Research Practices in Psychology and How We Communicate About Them

By

JULIA GODOY BOTTESINI
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Mijke Rhemtulla, Chair

Simine Vazire

Andrew Todd

Committee in Charge

2022

Acknowledgements

Amid the flurry of events leading up to the completion of this dissertation, I didn't realize I would get to thank everyone who helped me get here until it was too late to write a polished, properly spellchecked, PhD-serious sort of acknowledgements section. Finding the right words to thank everyone who helped me in my 7.5-year-long PhD journey in just a couple of days was an impossible task. Please know that, if I mentioned you here, I probably have a lot more thankfulness than I can express in a few words, and I am very grateful for your support.

First, I want to acknowledge that I was helped by so many people in my PhD journey, it would be impossible to list them all. For every obstacle I encountered, I also met at least as many wonderful people that were kind to me, invited me into their groups, taught me new things, and generally welcomed me with open arms. If that was you, thank you, and please continue to do that for other people, especially those who are starting anew thousands of miles away from anyone they know, like I was.

The dissertation you may or may not be on your way to reading would not have been possible without my mentor and the best advisor I could have hoped for, Simine Vazire. Despite her moving to another continent half way through my PhD, she was there for me every step of the way. When I wrote terrible first drafts of manuscripts, she had firm but always kind (and detailed) feedback. When I came up with wild research ideas, she helped me hone them into manageable projects. And when I was going through difficult personal stuff and couldn't make much progress in my research, she was generous with her time and worked with me to help me stay afloat. (Also, she sent me many great pictures of Hugo!) She'll know better than most that I am prone to hyperbole, but it is definitely not an exaggeration to say I would not have made it through this PhD without Simine's guidance and support. I feel incredibly lucky to have had the opportunity to learn from her for the past 5 years.

I was lucky enough to have not just one but two great advisors. After Simine's move, Mijke Rhemtulla took me in and welcomed me to her lab as if she had recruited me herself. Mijke encouraged me to widen my statistical and quantitative horizons, and despite my research being quite different from everyone else's in the lab, I always felt like my input was valued. Mijke was always willing to talk through a quanty problem with me, and had endless patience for my unorthodox questions. Her lab, and everyone in it, was a great source of support for me during my final PhD years, and I am very grateful to have been a part of it.

Next, I want to acknowledge the institutional support I received from UC Davis. This came in the form of financial support, but also excellent quantitative training, and faculty in the Social-Personality and Quant areas that welcomed my unusual research. In particular, Andrew Todd has been a source of thoughtful feedback and cool ideas ever since I roped him into being part of my IAC during my first year. Thanks, Andy, I really appreciate it! Of course, this paragraph would not be complete if I didn't recognize every graduate student's guardian angel, the appropriately named Angela Scully. Angela, you're a rock star, and your support, especially during the last year, has been instrumental in allowing me to complete this dissertation.

My path to getting into a PhD program was not at all clear when I first set out on it, and I was lucky enough to be able to get the training and opportunities I needed to succeed in Columbia University's

postbacc program. Thank you to all the graduate students, postdocs, and faculty there who helped me explore my interests and hone my research skills. Thank you in particular to Kaytee Turetsky, Hale Forster, and Travis Riddle, who embraced my enthusiasm for using new research methods and constantly encouraged me to learn more.

Doing metascience research in a Psychology department can sometimes be lonely, but finding other like-minded people in the broader open science and metascience community helped me feel less alone. In particular, SIPS has always been an organization that helped foster community for me and where I got to meet current and former lab mates, informal mentors, and other people with similar interests. My memories of SIPS conferences, both in person and online, are some of the fondest from my time in graduate school. If you were a part of that, whether in a big or small way, thank you.

Part of what made my PhD a great experience were all the other graduate students I met along the way. I was really lucky to start my PhD with a large cohort, and I enjoyed the sense of community that came with it. Despite the many changes brought on by the pandemic, I benefited immensely from always having senior and junior graduate students and lab mates that were willing to impart their wisdom (or data) on me. In particular, thank you to Jessie Sun, Sarah Schiavone, and Beth Clarke for your collaboration and support throughout the years, and Kailey Lawson, whose endless wisdom, patience, and kindness has helped guide me through some tough times.

Finally, I am very lucky to have had the support of a lot of family, both biological and chosen. In my chosen family, I want to specifically thank Alyssa Maness, Jordan Varney, and Sarah Sweigart. I literally wouldn't have made it through this PhD without them. They were with me through thick and thin, from beginning to end, and everything in between. I'll never be able to thank you enough for everything you've done for me. In my biological family, the biggest shout out goes to my wonderful Mom, Maria Cristina, who always pushed me to go further to pursue my dreams, even when that also meant geographically farther from her. We did it, Mom!

Table of Contents

Abstract

This dissertation attempts to examine research practices and the way we communicate about them in parts of the research process that may not always be at the forefront of people's minds. When researchers recruit participants for their studies, do we ever wonder what they think about how we treat their data? In Chapter 1, I examined psychology research participants' opinions about (mostly) common research practices in psychology, including questionable research practices (QRPs; e.g., *p*-hacking, HARKing) and practices to increase transparency and replicability. After running a study, researchers then write it up as a manuscript, which is how most research gets communicated to relevant stakeholders. But do different groups of researchers communicate their findings differently? In Chapter 2, I investigated which groups of researchers might be more or less prone to hedging their conclusions in their research articles, a first step towards better understanding when and why researchers make strong claims about their findings. Finally, when findings get disseminated to the public, which research practices are being rewarded with media attention? In Chapter 3, I explored what information science journalists use when evaluating psychology findings' trustworthiness and newsworthiness. By examining these often-forgotten aspects of research practices and their consequences, I hope to encourage more research on how we do and communicate psychological science.

Chapter 1

What Do Participants Think of Our Research Practices? An Examination of Behavioral Psychology Participants' Preferences

The content of this chapter has been previously published in the journal *Royal Society Open Science* under a Creative Commons BY license. Below is the citation for the corresponding published article.

Cite: Bottesini, J. G., Rhemtulla, M., & Vazire, S. (2022). What do participants think of our research practices? An examination of behavioural psychology participants' preferences. *Royal Society open science*, *9*(4), 200048.

Abstract

What research practices should be considered acceptable? Historically, scientists have set the standards for what constitutes acceptable research practices. However, there is value in considering non-scientists' perspectives, including research participants'. 1,873 participants from MTurk and university subject pools were surveyed after their participation in one of eight minimal-risk studies. We asked participants how they would feel if (mostly) common research practices were applied to their data: *p*-hacking/cherry-picking results, selective reporting of studies, Hypothesizing After Results are Known (HARKing), committing fraud, conducting direct replications, sharing data, sharing methods, and open access publishing. An overwhelming majority of psychology research participants think questionable research practices (e.g., *p*-hacking, HARKing) are unacceptable (68.3--81.3%), and were supportive of practices to increase transparency and replicability (71.4--80.1%). A surprising number of participants expressed positive or neutral views toward scientific fraud (18.7%), raising concerns about data quality. We grapple with this concern and interpret our results in light of the limitations of our study. Despite ambiguity in our results, we argue that there is evidence (from our study and others') that researchers may be violating participants' expectations and should be transparent with participants about how their data will be used.

Keywords: Research practices; Open Science; Scientific integrity; Informed consent

Introduction

What research practices should be considered acceptable, and who gets to decide? Historically, scientists — and as a group, scientific organizations — have set the standards and have been the main drivers of change in what constitutes acceptable research practices. Perhaps this is warranted. Who better to set the standards than those who know research practices best? It seems reasonable that decisions regarding those practices should be entrusted to scientists themselves. However, there may be value in considering non-scientists' perspectives and preferences, including research participants'.

The replicability crisis in psychology has demonstrated that scientists are not always good at regulating their own practices. For example, a surprisingly high proportion of researchers admit to engaging in questionable research practices, or QRPs (as described in John et al., 2012; see also Agnoli et al., 2017; Fox et al., 2018; Makel et al., 2019). These include things like failing to report some of the conditions or measures in a study, excluding outliers after seeing their effect on the results, and a wide range of other practices that can be justified in some instances but also inflate rates of false positives in the published literature (Simmons, Nelson, & Simonsohn, 2011). A large sample of social and personality psychologists reported engaging in these practices less often than "sometimes," but more often than "never" (Motyl et al., 2017).

To combat the corrupting influence of these practices on the ability to accumulate scientific knowledge, individual scientists and scientific organizations have led the push for making research practices more rigorous and open. In the case of funding agencies, the NIH's Public Access Policy dictates that all NIH-funded research papers must be made available to the public ("Frequently Asked Questions about the NIH Public Access Policy | publicaccess.nih.gov," n.d.)[1]. Some journals and publishers have also pushed

[1] To guarantee that future readers will have access to the content referenced here and in other non-DOI materials cited, we have compiled a list of archival links for those references (https://osf.io/26ay8/)

in the direction of more open scientific practices. For example, 53 journals, including some of the most sought-after outlets in psychology like *Psychological Science,* now offer open science badges, which easily identify articles that have open data, open materials, or include studies that have been preregistered ("Open Science Badges," n.d.). Although simply having badges doesn't necessarily mean the research is more open or trustworthy, there's evidence of significant increases in data sharing which may be attributable to the implementation of the badge system (Kidwell et al., 2016; Rowhani-Farid, Allen, & Barnett, 2017; c.f. Bastian, 2017).

How do scientists decide which practices are consistent with their values and norms? Currently, the norms in many scientific communities are in flux and are quite permissive regarding the use of both QRPs and open science practices. This approach of letting research practices evolve freely over time, without external regulation, tends to select for practices that produce the most valued research output. In the current system, what is most valued is often the quantity of publications in top journals, regardless of the quality or replicability of the research (Smaldino & McElreath, 2016). In short, scientists operate in a system where incentives do not always align with promoting rigorous research methods or accurate research findings. Thus, if we leave the development and evolution of research practices up to scientists alone, this may not select for practices that are best for science itself. Therefore, it may be a good idea to provide checks and balances on norms about scientific research practices, and these checks and balances should be informed by feedback from those outside the guild of science.

One way to obtain such feedback is to solicit the preferences and opinions of non-scientists, who can offer another perspective on the norms and practices in science, and are likely influenced by a different set of incentives than are scientists. One such group of non-scientist stakeholders are patients suffering from specific diseases, and their loved ones, who form organized communities to advocate for patients' interests. Some of these communities, called patient advocacy groups, have pushed for more efficient

use of the scarce data on rare diseases, including data sharing ("Patient Groups, Industry Seek Changes to Rare Disease Drug Guidance," n.d.). Other independent organizations, such as AllTrials, have also influenced scientific practices in the direction of greater transparency. With the support of scientists and non-scientists alike, AllTrials has championed transparency in medical research by urging researchers to register and share the results of all clinical trials (AllTrials, n.d.). In addition, non-scientist watchdog groups (e.g., journalists, government regulatory bodies) can call out problematic norms and practices, and push for new standards.

Another group of non-scientist stakeholders is research participants. While they have not traditionally formed communities to advocate for their interests (c.f., patient advocacy groups, Amazon Mechanical Turk workers' online communities), they are also a vital part of the research process and important members of the scientific community in sciences that rely on human participants. In fact, because they are the only ones who experience the research procedure directly, research participants can sometimes have information or insight that no other stakeholder in the research process has. As such, participants might have a unique, informative perspective on the research process.

A fresh perspective on research practices is not the only reason to care about what participants think. One practical reason to consider research participants' preferences is that ignoring their wishes risks driving them away. Most research in psychology relies on human participants, and their willingness to provide scientists with high quality information about themselves. Motivation to be a participant in scientific studies is varied, but besides financial compensation, altruism and a desire to contribute to scientific knowledge are common reasons people mention for participating (McSweeney et al., n.d.; Sanderson et al., 2016). If participants believe researchers are not using their data in a way that maximizes the value of their participation, they might feel less inclined to participate, or participate but

provide lower quality data. In addition, going against participants' wishes could undermine public trust in science even among non-participants, if they feel we are mistreating participants.

There are also important considerations regarding informed consent to take into account when thinking about research practices. Although informed consent is usually thought of in terms of how participants are treated within the context of the study, their rights also extend to how their data are used thereafter. This is explicitly acknowledged in human subjects regulations, but there has not been much attention paid to what this means for the kinds of research practices that have been the target of methodological reforms, beyond data sharing. Specifically, informed consent must contain not only a description of how the confidentiality and privacy of the subjects will be maintained, but also enough information in order for participants to understand the research procedures *and their purpose* (Protection of Human Subjects, 2009). There is some ambiguity in this phrase, but it could arguably encompass the types of questionable research practices scientists have been debating amongst themselves. For example, it is conceivable that participants might have preferences or assumptions about whether researchers will filedrawer (i.e., not attempt to publish or disseminate) results that do not support the researchers' hypothesis or theory. If we take informed consent to mean that participants should have an accurate understanding of the norms and practices that the researchers will follow, and should consent to how their data will be used, it is important to understand study participants' preferences and expectations.

What should we do with what we learn about participants' expectations and preferences about how we handle their data? If participants do have views about what would and would not be acceptable for researchers to do with their data, should scientists simply let those preferences dictate our research practices completely? Clearly not. Scientists are trained experts in how to conduct research, and many of our current research practices are effective and adequate. Moreover, it is probably unreasonable to

expect participants to understand all of the intricacies of data analysis and presentation. However, participants' expectations and preferences should inform our debates about the ethics and consequences of scientific practices and norms. Moreover, participants' expectations should inform our decisions about what information to provide in consent forms and plain language statements, to increase the chances that participants will be aware of any potential violations of their expectations.

There are several possible outcomes of investigating research participants' views about research practices. On the one hand, participants may feel that scientists' current research practices are acceptable. This would confirm that we are respecting our participants' wishes, and obtaining appropriate informed consent by treating participants' data in a way that is expected and acceptable to them. On the other hand, if participants find common research practices unacceptable, this may help us identify participants' misconceptions about the research process, and areas where there is a mismatch between their expectations and the reality of research.

If we do find that there is an inconsistency between participants' expectations and research practices, scientists have several options. First, they may want to listen to participants. Humans — of which scientists are a subset — are prone to motivated reasoning, and tend to have blind spots about their weaknesses, especially when they are deeply invested, a problem that a fresh perspective might alleviate. As outsiders who are familiar with the research, it is possible that participants may recognize those blind spots and areas for improvement better than researchers (particularly for "big picture" issues that do not require technical expertise). Second, researchers may decide not to change their practices completely, but to accommodate the principle behind participants' preferences. For example, if participants want all of their data to be shared publicly, in situations where this is not possible because of re-identification risk, researchers might make an effort to share as much of the data as possible. Finally, researchers may decide that a practice that is considered unacceptable by participants is still the

best way to go about doing research. In that case, better communication with participants may be needed to clarify why this practice is necessary and to honor the spirit of informed consent.

Any effort to take participants' preferences into account when engaging in research assumes participants do have preferences about the fate of their data. It is possible, however, that many participants have weak preferences or no preferences at all. This would still be useful for researchers to know, because it would increase researchers' confidence that they are not violating participants' preferences or expectations.

