [0.176]	Reward	0.040	0.015 [0.163]
Reveal (See in LIWC)	-1.816	-0.446 [0.182]*	Impersonal pronouns	0.073	0.127 [0.158]
Affiliation	-1.766	-0.056 [0.167]	Quotation marks	0.107	0.185 [0.167]
Third person plural	-1.533	-0.355 [0.178]*	Space	0.107	0.082 [0.163]
Achievement	-1.257	-0.160 [0.172]	Tentative	0.150	0.083 [0.162]
Male references	-1.045	-0.061 [0.188]	Certainty	0.172	0.509 [0.163]**
Adjectives	-1.004	-0.169 [0.156]	Present focus	0.209	0.067 [0.184]
Insight	-0.997	-0.038 [0.163]	Discrepancy	0.218	0.127 [0.162]
Home	-0.919	-0.067 [0.139]	Commas	0.239	0.015 [0.193]
Money	-0.902	-0.150 [0.212]	Dashes	0.250	0.088 [0.168]
Negation	-0.802	-0.083 [0.155]	Work	0.302	-0.201 [0.159]
Comparisons	-0.750	-0.107 [0.153]	First person plural	0.407	-0.081 [0.187]
Articles	-0.742	-0.051 [0.198]	Positive emotion	0.435	-0.198 [0.176]
Ingestion	-0.687	-0.097 [0.156]	Second person	0.482	0.240 [0.153]
Anger	-0.661	-0.225 [0.170]	Colons	0.533	-0.066 [0.151]
Apostrophies	-0.611	-0.130 [0.163]	Parentheses	0.576	0.116 [0.207]
First person singular	-0.606	-0.221 [0.166]	Adverbs	0.600	0.105 [0.160]
Religion	-0.559	-0.111 [0.179]	Risk	0.650	-0.235 [0.173]
Future focus	-0.539	-0.190 [0.157]	Body	0.657	0.025 [0.160]
Feel	-0.524	-0.422 [0.246]†	Periods	0.658	0.010 [0.160]
Causation	-0.442	-0.006 [0.159]	Female references	0.729	-0.040 [0.165]
Time	-0.376	-0.151 [0.167]	Word count	0.767	0.514 [0.273]†
Third person singular	-0.337	-0.177 [0.173]	Order (Power in LIWC)	0.971	0.036 [0.166]
Question marks	-0.290	-0.072 [0.162]	Leisure	1.256	0.247 [0.174]
Conjunctions	-0.125	-0.050 [0.158]	Words per sentence	1.259	0.267 [0.176]
Hear	-0.095	0.122 [0.158]	Differentiation	1.547	0.096 [0.155]
Semicolons	-0.059	-0.171 [0.175]	Assent	1.796	0.325 [0.182]†
Motion	-0.049	-0.017 [0.167]	Past focus	1.803	0.203 [0.193]
Abstraction	N/A	-0.333 [0.206]	Friends	1.907	0.048 [0.152]
Clout	N/A	-0.262 [0.190]	Quantifiers[2]	1.980	0.496 [0.188]**
Analytical thinking	N/A	-0.129 [0.151]	Prepositions[2]	-0.325	0.097 [0.178]
Readability	N/A	-0.080 [0.164]	Number[2]	-0.247	0.127 [0.173]
			Weak (Health in LIWC)[2]	-0.126	0.089 [0.151]
			Interrogatives	-0.044	0.225 [0.147]
			Words > 6 letters	-0.063	0.010 [0.173]
			Authenticity	N/A	0.026 [0.162]
			Common verbs	-	0.046 [0.161]
			Obfuscation	N/A	0.094 [0.179]
			Emotional tone	N/A	0.108 [0.162]
			Sadness	-	0.160 [0.165]

[2] Dictionaries that are associated with replicability in the study text.

