Research Article

Psychological Science
1–14

DOI: 10.1177/09567976241254037
www.psychologicalscience.org/PS

The Language of (Non)Replicable Social Science

Michal Herzenstein[1], Sanjana Rosario[2], Shin Oblander[3], and Oded Netzer[2]

[1]Lerner College of Business and Economics, University of Delaware; [2]Columbia Business School, Columbia University; and [3]Sauder School of Business, University of British Columbia

Abstract

Using publicly available data from 299 preregistered replications from the social sciences, we found that the language used to describe a study can predict its replicability above and beyond a large set of controls related to the article characteristics, study design and results, author information, and replication effort. To understand why, we analyzed 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, open data, open materials

Received 11/29/23; Revision accepted 4/18/24

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

We hypothesize that 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). Although 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 companies' stock returns and volatility, trading volume, fraud, and unexpected earnings (Loughran & McDonald, 2011). 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

Corresponding Author:
Michal Herzenstein, University of Delaware, Lerner College of Business and Economics
Email: michalh@udel.edu

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 article 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 capabilities. 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 that attempted to predict replicability, as Crockett et al. (2023) and Mottelson and Kontogiorgos (2023) have pointed 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 article'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 articles with only a moderate rate of success. We assembled a data set of 299 studies in psychology and behavioral economics with published replications not done by the original authors. Our data include information about the original article, focal study (the one that other labs attempted to replicate), authors, and the text used in the abstract, focal study, and the entire article, as well as information about the replication study.

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 because of the growing concern over 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' hunches 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.

Our first finding is that, controlling for a large set of metadata variables, the text in the focal study and in the entire article 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 data set, with different methods of text analysis, and with different underlying models. Our large set of metadata variables allows us to alleviate concerns that the text reflects the authors' characteristics, the study's design, its objective statistical power, how the paper was selected for replication (systematically or not), 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 utilized the Linguistic Inquiry and Word Count (LIWC 2015; www.liwc.app; Pennebaker et al., 2015) dictionaries, which are a set of 92 nested and context-free psychometric dictionaries. We also used measures of abstraction, obfuscation (Markowitz & Hancock, 2016), and readability (Flesch, 1948), along with narrative arcs that describe the structure of stories (Boyd et al., 2020). Controlling for the metadata, we found that the language in academic studies likely reflects their authors' intuition regarding veracity, which may explain the language's predictive ability of replicability.

Specifically, we found that replicable studies are often written in an elaborate and complex manner that reflects 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 pose a dilemma (or conflict) and then resolve 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 article's claims despite the possibly weaker evidence being presented (Dahlstrom, 2014). Although 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 Data Set

Using publicly available data on replication efforts of original studies in psychology and behavioral economics, we compiled a data set of 299 studies: 96 studies were replicated by Open Science Collaboration (2015; RPP), 49 by Many Labs (Ebersole et al., 2016; Klein et al., 2014, 2018, 2022; 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 preregistered, were published in well-regarded journals, and were not performed by the original authors. Table S1 in the Supplemental Material available online lists the replication efforts in our data set, and Table S2 presents the list of original articles included in our analyses and their published replication efforts.

We collected four types of measures on the original and replication papers: First, following Altmejd et al. (2019), our focal dependent variable was a binary indicator of whether a study is replicated, based on the assessment of the replication team. In most replication studies, this effectively means that the replication effort found a significant effect ($p \leq .05$) in the same direction as the original paper. Overall, 42% of the replication attempts in our data set were successful. Our second dependent variable was 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 the replication plan, and then traded contracts that pay \$1 if the study was successfully replicated and \$0 otherwise. Dreber et al. (2015) explained 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). Second, we collected metadata from the original papers—the paper's discipline (social psychology, cognitive psychology, or economics); 45 keywords or JEL codes (see the Supplemental Material for how we processed the keywords); publication year; information about the authors (number of authors, proportion of male authors, and proportion 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, or 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, and post hoc power). Third, metadata from the replication papers on whether the original authors advised the replication team, and indicators of the replication project—RPP, ML, EE, SSRP, or other. Forth, 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 data set of 2,420 articles from the same journal issues as the replicated articles, containing articles 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 data set 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 used these averaged embeddings as features in our predictive analysis. Second, we used the Linguistic Inquiry and Word Count to calculate the frequency of each LIWC dictionary in the text. We filtered and cleaned up the dictionaries to ensure they were meaningful in our context (see the 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 and the 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 article's replicability using a host of metadata variables and textual features, we evaluated 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 hyperparameters, and LIWC), different subsets of the metadata (excluding the variables capturing study results—effect size and *p* value—or using only these variables, following Yang et al., 2020), and two sizes of train–test split (80%–20% and 70%–30%). We calibrated the model tuning parameters (e.g., ridge penalty) using tenfold cross-validation within the training set and then estimated out-of-sample performance using the predicted values on the test set. Ridge models performed best, and the other variations were not meaningfully different. Hence, we present here results with ridge regressions, an 80%–20% train–test split, and the full set of metadata variables. We show a meaningful subset of other models in Tables S7 and S8.

