Review Article

Scientific Misconduct in Psychology

A Systematic Review of Prevalence Estimates and New Empirical Data

Johannes Stricker and Armin Günther

Leibniz Institute for Psychology Information, Trier, Germany

Abstract: Spectacular cases of scientific misconduct have contributed to concerns about the validity of published results in psychology. In our systematic review, we identified 16 studies reporting prevalence estimates of scientific misconduct and questionable research practices (QRPs) in psychological research. Estimates from these studies varied due to differences in methods and scope. Unlike other disciplines, there was no reliable lower bound prevalence estimate of scientific misconduct based on identified cases available for psychology. Thus, we conducted an additional empirical investigation on the basis of retractions in the database PsycINFO. Our analyses showed that 0.82 per 10,000 journal articles in psychology were retracted due to scientific misconduct. Between the late 1990s and 2012, there was a steep increase. Articles retracted due to scientific misconduct were identified in 20 out of 22 PsycINFO subfields. These results show that measures aiming to reduce scientific misconduct should be promoted equally across all psychological subfields.

Keywords: scientific misconduct, research practices, research integrity, article retractions

Cases of scientific misconduct undermine the credibility of published results and ultimately reduce the confidence in the value of scientific research as a whole (Fang, Steen, & Casadevall, 2012). The detection of some spectacular cases of scientific misconduct (e.g., the case of Diederik Stapel; Callaway, 2011) has contributed to concerns over the validity of published results in psychology, especially in social psychology (e.g., see Rovenpor & Gonzales, 2015). For instance, Carey (2011), referring to expert evaluation, stated in a New York Times article that "the [Stapel] case exposes deep flaws in the way science is done in a field, psychology, that has only recently earned a fragile respectability". Similarly, some psychological researchers themselves seem to be unsettled about the credibility of their field. For example, Motyl et al. (2017, p. 10) found that their sample of social and personality psychology researchers had the impression that "the field overall might be pretty rotten".

Scientific misconduct includes data fabrication, data falsification, plagiarism, and other serious and intentional practices that distort scientific results or lead to incorrect information about contributions to research (e.g., undisclosed competing interests; Hofmann, Helgesson, Juth, & Holm, 2015; Resnik, Neal, Raymond, & Kissling, 2015). Honest errors or differences of opinion do not qualify as scientific misconduct (Office of Research Integrity, 2011; Office of Science and Technology Policy, 2000). Besides the negative effect on the credibility of scientific research, there is a large number of additional adverse effects of scientific misconduct. These negative consequences include the misplacement of monetary investments (e.g., grant funding) and research capacity, misinformation of the public and policy makers, damage of the careers of colleagues and graduate students unknowingly involved in fraudulent projects, the delay of scientific progress, and costs associated with the investigation of misconduct cases (Michalek, Hutson, Wicher, & Trump, 2010; Stroebe, Postmes, & Spears, 2012).

While the toxic consequences of scientific misconduct are indisputable, the prevalence of these practices has been subject to debate (Gross, 2016; Marshall, 2000). This question is particularly relevant because reliable data on the occurrence of a phenomenon are crucial to understanding its causes and to developing prevention strategies. Many factors contributing to the engagement in scientific misconduct have been discussed. Those include the academic "publish-or-perish" culture (e.g., De Rond & Miller, 2005) and academic capitalism (Münch, 2014) leading to competitive and individualist norms (Louis, Anderson, & Rosenberg, 1995; Motyl et al., 2017). Many researchers experience significant pressure to publish significant and preferably surprising results in high-ranking journals to achieve tenure or promotion (Nosek, Spies, & Motyl, 2012), job security or financial rewards (Franzoni, Scellato, & Stephan, 2011). There is some evidence that this pressure has increased in the last decades (e.g., Anderson, Ronning, De Vries, & Martinson, 2007).

Quantification of Scientific Misconduct

Three different approaches have been used to estimate the prevalence of scientific misconduct:

(1) In survey studies, researchers anonymously indicate their involvement in scientific misconduct or estimate the involvement of their colleagues. A meta-analysis of survey studies (Fanelli, 2009) showed that a pooled weighted average of 1.97% of scientists from all scientific fields have admitted to have participated in fabricating, falsifying, or modifying data. 14.12% reported that they believed that their colleagues were involved in such practices. Survey studies on the prevalence of scientific misconduct have been criticized for providing varying estimates due to differences in item wording, survey distribution method, social desirability and other factors (Fanelli, 2009; Fiedler & Schwarz, 2016).

(2) Through statistical (re)analyses of reported findings, some researchers attempt to identify statistical inconsistencies in published studies (e.g., inconsistencies between a reported *p* value and its test statistics) indicating scientific misconduct or questionable research practices (QRPs; e.g., inappropriately "rounding down" *p* values just over .05; e.g., Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Yet, a considerable proportion of statistical inconsistencies may be a result of inadvertent honest errors rather than scientific misconduct (Bakker & Wicherts, 2011). Thus, studies based on statistical (re)analyses might strongly overestimate the prevalence of scientific misconduct.