Panel C: Abstract text

Text features associated with nonreplicability	LASSO coefficient	One-at-a-time coefficient [SE]	Text features associated with replicability	LASSO coefficient	One-at-a-time coefficient [SE]
Adjectives	-0.063	-0.207 [0.165]	Certainty	0.208	0.441 [0.174]*
Abstraction	N/A	-0.374 [0.182]*	Quantifiers	0.183	0.375 [0.162]*
Power	-	-0.263 [0.174]	Past focus	0.056	0.374 [0.180]*
Risk	-	-0.199 [0.175]	Emotional tone	N/A	0.286 [0.170]†
Differentiation	-	-0.194 [0.163]	Achievement	-	0.187 [0.182]
Obfuscation	N/A	-0.178 [0.167]	Motion	-	0.172 [0.165]
Causation	-	-0.175 [0.169]	Interrogatives	-	0.139 [0.153]
Third person plural	-	-0.170 [0.170]	Positive emotion	-	0.135 [0.170]
See	-	-0.162 [0.167]	Commas	-	0.134 [0.176]
Clout	N/A	-0.146 [0.175]	Reward	-	0.131 [0.168]
Comparisons	-	-0.139 [0.162]	Word count	-	0.127 [0.176]
Tentative	-	-0.135 [0.162]	Prepositions	-	0.117 [0.172]
Analytical thinking	N/A	-0.129 [0.161]	Auxiliary verbs	-	0.116 [0.161]
Authenticity	N/A	-0.121 [0.161]	Periods	-	0.103 [0.151]
Future focus	-	-0.102 [0.164]	Words > 6 letters	-	0.098 [0.167]
Time	-	-0.085 [0.160]	Articles	-	0.089 [0.185]
Space	-	-0.076 [0.160]	Discrepancy	-	0.070 [0.177]
Insight	-	-0.067 [0.167]	Words per sentence	-	0.069 [0.160]
Present focus	-	-0.066 [0.169]	Dashes	-	0.062 [0.161]
Adverbs	-	-0.028 [0.168]	Common verbs	-	0.059 [0.163]
Affiliation	-	-0.014 [0.168]	Conjunctions	-	0.042 [0.162]
Readability	N/A	-0.014 [0.161]	Parentheses	-	0.037 [0.163]
Number	-	-0.013 [0.171]	Negation	-	0.029 [0.155]
Impersonal pronouns	-	-0.012 [0.160]			
Work	-	-0.008 [0.159]			

Note (for all panels of Table S9): The metadata variables are unregularized (i.e., they do not have a LASSO penalty) so as to ensure they are fully controlled for. The LASSO penalty for the textual variables is selected via 10-fold cross-validation. To mitigate collinearity, we include only low-level LIWC dictionaries (i.e., excluding high-level dictionaries and summary variables) in the LASSO model, reporting their coefficients only in the one-at-a-time regressions where collinearity is not a concern. We note that the full text LASSO has worse collinearity than the abstract and study LASSO models since in the full text fewer dictionaries are filtered out in the pre-processing stage (i.e., removing dictionaries that occur in fewer than 50% of documents). As a result, the full text LASSO has more unstable coefficients than the other two models (larger magnitude coefficients that can change substantially depending on the LASSO penalty, whereas the abstract and study LASSO models give smaller coefficients that are not sensitive to the LASSO penalty).

Table S10: Top words for each LIWC low-level dictionary based on all study texts in our corpus

