THE PSYCHOLOGY OF SCIENTIFIC FRAUD

by

Benjamin J. Zubaly

Thesis
Submitted in Partial Fulfillment of the Requirements for
the Degree of

Bachelor of Arts (Honours) in Psychology

in the

Department of Psychology

Trinity Western University

April 2024

Table of Contents

Abstract

Scientific fraud disrupts the pursuit of truth and leads to many monetary and social costs. However, psychology's potential contribution to solving this problem has largely been neglected due to methodological difficulties. The present study used text analysis to compare characteristics of matched fraudulent and genuine scientific papers (N = 88), across single- and multi-author papers. Results indicated no statistically significant differences between fraudulent and genuine papers on use of linguistic obfuscation, language suggesting certainty, or number of references across single- and multi-author papers. Fraudulent authors were also not more likely to be the corresponding author. This investigation provides an improved methodological framework with novel hypotheses that should be tested more robustly by future research.

Introduction

In modern science high-profile cases of scientific fraud are routinely discovered. These cases often make headlines, stoking distrust in the institutions and reliability of science, both for scientists and the general public. Such cases continue to be revealed, creating frustrating for the vast majority of stakeholders who vehemently disapprove of such moral transgressions and the numerous reformers that have proposed potential solutions. Psychology has a unique role to play in this story. To understand and mitigate scientific fraud, we can address the psychology of the fraudulent scientist.

In the early 2010s, psychologists increasingly began to acknowledge that there were insufficiencies in their shared ways of doing science (Nelson et al., 2018). The quality of psychology as a science seemed to suffer from a broad set of problems in the ways that it was being conducted and institutionalized, including publication bias, p-hacking, and a range of methodological and statistical insufficiencies. For example, Simmons et al. (2011) demonstrated that through exploiting "researcher degrees of freedom"—common flexibilities in methodological and analytic decisions—false-positives can become more likely than true effects even at $p < .05$. Large-scale replication efforts have subsequently ushered the *replication crisis* into being. For example, the Open Science Collaboration (2015) attempted to replicate 100 studies from top psychology journals, with only approximately one-third to one-half showing the same effect as the original study, depending on the criterion used to evaluate the replication attempts.

Many reasons for the replication crisis have been discussed, but most have centered around methodological or analytic practices (e.g., questionable research practices) and poor incentives, with implemented solutions including preregistration and open-science, and there has

been significant progress in these areas (Nelson et al., 2018). However, interest in methods and replication has perhaps overshadowed another phenomenon that threatens the credibility of psychology and other scientific disciplines: fraud.

Definitions

For the purposes of this paper, research misconduct will be defined as it is by the U.S. Public Health Service (Caron et al., 2023). Research misconduct includes falsification, fabrication, or plagiarism that is committed intentionally, knowingly, or recklessly. Scientific fraud will have the same meaning but will refer specifically to the scientific context (e.g., as opposed to the arts or humanities). Data fraud will refer to intentional, known, or reckless falsification or fabrication of research data (i.e., research misconduct not involving plagiarism). Fraud and research fraud will generically refer to research misconduct, scientific fraud, and/or data fraud.

Problem Scope

The influence of fraud on the quality of scientific findings may be greater than most appreciate. Yet, unlike other influences on the reliability of science, studying fraud proves to be uniquely difficult, given the sparse and low quality available data (Zuckerman, 2020). Anecdotally, there have been cases of authors fabricating data for dozens of papers. For example, Deidrick Stapel amassed over 50 retractions, primarily for data fraud (Palus, 2015). Before much of his work was outed by his students as fraudulent in 2011, Stapel enjoyed a celebrated career. His work was published in top journals, and he was invited to give prestigious talks regarding research built on his made-up data. Although anecdotes such as this abound and are informative for understanding what fraud can look like in practice, there is also a need for continued formal research to better understand the scope of and processes related to fraud.

Retraction and Fraud

Although journals use retraction to rid the scientific literature of fraudulent papers, this process relies on misconduct being caught and reported, and by the time a decision to retract a paper is made, often the damage has already been done. Retraction seems to occur at an average of around 2 years after publication (Furman et al., 2012), and some scholars have argued that this process is too slow (for example, Loadsman, 2019). There is a distinct drop in citations after a paper is retracted (Candal-Pedreira et al., 2020; Dinh et al., 2019; Furman et al., 2012; Sotudeh et al., 2022), and the citations that do happen also tend to be less "mainstream"—often coming from outside of the geographical area and discipline of the original retracted paper (Dinh et al., 2019; Pfeifer & Snodgrass, 1990). However, this decrease in impact does not occur in all cases; for example, comparing Scott S. Reuben's retracted articles with his nonretracted articles—matched by chronological publication order and paper type—revealed that his retracted articles were cited 92% *more often* over the subsequent 10 year period (Szilagyi et al., 2022). Furthermore, prominent and influential cases of academic fraud can go undetected for decades, calling into question whether modern science is truly self-correcting (Stroebe et al., 2012).

Misconduct is often found to be the most common reason for retraction of research publications (e.g., Armond et al., 2021; Chambers et al., 2019; Craig et al., 2020; Fang et al., 2012; Gaudino et al., 2021; Li & Mol, 2019; but see Steen, 2011 and Wang & Su, 2022), and fraudulent papers are also more likely to be published in high impact journals (Steen, 2011b). Although attention—operationalized by citation count—is a key predictor of paper retraction (Furman et al., 2012), papers retracted from more eminent journals (e.g., Nature, Science, or Cell) seem to see more citations post-retraction than papers retracted from journals with less prestige (Wang & Su, 2022), and uncertainty remains as to how many less eminent cases of

scientific misconduct scatter the scientific literature, yet to be discovered. Perhaps more directly alarming, retracted literature may exert a distinct influence on public policy; the rate of retracted paper citations in the policy literature is similar to that of non-retracted research (and it is around twice as likely to be cited positively than negatively; Malkov et al., 2023).

Prevalence of Fraud

Fraud is inherently secretive, so precisely estimating its prevalence is next to impossible. Nevertheless, a number of anonymous surveys have been administered to scientists to assess rates of scientific misconduct and data fraud. Fanelli (2009) reviewed and meta-analyzed these studies. Across seven surveys, the pooled weighted average rate of scientists who admit that they have "fabricated, falsified, or modified data or results at least once" was 1.97 percent (95% CI = .86 to 4.45; p. 1). By including only the four surveys with questions that used the words "fabrication" or "falsification" (aligning with our definition of data fraud), the estimate fell slightly to 1.06 percent (95% CI = .31 to 3.51). As the authors noted, however, these data are likely to be confounded by socially desirable responding, so the actual rate may be substantially higher.

Although research misconduct is present in all areas of study (Grieneisen & Zhang, 2012), rates differ widely between disciplines. Fanelli (2009) showed that surveys assessing clinical, medical, and pharmacological researchers yield higher rates of self-reported research misconduct than other disciplines or mixed samples. Within medicine, a study analyzing 134 medical journals across 10 disciplines found rates of retraction to be highest among anesthesiologists and lowest among radiologists (Sebo et al., 2023). Those with the highest number of retractions also tend to be anesthesiologists ("The Retraction Watch Leaderboard,"

2015). As medical research is often applied in settings where life is on the line, the elevated rates in these disciplines are worrying.

Negative Effects of Fraud

Beyond contributing to replicability difficulty, fraudulent research has a host of negative effects. Perhaps most obviously, and most generally, research is done to produce true knowledge that is often intended for practical application to accomplish human goals. Implementing or believing false research findings will disproportionately produce unintended, undesired outcomes. Fraudulent research also puts future research participants at risk. Steen (2011a) analyzed citations of retracted primary research reports between 2000 and 2010, finding, conservatively, that over 70,000 patients were treated in clinical studies that cited retracted research, suggesting that many patients are put at risk by flawed research. Retracted research is also cited by policy-related literature at similar rates to non-retracted research (Malkov et al., 2023). Although science is generally considered a self-correcting system, fraudulent research, at the very least, slows down this process.

Fraudulent research decreases public trust in science, as well as scientists, and wastes resources. Survey research shows that people are almost unanimously morally opposed to data falsification and fabrication, with 90% of the public supporting the notion that scientists caught in the act should be fired and banned from government funding (Pickett & Roche, 2018). Other scientists also experience a loss of trust; perhaps due to generalized stigma against fraud or lower confidence in research areas where fraud is found, previous collaborators of fraudulent scientists experience a significant drop in subsequent citations compared to controls (Hussinger & Pellens, 2019; Mongeon & Larivière, 2016). Because funding for science very frequently comes from public sources, any individual that recognizes the value of the scientific enterprise has a stake in

preventing fraudulent research. Public research funds must be efficiently allocated to critical research domains to facilitate scientific progress, but labs that conduct fraudulent science are often supported by public funds. Further, researchers that build upon fraudulent science waste their time, money, and energy following misleading "findings." Outside of laboratory spending, investigation into research misconduct involves the investment of millions of dollars of public funding every year (Gammon & Franzini, 2013; Stern et al., 2014). Public funding for science is only properly allocated if it supports the discovery and communication of scientific truths. Fraud bastardizes truth, and if truth is not held as a sacred value in the scientific enterprise, its fruits are worth nothing.

Reasons For and Solutions to Research Fraud

There have been several suggestions regarding the reasons that scientists commit research fraud. For example, Sovacool (2008) reviews three common narratives that attempt to account for the phenomenon (and associates them with different solutions): individual impurity (resolved by self-regulation), institutional impropriety (resolved by external control), and structural crisis (resolved by a restructuring of values and practices). As reviewed in Biagioli et al. (2019), researchers often proposed reasons such as poor institutional culture and incentive systems. The measures (e.g., h-index) that are used to reward scientists with promotion, grant funding, and awards can be improved through publishing more research, and research misconduct can be used to publish research without the investment involved in legitimate research. Furthermore, scientists understand they must either "publish or perish." Failing to publish, especially for early career researchers, will eventually put them out of their position and strip them of their livelihood.

There are also a number of different views on how to address the issue of academic fraud, ranging from preregistration (Craig et al., 2020), to changes in incentives (Wilson, 2020), to regular data auditing (Shamoo, 2013), and criminalization of fraudsters and protection of whistleblowers (Sovacool, 2005). One particularly interesting development is the potential for utilizing Benford's law to detect fraudulent data in a collection of publications by comparing expected values with those reported by researchers (Horton et al., 2020). However, there seems to be little progress in implementing reforms to reduce the frequency of academic misconduct, with whistleblowers being the most likely avenue of unearthing fraud (Stroebe et al., 2012).

Psychology's Potential Contribution

Psychology has a unique contribution to make to the problem of scientific fraud. Understanding fraud is a step towards combatting it and reducing its negative influences. That is, if we understand the psychology of the fraudulent scientist, we may be able to formulate interventions to prevent fraudulent behavior, such as training methods for budding researchers. Indeed, recently there has been increased attention to the potential for a new discipline, the "psychology of science," to help better understand how scientists operate, thereby providing clues for making science work better (Aczel, 2023). Despite psychology's potential contribution, little research has been done on the psychology of scientific fraud due to methodological issues. Fraud is virtually impossible to study experimentally, and, due to secrecy, existing data is generally hidden from sight. However, through archival analysis, analysis of the features of fraudulent research, and text analysis of fraudulent research, some answers are beginning to surface.

Archival Analysis

Two studies have systematically analyzed existing records of cases of research misconduct in pursuit of psychological causes. Utilizing multidimensional scaling techniques and cluster analysis, Davis et al. (2007) analyzed documents from case files of research misconduct from the Office of Research Integrity (ORI) in the United States, reducing the information to seven factors that were implicated in research misconduct: (1) personal and professional stressors, (2) organizational climate, (3) job insecurities, (4) rationalizations A, (5) personal inhibitions, (6) rationalizations B, and (7) personality factors such as impatience, amnesia, laziness, character flow, and personal need for recognition. The reports informing this study were formulated *post-hoc* and, therefore, are necessarily somewhat speculative; however, the categories derived from them align nicely with "the fraud triangle," a model of dishonest behavior used by auditors to detect fraud in professional settings which includes (1) the opportunity to cheat (e.g., factor 2), (2) the motivation to cheat (e.g., factors 1, 3, and 7), and (3) the ability to rationalize or justify dishonest behavior (e.g., factors 4, 6, and 7) (Houdek, 2017).