(3) The analysis of retracted articles and retraction notices has recently emerged as a main format for investigating scientific misconduct (for a review, see Hesselmann, Graf, Schmidt, & Reinhart, 2017). Analyses investigating scientific misconduct via retracted articles are mostly based on cases that after thorough investigations have been judged to be guilty of scientific misconduct. Yet, as it is often difficult to detect scientific misconduct (Stroebe et al., 2012), this approach provides an estimate only for the lower bound of the prevalence of scientific misconduct. Also, estimates derived from this approach are influenced by the quality of the monitoring systems implemented to detect scientific misconduct.

Taken together, all three approaches in the quantification of scientific misconduct possess unique strengths and weaknesses in their ability to investigate the prevalence, distribution, and development of scientific misconduct. Thus, findings from all three approaches should be integrated in a field addressing scientific misconduct.

The Present Study

The aim of the present study was to examine the prevalence and the development of scientific misconduct in psychology and its subfields. First, we conducted a systematic review of articles reporting quantitative prevalence estimates of scientific misconduct in psychology. Another concept that is linked to concerns about the validity of published psychological research are QRPs (e.g., Świątkowski & Dompnier, 2017). QRPs comprise practices that unambiguously qualify as scientific misconduct (e.g., falsifying data) and others that are less clear (e.g., failing to report all of a study's dependent measures; John, Loewenstein, & Prelec, 2012; Motyl et al., 2017; Stürmer, Oeberst, Trötschel, & Decker, 2017). Thus, there is some degree of overlap between scientific misconduct and some of the behaviors subsumed under the term "QRPs". Consequently, we also included prevalence estimates of QRPs in our review.

Second, we analyzed new empirical data on the prevalence and development of retractions due to scientific misconduct in psychology accounting for subfields of psychology, their size, and the number of unique authors responsible for scientific misconduct. A preliminary version of our data set was reported by Margraf (2015). This work did not take into account the retraction reasons (misconduct or not), nor psychological subfields or responsible authors. Our data, scripts for data analysis, and materials (for the systematic review and the empirical study of article retractions) are accessible via the PsychArchives repository https://doi.org/10.23668/psycharchives.872

Method

Systematic Review

We searched the databases PsycINFO and Scopus with the search-string "(prevalence OR incidence) AND ("scientific fraud" OR "research fraud" OR "scientific misconduct" OR "research misconduct" OR "scientific integrity" OR "data falsification" OR "data fabrication" OR plagiarism OR "research practices" OR "p-hacking" OR "HARKing" OR retract*)" in abstracts and titles (last update: June 2018). Results from Scopus were limited to the subject area "psychology". No other limits were set. Additionally, we conducted an exploratory literature search by entering our key words in Google Scholar and by following up references in the included studies. Our only inclusion criterion was that studies had to report quantitative prevalence estimates of scientific misconduct or QRPs in psychological research. In three studies, prevalence estimates of scientific misconduct were measured but not reported. We contacted the corresponding authors of these articles via e-mail and

received the relevant prevalence estimates from one article (Sacco, Bruton, & Brown, 2018). Studies addressing scientific misconduct in non-psychological research fields and in students (i.e., plagiarism and cheating in course work) were excluded.

Empirical Study

We used the search string "(retract*.ab. or retract*.ti.) and "01*".pt." (limit 1860–2017) to search PsycINFO for "retract*" in titles and abstracts of journal contributions (last update: January 2018). All records reporting that the respective article has been retracted or reporting the retraction of a previously published article were included in the analysis. Next, the original retraction notices were collected. Two independent raters categorized the retraction notices by reason for retraction (1. fraud, 2. plagiarism, 3. other misconduct, 4. author error, 5. publisher error, 6. other reason, 7. no explanation/justification). Categories 1, 2, and 3 were regarded as scientific misconduct. In case of scientific misconduct in multi-authored papers, responsible authors were identified based on the retraction notice. Coders were instructed to use the Retraction Watch Database (Center for Scientific Integrity, n.d.) to obtain additional information if needed. We used the articles' content classification in PsycINFO to allocate the retracted articles to the respective psychological subfield. For the calculation of the prevalence rate, we divided the number of retracted articles or responsible authors by the size of the field (i.e., the number of records with document type "Journal Article" in the respective field). In the case that there was no clear indication which author of a retracted paper was responsible for the scientific misconduct, the entire author collective was incorporated as a single responsible author in the analyses.

Results

Systematic Review

The literature search yielded 136 results from PsycINFO and 56 results from Scopus leading to 139 results after removing duplicates and retraction notices. In the first step, we evaluated the titles and abstracts. We excluded 121 articles in this step because no quantitative prevalence estimates of scientific misconduct were measured. The full text of the remaining 18 articles was examined resulting in the inclusion of four articles for the systematic review. The explorative literature search and suggestions from the review process of this article yielded 12 additional relevant articles. In the final database, there were 16 studies: six survey studies, nine studies with statistical (re)analyses and one study analyzing retracted articles. Methods and prevalence estimates of scientific misconduct and QRPs from all included studies can be found in Table 1.

Empirical Study

Searching PsycINFO for "retract*" in title and abstract yielded 2,302 records, including 402 retractions. 401 original retraction notices could be collected and were categorized for retraction reason by two independent raters. Interrater agreement (100 × (number of agreeing values/number of all coded values)) was 82.54%. Discrepancies were resolved by consulting the original retraction notice and by discussion. Of the 401 retractions, 260 (64.84%) were attributable to scientific misconduct (29.18% fraud, 26.68% plagiarism, 8.98% other misconduct). The overall retraction rate (1860–2017) due to scientific misconduct was 0.82 journal articles per 10,000 journal articles in PsycINFO. The development of retractions due to scientific misconduct since 1982 is shown in Figure 1. The rate of articles retracted due to scientific misconduct in psychological subfields can be found in Table 2.