Panel A: Dictionaries associated with replicability

Dictionary	Top words
Adverbs	when, about, there, also, only, however, how, such, so, where, very, probability, even, rather, respectively, again, therefore, well, finally, specifically, often, indeed, now, instead, still, just, relatively, here, though, simply, why, actually, immediately, probabilities, extremely, almost, generally, hence, clearly, completely, typically, never, somewhat, particularly, around, too, especially, yet, fully, already
Auxiliary verbs	were, was, is, are, be, would, had, did, can, should, will, may, could, has, been, do, being, might, does, having, cannot, done, must, become, becomes, let, doing, becoming, unable, am
Certainty	all, positive, completed, total, accuracy, fact, specific, specifically, correctly, complete, indeed, always, defined, positively, extremely, every, directly, confidence, explicitly, clear, clearly, completely, never, must, particularly, accurate, especially, exactly, necessary, apparent, absolute, nothing, accurately, necessarily, namely, certainty, precise, pure, obvious, confident, corrected, fundamental, perfectly, distinction, entirely, precision, ever, correction
Common verbs	were, was, is, are, be, one, would, had, have, did, see, used, asked, mean, should, will, given, may, using, could, following, has, been, do, made, being, found, use, might, showed, support, affect, does, described, left, learning, read, obtained, make, told, based, means, tested, appeared, followed, informed, seen, find, reading, making
Comparisons	as, than, more, after, same, different, higher, less, either, differences, greater, compared, most, before, further, lower, best, larger, particular, better, least, stronger, comparison, like, comparisons, faster, smaller, later, earlier, longer, neither, various, bottom, older, compare, similarly, top, equally, highest, comparing, similarity, closer, younger, equivalent, weaker, former, fewer, unique, lowest, middle
Conjunctions	and, as, when, if, but, also, then, whether, however, because, how, so, while, whereas, although, though, nor, until, otherwise, plus, nevertheless, whenever, unless
Differentiation	not, or, than, if, but, different, whether, however, either, differences, vs, whereas, rather, although, without, versus, others, differ, version, except, separate, alternative, instead, just, cannot, though, excluded, against, actually, nor, neither, differed, separately, differential, inequality, alternatives, otherwise, despite, adjusted, unlike, respective, really, differently, nevertheless, adjustment, excluding, else, exception
Discrepancy	would, if, should, could, rather, preferences, preference, need, problems, desirability, wanted, preferred, regardless, lack, must, desire, want, needed, desirable, desired, undesirable, impossible, unusual, lacking
Health	life, physical, weak, weaker, operation, live, health, weakly, diagnostic, pain, exercise, physically, bipolar
Interrogatives	which, when, who, whether, how, where, what, why, whose, whom, whenever, whatever
Leisure	play, music, played, games, parties, family, novel, playing, cards, express, plays, booklet, bar, exercise, books, pooled, weights, running, dramatically, dramatic
Male references	men, he, male, his, him, man, males, himself
Motion	following, behavior, increase, received, increased, change, followed, increases, actions, follows, faster, approach, behavioral, receive, changes, increasing, lead, receiving, run, roll, behaviors, quickly, leads, led, turn, entered, driven, brief, follow, put, move, car, step, approached, ran, removed, go, receives, slower, came, come, explore, comes, leading, walking, went, changed, fell
Negation	not, no, negative, without, cannot, nor, neither, none, never, negatively, nothing
Number	two, one, first, three, second, five, half, third, zero, single, once, ten, twice, fourth, double, fifth, sixth, twelve, twenty, thirty, twenty-four
Past focus	were, was, had, did, used, made, asked, given, been, tested, completed, previous, showed, included, left, obtained, told, provided, prior, appeared, followed, informed, seen, past, paid, explained, remember, played, affected, taken, earlier, differed, created, took, sat, accepted, wanted, felt, called, gave, viewed, done, began, help, supported, said, saw, believed, written, led
Power	high, low, higher, order, age, over, manipulation, strong, students, under, important, lower, above, dependent, up, below, best, punishment, large, rejection, larger, small, influence, power, stronger, competition, controlling, strongly, down, smaller, principal, judged, bottom, political, confidence,

	influenced, highest, controlled, allowed, manipulated, respect, lead, status, rejected, dominance, help, controls
Prepositions	of, in, to, for, with, as, on, by, from, than, at, between, about, after, across, vs, during, into, over, before, within, out, under, without, above, versus, below, since, except, like, including, regarding, down, against, along, until, near, around, via, away, despite, upon, off, outside, behind, unlike, throughout, plus, beyond, excluding
Present focus	is, are, be, have, see, can, mean, has, do, present, use, support, does, practice, make, provide, means, find, need, think, consider, work, now, follows, cannot, give, feel, explain, appear, trust, seems, determine, appears, take, look, include, provides, know, run, share, get, lack, believe, keep, vary, seem, describe, turn, start, become
Quantifiers	each, more, all, both, average, any, less, either, groups, most, some, total, much, sample, another, part, amount, many, least, single, section, series, percentage, approximately, few, remaining, various, none, every, equally, multiple, samples, somewhat, inequality, majority, added, fewer, piece, double, whole, amounts, variety, extra, sampling, adding
Sadness	low, lower, rejection, rejected, alone, failed, lowest, sadness, failure, sad, missing, rejecting, isolation, lost, lose, depressed, unhappy, empty, lowering, suffer
Space	in, on, at, both, high, low, higher, across, level, where, into, over, within, out, levels, further, under, lower, left, above, point, up, below, large, side, larger, full, small, section, separate, room, short, together, way, down, direct, length, smaller, positions, longer, place, long, bottom, top, dimensions, directly, central, highest, close, space

Note: top words in each dictionary are from the entire corpus (not just the papers that were replicated). Each word appears in at least 5 studies, and the dictionary should appear in 50% of the studies.