Results

Table 1 presents the predictive ability of the text (i.e., regarding replicability) in each section of the article separately—focal study, full text, and abstract—as well as the metadata variables. Predictive accuracy was measured as the area under the receiver operating characteristic curve (AUC) and was compared across six models: metadata only, text embeddings only, embeddings and metadata, text features only (LIWC dictionaries, arc of narrative, and Flesch readability), text features and metadata, and finally all three sets combined. We did not include obfuscation and abstractions in the text features analysis because they are nested within LIWC dictionaries. For the focal-study text and full text, we found that the model that includes metadata and all text (i.e., embeddings and textual features) predicts replication better than the model that includes only the metadata ($AUC_{metadata+all\ study\ text}$ = .725 and $AUC_{metadata+all\ full\ text}$ = .716 versus $AUC_{metadata}$ = .696). The $AUC_{metadata}$ is the same for the focal study and full text because the dependent variable is at the study level. These AUCs were averaged over 10,000 random train–test splits, which allowed us to assess the predictive improvement's reliability—the proportion of runs in which the model that includes the text and metadata predicted better than the model that included only metadata is 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 on the basis of text features alone perform similarly to the ones that use only the metadata ($AUC_{text\ embeddings}$ = .703 and $AUC_{text\ features}$ = .683 for the focal-study text and $AUC_{text\ embeddings}$ = .699 and $AUC_{text\ features}$ = .671 for the full text versus $AUC_{metadata}$ = .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 found that they also improved predictions above and beyond the rich set of metadata ($AUC_{metadata+study\text{-}text\ features}$ = .713 and $AUC_{metadata+full\text{-}text\ text\ features}$ = .702 versus $AUC_{metadata}$ = .696). The text in the abstract alone is not as informative about the paper's replicability (Table 1, section 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 ($AUC_{metadata+all\ study\ text}$ = .717 and $AUC_{metadata+all\ full\ text}$ = .709 versus $AUC_{metadata}$ = .686).

In practice, when reading academic articles, readers often use their experience with prior articles to predict the replicability of newer articles. Accordingly, we tested whether the text in older articles helps predict the replicability likelihood of newer ones. We split our data set into older articles (published before 2012) and newer articles (published in 2012 or later), resulting in a split of approximately 80%–20% for train and test samples. We found that textual information improved predictive ability even when split over time ($AUC_{metadata+all\ study\ text}$ = .820 and $AUC_{metadata+all\ full\ text}$ = .795 versus $AUC_{metadata}$ = .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 articles. Because the prediction-market outcomes were not used in training our models, we could treat them as

Table 1. Predicting Article Replicability in Held-Out Samples by Text Section

Section A: Study text

Train–test split	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
239-60[a]	All articles in data set	Average test AUC	.6961	.7028	.7205	.6827	.7130	.7245
		SD across splits	.0607	.0624	.0610	.0637	.0612	.0606
		Runs with improvement[b]		53.49%	68.32%	42.63%	64.51%	70.30%
220-55[c]	Psychology articles only	Average test AUC	.6861	.6913	.7116	.6749	.7054	.7169
		SD across splits	.0664	.0671	.0658	.0681	.0663	.0655
		Runs with improvement[b]		52.61%	66.69%	44.45%	64.04%	69.13%
236-63	Over time	Test AUC	.6729	.7816	.8226	.7384	.7805	.8204
299-99[d]	Market prediction	Correlation	.5129	.6054	.6037	.5794	.6179	.6147