Discussion

Systematic Review

This is the first systematic review synthesizing existing studies reporting quantitative prevalence estimates of scientific misconduct and QRPs in psychology. In survey studies, self-admission rates for data falsification ranged between 0.6% and 2.3%. Prevalence estimates for the involvement of other researcher in data falsification ranged between 9.3% and 18.7%. Self-admission rates for other QRPs that may or may not qualify as scientific misconduct such as inappropriately altering or "cooking" research data (e.g., 6%, Braun & Roussos, 2012) or "rounding down" p values just over .05 (e.g., 33%, Motyl et al., 2017) were more prevalent. There was criticism regarding the prevalence definition applied in some of the survey studies (e.g., John et al., 2012) because the percentage of researchers who admitted to have engaged in a QRP at least once was equated with the prevalence of the respective QRP (Fiedler & Schwarz, 2016). Also, the validity of researcher's estimates of their colleagues' involvement in QRPs is questionable (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Fiedler & Schwarz, 2016).

Studies reporting statistical (re)-analyses found gross inconsistencies (i.e., reported p value significant, computed p value non-significant or vice versa) in 12.4%–20.5% of the published studies. However, the proportion of studies in which inconsistencies are attributable to scientific misconduct, QRPs or honest errors remains unclear. In the only study that investigated retractions (Grieneisen & Zhang, 2012), the number of analyzed retracted articles from psychology was low (n = 32 for psychology and n = 169 for Neurosciences; numbers derived from Supplementary Material). Also, the proportion of articles in psychology that

Zeitschrift für Psychologie (2019), 227(1), 53–63

Table 1. Methods and prevalence estimates from all studies included in the systematic review

Reference	Method	Prevalence estimates of scientific misconduct	Prevalence estimates of other QRPs and potential scientific misconduct
Survey studies			
Braun and Roussos (2012)	257 psychotherapy researchers from various professional associations completed an online survey about their scientific misbehavior.	The self-admission rates were 6% for inappropriately altering or "cooking" research data (10% in Europe, 5% in North America, 3% in Latin America), 2% for making up research data (1% in Europe, 0% in North America, 4% in Latin America) and 2% for denying authorship to someone who has contributed substantively to a manuscript (2% in Europe, 0% in North America, 1% in Latin America).	The self-admission rates for other QRPs ranged from 5% for compromising the rigor of a study's design or methodology in response to pressure from a commercial funding source to 23% for conducting research involving human subjects without prior approval from an IRB or Ethics Committee.
John et al. (2012)[a]	2,155 psychological researchers from major US universities completed an online questionnaire about their involvement in 10 QRPs and the involvement of other research psychologists in these practices. For half of the participants, there was an incentive for truth-telling.	The self-admission rate for falsifying data was 0.6% (1.7% in a group with an incentive for truth-telling). The percentage of other psychologists who have falsified data was estimated at 9.33% (9.86% with an incentive for truth telling).	The self-admission rates for other QRPs ranged between 3.0% (4.5% with an incentive for truth telling) for claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do) and 63.4% (66.5% with an incentive for truth telling) for failing to report all of a study's dependent measures. The estimated percentage of other psychologists who ever engaged in other QRPs ranged between 18.72% (21.37% with an incentive for truth telling) for claiming that results are unaffected by demographic variables when one is actually unsure (or knows that they do) and 61.01% (62.70% with incentive for truth telling) for deciding whether to collect more data after looking to see whether the results were significant. The mean self-admission rates for all QRPs were 40% in Social Psychology, 37% in Cognitive Psychology, 35% in Neuroscience, 32% in Personality Psychology, 31% in Industrial Psychology, 31% in Developmental Psychology 30% in Health Psychology, 28% in Forensic Psychology, and 27% in Clinical Psychology.
Agnoli et al. (2017)	277 members if the Italian Association of Psychology (AIP) responded to an online survey about their involvement in 10 QRPs and the involvement of other Italian research psychologists in these practices using a translated version of the questionnaire by John et al. (2012).	The self-admission rate for falsifying data was 2.3%. The respondents estimated that 18.7% of the other Italian psychology researchers have falsified data at least once.	The self-admission rates for other QRPs ranged between 3.1% for claiming that results are unaffected by demographic variables when one is actually unsure (or knows that they do) and 53.2% for deciding whether to collect more data after looking to see whether the results were significant.
Motyl et al. (2017)	1,166 researchers in social and personality psychology responded to an online survey about their use of QRPs.	2% of the respondents admitted to have at least once falsified data (although most respondents seem to have misinterpreted this item). 33% admitted to have at least once rounded down *p* values that were just over .05. Independent raters rated 0% of the justifications for "falsifying" data and 6% of the justifications for rounding down *p* values as unacceptable. Regarding the frequency, respondents on average indicated that they rarely or never "falsified" data or rounded down *p* values.	The self-admission rates for other QRPs ranged between 16% (claiming results were unaffected by demographics when they were) and 84% (selectively reporting studies that worked).
Stürmer et al. (2017)	88 early-career researchers responded to an online-survey about the prevalence of 14 QRPs in the German social psychology community distributed via mailing-lists.	3.5% of the participants rated inventing data as "moderately" prevalent and 1.2% as "very" prevalent. 5.9% rated manipulating/faking data as "moderately" prevalent and 1.2% as "very" prevalent.	Other QRPs were rated as "fairly" or "very" prevalent by 22.9% (transforming data to yield the significance level) to 82.2% (conducting many studies, but reporting only those producing significant results) of the participants.