In a historiometric study comparing cases of research misconduct to cases of other ethical violations in research and medical malpractice (40 cases for each category), DuBois et al. (2013) found that cases of research misconduct more often than not involved repeated offenses (68%) of lone actors (90%) over multiple years (M = 3.8). However, the authors struggled to find differences between the case types for their hypothesized environmental influences.

Features of Fraudulent Research

The features of fraudulent research may also help us understand the settings or characteristics of fraudsters and their research groups. In an analysis of retractions in PubMed between 2000 and 2010, Steen (2011b) compared retractions due to data fraud with retractions due to error and concluded that features of fraudulent papers indicate deliberate deception.

Specifically, he found that fraudulent papers tend to have more authors (but see Markowitz and Hancock, 2014), suggesting an attempt to diffuse responsibility, and he found that fraudulent papers take longer to retract than erroneous papers, which indicates attempts to conceal or not be noticed. He also found that most fraudulent papers have a first author with a record of multiple other retracted papers, such that most fraudulent papers are authored by repeat offenders. He also found that fraudulent papers were more likely to appear in journals with higher impact factors. These features, he contends, are not what one would expect if data fraud were "naïve, feckless [or] inadvertent" (p. 113).

In alignment with the notion that research fraud is deliberate and calculated, Walsh, Lee, and Tang (2019) found that research teams with a greater division of labor—measured by the share of tasks performed by only one author from available contribution statements—have a higher probability of having their paper retracted, suggesting that the siloing of tasks facilitates concealment of fraudulent practices from collaborators. With regard to gender, evidence indicates that men are overrepresented as authors of retracted publications, both by rate (Sebo et al., 2023) and among those with the highest number of retractions ("The Retraction Watch Leaderboard," 2015). In summary, features of fraudulent research suggest that the prototype of a fraudulent researcher is a man deliberately committing fraud in a large research team that is segmented to restrict information flow that could reveal fraudulent activities.

Text Analysis

Text analysis involves quantitatively analyzing the usage of words (e.g., by frequency) that constitute psychologically meaningful categories. This method can elucidate psychologically relevant attributes of to the speaker or writer. For example, the computerized text analysis program Linguistic Inquiry and Word Count (LIWC) is a psychometrically rigorous instrument

that can assess attentional focus, emotionality, social relationships, thinking styles, and individual difference traits based on text alone (Boyd et al., 2022; Tausczik & Pennebaker, 2010). Critically, this methodology circumvents problems of data availability regarding the psychological processes involved in committing scientific fraud by leveraging data that is available in all cases: the text of the scientific paper. Because no other methodology can access data from during the time of a fraudulent event, linguistic analysis is particularly informative for understanding the psychology of scientific fraud.

Language and Deception. While text analysis of deceptive statements has produced a diverse set of linguistic effects (e.g., see Tausczik and Pennebaker, 2010), meta-analysis suggests that these effects are highly contingent on the particulars of the communication and context (Hauch et al., 2014). Therefore, to inform our discussion of the psychology of research fraud, the present review of linguistic indicators of deception will be restricted to studies of deceptive behavior in scientific publications and other professional contexts.

Deception in Business Writing. Although much of the research on linguistic indicators of deception focuses on spoken words or communication through informal mediums, writing in business settings—like scientific writing—is formalized for a professional context, which makes research in this area uniquely informative for generating hypotheses about deception in scientific writing. Scholars of business ethics have systematically studied linguistic indicators of deception in business-related documents, unearthing several linguistic indicators of deceptive reports. Many studies have found that deceptive financial reports tend to have higher levels of linguistic obfuscation—often measured by indicators of readability, complexity, or variability of word or sentence length (Courtis, 1998; Goel et al., 2010; Humpherys et al., 2011; F. Li, 2008). As deceptive financial reporting can be interpreted negatively by internal and external stakeholders,

this is thought to function to make it more difficult for readers to assess the information a firm is attempting to hide. However, although these operationalizations of obfuscation seem to yield consistent effects, Linsley and Lawrence (2007) found that the readability of risk reports from firms in the United Kingdom did not influence understandability, expressing dissent regarding the obfuscation hypothesis.

In another informative business ethics domain, research has investigated firms engaged in "decoupling" (saying one thing while doing another) in the context of sustainability reporting. Perhaps in contrast to the linguistic obfuscation hypothesis, these investigations show that decouplers tend to use less linguistically sophisticated or complex language, which they interpret as a lack of cognitive complexity (Conrad & Holtbrügge, 2021; Crilly et al., 2016). It may be that in the case of financial filing, obfuscation is used to hide information from internal and external stakeholders, while for decoupling there is less scrutiny from internal stakeholders. In this case, reports of sustainability would only need to avoid external scrutiny through more simple language, thereby avoiding attention.

Linguistic Indicators of Research Fraud. Although the domain of scientific fraud has not been studied as thoroughly as deception in business writing or other communication formats, there are three preliminary studies that investigate the psychology of scientific fraud through linguistic analysis of fraudulent or retracted papers. The first exploratory study of this sort took a case study approach, analyzing Deidrick Stapel's linguistic indicators of deception by comparing 24 of his fraudulent papers to 25 of his genuine papers, all of which he was first author of (Markowitz & Hancock, 2014). Stapel's fraudulent publications included higher rates of terms related to scientific methods and empirical investigation; more amplifying terms and fewer diminishers; more terms related to certainty; fewer adjectives; more emotional actions, states,

and processes; and fewer coauthors. The higher rate of scientific methods and empirical investigation words suggests that Stapel struggled to approximate how much detail to include while reporting on fake data, especially considering he was not more descriptive in his fraudulent writing overall (e.g., as measured by fewer adjectives). Ultimately, consistent with previous deception research in other contexts, Stapel's frequent amplifying terms, sparse diminishers, and overall higher certainty may indicate attempts to persuade readers of high research quality—a strategy which may inoculate them against considering malfeasance.

Following up this exploratory work, Markowitz and Hancock (2016) aimed to increase generalizability by comparing papers retracted for scientific fraud (n = 253) with matched non-retracted papers (i.e., by journal, publication year, and, where possible, keywords) and papers retracted for ethical issues unrelated to data fraud (n = 62). To investigate the obfuscation hypothesis, they compared the three groups of papers using their novel linguistic obfuscation index. Results showed that fraudulent papers had higher linguistic obfuscation as measured by their index (as well as each of its subcomponents), and this effect withstood tests of potential confounding moderators such as scientific discipline and geographical origin. Furthermore, fraudulent papers had more references, and the obfuscation index of fraudulent papers was positively correlated with the number of references per paper. Aligning with linguistic analysis in finance and theoretical approaches that emphasize the goal-oriented and strategic nature of deception, these findings are consistent with the hypothesis that fraudulent scientists obfuscate in order to make their paper more difficult and costly to assess. In addition, the finding that attention (i.e., citation count) is a predictor of retraction indicates that one mechanism of detecting fraud is close scrutiny from the scholarly community (Furman et al., 2012), which supports the plausibility of this explanation.

In an analysis comparing retracted and non-retracted biochemistry papers published between 2018 and 2020, Dehdarirad and Schirone (2023) were able to use exact matching on year, journal, and topic and covariance balancing on previous correlates of retracted papers (i.e., gender, country, and scientific impact). Although this study suffers from the limitation that retracted papers are not necessarily fraudulent, their sophisticated matching procedure and doubly robust estimation technique (regression combined with propensity score matching) helped to remove confounding effects, and their exclusive reliance on papers from biochemistry is at least a pragmatic strength, considering the dangers of implementing false-positive findings in clinical medicine. Their results showed that retracted articles had 16% lower odds of using positive terms in abstracts and titles, replicating Markowitz and Hancock's (2016) analysis of the positive emotion terms subcategory of their obfuscation index. They also found that retracted papers were 15% less probable to use certainty language when presenting scientific findings. Although this finding contradicts those from the case study of Deidrick Stapel (who utilized significantly more certainty language; Markowitz & Hancock, 2014), it is conceptually consistent with Markowitz and Hancock (2016). That is, uncertain language could function to render findings less comprehensible, thereby masking deception.

Inconsistent Effects and the Present Study. While this preliminary research on fraudulent and retracted research has shown parallels with analyses of financial filings, effects have not been entirely consistent. Regarding linguistic uncertainty, Dehdarirad and Schirone (2023) found that retracted papers were more likely to use uncertain language, which contradicts the findings from Deidrick Stapel's corpus. Whether fraudulent research is associated with more uncertainty stands to be confirmed in the present study.

Inconsistencies also exist for reference count. Markowitz and Hancock (2016) found that fraudulent papers contained more references than non-fraudulent papers—interpreting this as a method of increasing cost of assessment by readers. A recent unpublished thesis was also interested in whether fraudulent papers would have more references due to the finding that deceptive texting involves the use of more third-person singular pronouns (e.g., *he, she, they, them*; Hancock et al., 2008). They reasoned that although academic writing conventions would preclude the use of pronouns, references to other scientific papers are at least analogous to third-person pronouns. Comparing retracted (n = 314) and non-retracted papers (matched by journal), they did not find a significant difference between the groups for word count of reference sections (Schmidt, 2022). However, using the number of words as a proxy for the number of references may have introduced substantial measurement error, and the present study will assess whether fraudulent research tends to utilize more references.

Schmidt (2022) also found no difference between retracted and non-retracted papers for readability—a key metric of obfuscation in financial filing studies and within Markowitz and Hancock's (2016) obfuscation index—and lower rates of complex words in retracted papers. These findings are consistent with research on communication by firms engaged in decoupling, but they challenge the obfuscation hypothesis. The present study will help to clarify the conditions under which the obfuscation hypothesis does and does not hold.

Limitations and New Directions

Although previous literature has found support for psychological hypotheses of scientific fraud through linguistic effects in retracted or fraudulent papers, these effects have not been entirely consistent across studies. Furthermore, all studies thus far have relied on papers with multiple authors or did not differentiate between single- and multi-author papers. Without

verification of who wrote each paper or proposed concepts or content to the direct writer, considering linguistic markers as indicators of psychological states or processes is an interpretive leap. The present research will fill this gap by reassessing inconsistent effects and testing new hypotheses through comparison of single-author fraudulent publications (SAFP), single-author genuine publications (SAGP), multi-author fraudulent publications (MAFP), and multi-author genuine publications (MAGP).

Analyzing single-author papers allows one to draw more valid conclusions about psychological processes, and it also allows for comparisons to differentiate between alternative explanations for findings to date. Specifically, there is suggestive evidence that linguistic obfuscation in fraudulent research may be moderated by the presence of research collaborators and function to hide scientific fraud from *the research group* rather than *external readers*.

First, most scientific fraud is detected by whistleblowers with insider information, making the immediate research group a more pertinent threat to one's secrecy (Stroebe et al., 2012). DuBois et al. (2013) found that—compared to cases of medical malpractice and other forms of ethical breaches in research—cases of research misconduct are more likely to be reported by subordinates and by institutional peers (rather than external peers). Second, in surveys, reporting of personal knowledge of data fabrication or falsification is much higher than anonymous reporting of engaging in data fraud personally, suggesting that one's colleagues are likely to discover research fraud when it is occurring. Furthermore, these colleagues often report having taken action against the fraud or person who commits the fraud. In Fanelli's (2009) meta-analysis of survey data pertaining to the frequency of data fraud, a pooled weighted estimate of scientists willing to admit having engaged in data fabrication or falsification at least once is around 1.06 percent, whereas 12.37 percent of scientists report having personal knowledge of a

colleague who fabricated or falsified research data (self-report estimates have fallen over time whereas non-self-report estimates have been more consistent). Although the latter estimate may be inflated by group knowledge of an individual's act of misconduct, these statistics suggests that a researcher who commits data fraud is likely to be discovered by a *close colleague*. Fanelli (2009) also reviewed studies that assess actions taken in response to personal knowledge of research fraud. Respondents report taking some form of action against the misconduct between 12.1 percent (reporting to Institutional Review Board) and 55.5 percent (confronting the culprit) of the time, depending on the action being probed by survey items (see Table 2 in Fenelli, 2009). The primacy of the research group for detecting and correcting cases of fraud suggests a differential benefit to fraudsters from obfuscating research to deceive close colleagues, rather than readers.