Section B: Full text[e]

Train-test split	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
239-60[a]	All articles in data set	Average test AUC	.6961	.6989	.7140	.6713	.7022	.7159
		SD across splits	.0607	.0647	.0629	.0664	.0630	.0630
		Runs with improvement[b]		51.60%	63.62%	35.82%	55.37%	64.08%
220-55[c]	Psychology articles only	Average test AUC	.6861	.6910	.7053	.6673	.6909	.7090
		SD across splits	.0664	.0688	.0676	.0708	.0693	.0676
		Runs with improvement[b]		52.78%	62.18%	40.34%	53.79%	63.96%
236-63	Over time	Test AUC	.6729	.7688	.8171	.7633	.8016	.7949
299-99[d]	Market prediction	Correlation	.5129	.5809	.6046	.5328	.5844	.6144

Section C: Abstract text[f]

Train–test split	Prediction	Stat	Metadata	Text embeddings	Metadata + text embeddings	Text features	Metadata + text features	Metadata + text features + embeddings
208-52[a]	All articles in data set whose abstract ≥ 100 words	Average test AUC	.6752	.6462	.6753	.5715	.6651	.6660
		SD across splits	.0678	.0716	.0710	.0722	.0701	.0721
		Runs with improvement[b]		35.82%	50.26%	12.52%	43.50%	44.21%
196-49[c]	Psychology articles only, whose abstract ≥ 100 words	Average test AUC	.6544	.6377	.6586	.5990	.6484	.6497
		SD across splits	.0728	.0755	.0757	.0770	.0764	.0771
		Runs with improvement[b]		42.27%	52.60%	27.24%	46.77%	47.08%
206-54	Over time	Test AUC	.7670	.7415	.8086	.4730	.6991	.7840
260-84[g]	Market prediction	Correlation	.4781	.5078	.5402	.3589	.5516	.5622

Note: This table shows the results of logistic regressions with ridge regularization. The dependent variable is a binary indicator for replicability (1 = replicable). We compare six model specifications, which include different subsets of three classes of variables: "Metadata," referring to characteristics of the paper, authors, and study; "Text embeddings," referring to the text represented by the word2vec embedding space (the hyperparameters are continuous bag of words [CBOW], three-word windows, 100 dimensions, 50 epochs, and with stop words; alternative specifications are presented in Table S7); and "Text features," referring to Linguistic Inquiry and Word Count (LIWC), arc of narrative, and readability variables. We present the average holdout predictions from 10,000 replications of a random 80% calibration/20% validation split of the articles in our sample. To evaluate the models' performance, we used the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The models are based on three slices of the text—Section A presents the results using the text in the focal study, Section B the entire article, and Section C the abstract—as well as three different slices of the data: all articles, only psychology articles, and over time (training on articles published before 2012 and predicting articles published in 2012 or later). Market-prediction models calculate the Pearson correlation between our models' predictions and end prices in prediction markets. [a]This is the average across 10,000 splits. We note that because there are several articles with multiple studies being replicated, we split the sample by article to make sure multiple studies of the same article 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. [b]This represents the 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 only. [c]We dropped 24 studies in economics to get 275 studies in psychology only. [d]We trained on the entire data set to predict replication outcome; predictions from the trained model were then correlated with the market predictions for the 99 studies or articles from the prediction markets. [e]Analysis is still at the study level, though the text is for the full article; 22 articles had more than one study replicated. [f]39 studies' corresponding articles 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. [g]15 studies' corresponding articles 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-market analysis.

another form of prediction test. Inspired by Camerer et al. (2016), we calculated the correlation between our models' predicted probability that a study would replicate with the prediction-market ending prices. We found that the correlation improves with the addition of the text ($r_{\text{metadata+all study text}}$ = .615 and $r_{\text{metadata+all full text}}$ = .614 versus r_{metadata} = .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; Youyou et al., 2023), we found 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 the topic of the article or ease of replicability. We elaborate on these aspects next.