(Continued on next page)

Table 1. (Continued)

Reference	Method	Prevalence estimates of scientific misconduct	Prevalence estimates of other QRPs and potential scientific misconduct
Sacco et al. (2018)	136 researchers from various fields who held at least one grant from the US-National Institutes of Health of which eight indicated psychology as their field completed an online survey about their willingness to engage in 40 QRPs and about the prevalence of these QRPs in their field.	Fabricating data by adding data for participants who in fact did not participate and assigning participants to study conditions based on pre-screen data in a way that is intended to maximize the likelihood of treatment effects were rated as very uncommon or uncommon by six participants and as somewhat uncommon and neither common nor uncommon by one participant each. Using others' ideas, words, images, or other materials without citation was rated as very uncommon or uncommon by 5 participants and as neither uncommon nor common and as somewhat common by one participant each.	The other QRPs had higher prevalence ratings. For instance, failing to report all of a study's outcome measures was rated as somewhat common or common by 50% of the psychological researchers.
Statistical (re)analyses			
Bakker and Wicherts (2011)[b]	In study 1, *p* values from 281 articles in psychological high-impact and low-impact journals were checked for errors (inconsistencies with its test statistic and *df*s). In study 2, *p* values from 63 randomly selected psychological articles were checked.		12.4% of all articles contained at least one gross inconsistency (i.e., reported *p* value significant, computed *p* value non-significant or vice versa). There was no statistical difference in the proportion of gross errors between high-impact and low-impact journals. Moreover, the gross inconsistencies were more likely to render an insignificant effect significant than vice versa which showed that the gross errors were predominantly in favor of the researchers' hypotheses. For instance, all rounding errors around a *p* value of .05 were in line with the researchers' hypotheses. In study 2, 6.3% of all articles contained at least one gross inconsistency.
Mazzola and Deuling (2013)[c]	The percentages of supported and unsupported hypothesis were compared between 215 articles from six industrial–organizational (I–O) psychology (2010-2012) journals and a sample of 127 dissertations from 16 PhD programs in I–O psychology (2010–2012).		1,231 (73.10%) of the hypothesis in published I–O psychology studies and only 404 (32.93%) of the hypothesis in I–O psychology dissertations were supported.
Caperos and Pardo Merino (2013)	1,212 *p* values from 102 studies published in three Spanish psychology journals in 2011 and 2012 were checked for errors (inconsistencies with its test statistic and *df*s)		17.6% of all articles with complete information contained at least one gross inconsistency (i.e., reported *p* value significant, computed *p* value non-significant or vice versa).
Bakker and Wicherts (2014)[d]	2,667 statistical results from 153 articles in high ranking psychological journals from 2001 to 2010 were analyzed. The authors compared the median *p* value, sample sizes, and the prevalence of reporting errors between studies with and without removal of outliers.		No differences between studies that reported outlier removal and studies that did not report outlier removal were found regarding the median *p* value, the sample size or reporting errors. In 41% of the articles without reported outlier removal, there was a discrepancy between the reported degrees of freedom (*df*) of *t* tests and the reported sample sizes (reporting error).

(Continued on next page)

Table 1. (Continued)

Reference	Method	Prevalence estimates of scientific misconduct	Prevalence estimates of other QRPs and potential scientific misconduct
Veldkamp et al. (2014)[b]	8,105 statistical results were retrieved from articles published in six high ranking journals from different psychological subfields from January 2012 to October 2012 and were checked for errors.		20.5% of the articles contained at least one gross inconsistency (i.e., reported *p* value as significant and computed *p* value non-significant or vice versa). No journal differed significantly from any other journal in the prevalence of articles with at least one gross inconsistency.
Bosco, Aguinis, Field, Pierce, and Dalton (2016)[e]	In study 1, 247 effect sizes of the relations between job performance and nine other variables from two top ranking I/O psychology journals were analyzed. The authors tested whether the effect sizes were larger for hypothesized in comparison to nonhypothesized relations. Also, the HARKing self-admission rate was established by contacting the authors. In study 2, 281 hypothesized and nonhypothesized effect sizes from a meta-analysis of the relation between job satisfaction and job performance were compared.		In study 1, hypothesized relations (mean $r = .20$) were larger than the nonhypothesized relations (mean $r = .09$). Also, 38% of the responding authors reported that (at least) one hypothesis has changed after the completion of data collection. In study 2, hypothesized job satisfaction–job performance relations (mean $r = .22$) were larger than nonhypothesized job satisfaction–job performance relations (mean $r = .16$).
Franco, Malhotra, and Simonovits (2016)	32 published psychological studies for which the complete experimental design and the full set of measured variables was available from a competitive grant program in the United States were analyzed. The authors checked whether all experimental conditions and outcome variables were reported in the published manuscript.		41% of the investigated studies failed to reported all experimental conditions and outcome measures. 63% of the reported tests but only 23% of the unreported tests were significant at the $p < .05$ level. Also, the reported effect sizes were about twice as large as unreported effect sizes.
Nuijten et al. (2016)[b]	Over 250,000 *p* values from 30,717 articles published in eight major psychology journals from 1985 to 2013 were checked for errors (inconsistencies with its test statistic and *df*s).		12.9% of all articles with null-hypothesis significance testing (NHST) contained at least one gross inconsistency (i.e., reported *p* value significant, computed *p* value non-significant or vice versa). Overall, 1.4% (3,581) of the *p* values were grossly inconsistent. Between 1985 and 2013, the prevalence of articles with gross inconsistencies has declined.
Cortina, Green, Keeler, and Vandenberg (2017)[g]	784 structural equation models from 75 papers published in the Journal of Applied Psychology and the Academy of Management Journal from 2011 to 2013 and 1993 to 1995 were checked. The authors tested whether there were discrepancies between the reported *df* and the calculated *df* based on the model description in the article Introductions (i.e., whether the models that were tested in the manuscripts differed from the models that the manuscripts claimed to test)		There were discrepancies between the calculated and reported *df* in 38.38% of the models that reported *df* and provided sufficient information to calculate *df*. These discrepancies could be reconciled in only 14.91% of the cases.