Panel B: Dictionaries associated with nonreplicability

Dictionary	**Top words**
Achievement	first, obtained, best, better, efficiency, competition, goals, work, limited, opportunity, determined, ability, top, confidence, created, importance, lead, able, dominance, motivation, working, leads, driven, incentives, planned, rank, create, failed, gains, advantage, motivated, earn, successful, improvement, opportunities, try, earned, efficient, worked, failure, cheating, obtain, potentially, skills, achieved, losses, dominant
Adjectives	as, than, after, same, high, different, average, higher, mean, positive, less, either, differences, greater, single, compared, most, before, strong, further, next, lower, round, new, general, additional, female, relevant, best, specific, large, identical, larger, particular, better, white, full, small, final, least, random, common, stronger, comparison, emotional, simple, short, free, comparisons, subsequent
Affiliation	we, our, interaction, groups, social, cooperation, communication, members, love, message, us, together, relationships, partners, interactions, relation, games, association, cooperative, share, help, parties, membership, friend, parents, family, shared, associations, siblings, kind, meet, conversation, relations, interacted, encouraged, dating
Anger	made, punishment, critical, argument, dominance, arguments, angry, cheating, dominant, threatening, argue, lies, threatened, dominated, argued
Anxiety	aversion, pressure, anxiety, fear, avoidance, threatening, suspicion, upset, inhibition, inhibitory, threatened, vulnerable, worry, anxious, uncertainty, avoided, doubt
Articles	the, a, an
Causation	experiment, results, effects, used, because, how, using, responses, made, manipulation, stimuli, stimulus, use, affect, factor, therefore, make, based, since, change, influence, independent, produced, making, controlling, activation, basis, outcomes, reasons, why, affected, consequences, changes, influenced, created, produce, cause, controlled, resulting, allowed, reaction, manipulated, lead, attribute, purpose, hence, product, reasoning, motivation, attributes