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 studies' 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 article. 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 article. (The percentage of runs in which the model including the metadata and text features predicted replicability better than the model with metadata 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 *least absolute shrinkage and selection operator* (LASSO) logistic 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 identified did not merely reflect differences in the conventions of certain disciplines (e.g., some disciplines write more parsimoniously than others) and the subject matter (as captured by the article's keywords; note that in these analyses we used fewer keywords because of identifiability constraints, see Table S6 for that list), differences over time (e.g., older articles may document fewer results), or the result of a more or less experienced original research teams or replication teams. To remedy 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 the dictionary *negative emotions*, which is nested in the dictionary *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 and 3. For full results, see Tables S9 through S11 in the Supplemental Material.

Results

The language of replicability has markers of complexity and truth-telling. Table 2 and Figure 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.

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[a]	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[a]	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[a]	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)[b]	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[c]	0.047	0.142 (0.152)	And, as, when, if, but
	Markers of complex text:			
	• Longer sentences by word count[d]	0.201	0.083 (0.160)	
	• Weak (Health in LIWC)[a,b]	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 Linguistic Inquiry and Word Count (LIWC) low-level dictionaries (excluding punctuation) that were selected by least absolute shrinkage and selection operator (LASSO) regressions and are associated with replicability. The one-at-a-time coefficients' significance levels are †$p < .1$, *$p < .05$, and **$p < .01$. See full results for the one-at-a-time regressions in Table S9. [a]Although we do not interpret the full text because of its lower predictive ability of the outcome, we note that this dictionary is associated with replicability in a one-at-a-time analysis using the full text, but not in the LASSO analysis of the full text. [b]We changed the name of the LIWC dictionary to be more descriptive in our context. The original name is in parentheses. [c]This dictionary is not associated with replicability in the full text. [d]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 ($\beta = -0.242$, $p = .147$). Therefore, we do not draw conclusions from that association.

Comparing the text in academic articles from predatory versus real journals, Markowitz et al. (2014) found that articles in real journals use more quantifiers and prepositions (which we discuss next), suggesting that 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; this provides 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 the 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, whereas 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.

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[a]	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[a,b]	—	−0.289 (0.173)†	We, our, us
3. Positivity	• Reveal (See in LIWC)[c]	−0.278	−0.298 (0.160)†	See, revealed, showed, shows
	• Positive emotions[b]	−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[d]	−0.106	−0.278 (0.198)	Positive, scores, obtained, good, best, better
4. Other selected dictionaries	• Work[b]	−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[e]		$\beta = -0.364$, $SE = 0.172$, $p = .035$	
	• Staging arc		$\beta = -0.265$, $SE = 0.159$, $p = .096$	

Note: This table presents all Linguistic Inquiry and Word Count (LIWC) low-level dictionaries (excluding punctuation) that were selected by least absolute shrinkage and selection operator (LASSO) regressions and are associated with nonreplicability. The one-at-a-time coefficients' significance levels are †$p < .1$ and *$p < .05$. See full results for the one-at-a-time regressions in Table S9. [a]Missing LASSO coefficients means that although the text feature was not selected by LASSO (likely because of collinearity with the other text features), it is associated with nonreplicability in the one-at-a-time analysis. Summary variables (such as abstraction) that were not part of LASSO regression are listed as n/a. [b]Although we do not interpret the full text because of its lower predictive ability of the outcome, we note that this dictionary is associated with nonreplicability in a one-at-a-time analysis using the full text, but not in the LASSO analysis of the full text. [c]We changed the name of the LIWC dictionary to be more descriptive in our context. The original name is in parentheses. [d]This dictionary is not associated with replicability in the full text. [e]These are the 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 $p = .043$ in the full text (see all results in Table S11) and when run alone with the metadata ($p = .053$ for study text, $p = .041$ for full text).