Third, Markowitz and Hancock (2014) found that Deidrick Stapel's fraudulent papers had fewer authors than his genuine papers, perhaps due to salience of the potential for collaborators to discover his fraud (which they eventually did). Finally, fraudulent research tends to have a higher degree of division of labor (Walsh et al., 2019); authors may attempt to hide their fraudulent activities from collaborators through separation of research tasks. Such behavior is analogous to the hypothesized function of higher obfuscation in scientific writing and suggests the salience of the research group in the mind of the fraudster. This leads to the prediction that linguistic obfuscation will be highest in cases of fraud where collaborators are present (multi-author fraudulent papers), although still higher on average when the fraudster is operating alone (single-author fraudulent papers).

Finally, we also propose that in multi-author fraudulent papers, the fraudulent actor will be more likely to be the corresponding author. Regulating communication between the research

group and those external to the research group would allow fraudulent actors to monitor, manage, and deescalate suspicious inquiries. In summary, the present study will replicate and extend previous work by retesting three previous hypotheses and making three novel predictions, each of which are displayed below in *Table 1*.

Table 1

Study Hypotheses

Label	Hypothesis	Novelty
1a	Fraudulent research will be written with more linguistic obfuscation than non-fraudulent research.	Replication
1b	Single-author fraudulent research will be written with more linguistic obfuscation than non-fraudulent research but less linguistic obfuscation than multi-author fraudulent research.	Novel
2a	Fraudulent research will contain more references than non-fraudulent research.	Replication
2b	Single-author fraudulent research will contain more references than non-fraudulent research but fewer references than multi-author fraudulent research.	Novel
3	Fraudulent research will contain less certainty than non-fraudulent research.	Replication
4	For multi-author fraudulent papers, the fraudulent author will be the corresponding author more frequently than other authors.	Novel

Methods

This project was preregistered prior to data collection (see osf.io/a38jv), and all data and code to reproduce the analyses are available in the GitHub repository associated with the Open Science Framework project (see https://osf.io/vmq48/?view_only=24a59f4b9f984b68971437d6d943d4b9). This study was exempt from Human Research Ethics Board review because the data were available in the public domain without the expectation of privacy (see TCPS Article 2.2; Government of Canada, 2022). For this retrospective cohort study, we approximated a matched case control design. Previous literature has revealed several extraneous characteristics associated with fraudulent or retracted papers. Attempting to create parity across comparison groups based on these prespecified criteria allowed us to minimize potential confounds while retaining potentially lost power through statistical control.

Fraudulent papers were identified using the Retraction Watch Database (RWDB; 2024). This database is the most comprehensive available data source for retracted papers in the scientific literature, comprising 50,343 entries when we downloaded it on January 21, 2024. The blog Retraction Watch, an offshoot of The Center for Scientific Integrity, has recently sold the rights to this database to Crossref, making it a freely available resource (Hendricks et al., 2023). The RWDB contains bibliographic data and records the discipline, country, article type (e.g., research article, case report), and reason(s) for retraction for all possible entries. The article type "research article" is defined as "[p]ublished item describing a hypothesis, means of exploring the hypothesis, the results of the exploration, and the conclusions drawn from the results" ("RWDB User Guide Appendix A," 2018). All papers collected for our analysis were research articles, whether by their indexing in the RWDB (in the case of fraudulent papers) or by verification through skimming the paper before inclusion (in the case of comparative genuine papers).

Reasons for retraction are independently determined by the administrators of the RWDB on a case-by-case basis using information from, for example, retraction notices or news articles, and retractions corresponding to our definition of data fraud are entered as "Falsification/Fabrication of Data." Fraudulent papers collected for our analysis were originally considered fraudulent by this criterion. Then, although not originally a part of our preregistration, a later inspection of the "Reasons" field in the RWDB indicated that the reasons "Fabrication/Falsification of Image(s)" and "Manipulation of Images" also fit our definition of data fraud, and papers engaging in these practices were considered eligible for inclusion. All genuine papers were checked to ensure they were not fraudulent by first exporting each to Zotero (Center for History and New Media, 2023) with the Retraction Watch Integration installed (Stillman, 2019), which automatically flags retracted papers logged in the RWDB.

Matching Criteria

Previous research has found retracted papers to differ in aggregate from nonretracted papers on a variety of dimensions, and other potential confounds may influence the linguistic features of papers. We matched these variables across groups whenever possible to facilitate valid inferences from our comparisons. In cases where not all criteria could be matched—i.e., if there were no papers with exact matches across all criteria when collecting MAFP from the RWDB or SAGP and MAGP from journals—we prioritized criteria in the order they appear below. Ultimately, the idiosyncratic nature of each case where a paper could not be exactly matched forced the first author to make judgments about paper inclusion on a case-by-case basis, without the possibility of an algorithmic system. However, whenever possible matching criteria were optimized to ensure that the most important features, or the largest number of features, matched for each paper choice. Matching characteristics were then dummy coded to denote when

a paper deviated from its match in order to assess the degree of similarity between groups both in aggregate and at the paper level.

Topic

Differences in scientific writing due to content area and reporting practices were controlled by matching for topic. This was accomplished whenever possible by using the paper's journal as a proxy. In multiple cases, the RWDB did not contain any MAFP from the same journal as the SAFP in question, so these papers were matched by the RWDB "Subject(s)" data (deviating from our preregistration). This data denotes the field of study for each paper, and a given entry often has multiple subject categories assigned. When papers had multiple entries in the subject field, matching across all subject entries was preferred, but matching using as few as one subject entry was considered acceptable.

Publication Date

Papers may vary in writing style and other conventions across time. To control for this, papers were matched according to the year of the publication date whenever possible.

Gender

Research indicates that men are overrepresented among authors of retracted publications both by rate (Sebo et al., 2023) and among those with the highest number of retractions ("The Retraction Watch Leaderboard," 2015), and gender differences in language use have been found in other linguistic contexts (Newman et al., 2008). Therefore, to control for potential confounds, gender of first author (male, female, or other gender) was determined through online searches (e.g., university profiles, lab websites, social media, news reporting, etc.).

Institutional Prestige

Frequency of jargon use by graduate students in their dissertation titles differs according to the prestige of their research institution, indicating that institutional prestige is associated with linguistic styles (Brown et al., 2020). High prestige research institutions also seem to have a lower rate of retractions (Lievore et al., 2021). In accordance with Dehdarirad and Schirone (2023), we dichotomized institutional prestige into major research institutions and other research institutions according to whether they were present in the top 200 schools of the Times Higher Education World University Rankings (2024). In one case, no papers were available other than a paper published by a non-university research institution. This case was classified as lower prestige.

Country

Retraction rates and linguistic features may differ between countries and authors for whom English is their first language or is learned later. As a proxy to control for these issues, papers were matched by the country of residence of the first author, based on the location of their institutional affiliation when paper was published.

Procedure

For all papers (N = 88) full-text PDFs were downloaded during data collection, renamed with the DOI of the original publication, and stored for data preprocessing. All eligible SAFP in the RWDB (n = 22) were assembled and online searches were conducted to collect data for matching characteristics not already contained in the RWDB (i.e., gender and institutional prestige). Next, MAFP were initially identified from the RWDB through filtering by journal, year, and country, and removing filters in the opposite order when necessary. These papers were investigated sequentially to identify matches on our other criteria through internet searches, and

the RWDB data, as well as other matching data, were recorded. Once all fraudulent papers (SAFP and MAFP) were assembled, online searches for genuine papers (SAGP and MAGP) were conducted by sifting through articles published in the journal volume of the appropriate year. Papers were investigated in their chronological order of issues and the order in which they appear in those issues. The search stopped, and data was recorded, once optimal matching characteristics were discovered. For potential multi-author papers, we also considered the number of authors on the paper by aiming for fewer authors. This approach was intended to increase the likelihood that the fraudulent author exerted influence on the text of the paper. SAGP were matched with SAFP, MAFP were matched with MAGP, and SAFP were matched with MAFP; this facilitated optimal matching across all groups where our hypotheses predicted group differences.

Data Preprocessing

All automated data preprocessing procedures were conducted using Python (2023) and all Python scripts were executed in a *Jupyter Notebook* (2023). Scripts were developed with ChatGPT using the GPT-4 model (OpenAI, 2023) then edited to best suit our purposes, and all scripts are available in our data analysis plan or upon request (Zubaly & Arbeau, 2023). Stored full-text PDF versions of all papers were first converted to plain text files using the fitz module. In four cases (one from each paper group), the original PDF of the paper was a scanned copy, so the Optical Character Recognition tool within *Adobe Acrobat* (2023) was used to first generate readable text for the papers before rerunning our script to convert them into text files. Then, each file was opened manually to ensure the text was digitally read correctly, as well as to remove abstracts and other sections to be excluded from our text analysis procedure. Only main body text (i.e., introduction, methods, procedure, data analysis, results, discussion, and conclusion

sections, not including titles) was retained. Next, we used a custom script with the os and re modules to remove extraneous linguistic features from the texts. This script removed all periods that did not end sentences; converted all words to lowercase; removed all variants of the word "retract"; removed all variants of the word "withdraw"; removed common special characters for scientific and nonscientific writing; replaced newline characters with a space; removed all numbers; removed dates in common formats; and removed the terms "introduction," "methods," "procedure," and "discussion." These preprocessing procedures help to mitigate the miscounting of linguistic features during the text analysis procedure.

Text Analysis

Once data preprocessing procedures were complete, all plain text files were uploaded to text analysis programs for measurement of linguistic indicators. The resultant comma separated value file outputs were then cleaned manually, merged into one dataset, and merged with the main dataset based on their DOIs using a custom Python script that utilized the os and pandas modules.

Measures

Linguistic Obfuscation

Linguistic obfuscation was measured with the Obfuscation Index developed by Markowitz and Hancock (2016). This index (O) is composed of the difference between the sum of causal terms (C), the abstraction index score (A), and jargon (J) and the sum of positive emotion terms (P) and Flesch Reading Ease (F). If we let the elements in parentheses above represent their respective variables, *Equation 1* represents the linguistic obfuscation score for an observation i.

Equation 1

Mathematical model of the linguistic obfuscation index

$$O_i = (C_i + A_i + J_i) - (P_i + F_i)$$

Each variable was standardized before constructing the index. Although the reliability and validity of this index has yet to be formally assessed, it is composed of elements that have routinely been used in text analysis and to test the linguistic obfuscation hypothesis in the business ethics literature. Furthermore, Markowitz and Hancock (2016) showed that each subcomponent correlated with each other subcomponent in the expected direction for internal reliability in their sample, and each subcomponent of the abstraction index showed this same property (see Table 1 and Table 2, respectively, in Markowitz, 2016). They also showed that while the Obfuscation Index was associated with fraud in their sample, so was each subcomponent. All subcomponents of the index, with the exception of Flesch Reading Ease, were measured with Linguistic Inquiry and Word Count 2022 Version (LIWC; Boyd et al., 2022; Tausczik & Pennebaker, 2010). Causal terms (e.g., *how*, *because*, *make*) and positive emotion terms (e.g., *good*, *love*, *happy*) were measured using the standard LIWC dictionaries. Jargon was measured using the percentage of words *not indexed* by *any* LIWC dictionaries; the LIWC dictionaries tend not to contain words that are highly specialized for academic communication, so this provided an appropriate proxy for jargon use. The abstraction index was composed of the inverse of the sum of the standardized LIWC scores for articles, prepositions, and quantifiers, all of which have standard dictionaries in LIWC. Articles (e.g., *a*, *an*, *the*), prepositions (e.g., *in*, *on*, *at*), and quantifiers (e.g., *some, many, few*) each contribute to descriptiveness or concreteness. Taking the inverse of their sum makes more concrete language count negatively towards the total score, and this provides a good indication of abstraction. Flesch Reading Ease was measured

using a custom Python script that utilized the textstat module (Aggarwal, 2022). Flesch Reading Ease (Flesch, 1948) is a standard measure of readability, and readability is often measured to indicate obfuscation.