It is likely that at least some participants do have clear preferences about what we do with their data. On the subject of data-sharing, studies with genetic research or clinical trial participants suggest that, despite some concerns about privacy and confidentiality, a majority of participants support sharing of de-identified data, and are willing to share their own data, with some restrictions (Cummings, Zagrodney, & Day, 2015; Mello, Lieou, & Goodman, 2018; Trinidad et al., 2011).

There is also data on what participants think about selective reporting, that is, the practice of reporting only a subset of variables or studies performed when investigating a given question, and about data fabrication. In a series of studies, Pickett and Roche (2018) examined attitudes towards these practices among the general public in the United States — a population similar to research participants in many psychology studies — and among Amazon Mechanical Turk workers. Across both samples, there was high agreement that data fabrication is morally reprehensible and should be punished. Furthermore, in the Amazon Mechanical Turk sample, 71% of participants found selective reporting to be morally unacceptable, with over 60% saying researchers should be fired and/or receive a funding ban if they engage in selective reporting.

In addition to this empirical evidence, it seems intuitive that many participants would be surprised and disappointed if their data were being used in extremely unethical ways (e.g., to commit fraud, or further

the personal financial interests of the researchers at the expense of accurate scientific reporting). What is less clear is whether participants care, and what they think, about a wider set of questionable research practices and proposed open science reforms that are currently considered acceptable, and practiced by at least some researchers, in many scientific communities.

Study Aims

To further investigate this topic, we asked a sample of actual study participants, after their participation in another study, about how they would feel if some common research practices were applied to their own data . We did this using a short add-on survey (that we will refer to as the *meta-study*) at the end of different psychological studies (that we will refer to as the *base studies*). The meta-study asked participants to consider several research practices and imagine that they would be applied to the data they had just provided in the base study.

We asked participants about eight research practices, including questionable research practices (QRPs) and their consequences, and open science or proposed best practices, referred to here as *open science practices.* We followed two guidelines when choosing which practices to include. First, we sought to include the most common open science practices and every QRP from John et al. (2012) that is simple enough for participants to understand without technical expertise. Second, we selected those practices we judged as most directly impacting participants' contributions. For example, filedrawering could reduce participants' perceived value of their contribution because their data may never see the light of day; *p*-hacking (repeating statistical analyses several different ways but only reporting some of them) might distort the accuracy of reported findings and decrease the value of participant's contributions; posting data publicly could increase participants' concerns about privacy. Conversely, publishing the results in an open access format would enable participants to potentially access the results of research they have contributed to, which may be important to them.

The practices we asked participants about were: (1) *p*-hacking, or cherry-picking results, (2) selective reporting of studies, (3) HARKing (hypothesizing after the results are known), (4) committing fraud, (5) conducting direct replications, (6) sharing methods ("open methods"), by which we mean making the procedure of a study clear enough that others can replicate it, (7) publishing open access papers, and (8) sharing data ("open data").

What is the best way to present these research practices to participants? One option is to describe the practice (and, in some cases, its complement) without giving any explanation for why a researcher might engage in this practice. Another option is to explain the context, incentives, and tradeoffs that might lead a researcher to choose to engage in this practice. We carefully considered both options, and decided on the former in all but one case (data sharing, see Method below). While providing participants with context for these research practices may help them understand why scientists might engage in them, and the benefits and costs of doing so, we did not feel it would be possible to provide this context in a way that was not leading, without having participants take an hours-long course in research methods and scientific integrity. In addition, we felt that participants' naive reactions to these practices would be most informative for extrapolating to what a typical research participant thinks about these practices (i.e., without special insight or expertise into the technical, social, and political aspects of scientific research). In light of these considerations, we asked participants for their views about these practices without providing much information about the costs and benefits of each practice (with the exception of data sharing). As a result, participants' responses should be taken to reflect their spontaneous views about these practices, which might capture ideals rather than firmly-held expectations.

The goal of this study was to provide accurate estimates of research participants' views about these research practices. We had two research questions (though we did not have hypotheses about the results):

RQ1*: What are participants' views about questionable research practices (including p-hacking, selective reporting, and HARKing) and fraud?*

RQ2*: What are participants' views about open science practices (data sharing, direct replication, open methods, open access)?*

Scope

Because we did not have the time or resources to survey the full range of psychological science research studies, we limited our scope to minimal-risk psychology studies on English-speaking convenience samples that were run entirely on a computer or online, where all the data were provided by the participant in one session.

By including only this subset of studies, we expected to have minimal to no variance in study sensitivity, effort required for data contribution, and other characteristics of the studies which might affect participants' opinions of the examined research practices. Therefore, we recognize that we cannot explore any potential effects of these variables in this study, nor generalize the obtained results beyond the types of studies we included. However, we are able to generalize the results to other minimal-risk studies of the same kind, a common design that we believe represents a large proportion of psychology studies.

Pilot Studies

In order to help us develop the materials for the proposed study, we conducted three pilot studies. In the first study (Pilot Study A), we aimed to gauge participants' opinions about data sharing only. In the

second study (Pilot Study B), we added questions about all of the practices we planned to ask about in our proposed study, and changed the language of the data sharing question based on the results from Pilot Study A. In a third study (Pilot Study C), we fine-tuned the language used in the questions, which were almost identical to the proposed study. All materials and data for these pilots can be found at https://osf.io/bgpyc/.

With the notable exception of open access publishing[2], a majority of participants seemed to support using research best practices ("open science practices"). These preliminary results suggest that participants do have consistent opinions about these matters, which they are able to articulate. Participants overwhelmingly supported data sharing — over 70% for all versions of the question — including sharing publicly, and sharing so others can verify the claims being made or reuse the data. Sharing enough detail about the procedure of the study to allow others to replicate it (i.e., open methods) was also supported by a majority of participants. Finally, most participants (over 60% for all versions of the question) favored replication, even when it was presented as a trade-off between replicating the same study or moving on to a new study.

Furthermore, research participants seem to have strong preferences against the use of questionable research practices, with a majority of participants — over 75% for all questions and versions, with one exception[3] — disapproving of QRPs. In fact, the proportion of participants indicating that researchers should not *p*-hack, filedrawer studies, or HARK was similar to the proportion rating fraud as unacceptable. It is reassuring, however, that the distribution of answers was more extreme for fraud

[2] While participants still favored open access over publishing behind a paywall, a sizeable portion of participants selected the middle answer, indicating they were indifferent (Pilot B: 34.5%; Pilot C: 21.4%)

[3] Participants tended to see selective reporting of studies (i.e., filedrawering) less negatively when it was presented without explicitly saying the researchers reported only the results that came out the way they predicted (neutral version of the question; see https://osf.io/eyfcu/). In the UK sample (Pilot C), slightly under 70% of participants saw selective reporting as unacceptable.

(80.7% of participants in Pilot B and 92.8% in Pilot C selected the most extreme response for fraud, vs. 11.4-56.1% for the three QRPs mentioned here). Detailed results and figures for all three pilots can be found in the supplementary materials.

Registered Report Study

The present registered report study expands our pilot studies to investigate participants' opinions about fraud and questionable research practices (RQ1), and open science practices (RQ2), in a much larger sample. By including both research pool participants at multiple universities and Amazon Mechanical Turk ("MTurk") workers — two groups that make up a large proportion of psychology research participants — we can improve generalizability as well as explore any preference discrepancies between undergraduate participants and MTurkers. Based on the pilot results, we honed our questions to more adequately measure participants' preferences, with as little interference or bias as possible. Finally, including a larger selection of minimal risk base studies improves the generalizability of the results to other minimal risk studies.

Method

Participants

We aimed to collect data from both online platforms and undergraduate student populations. In computing our target sample size, we chose a simple target analysis — estimating proportions (e.g., proportion of participants who chose a response above the "Indifferent" midpoint on a given question). Specifically, we aimed for enough precision such that the width of our 95% confidence interval would be at least as narrow as +/- 3% when the proportion is equal for all categories (precision is higher for uneven proportions). To achieve this, our precision analysis suggests that our target sample size should be 1,317 participants — see https://osf.io/v68hu/ for R code and the supplementary materials for

details on this calculation. We aimed for a sample of (1) approximately 50% online participants and 50% undergraduate student participants (2) from at least 3 universities (for the student sample), and (3) at least 8 different base studies. However, it was difficult to be sure we would be able to achieve this breakdown at the subgroup level because we relied on cooperation with other researchers (see below). To ensure we would be able to compare online and undergraduate participants' views, we set a maximum of 60% of participants from either population. Although the exact breakdown of online vs. student participants might vary within this range, we planned to collect data from at least 1,600 participants before exclusions (see supplement on precision calculation for details on how we arrived at this number). Data was collected by base study (i.e., we continued to seek out new base studies and collect the full sample size agreed upon for that study) and we stopped seeking out new base studies when, after completing data collection for a base study, these targets had been reached. After that, we finished collecting the planned sample for any base studies that were still ongoing, but did not begin any new data collection. An explanation of how participants were compensated for their time can be found in the supplementary materials.

Study Selection

We used two main strategies to acquire base studies for our sample. First, we asked researchers whom we know personally or heard about who had extra time in their studies to add our questions to the end of their survey, as we did for Pilot Study A. Second, we offered to run an agreed upon number of participants ourselves using other people's base studies, either on MTurk or on the UC Davis student subject pool, in exchange for adding our questions to the end of their study. These scenarios could happen alone or in combination. That is, for some base studies, it is possible only the base-study researcher collected data, only our team collected data, or both teams collected data. To find these

researchers, we planned to use the Study Swap website (osf.io/meetings/studyswap/), social media, and personal contacts.

We decided which studies to include in our sample on a case-by-case basis. The base studies had to meet the following criteria: (1) a minimal-risk study where all the data would be collected in a single session, either online or on a local computer, (2) the study was run in English, and, if it used an undergraduate subject pool, it was run at a college or university where English is the primary language of instruction, (3) the participants were recruited from either college/university subject pools or online platforms and meet our inclusion criteria (see below), (4) feasibility constraints — e.g., whether we had the resources to run participants on our end, whether the IRB approval would be easy to obtain, etc.; (5) progress of our sample size goals — e.g., if we had met our goal for student or online participants, we stopped collecting data from that population; (6) time constraints — we would be able to complete data collection for the study within the time frame allotted for the project; (7) sample size — the study would provide a minimum of 50 participants; and (8) base study materials would be made publicly available.

Sample Selection Criteria

Participants had to speak English and be at least 18 years old. They also had to qualify for and complete the base study that preceded ours, so our study inclusion criteria included the inclusion criteria used by each base study to which we appended our meta-study. For example, if one of the studies selected only first-generation college students, or only women, this was also a criterion to participate in our meta-study for that subsample.

We had funds to collect data on MTurk and resources to collect data from the UC Davis undergraduate subject pool, so data collection conducted by us came from one of these two populations. For MTurk samples recruited by us, we planned for participants to meet the following criteria: (1) be located in the

United States; (2) have a Human Intelligence Task (HIT) approval rate of 90% or higher[4]; and (3) have at least 10 HITs approved. MTurk samples recruited by partner researchers running base studies would follow that team's criteria.

It was also possible that some data would be collected by the base study researchers, and these data could be collected from other colleges' or universities' subject pools, or online platforms other than MTurk, like Prolific (https://prolific.co/). In these cases, we planned for the selection criteria for participants (beyond the requirement that participants speak English and be at least 18 years old) to be decided by the base study researchers.

Meta-study

The meta-study asked participants to consider an anonymized version of the data they had just provided in the base study, and imagine a series of hypothetical situations in which researchers use different research practices on their data. Specifically, we asked them their opinions on the eight practices shown in Table 1. We honed the wording of these questions using the data and feedback from our pilot studies, which we describe in detail in the supplementary materials. The full text for the questions in Table 1 can be found at https://osf.io/p8n9w/.

Table 1.1. Description of Survey Questions.

Question Number	**Question Topic**	**Number of Versions**
1	*p*-hacking or cherry-picking results	2 versions
2	selective reporting of studies (filedrawering)	2 versions

[4] A HIT, or "human intelligence task", is a task available for Amazon Mechanical Turk workers. A workers' HIT approval rate is the proportion of tasks that have been approved by the requester. The authors consider 90% to be a reasonable cutoff to ensure high quality data.

3	HARKing	1 version
4	fraud	1 version
5	direct replication	1 version
6	open methods	2 versions
7	open access publication	1 version
8	data sharing	2 versions

Note. Each participant saw only one version of each question. See materials for a full description of the question wording, versions, and response options.

Our goal was to ask the questions in a way that is not leading. When we could not find a way to do this while still providing a clear description of the practice (i.e., for Questions 1, 2, and 8 — see table 1 for a list of which questions correspond to which research practice), we wrote two different versions of the question reflecting the tradeoff between providing a fuller but potentially leading description of the practice, and providing a vaguer but less valenced description of the practice. For Question 6, we also created two versions: the "positive" version of the question, asking participants how they would feel if the researchers shared enough details about their materials and procedures for others to conduct a replication study, or the "negative" version, which asks participants how they would feel if researchers did not share enough details. If the answers differ by version, which the pilot studies suggested might happen, we would have estimates of the distribution of responses to these practices for two different, but hopefully reasonable, ways to ask the same question. In other words, these two versions provide a kind of robustness check across variations that we hope capture similar phenomena. For questions 1, 2, 3, and 4, the research practices were described in simple terms, and participants were asked to rate each practice on a 5-point scale with anchors at -2 ("definitely not acceptable") through 0 ("Indifferent") to +2 ("definitely acceptable").

Question 5 asked participants their opinion about whether researchers should attempt to replicate a finding before publishing it or simply move on to a new project. With this question, we hoped to make the tradeoffs involved in conducting a direct replication (vs. not conducting one) clear, without leading participants towards one answer or the other. Participants answered on a 5-point scale with anchors being "strongly prefer that the researchers move on to their next project", "slightly prefer that the researchers move on to their next project", "indifferent", "slightly prefer that the researchers replicate the study", and "strongly prefer that the researchers replicate the study".

Questions 6, 7, and 8 asked participants to consider situations where researchers can choose to use open science practices. For Question 6 ("open methods"), which is about whether researchers should share their materials and procedures, participants were asked to rate this practice on a 5-point scale with anchors at -2 ("feel strongly that the researchers should **not** do this") through 0 ("indifferent") to +2 ("feel strongly that the researchers **should** do this"). There were two versions of this question. The positive version describes researchers providing all necessary information for replication, while the negative version (reverse scored) describes not providing enough information.

For Question 7, participants were asked whether they have a preference for where the article reporting the results of the base study should be published: an open access journal vs. a pay-walled journal. Participants answered on a 5-point scale with anchors being "strongly prefer that it cost about $30 to read the article", "slightly prefer that it cost about $30 to read the article", "indifferent", "slightly prefer that the article be free to read", and "strongly prefer that the article be free to read." The value of $30 dollars is typical of several top journals in Psychology. Based on feedback on our pilot materials, we also added a clarification statement so respondents understand that the value paid for the article does not go to the authors of the article — a reasonable but false assumption — but to the publisher.