(Continued on next page)

Table 1. (Continued)

Reference	Method	Prevalence estimates of scientific misconduct	Prevalence estimates of other QRPs and potential scientific misconduct
Retraction analyses			
Grieneisen and Zhang (2012)[f]	42 literature databases were used to locate retracted articles from 1928 to 2011 (n = 4,449) across the full spectrum of scientific disciplines. Ratios of retractions were calculated by dividing the number in each scientific field by the Web of Science 2010 records in this field. PsycINFO was not included as a data source.		Retraction rates were 0.16% in Psychology, Mathematical (1 article), 0.12% in Psychology, Social (4 articles), 0.11% in Psychology, Psychoanalysis (1 article), 0.09% in Psychology (8 articles), 0.09% in Psychology, Developmental (4 articles), 0.06% in Psychology, Experimental (5 articles), 0.06% in Psychology, Multidisciplinary (5 articles), 0.04% in Psychology, Biological (1 article), 0.04% in Psychology, Clinical (3 articles), 0% in Psychology, Applied (0 articles), 0% in Psychology, Educational (0 articles) and 0.38% in Neurosciences (169 articles).

Note. [a]This study has been criticized for overestimating questionable research practices (QRPs) due to methodological problems (e.g., ambiguous item wording and equating the prevalence of a QRP with the percentage of researchers who ever engaged in a QRP) by Fiedler and Schwarz (2016). [b]Actions that qualify as scientific misconduct (e.g., deliberately misreporting p values) provide an explanation for the reported gross statistical inconsistencies. Yet, a large proportion of the identified gross inconstancies might be attributable to honest error rather than scientific misconduct. [c]Mazzola and Deuling (2013) interpreted their findings as evidence for selective reporting and HARKing. Yet, the authors mention alternative interpretations (e.g., "file drawing" null result projects) that may be equally plausible. [d]Bakker and Wicherts (2014) regarded outlier removal as a potential indicator of significance chasing or p-hacking. Discrepancies between the reported sample sizes and df were interpreted as an indicator for the failure to report outlier removal or missing values. [e]Bosco et al. (2016) interpreted their findings as evidence of HARKing and its negative impact. 13 alternative explanations of the findings were tested and ruled out. [f]We were unable to obtain the number of psychological articles retracted due to scientific misconduct from this study. It is unlikely that all of the included retractions reflect cases of scientific misconduct. [g]There were various reasons for the detected inconsistencies. Some reasons seem to reflect insufficient reporting (e.g., not enough detail provided to know which items go into which parcels) whereas others seem to reflect QRPs (e.g., unreported freeing of measurement error covariances).

were retracted due to scientific misconduct was not reported.

Taken together, the existing studies show that the self-admission rates for scientific misconduct are lower than self-admission rates for other QRPs that are regarded as less severe (see Sacco et al., 2018). Also, the self-admission rates for scientific misconduct were considerably lower than prevalence estimates regarding the actions of other psychological researchers and lower than the percentage of gross statistical inconsistencies. Even between the survey studies, estimates varied strongly and might overestimate (e.g., because of difficulties in item interpretation; e.g., Motyl et al., 2017) or underestimate (e.g., because of social desirability; Edwards, 1957) the prevalence of scientific misconduct. Thus, additional empirical data were required to obtain a reliable lower bound prevalence estimate of scientific misconduct in psychology.

Empirical Study

This study was the first empirical investigation analyzing a large number of psychological articles retracted due to scientific misconduct. Our empirical analyses revealed that the percentage of retractions that was attributable to scientific misconduct (64.84% in PsycINFO) was similar to the biomedical and life-science literature (67.40% in PubMed; Fang et al., 2012) and similar to estimates derived from a variety of scientific disciplines and databases (47% "publishing misconduct" and 20% "research misconduct"; Grieneisen & Zhang, 2012). The overall rate of journal articles retracted due to scientific misconduct was somewhat higher in PsycINFO (0.82 per 10,000 journal articles) compared to Medline (0.56 per 10,000 journal articles; Wager & Williams, 2011). Importantly, all comparisons with other disciplines should be interpreted with caution due to differences in methods and covered time periods. For example, Fang et al. (2012) consulted further information in addition to the retraction notices to classify reasons for retractions whereas other authors did not (e.g., Wager & Williams, 2011).