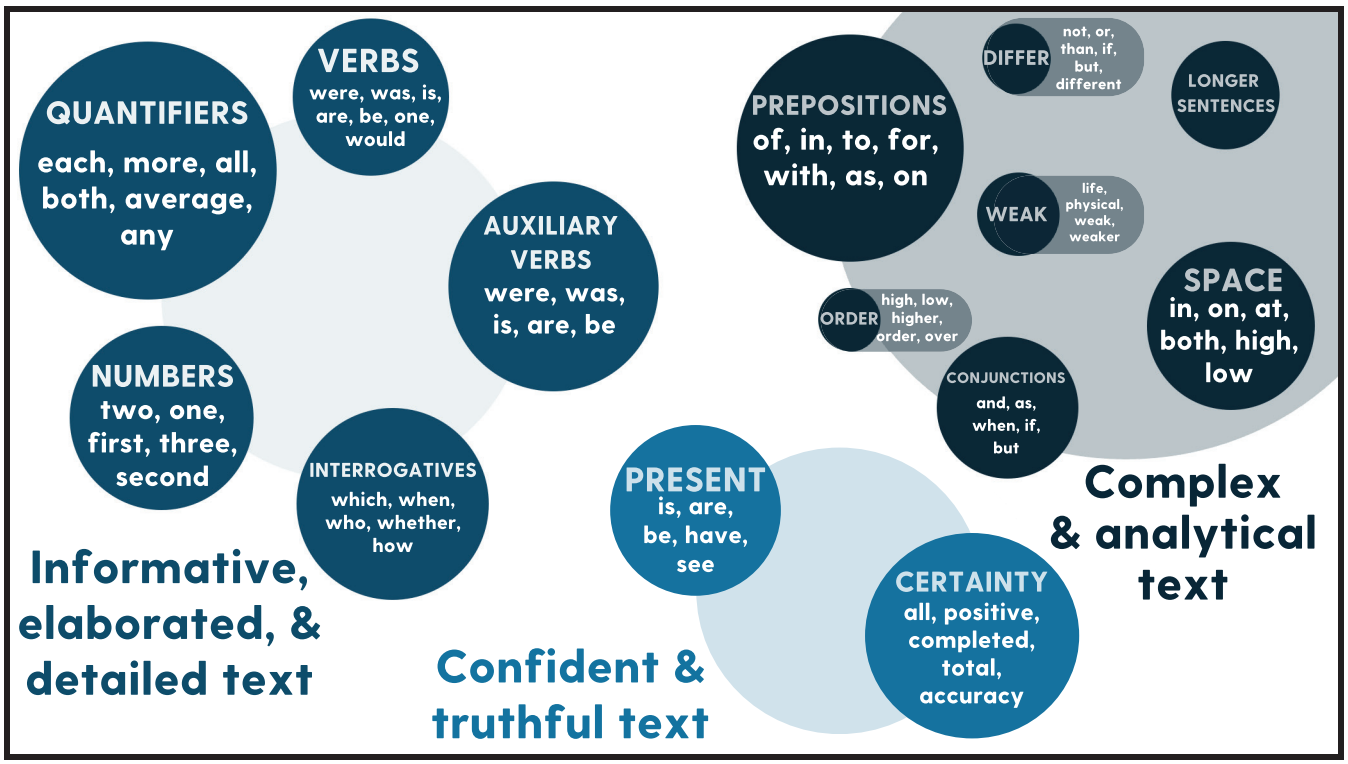

Fig. 1. Writing styles associated with replicable studies. The bubble size of each dictionary is based on the one-at-a-time coefficient from Table 2.

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. Indeed, words related to order ("higher," "over") and words related to 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 they compare and contrast 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 interpretation aligns well with words in the comparative dictionaries weak ("weak," "weaker") and differentiation ("not," "but") that are associated with replicability.

Replicable studies exude confidence, with authors commonly using certainty words ("all," "total"), whereas 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.

The language of nonreplicability has markers of deception and persuasion. Table 3 and Figure 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 that the text is vague, uncertain, and uncommitted (Larrimore et al., 2011), and they are common in fraudulent research (Markowitz & Hancock, 2016), predatory journals (Markowitz et al., 2014), and deceptive financial reporting (F. Li, 2008). Text written in the 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,"