For Question 8, we asked two versions of the question: one that explicitly stated reasons why a researcher may or may not want to share their data ("reasons provided"), and one that did not ("neutral"). The reasons-provided version spells out the main reasons for and against data sharing. We developed this list of reasons by consulting published work on researchers' stated reasons for sharing or not sharing data (Washburn et al., 2018). The neutral version of this question asks participants to consider potential reasons before answering, and makes it clear that valid reasons exist both for and against data sharing. Participants answered on a 5-point scale with anchors at -2 ("feel strongly that the researchers should **not** do this") through 0 ("indifferent") to +2 ("feel strongly that the researchers **should** do this").

For all of the questions, we used a 5-point response scale. This was changed from a 7-point scale in Pilots B and C, as we believe this better reflects the granularity of judgment that is reasonable to expect from research participants. Moreover, having fewer response options gives us more precision when estimating the proportion of people who choose each response option. We also changed the order and anchors for some of the questions. For questions where it makes sense to have a negative and positive end of the scale (e.g., "researchers should **not** do this" vs. "researchers **should** do this") we kept the numbering (-2 to +2) with anchors at the ends and midpoint. However, some of the questions represent trade-offs (e.g., whether to publish open access vs. behind a paywall) which have no clear "positive" or "negative" end. Therefore, we labeled all 5 points for Questions 5 (direct replication) and 7 (open access publishing) with words rather than numbers, to avoid inadvertently conveying that one end of the scale is more desirable than the other (e.g., higher numbers, or positive numbers).

For questions which have two versions, participants were randomly assigned to answer one or the other. Random assignment was independent between questions; e.g., a participant who was assigned to the neutral version of the data sharing question could be assigned to either the neutral or the

reasons-provided version of the selective reporting question. Furthermore, the order of the eight questions was randomized.

Finally, we asked additional questions for potential exploratory analyses. First, we asked about demographics, including gender, race and ethnicity, year of birth, education, proximity to science, and the number of psychology studies the participant participated in during the previous two weeks. We also measured trust in psychological science with three statements ("Findings from psychology research are trustworthy," "I have very little confidence in research findings from psychology" (reverse-scored), and "I trust psychology researchers to do good science") which participants were asked to rate on a 7-point scale from (1) "strongly disagree" to (7) "strongly agree." We also asked participants "Have you heard of the replication crisis in psychology?" (Yes/No). If participants answered yes, we then asked them "Please describe what you have heard about the replication crisis:" and provided an open-ended text box for their response. Although we did not have planned analyses that use these additional questions, they were collected to allow for exploratory analyses both by the authors and others who wish to reuse the data.

Data Exclusion Criteria

The survey included one open ended attention/comprehension check. We planned for these answers to be coded by an independent coder, who would be blind to how they related to the rest of the data, as "appropriate," "inappropriate," and "unclear." Only "inappropriate" answers would be excluded.

Results

Sample

The data were collected between January and October of 2021, yielding a total of 1,990 observations before exclusions from 8 different base studies. After performing the preregistered exclusions, we

obtained a final sample of 1,873 participants — 40% from participants in Amazon Mechanical Turk studies (5 studies) and 60% from university subject pool study participants (3 studies across 4 subject pools) — with the breakdown described in Table 2.

57.9 % of participants described themselves as female, 40.4% described themselves as male, 0.8% self-identified as non-binary or a third gender, 0.3% preferred to self-describe, and 0.5% preferred not to report their gender. The median year of birth for participants was 1999 (IQR = 14; range = 1946-2003). Participants could select multiple race and ethnicity categories; 51.8% identified as white, 29.2% as Asian, 13.5% as Hispanic or Latino or Chicano or Puerto Rican, 8.0% as Black or African American, 1.6% Middle Eastern or North African, 0.8% American Indian or Alaska Native, 0.7% Native Hawaiian or Pacific Islander, and 1.4% said they had another identity; 7.5% of participants declined to self-identify on race and ethnicity.

Table 1.2. Sample size, population, and short description of each base study.

Base Study	**Sample size**			**Population**	**Study description**
	Before exclusions	After preregistered exclusions only	After non-preregistered (strict) exclusions		
BS01	500	499	437	Sacramento State University and University of California, Davis Subject Pools	A study about individuals' reactions to marginalized individuals in positions of power
BS02	390	389	363	University of Pennsylvania and University of California, Davis Subject Pools	A study exploring the reasons that people overassess experts' abilities.
BS03	237	237	227	Princeton University Subject Pool	A study about friendship formation and related attitudes

BS04	162	145	93	Amazon Mechanical Turk Workers	A study about interviews in false confessions documentaries and how they influence laypeople's perceptions of a confession
BS05	252	201	115	Amazon Mechanical Turk Workers	A study about perspective taking of climate refugees
BS06	106	100	86	Amazon Mechanical Turk Workers	A study about the relationships between Dark Personality, Self-Control and Aggression.
BS07	130	123	99	Amazon Mechanical Turk Workers	A study testing interindividual variability in free- and cued-recall memory performance
BS08	213	179	117	Amazon Mechanical Turk Workers	A study about people's perceptions of the most moral, least moral, and morally average people they personally know.
Total	1,990	1,873	1,537	--	--

Deviations from Planned Recruiting Strategy

In the base-study recruiting phase, we communicated with several potential base studies, and received responses from many researchers willing to collaborate. Those not mentioned here did not meet our inclusion criteria, or the collaborating researcher later decided against following through for a variety of reasons, and we never reached the data collection stage with these studies. The one exception to this was a study for which we did not have enough information to realize it did not meet our inclusion

criteria until after data collection was completed, so although we do have the participants' data for this other study, we are not including it or its data here. One other slight deviation from our recruiting plan was that BS03 did not have an initial agreed-upon sample size, but an end date (October 31st) when the collaborating researchers had preregistered to check whether they had enough data to perform their analyses; this served as the stopping rule for our part of the study. Finally, some MTurk studies ended up with a few more observations than we aimed to collect, and BS02 had 10 fewer observations than agreed due to reaching the end of their semester.

Analyses

Our primary analyses examine the distribution of participants' responses, which we examined using descriptive statistics presented in Tables 3 and 4. We also present the corresponding visualizations of the distributions of responses in Figures 1 and 2. We first report the descriptives using the preregistered exclusion criteria, starting with results for the combined samples (Table 4 and Figure 1), then results for each question version separately (for questions that had more than one version; Table 5 and Figure 2). We also report the exploratory analyses we outlined in the stage 1 manuscript.

Despite our relatively strict preregistered exclusion criteria, we nevertheless found certain patterns of results suspicious, especially in the percentage of participants who expressed neutral or positive views of scientific fraud. Because of this, in our "exploratory analyses not described in the preregistration" section, we repeat most of the analyses with non-preregistered but stricter exclusion criteria, which we fully describe at the beginning of the section. With these stricter criteria, we aimed to provide an alternative test of our research questions, and we suspect that some readers might feel these are more appropriate results to interpret, given our possible data quality issues. We have clearly marked these results as exploratory. All the code used in the analyses and figures presented here can be found at

https://osf.io/34gbv/; this follows and expands the stage 1 preregistered analyses which can be found at https://osf.io/ytdek/.

Preregistered main analyses

For each question, we were interested in the proportion of participants that selected a negative, neutral, and positive response. Table 3 details the response scales for each question and its labels.

Table 1.3. Response scale anchors for each question.

Question Number	Question Topics	Response scale anchors
1, 2, 3, 4	*P*-hacking, filedrawering, HARKing, fraud	-2: Definitely not acceptable 0: Indifferent +2: Definitely acceptable
5	direct replication	-2, -1: [Strongly/Slightly] prefer that the researchers move on to their next project 0: Indifferent +2, +1: [Strongly/Slightly] prefer that the researchers replicate their study
6, 8	open methods, data sharing	-2: Feel strongly that the researchers should **not** do this 0: Indifferent +2: Feel strongly that the researchers **should** do this
7	open access publication	-2, -1: [Strongly/Slightly] prefer that it cost about $30 to read the article 0: Indifferent +2, +1: [Strongly/Slightly] prefer that the article be free to read

Note. Anchor numbers were not shown for questions 5 and 7.

Overall results: Preregistered exclusion criteria. As can be seen in Table 4 and Figure 1, for all eight questions, a clear majority of participants selected a response on one side of the neutral point. That is, between 68% and 81% of participants reported that *p*-hacking, filedrawering, HARKing, and fraud are not acceptable, that they prefer that researchers share their methods and data, that replication is preferable to moving on without replicating, and that publishing open access is preferable to publishing

behind a paywall. Fewer than 15% of participants selected the neutral option ("indifferent") for each question, except for the open access publishing question, for which 25% of participants selected the neutral option. Participants' preferences/opinions were most pronounced for replication and fraud, though a troubling percentage of participants (19%) expressed indifferent or positive attitudes about fraud. We return to this unexpected pattern of results below, in the non-preregistered section.

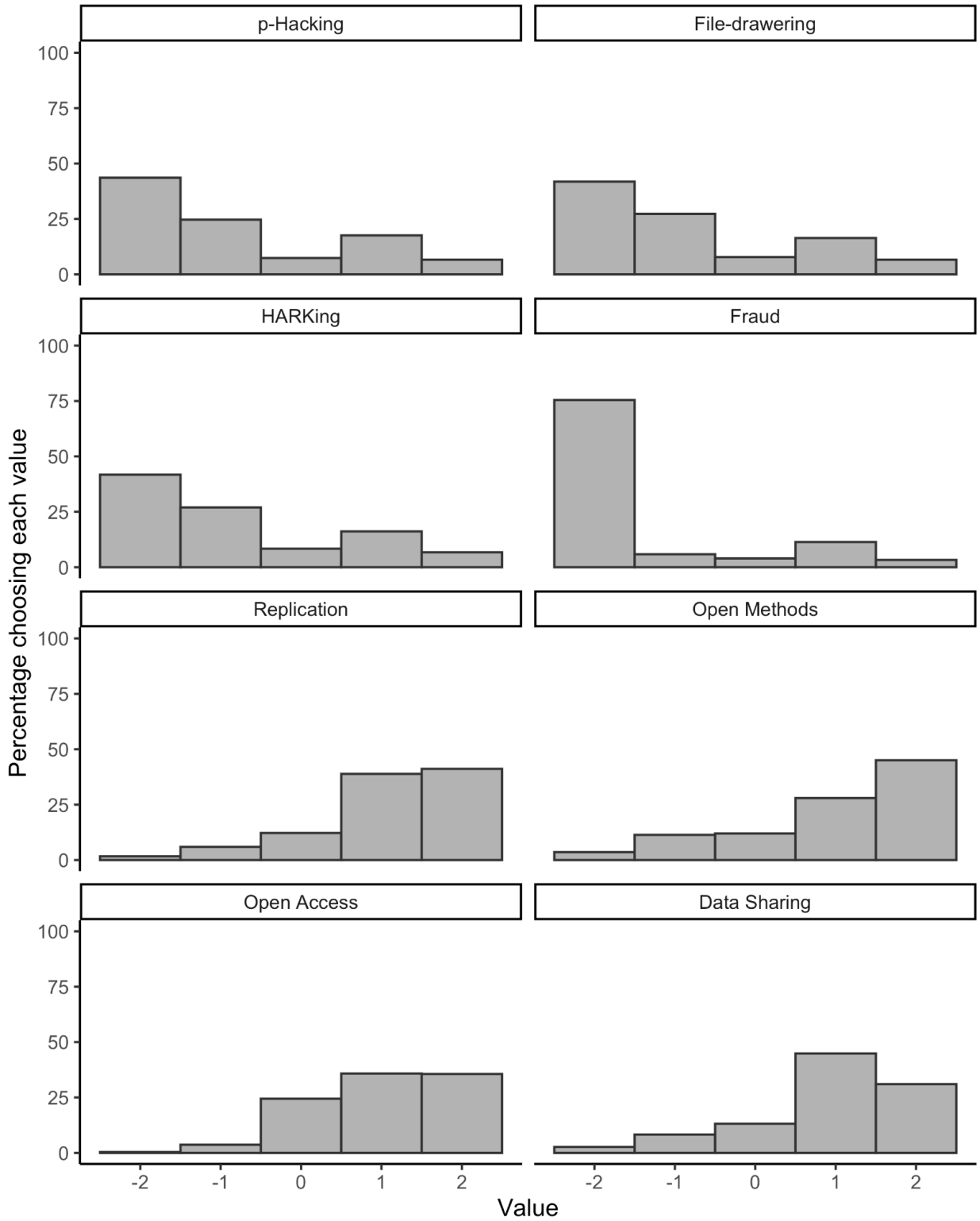

Figure 1.1. Distribution of participants' answers for each question. For the top four panels, negative numbers indicate that participants found the practice unacceptable while positive numbers indicate they found the practice acceptable. For the bottom four panels, higher numbers indicate more support for the practice. *N* = 1,873. See also Table 4.

Table 1.4. Descriptive statistics for each question, with preregistered exclusions, collapsing across question version.

Question	**Median (IQR)**	**Category**	**% [LL, UL]**
Question 1: *p*-hacking / cherry-picking results	-1 (1)	Not acceptable	68.3 [66.2, 70.5]
		Indifferent	7.42 [5.29, 9.56]
		Acceptable	24.2 [22.1, 26.4]
Question 2: selective reporting of studies / filedrawering	-1 (2)	Not acceptable	69.2 [67.1, 71.3]
		Indifferent	7.79 [5.71, 9.94]
		Acceptable	23.0 [20.9, 25.2]
Question 3: HARKing	-1 (2)	Not acceptable	68.7 [66.6, 70.9]
		Indifferent	8.38 [6.30, 10.5]
		Acceptable	22.9 [20.8, 25.1]
Question 4: fraud	-2 (0)	Not acceptable	81.3 [79.6, 83.1]
		Indifferent	4.00 [2.30, 5.75]
		Acceptable	14.7 [13.0, 16.4]
Question 5: direct replication	1 (1)	Move on	7.69 [5.98, 9.49]
		Indifferent	12.2 [10.5, 14.0]
		Replicate	80.1 [78.4, 81.9]
Question 6: open methods	1 (2)	Rs should not do this	14.9 [13.0, 17.0]
		Indifferent	12.0 [10.0, 14.0]

		Rs should do this	73.0 [71.1, 75.0]
Question 7: open access publication	1 (2)	Paywall	4.11 [2.08, 6.22]
		Indifferent	24.5 [22.4, 26.6]
		Free	71.4 [69.4, 73.5]
Question 8: data sharing	1 (1)	Rs should not do this	10.9 [9.08, 12.9]
		Indifferent	13.2 [11.3, 15.1]
		Rs should do this	75.9 [74.0, 77.8]

Note. *N* = 1,873 for all questions. Multinomial 95% confidence intervals [LL, UL] using the Sison-Glaz method. "Rs" in questions 6 and 8 refers to "researchers". Each response category except "Indifferent" collapses across two response options on the 5-point scales.

Results for question versions: Preregistered exclusion criteria. For the four questions with multiple versions, we examined the descriptive statistics and distribution of responses separately for each version (Table 5 and Figure 2). As preregistered, we did not conduct inferential tests comparing the two versions of each question, as we did not have hypotheses regarding the effect of version. Instead, we provide the results separately for each version to provide a sense of the robustness of results across question formats.