With regard to the temporal development, there was a steep increase in retractions due to scientific misconduct of journal articles in PsycINFO between the late 1990s and 2012. There were almost no retractions due to scientific misconduct in psychology before the late 1990s. Grieneisen and Zhang (2012) identified a similar trend in their study covering a wide range of scientific disciplines. This could either be explained by an increase in scientific misconduct or by changing mechanisms (e.g., plagiarism screening) and standards (e.g., journal policies) to detect and retract fraudulent articles. Fanelli (2013) argued that the increase in article retractions is attributable to improved detection and retraction systems. For instance, he found that the proportion of journals that retract articles has grown

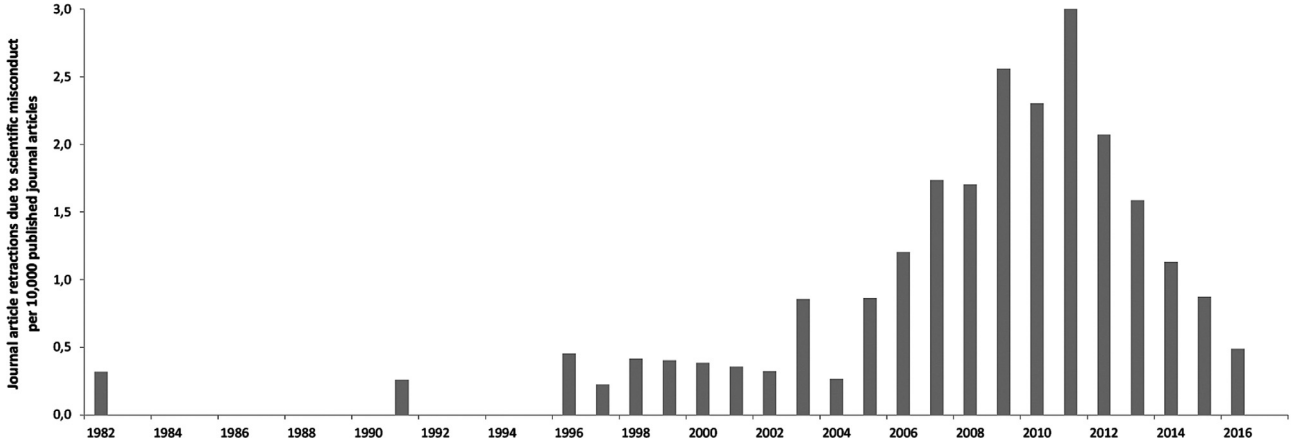

Figure 1. Development in number of journal articles retracted due to scientific misconduct per 10,000 published journal articles in PsycINFO from 1982 to 2017 by publication year of the retracted article.

Table 2. Number of article retractions due to scientific misconduct and number of authors responsible for scientific misconduct per 10,000 journal articles in PsycINFO subfields (1860–2017)

	Number of retracted articles[a]		Number of responsible authors[a]	
Subfield	*N*	Per 10,000 published articles	*N*	Per 10,000 published articles
Social psychology	31	3.80	7	0.86
Consumer psychology	9	2.52	7	1.96
Sport psychology & leisure	5	1.85	5	1.85
Engineering & environmental psychology	7	1.70	6	1.45
Personality psychology	17	1.64	5	0.48
Physiological psychology & neuroscience	57	1.48	54	1.40
Intelligent systems	4	1.25	4	1.25
Industrial & organizational psychology	11	0.77	11	0.77
Psychological & physical disorders	55	0.72	49	0.64
Health & mental health treatment & prevention	44	0.72	41	0.67
Human experimental psychology	15	0.69	8	0.37
Communication systems	3	0.61	3	0.61
Social processes & social issues	8	0.46	7	0.40
General psychology	1	0.38	1	0.38
Developmental psychology	7	0.38	5	0.27
Animal experimental & comparative psychology	3	0.34	3	0.34
Educational psychology	8	0.33	8	0.33
Forensic psychology & legal issues	1	0.30	1	0.30
Professional psychological & health personnel issues	3	0.27	3	0.27
Psychometrics & statistics & methodology	4	0.21	3	0.16
Psychology & the humanities	0	0.00	0	0.00
Military psychology	0	0.00	0	0.00

Note. To allocate the retracted articles to the respective psychological subfield, we used the articles' content classification in PsycINFO. [a]About 13% of the journal articles in PsycINFO were assigned to two subfields. Accordingly, in the determination of the subfield size, the number of retracted articles and the number of responsible authors, these articles are included in both subfields.

dramatically while the cases of misconduct identified by the US Office of Research Integrity have not increased. Interestingly, the trend that was identified for article retractions in psychology was not found for gross statistical inconsistencies in published psychological articles which are regarded as a potential indicator of scientific misconduct or QRPs (Nuijten et al., 2016). This finding supports Fanelli's (2013) notion that the increase in article retractions is mostly attributable to improved detection and retraction systems (also see Gross, 2016). In recent years,

the rate of articles retracted due to scientific misconduct seemed to decline. This is likely to be due to the time delay with which cases of scientific misconduct are usually detected (Fang et al., 2012).