Feel	round, feelings, feel, hand, weight, felt, press, feeling, cold, weighted, hard, pressing, warmth, hot, warm, pressed, pain, hands, feels, weights, harder
Female references	female, her, she, females, woman, mother, herself
First person plural	we, our, us
Future focus	then, will, may, predicted, expected, might, prediction, predictions, future, expectations, predicts, predictive, predictors, expectation, predicting, predictor, anticipated, potentially, wants, soon, going, plan, preparation, expecting, plans, prepared
Impersonal pronouns	that, this, other, which, it, these, who, those, what, another, its, itself, whose, things, whom, anything, everyday, whatever
Insight	effects, analysis, mean, information, figure, analyses, questions, found, revealed, question, related, learning, memory, evidence, perceived, category, beliefs, decision, findings, means, correlations, preferences, preference, correlation, correlated, finding, attention, informed, find, choices, choose, examined, acceptance, evaluation, categories, think, examine, consider, feelings, thought, reference, knowledge, unrelated, reasons, thoughts, identification, explained, remember, feel, decisions
Money	payment, prices, payoff, free, cost, earnings, paid, payoffs, costs, wealth, compensation, profits, economic, investigate, accounted, poor, cash, cent, earned, payments, buying, donated, accounts, store, spent, investigated, poorly, rich, euro, poorer, investigation, accounting
Positive emotion	positive, value, greater, strong, support, important, values, well, good, best, better, play, stronger, love, acceptance, strongly, free, interest, opportunity, happiness, played, determined, happy, trust, positively, desirability, confidence, accepted, importance, respect, interested, share, satisfaction, supported, attraction, giving, wealth, active, parties, freedom, care, original, easily, favor, benefit, incentives, shared, profits, supporting
Reward	positive, scores, obtained, good, best, better, goals, earnings, approach, opportunity, taken, positively, desirability, confidence, take, took, get, taking, added, accessibility, willing, benefit, scored, profits, approached, desire, gains, advantage, plus, desirable, accumulated, earn, successful, getting, scoring, opportunities, benefits, earned, obtain, achieved, takes, adding, great, success, confident, successfully, fulfillment, accessible
Risk	aversion, yielded, consequences, trust, problems, difficult, stop, difficulty, lack, bad, wrong, avoidance, concern, failed, yields, failure, losses, threatening, worst, consequently, suppression, secure, carefully, consequence, inhibition, inhibitory, lose, threatened, stopped, yielding, suppressed, safe, avoided, lacking
See	see, revealed, showed, shows, round, show, appeared, white, screen, seen, black, pictures, appear, appears, clear, red, displayed, view, blue, saw, images, light, displays, showing, green, search, visible, depicted, colors, reveals, looking, look, gray, sought, seeing, circle, photographs, hidden, brown, watch, watched, square, appearing, watching, monitor, eyes, views
Third person plural	they, them, themselves
Time	when, first, after, then, while, during, present, age, before, respectively, next, received, again, sequence, finally, due, times, prior, end, last, since, repeated, future, final, periods, years, past, often, never, now, always, subsequent, min, still, once, faster, constant, receive, immediately, repetition, previously, later, earlier, older, until, sat, stop, temporal, minutes, receiving
Work	test, analysis, performance, reported, analyses, studies, assigned, students, sessions, instructions, learning, associated, university, read, practice, tasks, instructed, tested, tests, presentation, performed, produced, competition, computer, payment, goals, payoff, work, required, agent, payoffs, class, psychology, association, computed, project, agents, undergraduates, laboratory, political, testing, recruited, reports, produce, analyzed

Note: top words in each dictionary are from the entire corpus (not just the papers that were nonreplicable). Each word appears in at least 5 studies, and the dictionary should appear in 50% of the studies.

Table S11: Binary logit regressions (1 = replicated) with arc of narrative variables

Variable	Study	Full Text	Abstract
Intercept	-2.786†	-3.262†	-1.368
	[1.664]	[1.706]	[1.912]
Staging	-0.265†	0.058	0.287
	[0.159]	[0.163]	[0.220]
Cognitive Tension	-0.364*	-0.328*	-0.318†
	[0.172]	[0.162]	[0.183]
Plot Progression	0.171	0.199	-0.305
	[0.161]	[0.161]	[0.215]
Power	-1.254	-1.338	-1.585
	[1.126]	[1.153]	[1.281]
P-value	-13.456	-13.594†	-11.215
	[8.316]	[8.133]	[8.658]
Effect size	3.978**	3.922**	4.155*
	[1.405]	[1.423]	[1.655]
Number of participants	0.633*	0.635*	0.696*
	[0.252]	[0.261]	[0.292]
Citations count	0.026	0.061	-0.058
	[0.144]	[0.145]	[0.169]
Publication year (relative to 2000)	-0.028	-0.022	-0.050†
	[0.019]	[0.019]	[0.026]
Discipline: economics	0.152	-0.191	-0.606
	[1.365]	[1.368]	[1.405]
Discipline: social	0.523	0.297	0.393
	[0.563]	[0.544]	[0.601]
Effect type: interaction	-0.848	-0.774	-0.408
	[0.669]	[0.664]	[0.720]
Effect type: main effect	0.233	0.162	0.410
	[0.590]	[0.589]	[0.636]
Study done in US	0.003	0.026	-0.351
	[0.349]	[0.348]	[0.388]
Subjects: community	0.855	0.854	-0.085
	[0.707]	[0.703]	[0.759]
Subjects: online	-0.459	-0.207	-0.781
	[0.941]	[0.959]	[0.988]
Subjects: students	-0.030	0.068	-0.597
	[0.497]	[0.494]	[0.553]
Number of studies	-0.148†	-0.118	-0.070
	[0.085]	[0.084]	[0.097]
Number of references	-0.007	-0.006	-0.011
	[0.007]	[0.008]	[0.009]
Number of tables	0.023	-0.001	0.073
	[0.066]	[0.064]	[0.073]
Number of figures	0.157*	0.154*	0.185*
	[0.075]	[0.076]	[0.087]