As shown in Table 5 and Figure 2, participants reported more extreme views about *p*-hacking and filedrawering when these practices were described as motivated (i.e., "only reported the results/studies that came out the way they predicted"), compared to participants who saw these practices described in a neutral manner (i.e., "only reported some of the results/studies"). However, in both versions, the majority of participants rated these practices as not acceptable. For open methods, the question framing ("provided a lot of details [...] other researchers could easily conduct a replication" vs. "did not

provide a lot of details [...] other researchers could not easily conduct a replication"; responses to the second version were reverse-scored) led to slightly different distributions, with more participants supporting closed methods in the second version. Finally, for the data sharing question, one version of the question did not present any reasons why researchers might choose to share or not to share their data, while the other version presented reasons for both choices. Across both versions, most participants selected responses in favor of data sharing, but participants who read about reasons for and against data sharing had slightly less extreme views in favor of data sharing than did participants who did not read reasons.

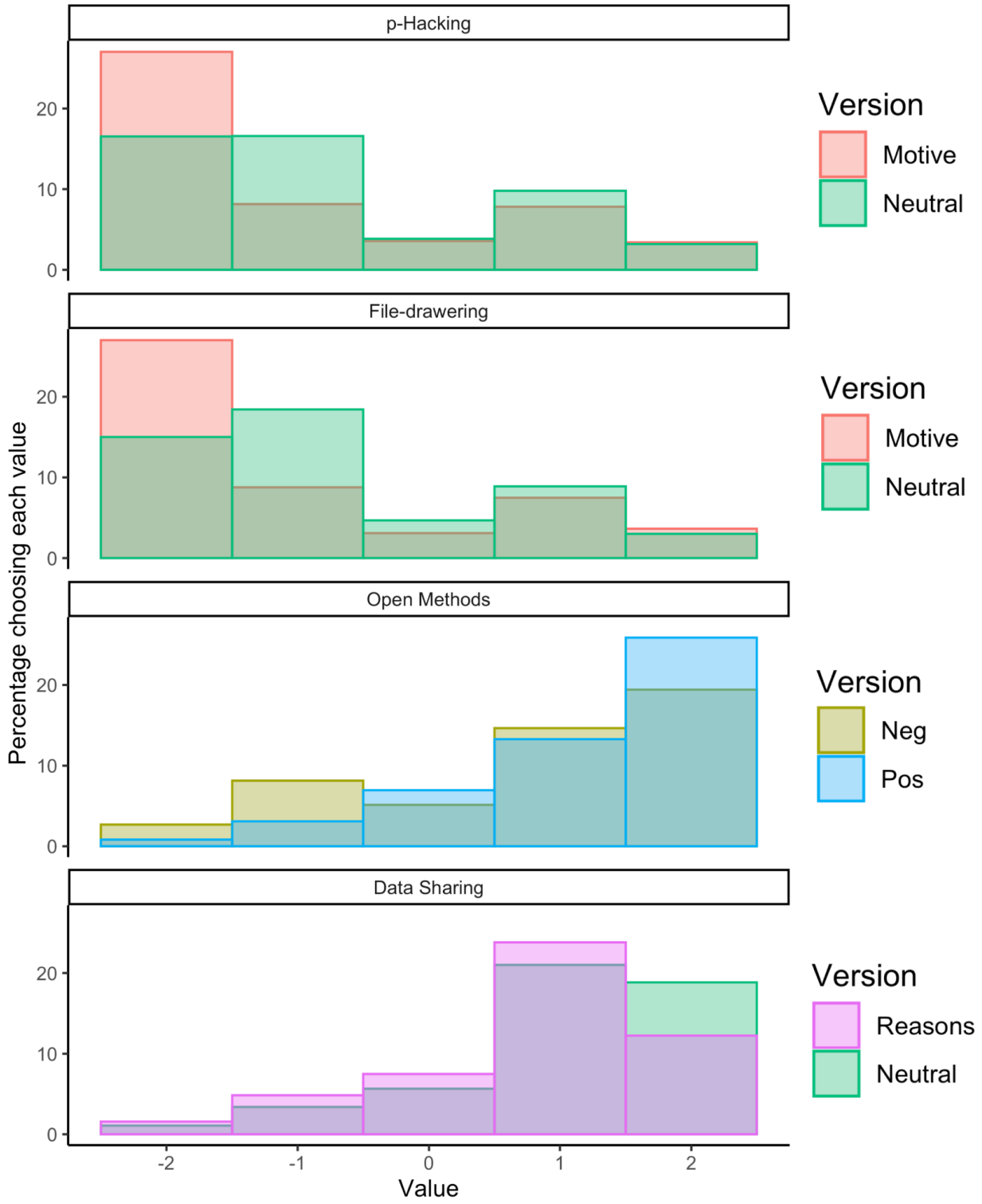

Figure 1.2. Distribution of participants' answers for each question, by question version, for the four questions with two versions. For the top two panels, negative numbers indicate that participants found the practice unacceptable while positive numbers indicate they found the practice acceptable. For the bottom two panels, higher numbers indicate more support for the practice. For *p*-hacking and filedrawering, the neutral version described the behavior only (i.e., researchers "only reported some of the results" or "did not report all of the studies they ran"), while the motive version implied motivated reasons behind the selective reporting of results or studies (e.g., "only reported

the results/studies that came out the way they predicted"). The positive and negative versions of the open methods question were phrased as "[researchers] provided [vs. did not provide] a lot of details about how they did the study. Therefore, other researchers could [vs. could not] easily conduct a replication…". The neutral data sharing question asked whether participants thought "researchers should share the dataset when they publish their results" while the reasons version asked the same but provided some reasons why researchers may or may not want to share their data (e.g., concerns about scooping or making it possible for others to verify their work). See Table 5 for more detailed results and sample sizes.

Table 1.5. Descriptive statistics for questions with two versions, with preregistered exclusions only.

Question	**Category**	**% [LL, UL]**	
Question 1: *p*-hacking / cherry- picking results		**Neutral (%)** (*Mdn* = -1, *IQR* = 3) n = 934	**Motive (%)** (*Mdn* = -2, *IQR* = 2) n = 939
	Not acceptable	69.5 [66.6, 72.7]	73.5 [70.7, 76.4]
	Indifferent	8.18 [5.23, 11.3]	7.05 [4.25, 9.99]
	Acceptable	22.3 [19.3,25.4]	19.5 [16.7, 22.4]
Question 2: selective reporting of studies/ filedrawering		**Neutral (%)** (*Mdn* = -1, *IQR* = 2) n = 950	**Motive (%)** (*Mdn* = -2, *IQR* = 2) n = 923
	Not acceptable	66.8 [63.9, 69.9]	71.6 [68.8, 74.6]
	Indifferent	9.37 [6.42, 12.5]	6.18 [3.36, 9.18]
	Acceptable	23.8 [20.8,26.9]	22.2 [19.4, 25.2]
Question 6: open methods		**Rs use open methods (%)** (*Mdn* = 2, *IQR* = 1) n = 907	**Rs use "closed" methods (%)** (*Mdn* = 1, *IQR* = 2) n = 966
	Rs should not do this	7.83 [5.29, 10.5]	68.1 [65.2, 71.1]
	Indifferent	13.9 [11.4, 16.6]	10.2 [7.35, 13.3]

	Rs should do this	78.3 [75.7, 81.0]	21.6 [18.7, 24.6]
Question 8: data sharing		**Neutral (%)** (*Mdn* = 1, *IQR* = 1) n = 926	**Reasons provided (%)** (*Mdn* = 1, *IQR* = 1) n = 947
	Rs should not do this	8.96 [6.48, 11.5]	12.9 [10.1, 15.8]
	Indifferent	11.3 [8.86, 13.8]	15.0 [12.2, 17.9]
	Rs should do this	79.7 [77.2, 82.2]	72.1 [69.4, 75.0]

Note. The median and interquartile range reported for Question 6, closed methods version, is after the question was reverse scored. Multinomial 95% confidence intervals [LL, UL] using the Sison-Glaz method. "Rs" in questions 6 and 8 refers *to "researchers".* Each response category except "Indifferent" collapses across two response options on the 5-point scales.

Exploratory Analyses

Exploratory analyses described in preregistration. In the preregistration, we described that we would also conduct exploratory analyses examining: 1) variance partitioning of responses to identify the proportion of variance that was between-studies, and 2) differences between the MTurk and subject pool participants' responses.

Variance partitioning. As shown in Table 6, the variance in each outcome that can be attributed to between-study variability ranged between 0.2 to 15.7% with preregistered exclusions. For questions 1-4, related to questionable research practices, between-study variance seemed to be higher, ranging from 11.1% to 15.7%, while it was lower for positive research behaviors like replication, open access publishing, and data sharing. The open methods question is a notable exception, although that might be related to the fact that, while the positive version of this question asks about a positive research behavior — making one's methods open — the negative version asks about a questionable behavior — making one's methods "closed". This pattern is less pronounced but still clearly visible after performing

non-preregistered strict exclusions (Table 6, right column), potentially indicating more consensus between studies for positive practices than questionable practices. We should note, however, that this between-study variance includes variance due to the different platforms (Amazon MTurk vs. university subject pools), as each study was conducted on only one of these platforms. Thus, between-study variance could be driven by between-platform variance. We examine differences between the responses from MTurk and subject pool participants in the next section.

Table 1.6. Variance partitioning. Proportion of between-study variance in each outcome variable.

Question	**Between-study variance** (after preregistered exclusions only)	**Between-study variance** (after stricter, non-preregistered exclusions)
Question 1: *p*-hacking / cherry-picking results	11.1%	4.7%
Question 2: selective reporting of studies/filedrawering	11.9%	3.5%
Question 3: HARKing	12.7%	4.1%
Question 4: fraud	15.7%	3.6%
Question 5: direct replication	1.6%	1.5%
Question 6: open methods	8.4%	3.7%
Question 7: open access publication	1.3%	2.1%
Question 8: data sharing	0.2%	0.7%

MTurk vs. subject pool. For all questions except data sharing (Question 8), there were statistically significant differences between the MTurk and subject pool samples (all *ps* < .001; see supplemental

materials Table S4 for detailed results of Pearson's Chi-squared tests). These results suggest that students may hold more extreme views than MTurk participants when it comes to research practices, more strongly disapproving of questionable practices and supporting open practices. The same pattern holds after implementing the non-preregistered, strict exclusions (which are described in detail below). The distribution of responses (after strict exclusions) is shown in Figure 3 and reported in Table 7. See Figure S17 for an analogous visualization to Figure 3 but with preregistered exclusions only.

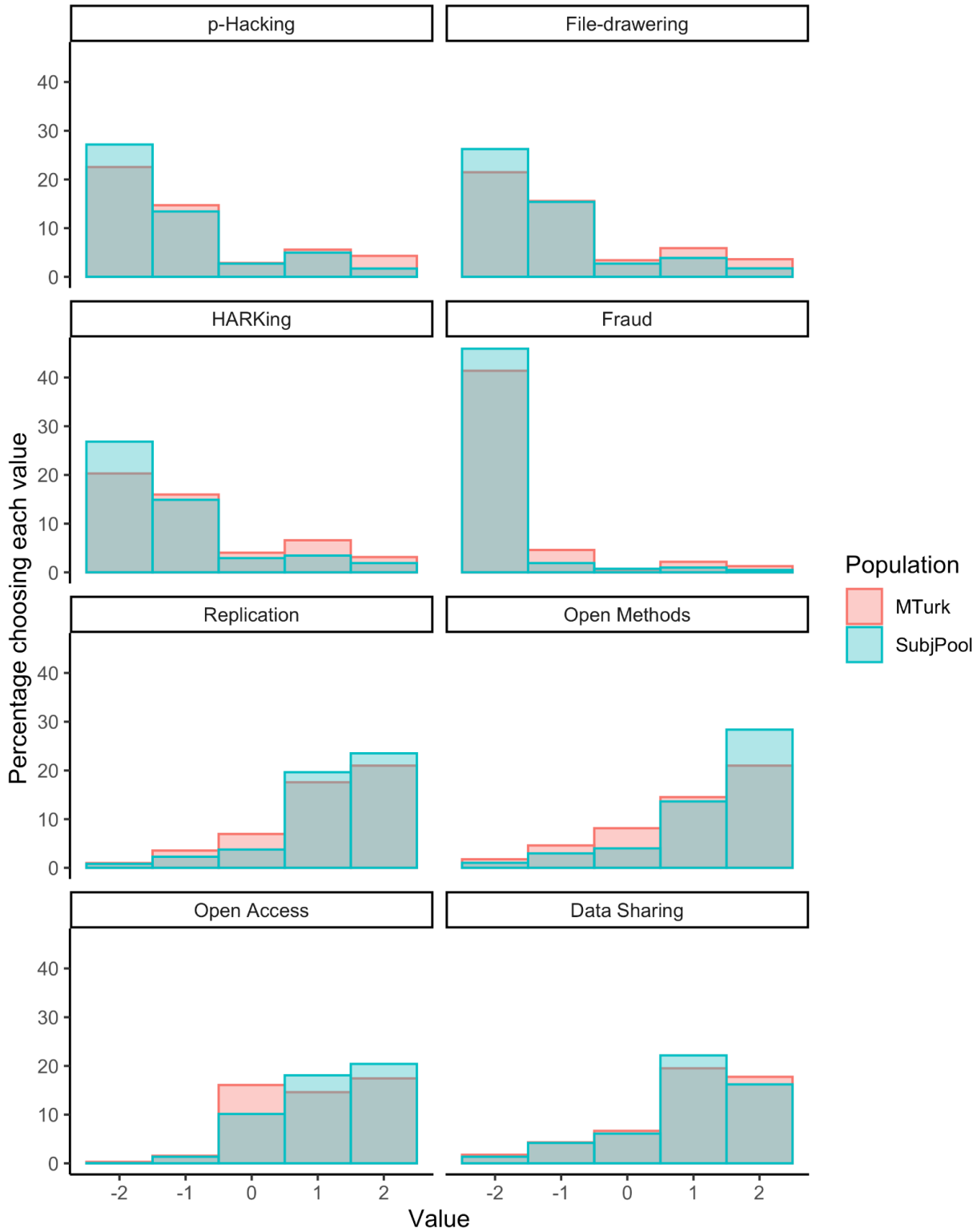

Figure 1.3. Distribution of university subject pool and MTurk participants' responses for all 8 questions, after non-preregistered, strict exclusions. For the top four panels, negative numbers indicate that participants found the practice unacceptable while positive numbers indicate they found the practice acceptable. For the bottom four panels, higher numbers indicate more support for the practice. N = 1,537. See also Table 7.

Table 1.7. Descriptive statistics for all eight questions, with non-preregistered (strict) exclusions, by population (university subject pool students vs. Amazon Mechanical Turk workers).