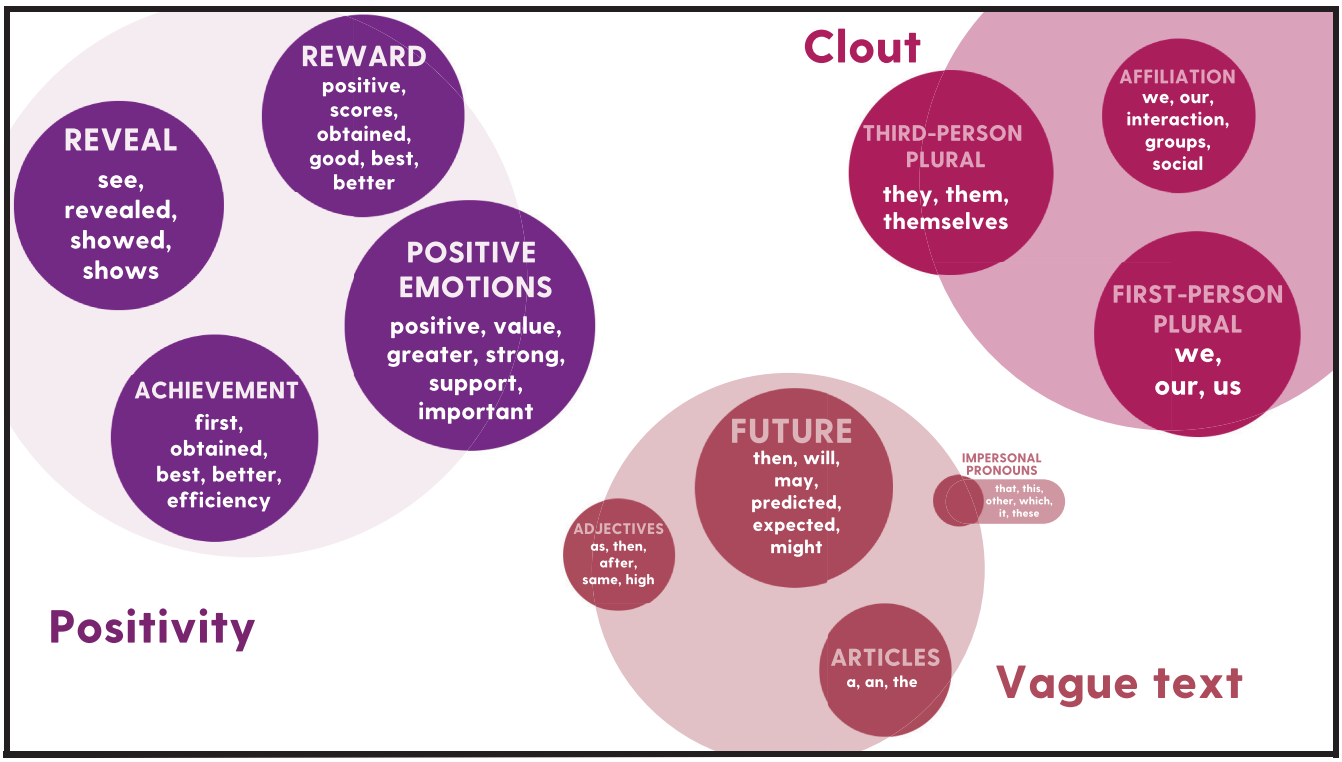

Fig. 2. Writing styles associated with nonreplicable studies. The bubble size of each dictionary is based on the one-at-a-time coefficient from Table 3.

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 this discussion that although we reference research on deception, replicability likelihood does not necessarily imply lying, because 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 data set were written by multiple authors, so the use of plural pronouns is not surprising (although we controlled 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 require the use of the writer's clout in delivering the ideas in the text (Moore et al., 2021). As such, the finding that clout lands on the opposite side of the replicability divide from certainty and differentiation mirrors past findings.

Authors of nonreplicable studies write more positively as evidenced by the following four dictionaries: reveal ("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 underperformance (Kang et al., 2018) and fake reviews (J. Li et al., 2014), because authors convey a level of optimism that 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. Although 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) showed 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). Because 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 ($\beta = -0.358$, $p = .037$), and staging is marginally so ($\beta = -0.266$, $p = .095$). Moreover, cognitive tension is consistently associated with nonreplicability—in the full text of the article ($\beta = -0.328$, $p = .043$); in the abstract, albeit marginally ($\beta = -0.318$, $p = .082$); and in "one-at-a-time" analyses that include all the metadata ($\beta_{study} = -0.324$, $p = .053$; $\beta_{full\ text} = -0.327$, $p = .041$). See the full set of results in Table S11. These findings suggest that although good storytelling is a desirable trait of the narrative, it may make it easier for readers to believe nonreplicable results (Dahlstrom, 2014).