Certainty

Certainty was operationalized as the certainty sentiment score indexed by The Lexical Suite (Rocklage et al., 2023). The Lexical Suite's certainty sentiment score improves on the measurement of certainty by other common text analysis programs (e.g., LIWC certainty) for three main reasons. First, the dictionary was developed using a data-driven approach with multiple sources of candidate words to approach saturation. Second, dictionary development including contributions of the BioScope corpus—a collection of biomedical texts that have been annotated for how speculative different n-grams within scientific texts are—for the initial word list, suggesting that the final dictionary more accurately measures certainty in a scientific context. Third, the dictionary was refined and certainty scores for each word were assigned by many raters, allowing for imputation of certainty scores for each dictionary word. These steps allow the final certainty sentiment score to avoid limitations of previous measures, providing greater validity, reliability, and specificity than previous text analysis measures of certainty (Rocklage et al., 2023). Examples of n-grams in the certainty sentiment dictionary include *very vague* (1.97) and *beyond any doubt* (8.8).

References

Previous research has attempted to quantify references by calculating the word count of reference sections (Schmidt, 2022). In contrast, in the present study references were counted manually using the original PDF of each paper. This approach allowed us a higher degree of

reliability, as the average length of references may be different between disciplines due to content matter and stylistic differences.

Fraudulent Corresponding Author

After the collection of all multi-author fraudulent papers, data regarding whether the fraudulent author was the corresponding author (n = 4/8) was collected using publicly available resources as evidence (e.g., Retraction Watch articles, retraction notices from journals, news articles, etc.).

Data Analysis

Data were cleaned and analyzed using R (*R*, 2023) in RStudio (RStudio Team, 2023), and all analyses were documented within a Quarto document (*Quarto*, 2023) or R script. Following data exploration, all hypotheses were tested using null hypothesis significance tests with an α = .05 criterion. Hypotheses 1a, 2a, and 3 were tested by comparing group means using Student's t-test, and, when assumptions were violated, a robust nonparametric alternative test or bootstrapped confidence intervals around the t statistic. Hypotheses 1b and 2b were tested by comparing group means using one-way analysis of variance (ANOVA), and, when assumptions were violated, with nonparametric alternatives or bootstrapped confidence intervals for the F statistic. Bootstrapping was performed with 1,000 resamples.

Hypothesis 4 was tested using a binomial test to compare the proportion of fraudulent authors that were corresponding authors with an expected frequency that reflected the aggregate probability that any one author would be the corresponding author. This analysis rested on two assumptions. *First, that all authors on a paper were equally likely to be the corresponding author.* Although this assumption oversimplifies norms of assigning contributors to be corresponding authors, in the absence of more information regarding, for example, author order,

author responsibilities, or laboratory status, it is the most accurate prediction we can make regarding the likelihood of an author being the corresponding author. That is, in the absence of other information, it is the baseline prediction. *Second, if fraudulent authors did not attempt to control information flow, they were equally likely to be the corresponding author as their coauthors were.* This may also oversimplify the norms of assigning scientists to be corresponding authors. For example, it is possible that fraudulent authors perform more data management, and it is possible that scientists that are responsible for data management are more likely to be corresponding authors. However, given the absence of information (e.g., regarding research group norms), this assumption is reasonable. These two assumptions allowed us to make a prediction regarding the baseline expected frequency that fraudulent authors would be the corresponding authors of fraudulent research papers if they did not attempt to control information flow. This baseline expected frequency was the inverse of the mean number of authors on all MAFP where the corresponding author and fraudulent author were identifiable. The binomial test compared our observed proportion to this value.

Results

Descriptive Statistics

Full Sample

The mean year of publication for the entire sample was mid-2009 (*SD* = 10.47), with a median of 2013. The oldest paper was published in 1980, and the most recent paper was published in 2022. The mean number of references per paper was 45.82 (*SD* = 23.47), and the mean for word count as measured by LIWC was 4,436.25 (*SD* = 2,248.24). In line with previous research, most papers were from first authors at low lower-prestige research institutions (*n* = 70, 79.4%), and most first-authors were male (*n* = 73, 83%). Although twenty-two countries were

represented among our sample, the majority (n = 49) had a first author who resided in one of three countries: the United States (n = 29, 33%), China and India (both n = 10, 11.4%). The fourth most commonly represented country was the Netherlands (n = 6, 6.8%), and no other country represented more than n = 4 papers. Most fraudulent research papers were retracted for fabrication/falsification of data (n = 27, 61.4%), followed by manipulated image(s) (n = 12, 27.3%) and fabricated/falsified image(s) (n = 3, 6.8%), and two papers were retracted for fabrication/falsification of data and images (4.5%).

Matching Characteristics

We attempted to match characteristics at the paper level across groups and documented situations where this was not possible through dummy coding. Because a mismatch between two papers in comparison groups may be cancelled out by the opposite mismatch for another pair of papers, it is possible for there to be many mismatches at the paper level of comparison but similar aggregate frequencies of matching characteristics when comparing groups. Thus, the following documents mismatches for most characteristics at the paper level and at the aggregate level between each of the groups for which our hypothesis tests make comparisons. Year, however, is only documented by the average year difference between the papers.

MAFP Compared with SAFP. Country of first author was the same for 14 pairs (64%) and different for 8 pairs (36%). Papers were most commonly from the United States in MAFP (n = 8, 36%) and SAFP (n = 7, 32%). Gender was the same for 18 pairs (82%) and different for 4 pairs (18%), and MAFP had a total of 18 (82%) male first authors and 4 (18%) female first authors, while SAFP had a total of 19 (86%) male first authors and 3 (14%) female first authors. For institutional prestige, there was a match between 21 (95%) pairs and a difference for 1 (5%) pair, and MAFP had a total of 17 (.77%) authors from lower-prestige research universities and 5

(22%) authors from higher prestige institutions, while SAFP had a total of 18 (82%) authors from lower-prestige schools and 4 (18%) authors from high prestige schools. The mean difference in years between the two groups was -1.36, indicating that the MAFP tended to be slightly more recent than the SAFP. Finally, MAFP and SAFP were the only groups that sometimes did not match for journal, as MAFP were collected from the RWDB, unlike MAGP and SAGP which were found directly in the volume of the journal for the paper they were being matched to. There were 11 (50%) papers between these groups that we could not match by journal, and in all cases we matched based on one or more entry in the "Subject(s)" field of the RWDB.

MAGP Compared with MAFP. Country was the same for 21 (95%) pairs and different for 1 (5%) pair, with papers most frequently coming from the United States for both MAFP (8, 36%) and MAGP (9, 41%). Gender was the same for 20 (91%) pairs and different for 2 (9%) of pairs, but both groups had a total of 18 (82%) male first authors and 4 (18%) female first authors. Institutional prestige was the same for 21 pairs (95%) and different for 1 pair (5%), and MAFP had a total of 17 (.77%) authors from major research universities and 5 (22%) authors from lower-prestige institutions, while MAGP had a total of 18 (82%) authors from high-prestige schools and 4 (18%) authors from lower prestige schools. The mean difference in years between the two groups was -.64, indicating that the MAGP tended to be slightly more recent than the MAFP.

SAGP Compared with SAFP. Country was the same for 10 (45%) pairs and different for 12 (55%) of pairs. Still, the most frequent country of first author was the United States, with 7 (32%) for SAFP and 5 (23%) for SAGP. Gender was the same for 19 (86%) pairs and different for 3 (14%) pairs, and SAFP had a total of 19 (86%) male first authors and 3 (14%) female first

authors, while SAGP had a total of 18 (82%) male first authors and 4 (18%) female first authors. Institutional prestige was the same for 21 (95%) pairs and different for 1 (5%) pair, and SAGP had a total of 17 (.77%) authors from major research universities and 5 (22%) authors from lower-prestige institutions, while SAFP had a total of 18 (82%) authors from high-prestige schools and 4 (18%) authors from lower prestige schools. The average difference in year was 0, indicating that SAGP and SAFP are exactly matched for the year of publication in aggregate.

Bivariate Correlations

Bivariate correlations were calculated between (1) subcomponents of the abstraction index, (2) subcomponents of the linguistic obfuscation index, and (3) our three numerical outcome variables, linguistic obfuscation, certainty sentiment, and references.

Abstraction Index

Table 2 shows the correlations between each of the variables that made up our abstraction index. Although the direction of each relationship is positive, none of the relationships are significant at the .05 level, and the effect sizes are notably smaller ($.03 < r < .10$, all $p > .05$) than those found by Markowitz and Hancock (2016; $.176 < r < .263$). This finding calls into question the reliability of the abstraction index as a subcomponent of the linguistic obfuscation index in our sample, and it does not provide evidence for the validity of each of the indicators of abstraction (articles, prepositions, or quantity words).

Linguistic Obfuscation Index

Table 3 shows the correlations between the subcomponents of the linguistic obfuscation index. Our mathematical model of the linguistic obfuscation index (Equation 1) can be used to make predictions about the bivariate relationships that should be positive and the bivariate relationships that should be negative for the index to be internally consistent. That is, we should

Table 2

Correlations for subcomponents of the abstraction index

Variable	*M*	*SD*	1	2
1. Articles	8.38	2.20		
2. Prepositions	15.43	1.68	.10 [-.12, .30]	
3. Quantity words	4.23	1.32	.06 [-.16, .26]	.03 [-.18, .24]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$.

expect that each of the elements within the aggregated variables (i.e., that are within parentheses with one another) should positively correlate. Therefore, we should expect the relationships between causal terms, the abstraction index, and jargon words to be positive. However, each of these relationships are small, negative associations, with the exception of the relationship between abstraction and jargon, which is very slightly positive. The relationship between positive emotion words and Flesch reading ease should also be positive, but it is also a small, negative association.

We should also expect, based on Equation 1, that elements composing the linguistic obfuscation index will be negatively correlated across the aggregates (i.e., each of the variables that are in opposite aggregates). Indeed, *Table 3* shows there is a moderate negative relationship between causal terms and Flesch reading ease. However, we should expect the relationship between causal terms and positive emotion terms to be negative, but it is a small, positive correlation. We should expect a negative relationship between the abstraction index and positive

emotion terms, but there is no relationship to speak of. We also should expect a negative relationship between the abstraction index and Flesch reading ease, but, again, this relationship is too small to consider. As we should expect, there is a small negative relationship between Jargon and positive emotion terms. Lastly, we should expect a negative association between jargon and Flesch reading ease, but there is instead a small, positive association between them.

The observed relationships here may be expected *a priori* in some cases (e.g., the relationship between jargon and Flesch reading ease may be expected to be a positive relationship because the longer sentences someone writes the more likely they may be to use nonstandard words, simply by volume). However, from an internal consistency perspective, most

Table 3

Correlations for subcomponents of the linguistic obfuscation index

Variable	*M*	*SD*	1	2	3	4
1. Causal terms	2.80	1.22				
2. Abstraction index	5.09	30.48	-.14 [-.34, .07]			
3. Jargon	33.33	7.47	-.21* [-.40, -.00]	.12 [-.09, .32]		
4. Positive Emotion Terms	0.09	0.22	.25* [.04, .44]	.03 [-.18, .24]	-.22* [-.41, -.01]	
5. Flesch Reading Ease	34.90	10.94	-.32** [-.50, -.12]	.07 [-.14, .28]	.18 [-.03, .37]	-.20 [-.40, .01]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$.

of the relationships (70%) here are either in the wrong direction, are too small to be considered, or are statistically unreliable (i.e., p > .05). This warrants serious consideration for whether any findings related to the linguistic obfuscation in our sample can be considered internally valid.