In 20 out of 22 psychological subfields, there were articles retracted due to scientific misconduct. Based on the number of retracted journal articles, the largest prevalence was identified for Social Psychology. However, 80.65% of these cases were attributable to one author (D. Stapel). Based on the number of different responsible authors, Consumer Psychology had the highest prevalence. This finding shows, that the perception of some psychological subfields as being more fraudulent than others might be attributable to spectacular cases in which single authors were responsible for a large number of fraudulent studies ("repeat offenders"; Grieneisen & Zhang, 2012).

General Discussion and Limitations

Our systematic review showed that scientific misconduct including data falsification, data fabrication and other severe forms of misconduct in psychology is relatively rare in comparison to other QRPs. As expected, our empirical study yielded a somewhat lower prevalence estimate of scientific misconduct in comparison to survey studies. This reflects that scientific misconduct is not always detected. Yet, scientific misconduct was prevalent across a variety of geographic regions (Agnoli et al., 2017; Braun & Roussos, 2012; John et al., 2012) and in almost all psychological subfields.

Even single incidents of scientific misconduct can have immense effects (Michalek et al., 2010). Consequently, we believe that it is important to promote measures which diminish the incentives and possibilities to engage in scientific misconduct equally across all psychological subfields. In our eyes, a promising approach lies in the advancement of open data and open materials (Tenopir et al., 2011) and in the improvement of systems for reporting suspected scientific misconduct (Crocker & Cooper, 2011). However, we do not believe that scientific misconduct can be entirely prevented through detection systems. Thus, fostering an ethical organization culture clearly communicating acceptable and unacceptable behavior in psychology departments and research groups (e.g., through rewards systems; Kish-Gephart, Harrison, & Treviño, 2010) seems equally important.

Our study has, of course, some limitations. First, the number of studies in the systematic review was relatively low. The heterogeneity in methods did not allow meta-analytic integration of the results. Similarly, the number of retracted articles in our empirical study was low for some subfields so that comparisons between subfields should be interpreted with caution. Second, retraction notices provide an estimate only of the lower bound prevalence estimate of scientific misconduct as many cases can remain unnoticed. In this point, the investigation of scientific misconduct is similar to the calculation of crime rates, because only reported offenses are in the statistics (Bechtel & Pearson, 1985). Third, our empirical method was designed to quantify convicted cases of scientific misconduct. Other, subtler but potentially equally damaging (Simmons, Nelson, & Simonsohn, 2011) QRPs were only covered in our systematic review.

Despite these constraints, the present study contributes to the understanding of scientific misconduct in psychology. Our study yielded reliable lower bound estimates of scientific misconduct which showed that scientific misconduct occurs across almost all psychological subfields. Also, the increasing retraction rate in comparison to the 1980s and 1990s shows that there are mechanisms which generally have the ability to detect scientific misconduct. Thus, initiatives to strengthen these systems (e.g., by increasing research transparency) should be promoted across all psychological subfields and not be restrained to fields with prominent cases of scientific misconduct.

References

*References marked with an asterisk were included in the systematic review.

*Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS One, 12*, e0172792. https://doi.org/10.1371/journal.pone.0172792

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics, 13*, 437–461. https://doi.org/10.1007/s11948-007-9042-5

*Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666–678. https://doi.org/10.3758/s13428-011-0089-5

*Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS One, 9*, e103360. https://doi.org/10.1371/journal.pone.0103360

Bechtel, H. K. Jr., & Pearson, W. Jr. (1985). Deviant scientists and scientific deviance. *Deviant Behavior, 6*, 237–252. https://doi.org/10.1080/01639625.1985.9967676

*Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*, 709–750. https://doi.org/10.1111/peps.12111

*Braun, M., & Roussos, A. J. (2012). Psychotherapy researchers: Reported misbehaviors and opinions. *Journal of Empirical Research on Human Research Ethics, 7*, 25–29. https://doi.org/10.1525/jer.2012.7.5.25

Callaway, E. (2011). Report finds massive fraud at Dutch universities. *Nature, 479*, 15. https://doi.org/10.1038/479015a

*Caperos, J. M., & Pardo Merino, A. (2013). Consistency errors in *p*-values reported in Spanish psychology journals. *Psicothema, 25*, 408–414. https://doi.org/10.7334/psicothema2012.207

Carey, B. (2011, November 2). *Fraud case seen as a red flag for psychology research* (pp. A3). New York, NY: New York Times.

Center for Scientific Integrity. (n.d.). *Retraction watch database*. Retrieved from http://retractiondatabase.org/RetractionSearch.aspx

*Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2017). Degrees of freedom in SEM: Are we testing the models that we claim to test? *Organizational Research Methods, 20*, 350–378. https://doi.org/10.1177/1094428116676345

Crocker, J., & Cooper, M. L. (2011). Addressing scientific fraud. *Science, 334*, 1182. https://doi.org/10.1126/science.1216775

De Rond, M., & Miller, A. N. (2005). Publish or perish: Bane or boon of academic life? *Journal of Management Inquiry, 14*, 321–329. https://doi.org/10.1177/1056492605276850

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*. Worth, TX: Dryden Press.