Language reflects the authors' intuition about their study's veracity. Although our analyses control for authors' characteristics, research topic, and other article metadata, we cannot guarantee that other aspects, not controlled for in our analyses, may be correlated with the text of the article. To attempt to hold almost all else constant, we focus on six articles in our data set 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. Although this is only a cursory analysis because of the small sample size ($n = 6$), which does not permit formal statistical testing, it still provides important insights about the role of the text. We found 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 article. Most of these text features (20/26 = 77%) were in the same direction as our main analysis, showing that text signals tend to be directionally consistent within and across articles (see Table S12). An analysis of the statistics from the sibling studies shows that the p values of replicable studies were lower (Cohen's $d = 0.769$; large) and effect sizes were higher (Cohen's $d = 0.47$; medium) than those of nonreplicable studies in the same article (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 articles' 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 allowed 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. This 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 studies' replicability likelihood in the way they write. 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 that 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, because of publication bias, published articles 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 the 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 ones, 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 do the results hold, why boundary conditions occur, and how findings relate to past results. Further, even when articles tell interesting stories, our results suggest that the review team should focus on the method 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. Although 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 articles are manually replicated. Finally, although older articles were able to predict the replicability of newer articles, this result might change in the future, perhaps because of 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.

Transparency

Action Editor: Mark Brandt
Editor: Patricia J. Bauer

Author Contributions

Michal Herzenstein: Conceptualization; Data curation; Writing – original draft; Writing – review & editing.
Sanjana Rosario: Data curation; Formal analysis; Writing – original draft; Writing – review & editing.
Shin Oblander: Data curation; Formal analysis; Methodology; Writing – original draft; Writing – review & editing.
Oded Netzer: Conceptualization; Methodology; Supervision; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the University of Delaware and Columbia University.

Open Practices

This study was not preregistered. The list of original articles and their replication papers is publicly available at https://osf.io/qy8ev/; the processed text and metadata are publicly available at https://osf.io/qy8ev/; and the code is publicly available at https://osf.io/qy8ev/. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Acknowledgments

All authors contributed equally to this research. We thank Matthew McGranaghan for comments on a previous version and Rachel Lapp for graphic design support.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976241254037

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), Article e0225826. https://doi.org/10.1371/journal.pone.0225826.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, *6*(32), Article 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-2107. https://doi.org/10.48550/arXiv.2107.06751

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (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., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A. Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten E., Hummer, L., Imai., T., . . .Wu., H. (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.). 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, USA*, *120*(33), Article e2307596120.

Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. *Proceedings of the National Academy of Sciences, USA*, *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, USA*, *112*(50), 15343–15347.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., . . . Nosek, B. A. (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., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, *75*, Article 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, USA*, *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., Hangsan Ahn., P., Brady, A. J., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J. T., Cromar, R., Gardiner, G., Gosnell, C. L., Grahe, J., Hall, C., Howard, I., . . . Ratliff, K. A. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, *8*(1), Article 35271.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (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, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (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), Article e105937. https://doi.org/10.1371/journal.pone.0105937

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* [Conference session]. 2014 ASEE Annual Conference & Exposition (paper ID No. 8614).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv:1301.3781. https://doi.org/10.48550/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, USA*, *120*(33), Article e2309496120.

Nabi, R. L., & Green, M. C. (2015). The role of a narrative's emotional flow in promoting persuasive outcomes. *Media Psychology*, *18*(2), 137–162.

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.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article 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* (pp. 201–210).

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

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count: LIWC2015*. Pennebaker Conglomerates.

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), Article e115844. https://doi.org/10.1371/journal.pone.0115844

Reboul, A. (2021). Truthfully misleading: Truth, informativity, and manipulation in linguistic communication. *Frontiers in Communication*, *6*, Article 62. https://doi.org/10.3389/fcomm.2021.646820

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. https://doi.org/10.1080/09515089.2022.2113771

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), Article 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*. 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, USA*, *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, USA*, *120*(6), Article 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.