Numerical Outcome Variables

Table 4 shows the zero-order correlations between our numerical outcome variables. Each relationship is small enough to be negligible, with wide confidence intervals containing zero.

Table 4

Correlations for numerical outcome variables

Variable	*M*	*SD*	1	2
1. Linguistic Obfuscation	-0.00	2.04		
2. Certainty Sentiment	6.17	0.25	.07 [-.14, .28]	
3. References	45.82	23.47	-.04 [-.25, .17]	-.13 [-.33, .08]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. * indicates $p < .05$. ** indicates $p < .01$.

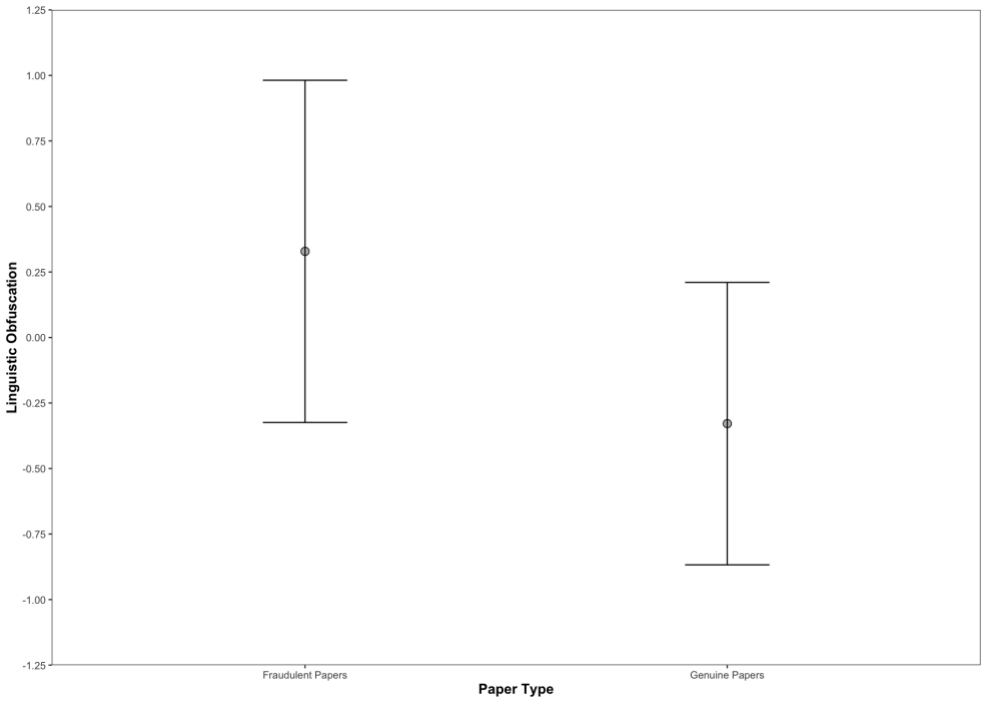
Testing Hypotheses

Hypothesis 1

Hypothesis 1a was tested using a t-test to compare linguistic obfuscation between fraudulent and genuine papers. Investigation of the QQ-plots and the results of a Shapiro-Wilk normality test for each group indicated that the assumption of normality was supported for both

Figure 1

Average linguistic obfuscation for genuine and fraudulent paper groups



Note. Error bars indicate the 95% confidence interval for each mean.

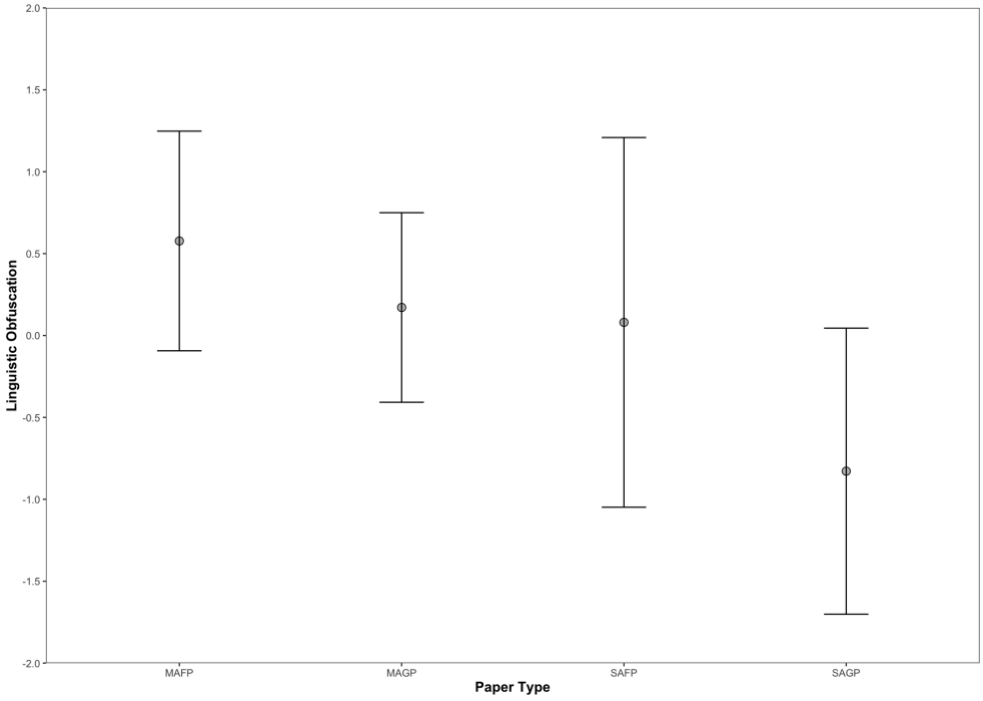
the genuine paper group (W = .97, p = .43, n = 44) and the fraudulent paper group (W = .97, p = .21, n = 44), and Levene's test indicated that we did not violate of the assumption of homogeneity of variance ($F(1, 86)$ = .91, p = .34). *Figure 1* shows the average levels of linguistic obfuscation in both the fraudulent and genuine paper groups with 95% confidence intervals. Results of the t-test showed that there was not a significant difference between the fraudulent papers (M = .33) and genuine papers (M = -.33) on linguistic obfuscation ($t(86)$ = 1.52, p = .13, Cohen's d = .32).

Hypothesis 1b was tested using a one-way ANOVA to assess for differences between the SAGP, SAFP, MAGP, and SAFP groups on linguistic obfuscation. The QQ-plot and Shapiro-Wilk test within each group indicated that the assumption of normality was supported for each of

the groups—MAFP (W = .96, p = .57, n = 22), MAGP (W = .93, p = .13, n = 22), SAFP (W = .96, p = .40, n = 22), and SAGP (W = .96, p = .45, n = 22)—and Levene's test indicated that we should accept the null that the variances are equal between the groups (F(3, 84) = 1.79, p = .16). *Figure 2* shows the means with 95% confidence intervals for linguistic obfuscation in each group. In line with our predictions, MAFP had the highest mean (M = .58), and SAFP (M = .08) was higher than SAGP (M = -.82). However, SAFP was very slightly lower than MAGP (M = .17). Ultimately, the result of the one-way ANOVA indicated that there were no significant differences between the four groups on linguistic obfuscation (F(3, 84) = 1.92, p = .13).

Figure 2

Average linguistic obfuscation for MAFP, MAGP, SAFP, and SAGP groups



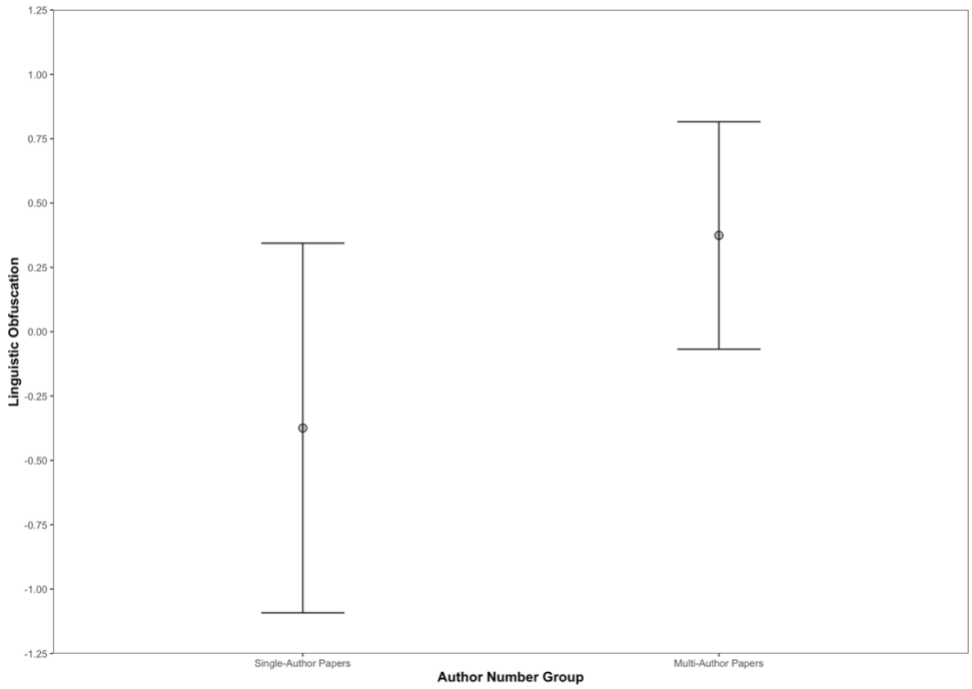
Note. Error bars indicate the 95% confidence interval for each mean.

Based on absolute differences between means, *Figure 2* suggests a main effect of fraud status, but it also suggests a potential main effect of author number independent of fraud. To

assess this, we conducted a follow up analysis comparing single-author papers to multi-author papers on linguistic obfuscation. Although QQ-plots and Shapiro-Wilk tests indicated that we should accept the assumption that single-author papers (W = .98, n = 44, p = .55) and multi-author papers (W = .98, n = 44, p = .65) were normally distributed for linguistic obfuscation, Levene's test indicated that we should reject the assumption of homogeneity of variances (F(1, 86) = 6.15, p = .02). Thus, we conducted a Welch's t-test to compare the mean levels of linguistic obfuscation between the groups. *Figure 3* shows the means for each group, with error bars representing 95% confidence intervals. Results of the Welch's t-test indicated that, while multi-author papers (M = .37) appear to have higher linguistic obfuscation than single-author papers (M = -.37), there was not a significant difference between the groups (t(71.48) = -1.74, p = .09).