Question	Category	% [LL, UL]	
		Subject Pool (%) n = 1,027	**MTurk (%)** n = 510
Question 1: *p*-hacking / cherry-picking results	Not acceptable	81.2 [79.0, 83.6]	74.5 [71.0, 78.4]
	Indifferent	5.45 [3.21, 7.83]	5.69 [2.16, 9.60]
	Acceptable	13.3 [11.1, 15.7]	19.8 [16.3, 23.7]
Question 2: selective reporting of studies/ filedrawering	Not acceptable	83.3 [81.1, 85.5]	74.1 [70.4, 77.9]
	Indifferent	5.45 [3.31, 7.68]	6.86 [3.14, 10.6]
	Acceptable	11.3 [9.15, 13.5]	19.0 [15.3, 22.8]
Question 3: HARKing	Not acceptable	83.4 [81.3, 85.6]	72.5 [68.8, 76.5]
	Indifferent	5.84 [3.70, 8.04]	8.04 [4.31, 12.0]
	Acceptable	10.7 [8.57, 12.9]	19.4 [15.7, 23.3]
Question 4: fraud	Not acceptable	95.6 [94.5, 96.8]	92.0 [89.8, 94.2]
	Indifferent	1.46 [0.399, 2.66]	1.18 [0.00, 3.39]
	Acceptable	2.92 [1.85, 4.12]	6.86 [4.71, 9.08]
Question 5: direct replication	Move on	6.23 [4.28, 8.25]	9.02 [5.69, 12.7]
	Indifferent	7.50 [5.55, 9.52]	13.9 [10.6, 17.6]
	Replicate	86.3 [84.3, 88.3]	77.1 [73.7, 80.8]
Question 6: open methods	Rs should not do this	7.98 [5.94, 10.2]	12.7 [9.02, 16.8]
	Indifferent	7.98 [5.94, 10.2]	16.3 [12.5, 20.3]

	Rs should do this	84.0 [82.0, 86.2]	71.0 [67.3, 75.0]
Question 7: open access publication	Paywall	2.82 [0.29, 5.41]	3.73 [0.00, 8.12]
	Indifferent	20.3 [17.7, 22.8]	32.2 [28.0, 36.5]
	Free	76.9 [74.4, 79.5]	64.1 [60.0, 68.5]
Question 8: data sharing	Rs should not do this	11.1 [8.67, 13.7]	12.2 [8.63, 16.0]
	Indifferent	12.2 [9.74, 14.8]	13.3 [9.80, 17.2]
	Rs should do this	76.7 [74.3, 79.3]	74.5 [71.0, 78.3]

Note. All differences between MTurk & subject pool are significant (Pearson's Chi-squared test) at $p < .001$ for all questions except question 8 (not significant). See supplemental materials Table S4 for details. Multinomial 95% confidence intervals [LL, UL] using the Sison-Glaz method. "Rs" in questions 6 and 8 refers to "researchers". Each response category except "Indifferent" collapses across two response options on the 5-point scales.

Exploratory analyses not in the preregistration. As mentioned above, the finding that a surprisingly high proportion of participants reported having neutral (4%) or positive (15%) views about fraud concerned us. To be clear, the item wording did not leave much room for misinterpretation. Specifically, we asked participants to "Imagine that the researchers changed the data to make the results come out the way they predicted (in other words, committed scientific fraud). Do you think this would be acceptable?" In our pilot studies, we found that only around 2 to 10% of participants expressed positive or neutral attitudes towards fraud. Similarly, Pickett and Roche (2018) found that 96% of a sample of MTurk participants believe data fabrication and fraud are morally unacceptable, and 91% of a representative US sample believe data fabrication and fraud should be a crime. Based on these previous results and on common sense, we were skeptical that 19% of participants truly have a neutral or positive view of research fraud.

Even more alarming, when we look at the distribution of responses for the MTurk and subject pool participants separately, we find that a whopping 32.8% of MTurk participants expressed neutral or

positive attitudes towards fraud. In the subject pool subsample, only 9% of participants expressed neutral or positive attitudes towards fraud, in line with our pilot studies and expectations. Thus, we suspect that the exclusion criteria we preregistered were not sufficiently strict, and that many non-serious responses remained in the MTurk samples, even after the preregistered exclusions.

Table 1.8. Percentage of participants expressing neutral or positive attitudes towards scientific fraud, divided by population (MTurk vs. university subject pool), after preregistered exclusions only and after non-preregistered (strict) exclusions.

Population	**% reporting neutral or positive attitudes towards scientific fraud**	
	After prereg exclusions only	After non-prereg, strict exclusions
Subject Pool	9.3	4.4
MTurk	32.8	8.0

Of course, estimating the proportion of participants who believe fraud is acceptable was one of the aims of the current study, so deciding to change our exclusion criteria after seeing the results seriously increases the risk of bias of any subsequent analyses. We must seriously consider the possibility that a non-trivial fraction of participants believe it is acceptable for researchers to commit scientific fraud, at least in the context of simple, low-risk psychology studies, and we address this possibility in the discussion. Nevertheless, if we are correct that many of the participants expressing neutral or, especially, positive views of fraud are not responding seriously, including them in our analyses affects the accuracy of all other estimates reported here. Thus, we believe it is prudent to also explore stricter exclusion criteria and examine the results for all 8 questions in this smaller subsample.

In exploring other possible ways to identify low quality or non-serious responders, in addition to the preregistered exclusion criteria, we examined our data from many different angles. We considered how

to use the open-ended responses to filter out non-serious responders; we examined the distributions to identify unexpected bumps (e.g., we found that, before reverse-scoring, all questions and versions showed an unexpected bump in the distribution of responses for the response option corresponding to "+1" on the -2 to +2 scale, but only in the MTurk samples); and we examined the effects of various exclusion criteria on the number of participants excluded, and the proportion of remaining participants reporting neutral and positive attitudes towards fraud. That is, our process was iterative and very much data-dependent, and thus the results of all analyses after applying these non-preregistered exclusion criteria should be taken as susceptible to our biases. Limitations related to these decisions are discussed below.

We settled on the following strict exclusion criteria. First, the first author went through each participant's three open-ended questions (see materials for question wording). The first author used her judgment to mark participants as "suspicious" or not. Importantly, the first author only had access to these three answers and none of the participants' other responses. Examples of "suspicious" answers (beyond the preregistered criteria) include those using ungrammatical English, answered in all capital letters, nonsensical comments (e.g., "GOOD"), or answers that were copy-pasted from internet search results. This criterion excluded 181 participants (9.7%). Second, we calculated the standard deviation for each person's answers to the 8 main questions (before reverse-scoring). A person who gave the same answer to each question would have a SD of 0, while the most extreme response possible would have a SD of 2.14,. The median SD for observations in our dataset was 1.51. We chose a cut-off of 0.8 and excluded participants with smaller SDs, which excluded 247 participants (13.2%). Finally, we also excluded participants who took less than 2 minutes to complete the whole study, our subjective judgment of the minimum amount of time it would take someone to skim and click through the survey while still possibly providing meaningful answers. The time criterion mostly applied to MTurk

participants, as we did not have the time participants spent on just the meta-study portion of their study for subject pool participants (and for one MTurk study). This criterion excluded 25 participants (1.3%).

When combined, all three criteria excluded a total of 336 participants, or 17.9% of the total sample after our preregistered exclusions, with 12.7% being from MTurk and 5.2% being from university subject pools. The resulting sample sizes for each base study after applying these strict exclusion criteria can be found in Table 2, and the proportion of participants reporting neutral or positive opinions of fraud before and after exclusions can be seen in Table 8.

The results for our main research question, exploring the distribution of responses for each of the 8 questions after these strict exclusions, can be found in Table 9 and Figure 4. As these results show, the distributions look similar to those found with the less strict, preregistered exclusion criteria, but they tend to be more extreme (greater consensus). Interestingly, even after applying these strict exclusion criteria (which excluded mostly MTurk participants), the responses from the subject pool participants were still more extreme than those of MTurk participants.

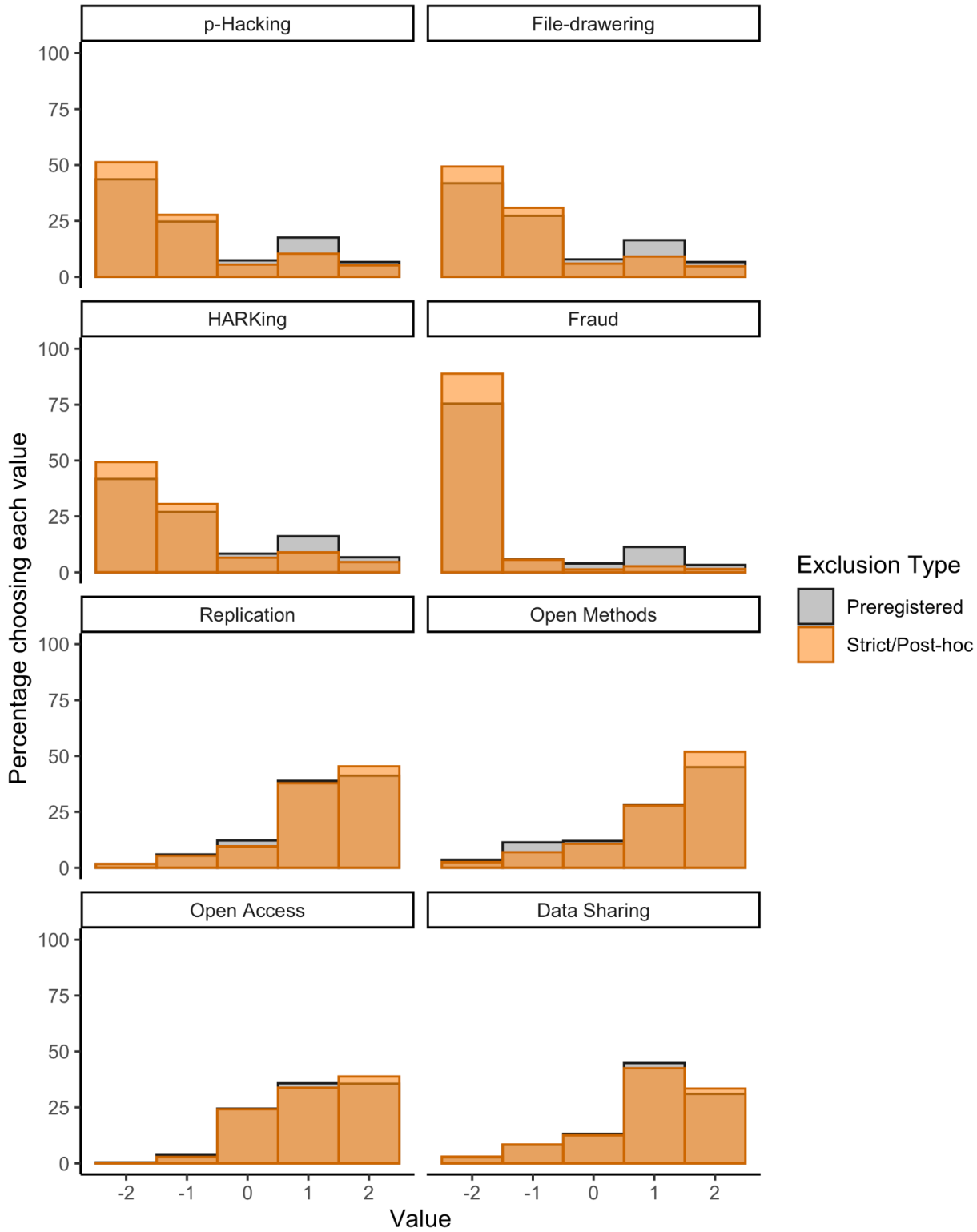

Figure 4. Distribution of participants' answers for each question with non-preregistered, strict exclusions (orange), overlayed on the same distribution with preregistered exclusions only (gray), presented in Figure 1. For the top four panels, negative numbers indicate that participants found the practice unacceptable while positive numbers indicate they found the practice acceptable. For the bottom four panels, higher numbers indicate more support for the practice. See Table 9 for additional results and sample sizes.

Table 1.9. Descriptive statistics for each question, with non-preregistered (strict) exclusions, collapsing across question version.

Question	Median (IQR)	Category	% [LL, UL]
Question 1: *p*-hacking / cherry-picking results	-2 (1)	Not acceptable	79.0 [77.0, 81.0]
		Indifferent	5.53 [3.58, 7.57]
		Acceptable	15.5 [13.5, 17.5]
Question 2: selective reporting of studies	-1 (1)	Not acceptable	80.2 [78.3, 82.2]
		Indifferent	5.92 [4.03 7.91]
		Acceptable	13.9 [12.0, 15.8]
Question 3: HARKing	-1 (1)	Not acceptable	79.8 [77.9, 81.8]
		Indifferent	6.57 [4.68, 8.58]
		Acceptable	13.6 [11.7, 15.6]
Question 4: fraud	-2 (0)	Not acceptable	94.4 [93.4, 95.5]
		Indifferent	1.37 [0.33, 2.45]
		Acceptable	4.23 [3.19, 5.32]
Question 5: direct replication	1 (1)	Move on	7.16 [5.40, 8.98]
		Indifferent	9.63 [7.87, 11.4]
		Replicate	83.2 [81.5, 85.0]
Question 6: open methods	2 (1)	Rs should not do this	9.56 [7.68, 11.6]
		Indifferent	10.7 [8.85, 12.7]
		Rs should do this	79.7 [77.8, 81.7]

Question 7: open access publication	1 (2)	Paywall	3.12 [0.91, 5.41]
		Indifferent	24.2 [22.0, 26.5]
		Free	72.7 [70.5, 75.0]
Question 8: data sharing	1 (1)	Rs should not do this	11.5 [9.37, 13.6]
		Indifferent	12.6 [10.5, 14.7]
		Rs should do this	76.0 [73.9, 78.1]

Note. N = 1,537 for all questions. Multinomial 95% confidence intervals [LL, UL] using the Sison-Glaz method. "Rs" in questions 6 and 8 refers to "researchers". Each response category except "Indifferent" collapses across two response options on the 5-point scales.

Discussion

Do people who participate in research have preferences about what scientists do with the data they have provided, and if so, what are those preferences? We attempted to provide an answer to these questions by directly asking participants. Specifically, people who had just participated in a variety of minimal-risk psychology studies self-reported their views about how researchers should treat their data in relation to 8 research practices.

Our results show that an overwhelming majority of psychology research participants in these types of studies think the questionable research practices (QRPs) presented here are unacceptable (though, surprisingly, participants did not have much more extreme views about fraud than about QRPs). Additionally, they were very supportive of practices to increase transparency and replicability, such as conducting direct replications of studies, openly sharing methods (e.g., materials, code, etc.) and data, and publishing in an open access format. For most questions, 5-30% of participants had a different view from the majority. Although an "indifferent" option was offered for every question (and labeled as such), not many people were indifferent, with values ranging from 1-15% for all questions but one; the

open access vs. paywalled publishing question was an exception, with about a quarter of participants reporting being indifferent. The similarity in response distributions for different versions of the same question indicates that, although responses can be pushed around by changes in wording or differences in framing, the overall pattern of results seems robust to such variations.

These results, other than participants' views about fraud, are consistent with Pickett and Roche (2018), who found that 71% of MTurk participants surveyed report that selective reporting of research findings is morally unacceptable. Indeed, Pickett and Roche found that most participants reported that researchers should be punished (fired and/or banned from receiving funding) for engaging in selective reporting. Given the consistent consensus about questionable research practices in our study and in Pickett and Roche's study, we first discuss what our results would mean if taken at face value and assuming they are accurate estimates of the views of participants in minimal-risk psychology studies. Then, we discuss reasons why our results may be inaccurate or why such conclusions may be premature.