Supplemental materials	-0.308	-0.401	-0.387
	[0.520]	[0.523]	[0.592]
Number of Authors	-0.072	-0.028	-0.164
	[0.121]	[0.121]	[0.141]
Full professor ratio	0.085	0.108	-0.183
	[0.515]	[0.513]	[0.591]
Male authors ratio	0.612	0.587	0.500
	[0.538]	[0.531]	[0.576]
Replication project: ML	0.362	0.523	0.271
	[0.495]	[0.484]	[0.548]
Replication project: EE	0.159	0.448	-0.167
	[1.619]	[1.601]	[1.777]
Replication project: SSRP	1.326†	1.356†	1.741†
	[0.800]	[0.801]	[0.92]
Replication project: RPP	0.111	0.198	-0.207
	[0.492]	[0.483]	[0.563]
Original authors involved	-0.488	-0.353	-0.311
	[0.370]	[0.374]	[0.387]
keyword: anchoring	-0.499	-0.414	-1.707
	[1.133]	[1.116]	[1.377]
keyword: attention	-0.823	-0.938	-1.024
	[0.881]	[0.886]	[0.927]
keyword: attitude	-0.729	-0.932	-0.899
	[0.810]	[0.822]	[0.876]
keyword: bias	-0.281	-0.466	-0.598
	[1.013]	[1.022]	[1.140]
keyword: cognitive processes	-0.360	-0.707	-1.110
	[0.759]	[0.766]	[0.831]
keyword: construal level	-0.840	0.125	-1.031
	[1.394]	[1.369]	[1.475]
keyword: consumer economics[1]	-0.820	-0.963	
	[1.610]	[1.526]	
keyword: emotion	-1.117	-0.999	-0.824
	[1.086]	[1.098]	[1.198]
keyword: individual behavior	-0.438	-0.139	1.284
	[1.169]	[1.164]	[1.405]
keyword: learning	-0.059	-0.292	0.060
	[0.906]	[0.904]	[0.911]
keyword: memory	-0.743	-0.880	-1.045
	[0.587]	[0.603]	[0.662]
keyword: money	0.207	0.437	0.530
	[1.386]	[1.312]	[1.281]
keyword: moral	-0.788	-0.770	-0.487
	[0.796]	[0.787]	[0.862]
keyword: motivation	-0.139	0.034	-0.087
	[1.099]	[1.103]	[1.152]
keyword: perception	-1.529	-1.424	-1.220

	[0.988]	[0.991]	[1.167]
keyword: prejudice	-0.871	-0.304	-0.922
	[1.304]	[1.219]	[1.322]
keyword: similarity	-2.614*	-2.569*	-2.346
	[1.295]	[1.21]	[1.557]
keyword: social other	-1.578*	-1.438†	-1.512†
	[0.747]	[0.734]	[0.807]
keyword: the self	-1.068	-1.229	-1.207
	[1.076]	[1.099]	[1.179]

Note: Each column presents the results of a different regression, based on the part of the text in the column title.

[1] The keyword "consumer economics" is missing in the abstract text regression because it was collapsed into non-cooperative games. This is because some papers are missing an abstract, and some abstracts are shorter than 100 words, which is the minimum required for this analysis. Hence they were removed from the analysis, which made the keyword consumer economics unidentifiable.

** $p < 0.01$, * $p < 0.05$, † $p < 0.1$

To supplement Table S11, we ran the arc of narrative measures in a one-at-a-time analysis with all the metadata. Here are the coefficients [SE] from these analyses (each number represents a separate regression):

	Study	Full text	Abstract
Cognitive tension	-0.324 [0.167]†	-0.327 [0.161]*	-0.243 [0.175]
Staging	-0.212 [0.152]	0.095 [0.155]	0.075 [0.183]
Plot progression	0.058 [0.152]	0.217 [0.156]	-0.143 [0.184]