Figure 3

Average linguistic obfuscation for single- and multi-author paper groups



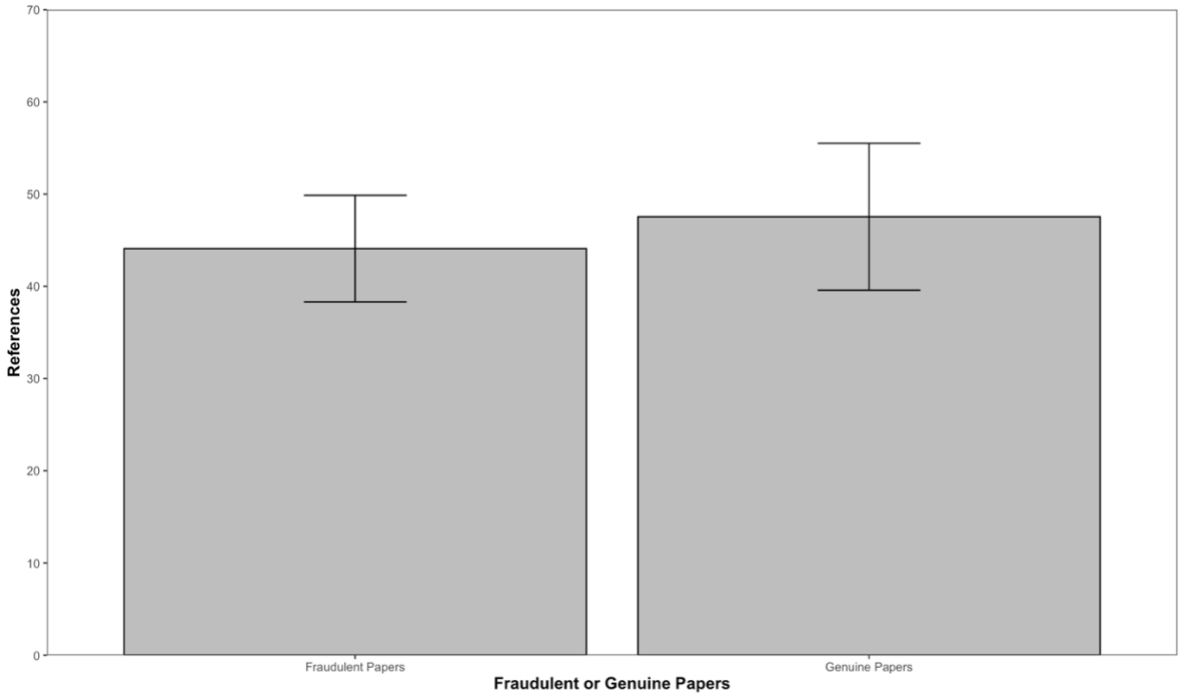
Note. Error bars indicate the 95% confidence interval for each mean.

Hypothesis 2

Hypothesis 2a was tested by comparing the number of references between fraudulent and genuine papers. Investigation of QQ-plots and the results of a Shapiro-Wilk normality test for each group indicated that the assumption of normality was violated for the genuine paper group ($W = 0.86$, $p < .01$, $n = 44$) and the fraudulent paper groups ($W = .92$, $p < .01$, $n = 44$). Thus, we conducted a Mann-Whitney U test to compare the rank sum of references between the groups, but *Figure 6* also shows the mean number of references in both the fraudulent and genuine paper groups with 95% confidence intervals. In contrast to our hypothesis, genuine papers ($M = 47.55$) had higher mean references than fraudulent papers ($M = 44.09$) in absolute terms, and results of the Mann-Whitney U test indicated that the number of references did not differ between the two groups ($W = 934.5$, p = .78).

Figure 6

Average number of references for genuine and fraudulent paper groups



Note. Error bars indicate the 95% confidence interval for each mean.

Due to the possibility that our results for Hypothesis 2a could have been confounded by word count, we conducted an unplanned follow-up analysis using the ratio of references to word count as the dependent variable. QQ-plots and Shapiro-Wilk tests indicated that for both the fraudulent group (W = .95, p = .048, n = 44) and the genuine group (W = .91, p < .01, n = 44) the assumption of normality for the ratio of references to words should be rejected. Thus, we conducted an independent samples t-test with bootstrapped confidence intervals to compare the references to word count ratio between the fraudulent and genuine paper groups. Results showed that the fraudulent paper group (M = .01) and genuine paper group (M = .01) did not differ significantly, as the bootstrapped confidence interval contained zero (t(86) = -.38, 95% BCI = -2.20 to 1.66, p = .70, Cohen's d = -.08).

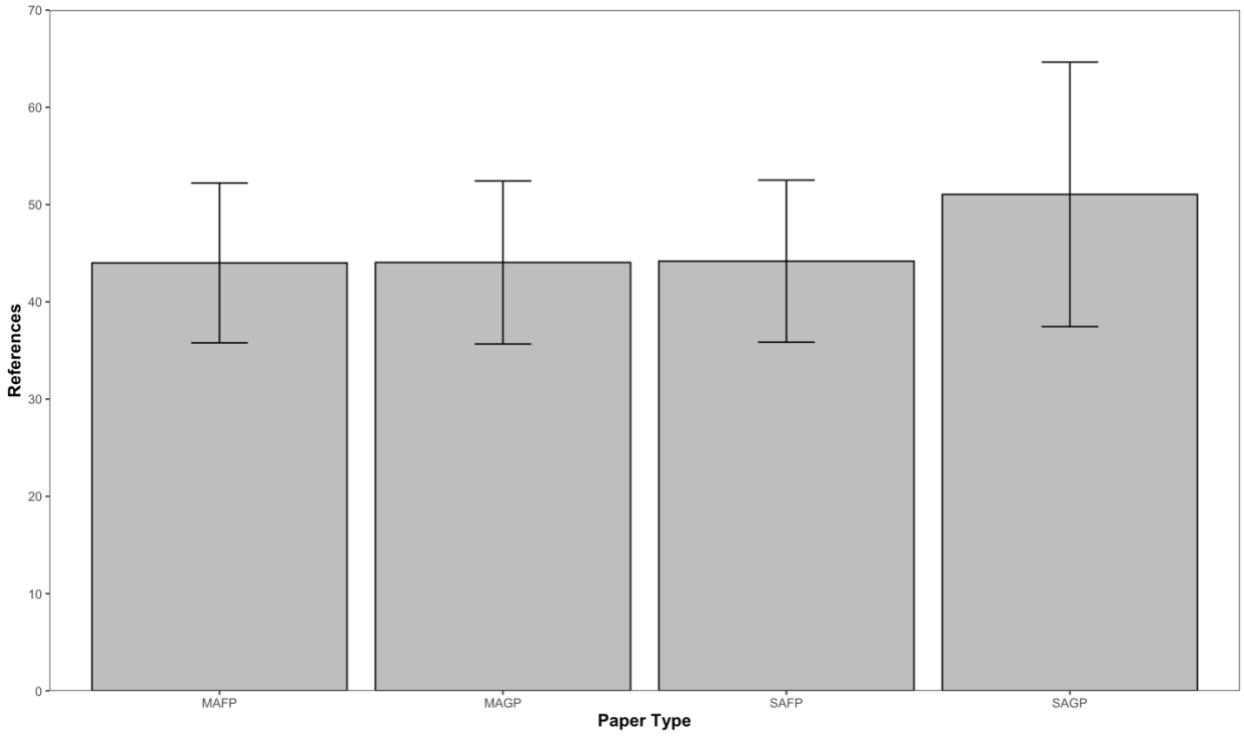
Hypothesis 2b was tested using a Kruskal-Wallis rank sum test to compare the number of references between the SAFP, SAGP, MAFP, and MAGP groups to ensure a robust method for count data after QQ-plots and Shapiro-Wilk tests indicated that for the SAGP (W = .86, p < .01, n = 22), SAFP (W = .89, p = .02, n = 22), and MAGP (W = .90, p = .04, n = 22) groups we should not assume normality of variance. Results showed there were no differences between the rank sums of each group (H(3) = .50, p = .92). *Figure 6* presents the means for references between each of the four groups, showing that—in contrast to our hypothesis—each group is roughly equal for average references with SAGP having higher references (although not statistically significant based on rank-sums).

Similar to Hypothesis 2a we conducted an unplanned follow-up analysis to reassess Hypothesis 2b using the ratio of references to word count as the dependent variable, thus eliminating the potential confounding effect of word count. QQ-plots and Shapiro-Wilk tests indicated that the assumption of normality should be rejected for SAGP (W = .91, p = .047, n =

22). Thus, we computed bootstrapped confidence intervals for our *F* statistic to assess the significance of our one-way ANOVA with the ratio of references to word count as the dependent variable. Results showed that there were no differences between the groups (*M* = .01 for all

Figure 6

Average number of references for MAFP, MAGP, SAFP, and SAGP groups



Note. Error bars indicate the 95% confidence interval for each mean.

groups), as the bootstrapped confidence interval for our *F* statistic contained 1 (*F*(3, 84) = .60, 95% *BCI* = .01 to 1.60, *p* = .618).

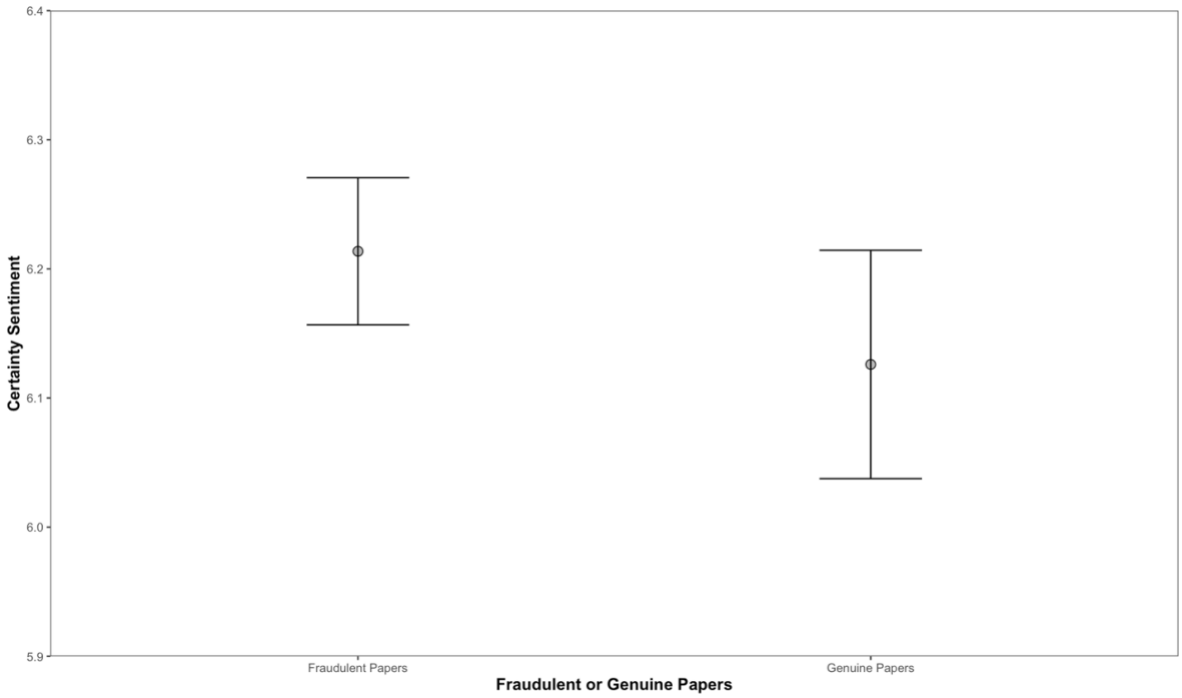
Hypothesis 3

Hypothesis 3 implemented Welch's t-test to test for differences between fraudulent and genuine papers on certainty sentiment because a Levene's test indicated that we should reject the assumption of homogeneity of variance (*F*(1, 86)= 4.02, *p* = .048). *Figure 6* shows that, contrary to our hypothesis, mean levels of certainty sentiment for fraudulent papers (*M* = 6.21) were

higher than genuine papers (M = 6.13), although this result was not significant ($t(76.47) = 1.63$, p = .11, CI = -.02 to .19, Cohen's d = .35).

Figure 6

Average certainty sentiment for fraudulent and genuine paper groups



Note. Error bars indicate the 95% confidence interval for each mean.

Hypothesis 4

Hypothesis 4 was tested with a binomial test to compare the observed likelihood that the fraudulent author is the corresponding author to the expected probability based on the inverse of the mean number of authors for MAFP (M = 3.125; $1/M$ = .32). In our sample, determining whether the fraudulent author was the corresponding author was only possible in n = 8 cases within the MAFP group. In n = 4 of these cases the fraudulent author was the corresponding author, so our observed likelihood was .5. The binomial test indicated that fraudulent authors were not significantly more likely to be corresponding authors than by chance (p = .28).