What should psychologists running minimal-risk research studies do with these findings? First, researchers may want to listen to participants' preferences more. Despite being provided with an opportunity to report being indifferent to what researchers did with their data, participants used this option relatively rarely, suggesting that most participants have opinions about what is acceptable to do with their data. These opinions may reflect not just what they wish would be done with their data, but also how they expect researchers to act. Going directly against participants' expectations might result in less cooperation or in unwillingness to provide high quality data.

At the extreme, it could become an ethical issue; if we continue to engage in practices that we know participants consider unacceptable — and therefore likely expect us not to engage in — we cannot say that participants are providing informed consent to participate in research. Clearly, participants should not be the only ones deciding what research practices are acceptable — highly trained researchers have

more information and knowledge to make these decisions. However, if we decide to continue to engage in practices that most research participants consider unacceptable, we should make that explicit in the consent process. For example, in the same way that we warn participants that their anonymized data may be shared with other researchers, we should also let them know that their data may not be shared or published at all, if we continue to selectively report studies or results.

What would it mean if we took the results of our preregistered analyses regarding fraud — namely, that 19% of participants have neutral or positive attitudes towards fraud — at face value? First, this would be very inconsistent with Pickett and Roche's (2018) findings from their Study 1, which was also conducted on MTurk and found that 96% of participants expressed the view that fraud is morally unacceptable (even though the word "fraud" was not used in their questions). Indeed, in their study, 96% of participants also believed that researchers who commit fraud should be fired, and 66% believed fraud should be a crime (in a later study with a representative sample, this view was even more prevalent). Thus, if we are to believe the results of our own preregistered analyses regarding fraud, this would suggest that there are important moderators of participants' views about scientific fraud. There are a number of plausible differences between ours and Pickett and Roche's study that could suggest moderator hypotheses. For example, our participants were asked about a scenario where researchers committed fraud on the data that the same participants had just provided in the base study, whereas participants in Pickett and Roche's study were asked about the abstract idea of fraudulent practices, and fraudulent practices in two hypothetical scenarios. Perhaps participants are less bothered by potential fraud when they have participated in the study themselves and can judge how (in)consequential fraudulent practices would be. However, as we explain below, we should also seriously consider the possibility that the results of our preregistered analyses regarding fraud are inaccurate and should not be taken at face value, particularly given the inconsistencies with Pickett and Roche's results.

Limitations

There are several reasons to be cautious in interpreting our results. One important limitation of this study is the potential for data quality issues, most obvious in the non-trivial proportion of people expressing positive or neutral views about scientific fraud. Notably, this proportion is much higher for MTurk than subject pool participants when using only our preregistered exclusion criteria (32.8% vs. 9.3%; see Table 8). While the proportion in the subject pool data is consistent with what we saw in our pilot data (around 2 to 10%), the results in the MTurk population are quite alarming, and at odds with another recent MTurk study (Pickett & Roche, 2018). We believe this may indicate data quality problems that need to be taken into account when interpreting our results. One implication of low data quality is that our results may be inaccurate. If non-serious responders were responding randomly, or frequently selecting the midpoint, this would add noise to our results and suggest that participants' true attitudes are even more extreme than our results reflect. However, we cannot rule out the possibility that non-serious responders responded in ways that exaggerated the consensus or extremity in our sample's responses.

We attempted to use non-preregistered strict exclusions to reduce the influence of non-serious responders, and although this serves as a robustness check, these exploratory estimates have their own limitations. First, our decisions were data-driven and we explored several ways of excluding participants, many of which we do not report here. This was a subjective process and one indicator we used to decide when we had reached a good set of exclusion criteria was the lower rate of participants reporting that fraud was acceptable. There are two important consequences of this process. First, the fraud estimates from these exploratory analyses are uninformative as our decisions about exclusions were driven in part by our preconceptions about what these levels should be. Second, the results with strict exclusions for all other questions probably underestimate the proportion of truly indifferent participants, because

someone who was indifferent to most things would likely have been excluded when we applied our strict exclusion criteria.

Another limitation relates to how we worded the questions. Although we spent a considerable amount of time writing and rewriting them to be as clear and unbiased as possible, our own opinions about these research practices are certainly reflected in the final wording, and likely had some influence on how participants responded to the questions. In fact, we see evidence that participants' opinions can be moved around by question wording: participants reported more extreme opinions when they read the version of the *p*-hacking or filedrawering questions that implied a motive for not reporting every result or study than when they read a neutral version of the same question. Similarly, for the data sharing question, participants reported less extreme views about data sharing after reading about the pros and cons of data sharing, and some of the reasons researchers may or may not want to share their data, compared to participants who were presented with the same question but without the explicit pros and cons. However, those same results provide some constraint around the plausible effects of question wording. Although changing the question wording affected how extreme the responses were, the proportion of participants who approve vs. disapprove of each practice remained relatively stable (compare Tables 3 and 8). It would be difficult to imagine a way in which we could ask the same question that would sway participants enough to change the general consensus we see across participants for most of the questions.

Another limitation of our study is that it is not clear what importance participants place on the views they have expressed here. Do participants have pre-existing views about the acceptability of these practices, or did they formulate these views on the spot in response to our questions? Either way, how important is it to participants that researchers behave in accordance with participants' expectations and views of what is acceptable? Here again, the findings of Pickett and Roche (2018) are relevant, as

participants in Study 1 reported their views on several potential punishments for researchers who engage in selective reporting. Their findings suggest that most participants believe selective reporting (similar to the *p*-hacking and filedrawering questions in our study) is quite serious, and should be punished. 63% of MTurk participants in Pickett and Roche's Study 1 reported that researchers who engage in selective reporting should be fired. However, participants in that study were given two scenarios as examples, only one of which was a minimal-risk psychology study (the other was a study about blood pressure medication). We suspect that participants view questionable research practices in the context of minimal-risk psychology research as less serious than in the context of medical research. Thus, it is an open question how serious participants believe questionable research practices to be in the context of minimal-risk psychology research.

In our opinion, the most important follow-up questions regarding the importance that participants place on these practices are: Would participants still choose to participate if they were aware of the (questionable and open) practices that researchers routinely engage in with their data? Would knowing how researchers are planning to use their data affect the quality of participants' responses? Would it affect their views of the credibility and importance of minimal risk psychology research, and their support for public funding of such research? Our findings suggest that these questions are urgent and worth studying, but we do not yet know the answers.

Finally, another important limitation of our study is that there are serious constraints on the generality of our findings. We believe our findings can be generalized beyond the current sample to some extent. Specifically, although we only had 8 base studies, we believe these base studies are fairly representative of other minimal-risk, online, cross-sectional psychology studies. Therefore, we believe the results of this study accurately represent the reported opinions of the typical research participant in minimal-risk,

online, cross-sectional psychology studies, and may apply to similarly simple online studies in other social and behavioral sciences. However, these results cannot be generalized further than that.

Specifically, we do not believe that our results would generalize to participants' views of how their data should be treated in studies with more intensive designs (e.g., longitudinal designs, field studies), higher-risk studies (e.g., studies collecting personal health information, recordings of private behavior), or studies on more obviously consequential topics (e.g., clinical trials). Elements of these studies may affect how much participants are invested in the research process, and could produce very different results. We can imagine these features shifting attitudes in various directions. Participants may feel even more strongly that their data should be handled with as little bias (less tolerance for questionable practices) and as much transparency (stronger endorsement of open practices) as possible when the study asked more of them or when the topic is perceived as more important. On the other hand, participants may be less enthusiastic about data sharing when the data they provided are more personal, and they may be more tolerant of publishing without replication when the topic is considered urgent and important. However, as mentioned earlier, studies in higher-risk contexts suggest that, despite some concerns about privacy and confidentiality, a majority of participants support sharing of de-identified data, and are willing to share their own data, with some restrictions (Cummings, Zagrodney, & Day, 2015; Mello, Lieou, & Goodman, 2018; Trinidad et al., 2011).

It is also unclear whether these results would generalize to other types of participants. First, differences between the general population and the typical research participant in opt-in samples have been well documented (MacInnis et al., 2018). Second, our participants were living (as far as we know) exclusively in the United States. It is possible that other countries or cultures may differ in their opinions of research practices, even for minimal-risk studies.

Conclusion

Our findings are more ambiguous than we would have hoped, due to data quality concerns raised by the surprising distribution of responses to our question about fraud. Nevertheless, we believe the findings paint a fairly clear picture of participants' views about questionable and open research practices: most participants in online, minimal-risk, simple, cross-sectional psychology studies would not approve of their data being used to *p*-hack, filedrawer, or HARK, and would prefer that the research findings be subjected to replication attempts and shared transparently and openly. These findings are in line with those in the literature.

Our findings add to a growing body of evidence suggesting that researchers may routinely violate participants' expectations about how their data will be used, assuming that participants do not expect researchers to act in ways that they (the participants) find unacceptable. If we want to honor participants' expectations, we have several choices. We can: 1) align our practices with participants' expectations, 2) change participants' expectations by educating participants and the public about why practices that they initially disapprove of may be necessary or beneficial for science, 3) do more research to understand the reasons and principles behind participants' expectations and look for ways to simultaneously honor participants' and researchers' values, or 4) transparently inform participants about how we will handle their data and accept that some may drop out or provide low quality data. While further research is necessary to understand the breadth of this problem, and what the consequences might be, in the meantime we should, at a minimum, communicate our plans more transparently to participants, so that they can make a more informed decision about participating in our research.

Ethics

Permission to perform this study was granted by the University of California Institutional Review Board (IRB), IRB IDs1423371-2, 1787646-1, and 1744965-1. Permission to perform this study (and accompanying base studies) at other universities was granted by the Sacramento State Institutional Review Board (IRB), IRB ID Cayuse-20-21-240; the Princeton University Institutional Review Board (IRB), IRB ID 13508-04; and the University of Pennsylvania Institutional Review Board (IRB), IRB IDs 844186 and 844870.

Data Accessibility

All data for the pilots is available at the OSF page for this project (https://osf.io/bgpyc/). Data for the main study can be found at https://osf.io/zr29g/. The registration for the stage 1 manuscript for this report can be found at https://osf.io/8anxu, and the corresponding stage 1 manuscript can be directly accessed at https://osf.io/re5uf/.

Author Contributions

JB and SV developed the study idea, design, and materials. JB and MR developed and wrote code for the planned analyses. JB ran pilot study B and coordinated with colleagues who ran pilot studies A and C. JB performed all pilot data analyses. JB did most of the data collection and coordinated with colleagues who did the rest of the data collection at their institutions. JB performed all stage 2 data analyses. JB drafted most of the first draft manuscript, SV drafted some parts. JB and SV made extensive revisions to the manuscript. All authors made minor edits and approved the final version.

CRediT taxonomy

J.B.: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing - original draft, and Writing - review & editing.

M.R.: Methodology, Supervision, and Writing - review & editing.

S.V.: Conceptualization, Investigation, Methodology, Supervision, Writing - original draft, and Writing - review & editing.

Competing Interests

We have no competing interests.

Funding

Funding for this study is provided by university research funds to Simine Vazire and Mijke Rhemtulla.

Acknowledgments

We thank Hale Forster, Oliver Clark, Jessie Sun, Gerit Pfuhl, Eric Y. Mah, D. Stephen Lindsay, Yeji Park, Kate M. Turetsky, Kevin Reinert, Samuel H. Borislow, Jasmin Fernandez Castillo, Greg M. Kim-Ju, Jeremy R. Becker, Kate Hussey, and Fabiana Alceste for agreeing to provide us with base studies. We also thank Hale Forster and Oliver Clark for running data collection for Pilots A and C; Jessie Sun, Yeji Park, Gerit Pfuhl, Jasmin Fernandez Castillo, Samuel H. Borislow, and Jack Friedrich for running data collection for parts of the main study; and Beth Clarke for comments on the manuscript.

Chapter 2

Which authors make bolder claims? An analysis of hedging and boosting words in a large sample of social and personality psychology articles.

The content of this chapter is currently in preparation for publication. Below is the citation for the corresponding manuscript.

Cite: Bottesini, J. G., Freeman, V., Schiavone, S. R., & Vazire, S. (in prep). Which authors make bolder claims? An analysis of hedging and boosting words in a large sample of social and personality psychology articles.

Introduction

Most of the discourse surrounding the replication crisis in the last decade has focused on the quality of the evidence being produced in psychology (e.g., Open Science Collaboration, 2015; Flake et al., 2017). But how we communicate about the evidence we have, and how calibrated our claims are to the strength of that evidence, is also relevant to the credibility of our science (Hoekstra & Vazire, 2021; Pashler & De Ruiter, 2017). Extravagant claims based on evidence that cannot support them can lead to loss of confidence.

While scientific communication comes in many shapes and forms, the main way scientists in psychology communicate their findings to one another and other interested stakeholders (e.g., the public, policy makers) is through peer-reviewed publications, or articles. Therefore, it is important to understand how the evidence being presented in articles matches the strength of the claims being made about said evidence. Are our articles making big claims our evidence cannot support, or are we making more circumscribed claims that better match the strength of our evidence? To answer this question, we need to measure two constructs: the strength of the evidence and the strength of the claims. While a good deal of literature has examined ways to measure strength of evidence (p-curve, p-uniform, z-test, PET-PEESE, and of course all of the literature on meta-analysis), there is relatively little attention to measuring the strength of author's claims. In this paper, we take a first step towards measuring the strength of claims made in articles by investigating the frequency with which psychology authors use hedging and boosting words in their discussion sections. We do not claim that these measures are valid; our aim is only to present preliminary work on these measures. Much more work needs to be done to examine their validity, and we take only the first few steps here. We also explore three potential correlates of hedging and boosting, all characteristics of the authors and their affiliations, to better understand what the hedging and boosting language might be capturing.

Hedging and boosting are linguistic devices used to decrease or increase the certainty of a statement, respectively. For example, the statement “it will rain tomorrow” could be hedged by adding words like “probably”, or replacing the auxiliary verb “will” with “might” (e.g., “it might rain tomorrow”). The same sentence can be boosted by adding words like “certainly” or “definitely” (e.g., “it will definitely rain tomorrow”). Therefore, one might reasonably expect stronger claims to include more boost words and weaker claims to include more hedge words.

Here, we operationalize the strength of the claims as the proportion of hedged and boosted sentences in the discussion section of articles — an approach heavily inspired by Riddle (2017). Riddle used a similar method to examine the relationship between hedging and boosting in sentences and article-level characteristics (each article’s number of citations, the presence of a statistical reporting error in the article, or the presence of statistical overfitting in the article), as well as first-author characteristics (their gender and institutional ranking. Overall, Riddle found an association between hedging and gender such that male authors tended to hedge less, and associations with the presence of an error, such that articles with at least one statistical reporting error tended to boost more, but also hedge more. None of the other examined variables showed a systematic relationship with hedging or boosting.