Table S12: A cursory analysis of the text in “sibling” studies

	LIWC measure	|Cohen’s d|
Dictionaries associated with nonreplicability in the entire corpus and in the sibling studies	Analytical thinking	1.086
	Staging	1.057
	Future focus	1.046
	Causation	0.781
	Reward	0.639
	Dashes	0.445
	Work	0.406
	Achievement	0.343
	Positive emotion	0.334
	Risk	0.306
	First person plural	0.243
	Time	0.176
	Cognitive tension	0.155
	Abstraction	0.140
	Anger	0.094
	Other punctuations	0.023
	Emotional tone	0.004
	Anxiety	0.003
	Impersonal pronouns	0.001
Dictionaries associated with replicability in the entire corpus and in the sibling studies	Male references	0.877
	Differentiation	0.787
	Readability	0.667
	Past focus	0.655
	Negation	0.545
	Common verbs	0.527
	Discrepancy	0.448
	Interrogatives	0.426
	Conjunctions	0.423
	Leisure	0.359
	Space	0.267
	Words per sentence	0.241
	Authenticity	0.210
	Present focus	0.174
	Parentheses	0.148
	Prepositions	0.141
	Comparisons	0.131
	Quantifiers	0.127
	Auxiliary verbs	0.125
	Adverbs	0.079
	Order (Power in LIWC)	0.055
	Motion	0.020
	Number	0.015

	LIWC measure	|Cohen's d|
Dictionaries associated with nonreplicability in the entire corpus, but with replicability in the sibling studies	Feel	0.836
	Money	0.678
	Third person plural	0.479
	Clout	0.276
	Tentative	0.273
	Colons	0.264
	Affiliation	0.182
	Reveal (See in LIWC)	0.181
	Insight	0.176
	Adjectives	0.072
	Semicolons	0.035
	Articles	0.026
	Female references	0.002
Dictionaries associated with replicability in the entire corpus but with nonreplicability in the sibling studies	Obfuscation	0.691
	Certainty	0.504
	Weak (Health in LIWC)	0.390
	Sadness	0.290
	Word count	0.225
	Commas	0.140
	Periods	0.095

Note: This table presents the results of paired comparisons between "sibling" studies that come from the same paper, one whose replication was successful and one whose replication failed. Two studies each come from paper IDs 36, 306, 313; three studies each from 207 and 217; and four studies each from 248. When more than one study from the same paper had the same outcome (e.g., two studies failed replication from the same paper), we aggregate by averaging the text variables over the studies with the same outcome from the same paper. By comparing the sibling studies we implicitly control for almost all of the metadata. We do not control for study metadata in this analysis (see separate analysis below).
Given the small sample size (n=6 paired comparisons), we present results in Cohen's *d* effect size (in absolute values). When calculating Cohen's *d*, we use the entire dataset of 299 studies to calculate the standard deviation of each variable. We do the same in Table S13.

Table S13: Comparing study statistics in the "sibling" studies

	Successful replication	Unsuccessful replication	Sample SD	|Cohen's d|
p-value	0.014	0.056	0.054	0.769
Effect size (r)	0.365	0.265	0.214	0.470
Number of participants	113.429	132.778	13,342.312	0.001

E. References

Altmejd, A., Dreber, A., Forsell, E., Huber, J., Imai, T., Johannesson, M., ... & Camerer, C. (2019). Predicting the replicability of social science lab experiments. *PloS one*, 14(12).

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436. (**EE**)

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644. (**SSRP**)

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., ... & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50), 15343-15347.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Brown, E. R. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82. (**ML3**)

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., ... & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, 75, 102117.

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., ... & Ratliff, K. A. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, *8*(1), 35271. (**ML4**)

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, 45(3), 142-152. (**ML1**)

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., ... & Batra, R. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490. (**ML2**)

Markowitz, D. M., & Hancock, J. T. (2016). Linguistic obfuscation in fraudulent science. *Journal of Language and Social Psychology*, 35(4), 435-445.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. (**RPP**)

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. Science advances, 7(21), eabd1705.

Yang, Y., Youyou, W., & Uzzi, B. (2020). Estimating the deep replicability of scientific findings using human and artificial intelligence. Proceedings of the National Academy of Sciences, 117(20), 10762-10768.

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, 25(5), 1968-1972. (**PN**)