Discussion

Despite an abundance of null findings, the present study provides novel hypotheses and an improved methodological framework for future work on the psychology of scientific fraud. By leveraging the only source of data that is ubiquitously produced in cases of scientific fraud—the text of the scientific papers—previous research has found that linguistic features differ between fraudulent and genuine scientific papers. However, these findings relied on datasets that did not differentiate between single-author and multi-author cases of fraud and did not require that a fraudulent author wrote the manuscript. By analyzing linguistic differences between teams of researchers without these distinctions, the psychological inferences from such results remain tenuous and preliminary.

To address this issue, we compared fraudulent and genuine papers with single-authored and multi-authored scientific papers. This method affords stronger psychological inferences for effects found in single-author papers, but it also allows us to make a preliminary assessment of author number as a confound of linguistic effects. That is, because fraudulent research tends to have more authors (Steen, 2011b), a higher number of authors may underly linguistic effects of fraud. Lastly, differentiating between single-author and multi-author papers allowed us to test for the effect of working within a research group on linguistic effects, as we hypothesized that the very goal-oriented behavior of not getting caught by the *research group* may underly linguistic obfuscation and references effects suggested by previous research.

In addition to taking author number into account, our study implemented a number of other methodological improvements. By exporting all papers presumed to be genuine during our search to Zotero with the Retraction Watch Integration installed, we checked each paper for whether it was included in the largest existing database of retracted publications available. This

allowed us to validate our categorization of papers as genuine or fraudulent. Next, by approximately matching the groups on a comprehensive list of criteria that have been demonstrated to differ between fraudulent and genuine papers—as well as others that may be expected *a priori*—we minimized the potential for confounding from these threats to internal validity. Indeed, the meta-data for our sample of papers aligns with previous research suggesting that fraudulent research is mostly produced by men (Sebo et al., 2023) at lower-prestige research institutes (Lievore et al., 2021). Finally, we implemented more precise measures. Specifically, by counting references directly instead of using word count of references sections as a proxy, we ensured the accuracy of these counts, and by using the certainty sentiment measure developed by Rocklage et al. (2023), we ensured greater validity, reliability, and specificity than previous measures of certainty. Their measure is also more appropriate for measurement of linguistic features in scientific publications because they took a data driven approach to its formulation, included the BioScope corpus in the original candidate *n*-gram list, and provided the opportunity for imputation of scores based on many raters' judgements.

We were also limited by some methodological issues. First, our sample size was extremely limited compared to prior studies of a similar nature. This was due to the limiting factor of single-author fraudulent papers in the RWDB—only $n = 22$ were available at the time when we downloaded the data. Thus, our sample size was limited to $N = 88$ papers, and our power to detect our observed effect sizes was severely limited—as we will review below. Next, although we accounted for a plethora of empirically and theoretically determined matching criteria to avoid their confounding, we were unable to make these matches across all papers. Principally, for half of our single-author papers we were unable to match by journal, instead electing to make up for this with the Subject(s) meta-data in the RWDB. Although the vast

majority of features were able to be matched, our actual matching outcomes were certainly less than optimal.

Despite a lack of statistical reliability, our investigation brings novel preliminary insights to the linguistic obfuscation hypothesis of scientific fraud. Keeping in mind that these findings were not statistically significant, our results suggest three potential effects. First, in our sample fraudulent papers had slightly higher linguistic obfuscation than genuine papers (see *Figure 1*), aligning with the findings of Markowitz and Hancock (2016). This suggests a main effect of fraud and that authors attempt to make their work less accessible in order to avoid fraud detection. Second, in our follow-up analysis looking for a difference between single-author and multi-author papers (see *Figure 3*), the multi-author papers in our sample have slightly higher levels of linguistic obfuscation, suggesting a main effect of the presence of collaborators. This could simply be due to the process of combining heterogenous ideas from different people during the writing process, which may inadvertently increase the difficulty of understanding the text independent of the effect of fraud. If fewer people are writing a manuscript than people are contributing content, the ideas may be represented inefficiently and in a less clear way due to a partial understanding by the writer in comparison to the originator of the idea. Third, as shown in *Figure 2*, the highest levels of linguistic obfuscation were found in multi-author fraudulent papers, and the lowest levels of linguistic obfuscation were found in single-authored genuine papers. This suggests that there may be an interaction effect between the presence of research collaborators and fraud, consistent with our Hypothesis 2b. That is, fraudulent authors may obfuscate their communication more due to the presence of research collaborators because collaborators are salient in the mind of fraudsters as potential whistleblowers. Thus, future work should investigate whether fraud effects on linguistic obfuscation are confounded by the number

of authors on a paper or, in accordance with our hypothesis, are moderated by the presence of a research group working on the project.

These three possibilities regarding linguistic obfuscation in fraudulent science should be considered in future work. The two main effects which we suggest—of fraud and of the presence of collaborators—should be verified and disentangled. By considering and measuring the presence of research collaborators, future work can determine whether the presence of collaborators confounds effects of fraud on linguistic obfuscation, independently effects linguistic obfuscation, or moderates the relationship between fraud and linguistic obfuscation.

While our linguistic obfuscation results provide fruitful avenues for future investigation, they should only be taken as suggestions for future hypotheses. This is because there are severe limitations to our analysis here. First, in our sample neither the linguistic obfuscation index or the abstraction index showed internal consistency, in contrast to the results of its developers, Markowitz and Hancock (2016). Although this may be due to stochastic variability, given our comparatively small sample, it calls into question any inferences that might be made based on our results. In addition, this investigation was underpowered to detect the observed effects. Indeed, we used the pwr package (Champely et al., 2020) to conduct *post-hoc* analyses for our parametric hypothesis tests given an alpha level of .05 and our sample size. For the t-test of Hypothesis 1a, with we had only a 31.7% chance of detecting a statistically significant effect with our observed Cohen's *d* of .32. Similarly, for our one-way ANOVA of Hypothesis 2a, we had only a 50% chance of detecting a statistically significant effect with our calculated Cohen's *f* of .26, and our follow up analysis on the effect of the presence of research collaborators we had a 40% chance of detecting a statistically significant effect with our Cohen's *d* of .37. Future work interested in the possibility that the presence of collaborators is an important factor in the

psychology of scientific fraud will need to either wait for more data in the RWDB or sample directly from the scientific literature, although the latter option may be quite difficult in practice given the arduous and unreliable process of detecting scientific fraud.

Regarding reference use as an analogous measure to use of third-person pronouns, we did not find that fraudulent papers used more references (Hypothesis 2a), and, in accordance with analysis of Deidrick Stapel's corpus of papers (Markowitz & Hancock, 2014), we actually found a higher mean number of references in our genuine papers, although this result was not significant. Regarding Hypothesis 2b, which held that references should be particularly high in the multi-author fraudulent paper group, there was a striking similarity between the means of all paper groups except for single-author genuine papers, which had a much higher variance as well due to an outlier with 146 references (although this should not have affected our statistical test, as we compared rank sums rather than means). These null findings were consistent despite taking word count into account as a potential confounder in follow-up analyses. In all, this suggest that fraudulent scientists do not use more references as a way of diverting attention away from their own work or increasing the cost of assessing its credibility, as Schmidt (2022) also concludes. However, Schmidt (2022) was unable to count references directly from papers and used word count of reference sections instead, introducing the possibility of measurement error. Ultimately, our null finding too should be taken as preliminary, due to our low power, and in future work, references should be counted directly to ensure accuracy.

We also predicted that fraudulent papers would show lower levels of linguistic certainty based on Dehdarirad and Schirone's (2023) notion that uncertain language may mask deception and fraud through making the work less comprehensible, similar to the proposed function of linguistic obfuscation. In contrast to their findings, however, the fraudulent papers included more

certainty language, although this result was not statistically significant. *Post-hoc* power analysis revealed that for our Welch's t-test we had only a 37% chance of detecting a statistically significant effect for our observed Cohen's *d* of .35. Although this is the largest effect size in our study, it is clear that the test was underpowered. However, it is possible that the opposite direction of our effect, compared to Dehdarirad and Schirone's (2023), reflects a true effect as measured with the certainty sentiment index from the lexical suite (Rocklage et al., 2023). To test this more thoroughly, future work should investigate the certainty hypothesis using this more valid, reliable, and precise measure of certainty language.

Finally, the hypothesis that fraudulent authors strategically attempt to control information flow stems directly from the fact that regulating communication between those in their research group and those out of their research group would allow one to deflect suspicious inquiries. Based on this prediction, we tested whether fraudulent authors were more likely than their coauthors to be the corresponding author on their fraudulent papers. While in the raw data the proportion of fraudulent authors that were corresponding authors was higher than what one would expect *a priori*, this effect was not statistically reliable due to so few cases in our sample where the fraudulent author and corresponding author were both known. Our power analysis indicated that given our proportions to compare and our total number of cases we only had an 11% chance of detecting a statistically significant effect. Therefore, this hypothesis warrants further consideration and more robust statistical testing to rule out.

In sum, the present study of the psychology of scientific fraud is consistent with understandings of fraud as a deliberate form of goal-oriented behavior (Markowitz and Hancock, 2016), but our limitation of sample size due to availability of single-authored papers left us with

insufficient power to test our hypotheses rigorously. Nevertheless, we have outlined multiple hypotheses for future work, and these should be tested when more data becomes available.

References

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F., & Pen, M. (2015). Estimating the reproducibility of psychological science. *Science, 349*. https://doi.org/10.1126/science.aac4716

Aczel, B. (2023). Why we need a 'psychology of science.' *Nature Human Behaviour*, 1–2. https://doi.org/10.1038/s41562-023-01786-4

Adobe Acrobat (2023.008.20555). (2023). [Computer software]. Adobe Systems Incorporated. https://www.adobe.com

Aggarwal, S. B., Chaitanya. (2022). *textstat: Calculate statistical features from text* (0.7.3) [Python]. https://github.com/shivam5992/textstat

Biagioli, M., Kenney, M., Martin, B. R., & Walsh, J. P. (2019). Academic misconduct, misrepresentation and gaming: A reassessment. *Research Policy, 48*(2), 401–413. https://doi.org/10.1016/j.respol.2018.10.025

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 1–47.

Brown, Z. C., Anicich, E. M., & Galinsky, A. D. (2020). Compensatory conspicuous communication: Low status increases jargon use. *Organizational Behavior and Human Decision Processes*, *161*, 274–290. https://doi.org/10.1016/j.obhdp.2020.07.001

Candal-Pedreira, C., Ruano-Ravina, A., Fernández, E., Ramos, J., Campos-Varela, I., & Pérez-Ríos, M. (2020). Does retraction after misconduct have an impact on citations? A pre–

post study. *BMJ Global Health*, *5*(11), e003719. https://doi.org/10.1136/bmjgh-2020-003719

Caron, M. M., Dohan, S. B., Barnes, M., & Bierer, B. E. (2023). Defining "recklessness" in research misconduct proceedings. *Accountability in Research*, *0*(0), 1–23. https://doi.org/10.1080/08989621.2023.2256650

Center for History and New Media. (2023). *Zotero* [Computer software]. https://www.zotero.org/

Chambers, L. M., Michener, C. M., & Falcone, T. (2019). Plagiarism and data falsification are the most common reasons for retracted publications in obstetrics and gynaecology. *BJOG: An International Journal of Obstetrics and Gynaecology*, *126*(9), 1134–1140. https://doi.org/10.1111/1471-0528.15689

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). *pwr: Basic Functions for Power Analysis* (1.3-0) [Computer software]. https://cran.r-project.org/web/packages/pwr/index.html

Conrad, M., & Holtbrügge, D. (2021). Antecedents of corporate misconduct: A linguistic content analysis of decoupling tendencies in sustainability reporting. *Business Ethics, the Environment & Responsibility*, *30*(4), 538–550. https://doi.org/10.1111/beer.12361

Courtis, J. K. (1998). Annual report readability variability: Tests of the obfuscation hypothesis. *Accounting, Auditing & Accountability Journal*, *11*(4), 459–472. https://doi.org/10.1108/09513579810231457