Building on the work by Riddle, we aim to examine two slightly different measures of hedging and boosting (similar to Riddle’s, but not identical). We examine the distribution of hedging and boosting in a set of psychology articles, and investigate how hedging and boosting language in social and personality psychology articles relates to meta-features of those articles. Specifically, the meta-features we look at are: (1) the first author’s presumed gender, (2) the prestige of their affiliated institution, and (3) the majority spoken language in the country where that institution is located. As described below, we selected these three variables because they have the potential to shed light on theoretical and practical issues, and because they are relatively easy to extract from each article’s meta-data. That two of these

variables were also examined by Riddle in a different set of psychology articles is a bonus, as it allows us to compare and contrast our results with Riddle's.

The boldness of claims made in an article has the potential to influence how it is perceived. Articles with bolder claims might be perceived as more persuasive, or their claims might be perceived as being more certain. More persuasive or convincing articles might attract less scrutiny from peer reviewers, get published more often, or be cited more. Conversely, it is possible that bold claims with little support from the evidence have the opposite effect — they could invite more scrutiny from reviewers, and articles with unsupported claims might get published less often, and when they are published, be cited less. It is even possible that both of these scenarios are true at the same time, but at different journals; some journals might value bold claims more than others. If the boldness of claims could affect how an article is received, it is important to understand possible differences in boldness between different authors. We were especially interested in whether boldness (hedging and boosting) might differ among authors based on three characteristics: the author's presumed gender, institutional prestige, and majority spoken language in their country.

While there are reasons to wonder about how the boldness of authors' claims is associated with each of these characteristics (author's gender, institutional prestige, and their country's dominant language), we want to emphasize that our study cannot disentangle the many possible causal mechanisms that could lead to such associations. Our aim is simply to examine what these associations look like, as we believe this will shed light on what our measures of hedging and boosting might be capturing. Moreover, if the measures are valid, these findings will help inform future theorizing about how author characteristics are related to writing style. We review some theories and reasons to expect such associations below.

The "gender gap in science" has received a lot of attention, with some of it focused specifically on gender bias in publication: women tend to be underrepresented in most fields, publish less, and be cited

less frequently than their male peers (e.g., Holman et al., 2018; Bendels et al., 2018). While much of this difference may be driven by direct gender bias, where evaluators know the gender of the author (Knobloch-Westerwick, 2013), it is possible that more subtle differences in how these groups communicate their findings also have an effect on how their articles are perceived. For example, if men make bolder claims than women, and this increases the success of their article, even steps designed to prevent gender bias, like implementing masked review, would not help to resolve this difference.

The logic is similar for any differences in hedging and boosting we might observe for our two other correlates. First, high prestige institutions tend to produce the overwhelming majority of research that is published in high prestige journals. If there is something about the training or mentoring that early career scientists receive at high-prestige institutions that encourages bolder claims, this could be creating an advantage that is not based on the quality of the study but on writing style. These differences, if they exist, might create a self-reinforcing cycle in which the style of high prestige institutions is regarded as "more appropriate" or somehow "better" for reporting study results.

Similarly, there may be something about how writers in countries where the majority language is English learn to communicate their findings in writing that is different from how other authors communicate. If we find differences in hedging and boosting are associated with authors being in a majority anglophone country vs. not, this could point to a quirk of English-language writing in psychology, or a broader cultural difference.

Following Riddle's (2017) approach, we operationalized the boldness of claims being made in each article by looking at hedging and boosting, and specifically, by using a dictionary approach. Broadly, the dictionary approach consists in creating or procuring a list of words that represent a given construct, and counting the number of times any of the words in the list appears in a piece of text. More frequent use of such words is thought to correlate with higher levels of that construct in the text. For example, a

dictionary approach could be used to measure anger in Twitter posts by examining the frequency of "angry" or "aggressive" words in the posts, like insults or negative words. This is a method often employed in psychology research, the most well-known representative of it being the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2015).

In this study, we created our hedging and boosting dictionaries by looking at and adapting other similar dictionaries, and looking through discussion sections of articles to identify candidate words. We aimed to include words that increase or decrease the uncertainty of statements, and that do not have a common second use or meaning. The proportion of sentences in the article's discussion in which the authors employ a hedging or boosting word can then be taken as a proxy for the strength of the claims being made.

As with all operationalizations of hard-to-measure constructs, this approach has both positive and negative aspects, all of which stem from its simplicity. First, as text analysis approaches go, a dictionary approach is one of the most transparent. It can be understood even by researchers with no expertise in computational methods. The dictionaries used — which are just lists of words or word stems — can be created, examined, and modified to make it as complete and accurate as possible, and can also be easily shared. This makes this method much more accessible than other natural language processing methods. Another positive consequence of this method's simplicity is that we can easily examine its face validity and make modifications to ensure it is *at least* face valid.

The simplicity of this method also creates its most obvious disadvantage: it is an extremely noisy measure. Text, even text as structured and formulaic as that found in research articles, is complex. Words have multiple meanings and the correct meaning for any given word can only be fully understood when taking the context into account. The dictionary method we employed does not use any other information about the text or sentence, so any word that is in the hedging or boosting lists will be

counted as a hedge or a boost, independent of whether it was intended as such in the text. For example, a sentence that includes the word *might* will be counted as a hedged sentence even if its meaning in that particular sentence is clearly *strength* or *power*. Although this is a significant disadvantage, it is one that can be somewhat overcome by having large amounts of data, which is the case in this study. We further examine how this trade-off might affect our results and their interpretation in our discussion.

Another potential disadvantage is that the dictionary approach might be biased. While noise can be overcome with large amounts of data, bias is not as easy to address. A dictionary measure is likely to be biased if there are systematic sources of variance in the appearance of words in the dictionaries that are not driven by the construct of interest (i.e., the boldness of claims). For example, if a certain subarea of psychology studies perceptions of uncertainty, the discussion sentences in those articles might include the word "uncertainty" – a word in our hedging dictionary – and related words with a higher frequency than other subareas, giving an impression that there are more hedged sentences in those articles' discussion sections, when in fact these words related to uncertainty are not being used to hedge the sentences but to describe the substance of the study.

These important disadvantages must be kept in mind. However, if this method proves to have reasonable validity (which we will not be able to establish here), it would be easy to scale up and apply to a wide range of texts, and many different research questions.

The Present Study

The current study was designed as an honors project, and thus focused on three substantive research questions that presume the validity of the hedging and boosting measures. For the honors project, we preregistered our research questions, methods, decisions, and planned analyses, all of which can be found at https://osf.io/s2x6v. We preregistered the three research questions below, which we will report in this paper. However, given the uncertainty about the validity of the measures of hedging and

boosting, we will also report and reflect on more basic descriptive results, such as the distributions of hedging and boosting scores in our sample. The preregistered research questions are:

RQ1: How is first author gender related to the use of hedging (1a) and boosting (1b) in the discussion sections of social and personality psychology journal articles?

RQ2: How is the prestige of first authors' institutions related to the use of hedging (2a) and boosting (2b) in the same population?

RQ3: How is the majority spoken language in the first author's country related to the use of hedging (3a) and boosting (3b) in the same population?

Method

Sample

The present study is part of a larger metascience project — the *Surveying the Past and Present State of Published Studies in Social and Personality Psychology* project, hereafter SPPSPSSPPP — investigating the methods and practices of social and personality psychologists in the last decade through the published literature (Schiavone & Vazire, 2022). Therefore, our sampling decisions were guided by the articles already available in the SPPSPSSPPP dataset.

The full corpus of SPPSPSSPPP includes over 8,000 articles published between 2010 and 2020 in *PLOS One*, *Collabra: Psychology*, *Journal of Experimental Social Psychology* (JESP), *Journal of Personality and Social Psychology* (JPSP), *Personality and Social Psychology Bulletin* (PSPB), *Psychological Science*, and *Social Psychological and Personality Science* (SPPS), as well as thousands of PsyArXiv preprints. These journals were selected for SPPSPSSPPP because they cover a variety of research areas within the subfield of personality and social psychology, vary in their impact level and selectivity, and primarily

publish empirical research. Basic metadata (e.g., DOI, title, publishing year, list of authors and their affiliations) is available for nearly every article in this sample, and — most relevant for this study — a large part also includes the full text of each article, in a PDF and/or an XML format.

Exclusion criteria

For the purposes of this study, articles that were labeled as non-empirical or meta-analyses, that used non-human subjects, that were not published in English, or that were preprints on PsyArXiv were excluded from the sample. For practical reasons that we detail further down, articles for which we did not have the full text in an XML format or for which we could not automatically detect a discussion section were also excluded from all analyses. After these exclusions, the total sample for the present study was 7,534 articles, split between a pilot study (N = 763) and main study (N = 6,771). All of these exclusions were preregistered.

Pilot vs. Main Study

Although we had access to all the articles included in this study's dataset before beginning this project, much data processing and coding was necessary to extract the variables used in our analyses. Due to the unstructured nature of the data, it was impossible to make some analytic decisions without exploring the data. Therefore, to minimize the risk that we would make data-driven decisions that impacted our results, we randomly split our full dataset into a pilot dataset (10%; N = 763) and a main study dataset (90%; N = 6,771). We preregistered our plans for each study.

The purpose of the pilot study was threefold. First, it allowed us to examine the impact of various operationalizations on the accuracy and distribution of the variables, letting us select the best methods and thresholds for operationalizing each variable (e.g., the confidence level to use as a threshold for categorizing authors' presumed gender). Second, we were able to verify and, when necessary, improve

the quality of the hedging and boosting dictionaries. Third, using the pilot data helped us determine the most appropriate analysis plan for the main study given each variable's distribution, resulting in a better main study preregistration.

Both the main study (https://osf.io/cz4h2) and pilot (https://osf.io/mx6gb) were preregistered. The results presented below are based exclusively on the main study sample, consisting of N = 6,771 articles from seven journals published between 2010 to 2020 (see Figure 1 for the journals and range of years included).

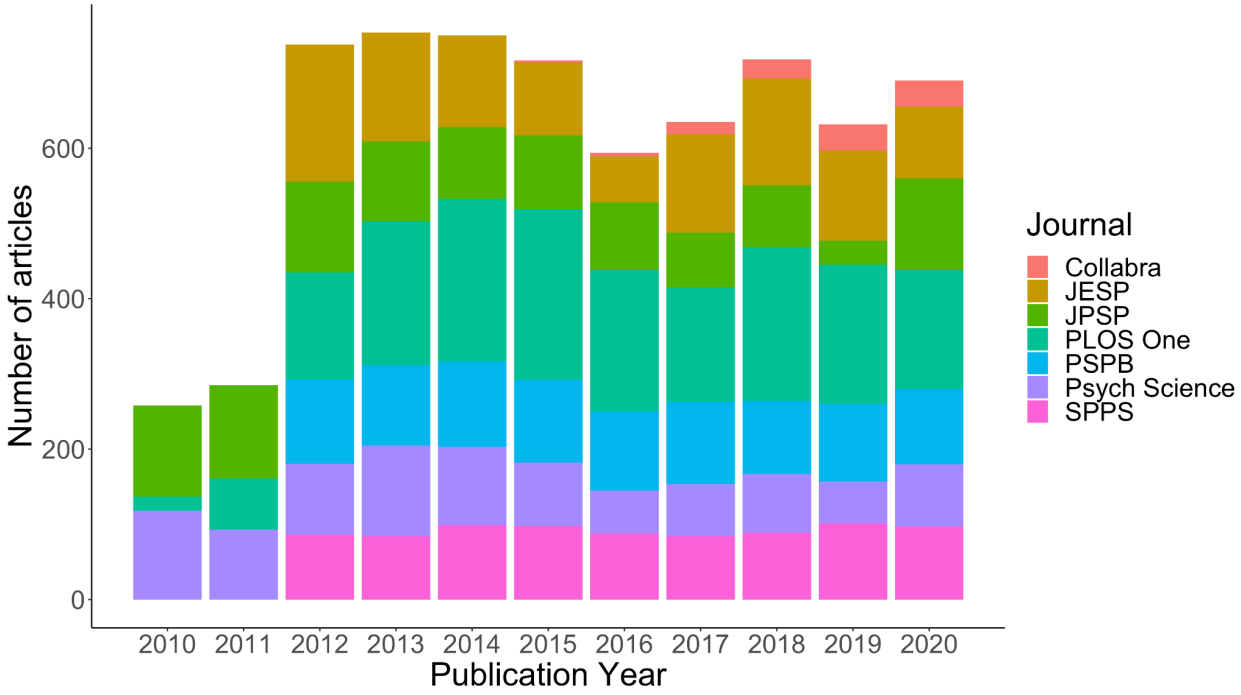

Figure 1: Proportion of articles by journal and publication year in the main study sample.

Power

A sensitivity power analysis using simulated data (available at https://osf.io/kgrhy) suggests that the main study's sample size allows approximately 80% power, with an alpha of .05, two-tailed, to detect

any effect at least as large as a 1% group difference for the main analyses (e.g., if men hedge in 49% of sentences and women hedge in 50% of sentences, we had 80% power to detect this difference).

Text Extraction

Since we were interested in the claims being made in the literature, we focused on the "discussion" part of each article, which we believe contains most of the article's claims. By relying on the XML structure of the articles, we were able to use R (R Core Team, 2020) to automatically extract the text from all sections labeled with the words "discussion", "concluding", or "conclusion", except if that heading also included the word "result" or "results". We opted to exclude sections for which the heading includes "results" because "Results and Discussion" is a frequent section label for multi-study articles, and these sections include several sentences describing a study's results rather than their interpretation. All the discussion sentences were combined into a single text for each article, after excluding the headings.

This process took advantage of the tree-like XML format, such that the text in any subsections of the selected sections would automatically be extracted as well. For example, if a section labeled "Discussion" included a subsection labeled "Limitations", both the text under the *Discussion* heading and the subsection heading *Limitations* would be extracted.

Because this extraction method relies heavily on the structure of the XML files being accurate, unexpected errors may occur. To verify that the text had been extracted correctly, we performed 100 random spot-checks comparing the published article discussion text to the extracted text. This indicated that the extracted section text generally included all the sections and subsections we expected. Even so, we come back to a few limitations of this method in our discussion section.

After extracting the text, we parsed it into sentences using the *tidytext* R package (Silge & Robinson, 2016). Although it performs quite well, the fact that academic texts often contain citations and other

unexpected symbols in their sentences created a bit of a challenge for the parsing algorithm, resulting in some sentence fragments. To prevent sentence fragments from artificially inflating the number of sentences in a given discussion, we excluded all "sentences" that contained 25 characters or fewer after removing numbers, symbols, and punctuation. Visual inspection confirmed that this was a reasonable cut-off to prevent the exclusion of full sentences.

Coded Variables

As we detail below, all variables in the present study were created through a variety of methods, in accordance with our preregistration, which can be found at https://osf.io/mhctz. Any deviations from the preregistration are clearly indicated.