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One, 4*, e5738. https://doi.org/10.1177/1056492605276850

Fanelli, D. (2013). Why growing retractions are (mostly) a good sign. *PLoS Medicine, 10*, e1001563. https://doi.org/10.1371/journal.pmed.1001563

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 of the United States of America, 109*, 17028–17033. https://doi.org/10.1073/pnas.1212247109

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*, 45–52. https://doi.org/10.1177/1948550615612150

*Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science, 7*, 8–12. https://doi.org/10.1177/1948550615598377

Franzoni, C., Scellato, G., & Stephan, P. (2011). Changing incentives to publish. *Science, 333*, 702–703. https://doi.org/10.1126/science.1197286

*Grieneisen, M. L., & Zhang, M. (2012). A comprehensive survey of retracted articles from the scholarly literature. *PLoS One, 7*, e44118. https://doi.org/10.1371/journal.pone.0044118

Gross, C. (2016). Scientific misconduct. *Annual Review of Psychology, 67*, 693–711. https://doi.org/10.1371/journal.pone.0044118

Hesselmann, F., Graf, V., Schmidt, M., & Reinhart, M. (2017). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current Sociology, 65*, 814–845. https://doi.org/10.1177/0011392116663807

Hofmann, B., Helgesson, G., Juth, N., & Holm, S. (2015). Scientific dishonesty: A survey of doctoral students at the major medical faculties in Sweden and Norway. *Journal of Empirical Research on Human Research Ethics, 10*, 380–388. https://doi.org/10.1177/1556264615599686

*John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*, 524–532. https://doi.org/10.1177/0956797611430953

Kish-Gephart, J. J., Harrison, D. A., & Treviño, L. K. (2010). Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work. *Journal of Applied Psychology, 95*, 1–31. https://doi.org/10.1037/a0017103

Louis, K. S., Anderson, M. S., & Rosenberg, L. (1995). Academic misconduct and values: The department's influence. *The Review of Higher Education, 18*, 393–422. https://doi.org/10.1353/rhe.1995.0007

Margraf, J. (2015). Zur Lage der Psychologie [On the state of psychology]. *Psychologische Rundschau, 66*, 1–30. https://doi.org/10.1026/0033-3042/a000247

Marshall, E. (2000). Scientific misconduct–How prevalent is fraud? That's a million-dollar question. *Science, 290*, 1662–1663. https://doi.org/10.1126/science.290.5497.1662

*Mazzola, J. J., & Deuling, J. K. (2013). Forgetting what we learned as graduate students: HARKing and selective outcome reporting in I–O journal articles. *Industrial and Organizational Psychology, 6*, 279–284. https://doi.org/10.1111/iops.12049

Michalek, A. M., Hutson, A. D., Wicher, C. P., & Trump, D. L. (2010). The costs and underappreciated consequences of research misconduct: A case study. *PLoS Medicine, 7*, e1000318. https://doi.org/10.1371/journal.pmed.1000318

*Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., … Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology, 113*, 34–58. https://doi.org/10.1037/pspa0000084

Münch, R. (2014). *Academic capitalism: Universities in the global struggle for excellence*. New York, NY: Routledge.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631. https://doi.org/10.1177/1745691612459058

*Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods, 48*, 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Office of Research Integrity. (2011). *Definition of research misconduct*. Rockville, MD: US Department of Health and Human Services. Retrieved from http://ori.hhs.gov/definition-misconduct

Office of Science and Technology Policy (OSTP). (2000). Federal policy on research misconduct. *Federal Register, 65*, 76260–76264. Retrieved from https://ori.hhs.gov/federal-research-misconduct-policy

Resnik, D. B., Neal, T., Raymond, A., & Kissling, G. E. (2015). Research misconduct definitions adopted by US research institutions. *Accountability in Research, 22*, 14–21. https://doi.org/10.1080/08989621.2014.891943

Rovenpor, D. R., & Gonzales, J. E. (2015). Replicability in psychological science: Challenges, opportunities, and how to stay up-to-date. *Psychological Science Agenda, 29*(1). Retrieved from www.apa.org/science/about/psa/2015/01/replicability.aspx

*Sacco, D. F., Bruton, S. V., & Brown, M. (2018). In defense of the questionable: Defining the basis of research scientists' engagement in questionable research practices. *Journal of Empirical Research on Human Research Ethics, 13*, 101–110. https://doi.org/10.1177/1556264617743834

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*, 1359–1366. https://doi.org/10.1177/0956797611417632

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*, 670–688. https://doi.org/10.1177/1745691612460687

*Stürmer, S., Oeberst, A., Trötschel, R., & Decker, O. (2017). Early-career researchers' perceptions of the prevalence of questionable research practices, potential causes, and open science. *Social Psychology, 48*, 365–371. https://doi.org/10.1027/1864-9335/a000324

Świątkowski, W., & Dompnier, B. (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology, 30*, 111–124. https://doi.org/10.1027/1864-9335/a000324

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., ... Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS One, 6*, e21101. https://doi.org/10.1371/journal.pone.0021101

*Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One, 9*, e114876. https://doi.org/10.1371/journal.pone.0114876

Wager, E., & Williams, P. (2011). Why and how do journals retract articles? An analysis of Medline retractions 1988–2008. *Journal of Medical Ethics, 37*, 567–570. https://doi.org/10.1136/jme.2010.040964

History
Received February 28, 2018
Revision received October 17, 2018
Accepted October 18, 2018
Published online March 29, 2019

Armin Günther
Leibniz Institute for Psychology Information
Universitätsring 15
54296 Trier
Germany
ague@leibniz-psychology.org