Craig, R., Cox, A., Tourish, D., & Thorpe, A. (2020). Using retracted journal articles in psychology to understand research misconduct in the social sciences: What is to be done? *Research Policy*, *49*(4), 103930. https://doi.org/10.1016/j.respol.2020.103930

Crilly, D., Hansen, M., & Zollo, M. (2016). The Grammar of Decoupling: A Cognitive-Linguistic Perspective on Firms' Sustainability Claims and Stakeholders' Interpretation. *Academy of Management Journal*, *59*(2), 705–729. https://doi.org/10.5465/amj.2015.0171

Davis, M. S., Riske-Morris, M., & Diaz, S. R. (2007). Causal Factors Implicated in Research Misconduct: Evidence from ORI Case Files. *Science and Engineering Ethics*, *13*(4), 395–414. https://doi.org/10.1007/s11948-007-9045-2

Dehdarirad, T., & Schirone, M. (2023). Use of positive terms and certainty language in retracted and non-retracted articles: The case of biochemistry. *Journal of Information Science*, 01655515231176650. https://doi.org/10.1177/01655515231176650

Dinh, L., Sarol, J., Cheng, Y.-Y., Hsiao, T.-K., Parulian, N., & Schneider, J. (2019). Systematic examination of pre- and post-retraction citations. *Proceedings of the Association for Information Science and Technology*, *56*(1), 390–394. https://doi.org/10.1002/pra2.35

DuBois, J. M., Anderson, E. E., Chibnall, J., Carroll, K., Gibb, T., Ogbuka, C., & Rubbelke, T. (2013). Understanding Research Misconduct: A Comparative Analysis of 120 Cases of Professional Wrongdoing. *Accountability in Research*, *20*(5–6), 320–338. https://doi.org/10.1080/08989621.2013.822248

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, *4*(5), e5738. https://doi.org/10.1371/journal.pone.0005738

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences*, *109*(42), 17028–17033. https://doi.org/10.1073/pnas.1212247109

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

Furman, J. L., Jensen, K., & Murray, F. (2012). Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine. *Research Policy*, *41*(2), 276–290. https://doi.org/10.1016/j.respol.2011.11.001

Gammon, E., & Franzini, L. (2013). Research misconduct oversight: Defining case costs. *Journal of Health Care Finance*, *40*(2), 75–99.

Goel, S., Gangolly, J., & Faerman, S. (2010). Can Linguistic Predictors Detect Fraudulent Financial Filings? *Journal of Emerging Technologies in Accounting*, *7*. https://doi.org/10.2308/jeta.2010.7.1.25

Government of Canada. (2022). *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans–TCPS 2 (2018)–Chapter 2: Scope and Approach*.

Grieneisen, M. L., & Zhang, M. (2012). A Comprehensive Survey of Retracted Articles from the Scholarly Literature. *PLoS ONE*, *7*(10), e44118. https://doi.org/10.1371/journal.pone.0044118

Hancock, J., Curry, L., Goorha, S., & Woodworth, M. (2008). On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes*, *45*, 1–23. https://doi.org/10.1080/01638530701739181

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2014). Are Computers Effective Lie Detectors? A Meta-Analysis of Linguistic Cues to Deception. *Personality and Social Psychology Review*. https://doi.org/10.1177/1088868314556539

Hendricks, G., Lammey, R., Ofiesh, L., Bilder, G., & Pentz, E. (2023, September 12). *News: Crossref and Retraction Watch* [Website]. Crossref. https://www.crossref.org/blog/news-crossref-and-retraction-watch/

Horton, J., Krishna Kumar, D., & Wood, A. (2020). Detecting academic fraud using Benford law: The case of Professor James Hunton. *Research Policy*, *49*(8), 104084. https://doi.org/10.1016/j.respol.2020.104084

Houdek, P. (2017). Professional Identity and Dishonest Behavior. *Society*, *54*(3), 253–260. https://doi.org/10.1007/s12115-017-0132-y

Humpherys, S. L., Moffitt, K. C., Burns, M. B., Burgoon, J. K., & Felix, W. F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, *50*(3), 585–594. https://doi.org/10.1016/j.dss.2010.08.009

Hussinger, K., & Pellens, M. (2019). Guilt by association: How scientific misconduct harms prior collaborators. *Research Policy*, *48*(2), 516–530. https://doi.org/10.1016/j.respol.2018.01.012

Jupyter Notebook (6.5.4). (2023). [Computer software]. https://jupyter.org

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2), 221–247. https://doi.org/10.1016/j.jacceco.2008.02.003

Li, W., & Mol, B. W. (2019). Re: Plagiarism and data falsification are the most common reasons for retracted publications in obstetrics and gynaecology. *BJOG: An International Journal of Obstetrics & Gynaecology*, *126*(10), 1289–1289. https://doi.org/10.1111/1471-0528.15829

Lievore, C., Rubbo, P., dos Santos, C. B., Picinin, C. T., & Pilatti, L. A. (2021). Research ethics: A profile of retractions from world class universities. *Scientometrics, 126*(8), 6871–6889. https://doi.org/10.1007/s11192-021-03987-y

Linsley, P., & Lawrence, M. (2007). Risk Reporting by the Largest UK Companies: Readability and Lack of Obfuscation. *Accounting, Auditing & Accountability Journal*, *20*, 620–627. https://doi.org/10.1108/09513570710762601

Loadsman, J. A. (2019). Why does retraction take so much longer than publication? *Anaesthesia*, *74*(1), 3–5. https://doi.org/10.1111/anae.14484

Malkov, D., Yaqub, O., & Siepel, J. (2023). The spread of retracted research into policy literature. *Quantitative Science Studies*, *4*(1), 68–90. https://doi.org/10.1162/qss_a_00243

Markowitz, D. M., & Hancock, J. T. (2014). Linguistic Traces of a Scientific Fraud: The Case of Diederik Stapel. *PLoS ONE*, *9*(8), 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. https://doi.org/10.1177/0261927X15614605

Mongeon, P., & Larivière, V. (2016). Costly collaborations: The impact of scientific fraud on co-authors' careers. *Journal of the Association for Information Science and Technology*, *67*(3), 535–542. https://doi.org/10.1002/asi.23421

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Newman, M., Groom, C., Handelman, L., & Pennebaker, J. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes - DISCOURSE PROCESS*, *45*, 211–236. https://doi.org/10.1080/01638530802073712

OpenAI. (2023). *ChatGPT* [GPT-4]. OpenAI. https://chat.openai.com

Palus, A. S. (2015, December 8). Diederik Stapel now has 58 retractions. *Retraction Watch*. https://retractionwatch.com/2015/12/08/diederik-stapel-now-has-58-retractions/

Pfeifer, M. P., & Snodgrass, G. L. (1990). The continued use of retracted, invalid scientific literature. *JAMA, 263*(10), 1420–1423.

Pickett, J. T., & Roche, S. P. (2018). Questionable, Objectionable or Criminal? Public Opinion on Data Fraud and Selective Reporting in Science. *Science and Engineering Ethics*, *24*(1), 151–171. https://doi.org/10.1007/s11948-017-9886-2

Python (3.11.5). (2023). [Computer software]. Python Software Foundation. https://www.python.org/

Quarto (1.3.450). (2023). [Computer software]. Posit. https://quarto.org

R: A language and environment for statistical computing. (2023). [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Retraction Watch Database User Guide Appendix A: Fields. (2018, October 23). *Retraction Watch*. https://retractionwatch.com/retraction-watch-database-user-guide/retraction-watch-database-user-guide-appendix-a-fields/

Rocklage, M. D., He, S., Rucker, D. D., & Nordgren, L. F. (2023). Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language. *Journal of Marketing Research*, *60*(5), 870–888. https://doi.org/10.1177/00222437221134802

RStudio Team. (2023). *RStudio: Integrated Development Environment for R.* [R]. Posit. http://www.rstudio.com/

Schmidt, E. (2022). *Can linguistic features unmask fraudulent research? A study that builds an NLP classifier to distinguish retracted papers from non-retracted papers based on text*

and linguistic features. [Master Thesis]. https://studenttheses.uu.nl/handle/20.500.12932/42411

Sebo, P., Schwarz, J., Achtari, M., & Clair, C. (2023). Women Are Underrepresented Among Authors of Retracted Publications: Retrospective Study of 134 Medical Journals. *Journal of Medical Internet Research*, *25*(1), e48529. https://doi.org/10.2196/48529

Shamoo, A. E. (2013). Data Audit as a Way to Prevent/Contain Misconduct. *Accountability in Research-Policies and Quality Assurance*, *20*(5–6), 369–379. https://doi.org/10.1080/08989621.2013.822259

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sotudeh, H., Barahmand, N., Yousefi, Z., & Yaghtin, M. (2022). How do academia and society react to erroneous or deceitful claims? The case of retracted articles' recognition. *Journal of Information Science*, *48*(2), 182–198. https://doi.org/10.1177/0165551520945853

Sovacool, B. K. (2005). Using Criminalization and Due Process to Reduce Scientific Misconduct. *The American Journal of Bioethics*, *5*(5), W1–W7. https://doi.org/10.1080/15265160500313242

Sovacool, B. K. (2008). Exploring Scientific Misconduct: Isolated Individuals, Impure Institutions, or an Inevitable Idiom of Modern Science? *Journal of Bioethical Inquiry*, *5*(4), 271–282. https://doi.org/10.1007/s11673-008-9113-6

Steen, R. G. (2011a). Retractions in the medical literature: How many patients are put at risk by flawed research? *Journal of Medical Ethics*, *37*(11), 688–692. https://doi.org/10.1136/jme.2011.043133

Steen, R. G. (2011b). Retractions in the scientific literature: Do authors deliberately commit research fraud? *Journal of Medical Ethics*, *37*(2), 113–117.

Stern, A. M., Casadevall, A., Steen, R. G., & Fang, F. C. (2014). Financial costs and personal consequences of research misconduct resulting in retracted publications. *eLife*, *3*, e02956. https://doi.org/10.7554/eLife.02956

Stillman, D. (2019, June 14). *Retracted item notifications with Retraction Watch integration*. https://www.zotero.org/blog/retracted-item-notifications/

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, *7*(6), 670–688. https://doi.org/10.1177/1745691612460687

Szilagyi, I.-S., Schittek, G. A., Klivinyi, C., Simonis, H., Ulrich, T., & Bornemann-Cimenti, H. (2022). Citation of retracted research: A case-controlled, ten-year follow-up scientometric analysis of Scott S. Reuben's malpractice. *Scientometrics*, *127*(5), 2611–2620. https://doi.org/10.1007/s11192-022-04321-w

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

The Retraction Watch Database. (2024). [dataset]. http://retractiondatabase.org/

The Retraction Watch Leaderboard. (2015, June 16). *Retraction Watch*. https://retractionwatch.com/the-retraction-watch-leaderboard/

Times Higher Education (THE) World University Rankings. (2024). https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking#!/length/-1/sort_by/rank/sort_order/asc/cols/stats

Walsh, J. P., Lee, Y.-N., & Tang, L. (2019). Pathogenic organization in science: Division of labor and retractions. *Research Policy*, *48*(2), 444–461. https://doi.org/10.1016/j.respol.2018.09.004

Wang, P., & Su, J. (2022). Expert-recommended biomedical journal articles: Their retractions or corrections, and post-retraction citing. *Journal of Information Science*, 01655515221074329. https://doi.org/10.1177/01655515221074329

Wilson, P. (2020). Academic Fraud: Solving the Crisis in Modern Academia. *Exchanges: The Interdisciplinary Research Journal*, *7*(3), Article 3. https://doi.org/10.31273/eirj.v7i3.546

Zubaly, B., & Arbeau, K. (2023). *Psychology of Scientific Fraud Data Analysis Plan*. https://benjaminjzubaly.github.io/The-Psychology-of-Scientific-Fraud-Thesis/

Zuckerman, H. (2020). Is “the time ripe” for quantitative research on misconduct in science? *Quantitative Science Studies*, *1*(3), 945–958. https://doi.org/10.1162/qss_a_00065