Hedging & Boosting

A dictionary approach was used to measure the proportion of sentences that contained hedging and boosting words or phrases. A dictionary approach consists of comparing a given text to a list of words and phrases of interest, and then counting the number of times those words appear in the text. For this study, we used adapted versions of the dictionaries used by Riddle (2017), which we provide in Appendices A and B of our preregistration document: https://osf.io/cz4h2/

Although it might be tempting to think of hedging and boosting as two sides of the same coin, we could not find a sufficiently compelling reason to combine them, and therefore measure them separately here. Each sentence extracted from the discussion section of articles was coded as being hedged (or boosted) in a binary way: if it contained at least one word or phrase from the hedging (or boosting) dictionary, it was considered hedged (or boosted). We then calculated the total proportion of hedged sentences and boosted sentences to create article-level variables. If 0% of discussion section sentences in an article included at least one hedge (or boost) word, this proportion would be 0 for that article,

while if 100% of discussion section sentences included at least one hedge (or boost) word, this proportion would be 1 for that article.

As explained in the introduction, we opted to use a dictionary approach over other possible approaches (e.g., manually coding instances of hedging and boosting in the corpus, or more sophisticated machine learning methods to detect uncertainty) due its ability to be scaled to a large sample while still being feasible and easily comprehensible by the average psychology researcher. Additionally, an important factor that affects the validity of dictionary approaches is how well the dictionaries capture the relevant constructs. To help increase the chances that our dictionaries accurately capture hedging and boosting in academic articles, we used as our starting point hedging and boosting dictionaries that had been developed and used specifically on a corpus of scientific articles, as we describe below.

The hedging dictionary was created by adapting and building on the hedging dictionaries used in two previous studies: Riddle (2017), who adapted it from Prokofieva and Hirschberg (2014), and Mina and Biria (2017) who used Hyland's (2005) interpersonal taxonomy to classify hedges in their corpus of social and medical science articles. To identify potential additional hedging words, we decided to also read over the discussion sections of a set of articles that we expected to include a great deal of hedging: failed replication studies. We expected these articles to contain a greater frequency (and variety) of hedges due to authors' caution when interpreting findings. Additionally, we verified that each word and phrase in our dictionaries had been appropriately categorized as a hedge using Merriam-Webster's Dictionary (https://www.meriam-webster.com), and no common synonyms had been omitted. Finally, in the pilot study, we looked at the hits for the words in the hedge dictionary to see, based on the context in which they appeared, whether they resembled hedges. Based on pilot study results, we excluded the words "feel", "feels", and "felt" from the hedging dictionary due to their greater prevalence in the

emotion literature which often did not constitute hedging (e.g., “participants rated felt emotions”). The final hedging dictionary contained 78 items.

The boosting dictionary was adapted from Hyland (1998; see Appendix D in the preregistration). We followed a similar procedure for piloting and revising the boosting dictionary as described for the hedging dictionary in the previous paragraph, though we did not examine failed replication articles (as we did not expect these to contain an exceptionally high volume or diversity of boosting words). After piloting, we excluded the words “think”, “thinks”, and “thought” due to duplication in the hedging dictionary. The final boosting dictionary contained 39 items.

Author presumed gender

The first author’s presumed gender was measured by extracting their first name from article metadata, and using the GenderGuesser package — a wrapper for the genderize.io API — to classify the name as belonging to a man or woman. The package also returned a probability for each classification, and the number of samples this probability was based on. Names with a probability equal to or below .75, or with a sample number lower than six were deemed too unreliable, and therefore were coded as “unknown” and treated as missing data. These thresholds were determined by examining the pilot data; the .75 threshold was optimal for the main study because it eliminated only 9% of names in the database while returning mostly accurate inferences, to the authors’ knowledge. In the main study, n = 3,395 (50%) authors were presumed men based on their first names, and n = 2,710 (40%) authors were presumed women based on their first names. A further n = 666 (10%) of authors’ first names were categorized as unknown.

Institutional prestige

To measure institutional prestige, we extracted each first author's institutional affiliation from the article metadata. We then attempted to automatically match each entry to the corresponding institution listed in the Times Higher Education (2020) World University Rankings (THEWUR) using a conservative algorithm to avoid false positive matches. Approximately 30% of institutions could not be automatically matched by this algorithm and had to be manually coded.

Institutions with a match in the THEWUR list were then assigned the institution's corresponding score, which could range from 0 to 100, which higher scores indicating superior performance in these metrics. This score, named *Score_Result* in the dataset we used, was a composite of institutional performance indicators across five domains, including teaching (learning environment), research (volume, reputation, and income), citations (research impact), international outlook (international collaboration, faculty, and students), and industry income (THEWUR, 2020). Institutions that were not listed on the THEWUR (2020) list were treated as missing data.

Majority spoken language in the author's country

We attempted to automatically extract the country where the first author's institution was located from their institutional affiliation in the article metadata. For approximately 20% of authors, their institutional affiliation needed to be manually coded by looking up the institution to find out where it was located.

To classify each country as being either majority native English speaking or not, we used the categorization from the U.K. government for Visa purposes (University of Sheffield, 2021; Appendix C of our preregistration). For the main study data, n = 4,314 (64%) of first authors' institutions were in majority native English-speaking countries, while n = 2,434 (36%) authors' institutions were in countries with majority spoken languages other than English.

Analyses

The main analyses were performed in R 3.6.3 (R Core Team, 2020) using the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) packages, allowing us to run multilevel models with articles (N = 6,771) nested within journals (N = 7). All predictors were on the article level. Articles with missing data on a variable were excluded from analyses that contained that variable (see Table 1 for Ns for each variable). Journals were modeled as a random effect to account for potential clustering.

All initial models included a random intercept and a random slope for each journal, but when that model generated a warning, we defaulted to removing the random slopes and leaving only a random intercept for each journal, as preregistered (https://osf.io/s2x6v). We have indicated where this occurred in the Results section by explicitly stating whether the model used for each analysis included only a random intercept or a random slope and intercept.

We ran separate models for each research question, or six models in total. In each model, the proportion of sentences that were hedged (or boosted) was regressed on one of the three predictor variables: presumed gender of the first author, majority spoken language, and institutional prestige. We also generated bootstrapped confidence intervals for each model. Our analysis code is available at https://osf.io/mhctz.

Results

Descriptives

Descriptive statistics for the main study are displayed in Table 1.

Table 2.1: Descriptive Statistics for Main Study

Continuous Variable	M	SD	Range	Skew	Kurtosis	N

Proportion of Hedged Sentences	.51	.14	0-1	-0.16	3.45	6,771
Proportion of Boosted Sentences	.17	.10	0-.88	1.16	6.32	6,771
Institutional Prestige	62.30	18.50	13.20-95.40	-0.07	2.30	5,643

Categorical Variable	Proportion			N
	Men	Women	Unknown	
First Author Presumed Gender	.50	.40	.10	6,771

	Proportion		N
	English	Other	
Majority Spoken Language	.64	.36	6,748

There was a good amount of variance in hedging across articles (see Figure 2), with the typical article containing at least one hedge in about half of sentences in discussion sections, and many articles containing a substantially higher or lower proportion of hedged sentences. There was less variance for boosting, with the typical article containing one or more boosts in about 17% of sentences, and few articles containing more than 25% of boosted sentences in their discussion sections. Furthermore, while hedging and boosting were operationalized as proportions (i.e., not continuous), the pilot data indicated that they were approximately normally distributed, and therefore suitable to treat as continuous outcome variables in our analyses. Additionally, the majority native English-speaking countries group was considerably larger than the non-English majority- speaking countries group (approximately 64% compared to 36%, respectively). This was expected given the present study's focus on English-language social and personality psychology journals, and we did not consider the latter group to be so small that it

may pose a problem for subsequent analyses. We also noted that there was approximately 10% of articles missing data on author presumed gender that were not categorizable by our measure. However, due to the relatively even groups of presumed men and women authors in our corpus (approximately 50% and 40%, respectively) and large sample size, we did not anticipate any problems related to loss of statistical power in our analyses.

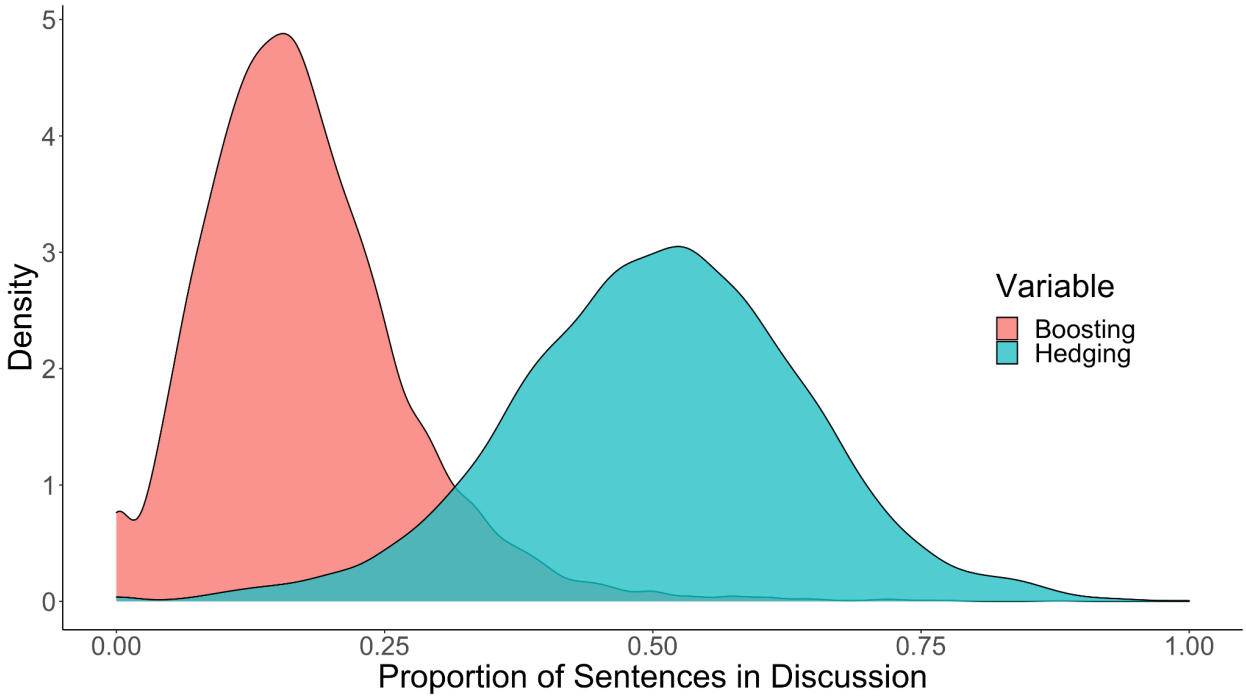

Figure 2.2: Proportion of hedged and boosted sentences in article discussion sections

Main analyses

RQ1: First author's gender

To examine the association between first author gender and hedging, we fit a model with only random intercepts. We found evidence of a significant association between author presumed gender and the proportion of hedged sentences in article discussion sections, such that male authors hedged less than female authors, B = –0.01, SE = 0.0034, p = .004, 95% CI [–0.02, –0.0034]. In other words, the average

woman hedged 51.6% of sentences, while the average man hedged 50.7% of sentences in discussion sections (see Figure 3).

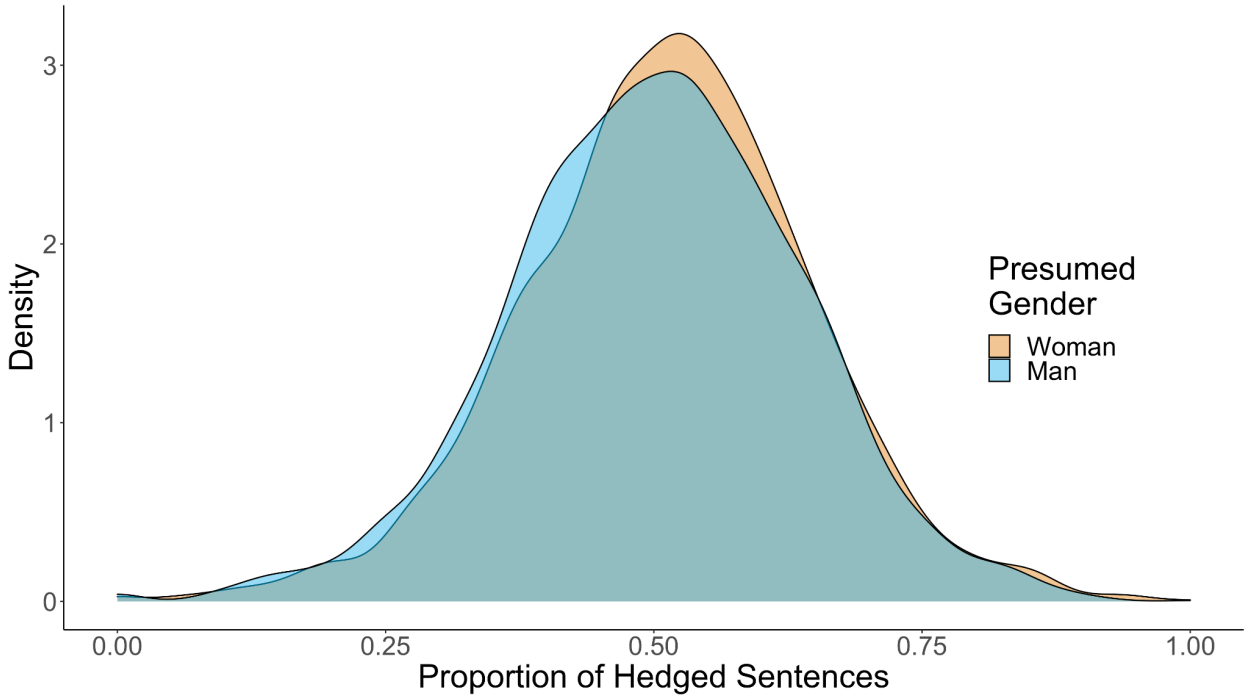

Figure 2.3: Proportion of hedged sentences by gender in article discussion sections

We then examined possible gender differences in boosting using a model with both random intercepts and random slopes for journals. We did not observe an association between author presumed gender and the proportion of boosted sentences in article discussion sections. Men and women first authors did not significantly differ on the average proportion of sentences that were boosted in the discussion sections, B = 0.0044, SE = 0.0035, p = .28, 95% CI [–0.0020, 0.01], with men boosting an average of 17.7% of sentences and women 17.3% of sentences in discussion sections. Note that we did not preregister how we would interpret non-significant effects, or what our smallest effect size of interest was. Deviating from our preregistration, we made a very post hoc decision to use the point estimate for the effect size for the gender difference in hedging (|B| = .01) as a threshold for a practically meaningful

effect. In this case, because the 95% confidence interval for gender differences in boosting included this effect size, we cannot rule out the possibility that there was a practically meaningful effect which we failed to detect.

RQ2: Institutional prestige

To examine the association between institutional prestige and hedging, we fit a model with only random intercepts. We found evidence of a significant positive association between the prestige of institutions and the proportion of hedged sentences in article discussion sections, B = 0.00034, SE = 0.000097, $p < .001$, 95% CI [0.00016, 0.00055]. Because a one-unit increase in this model represents a very small increment (i.e., a single point on the 100-point scale of institutional prestige), it may be more meaningful to compare two common scores, 50 and 90, representing institutions with moderate and high prestige, respectively. According to our results, authors from high prestige institutions (score of 90) hedged 1% more sentences (52.2%) in article discussion sections on average than those from moderate prestige institutions (score of 50; 50.9% of sentences hedged).

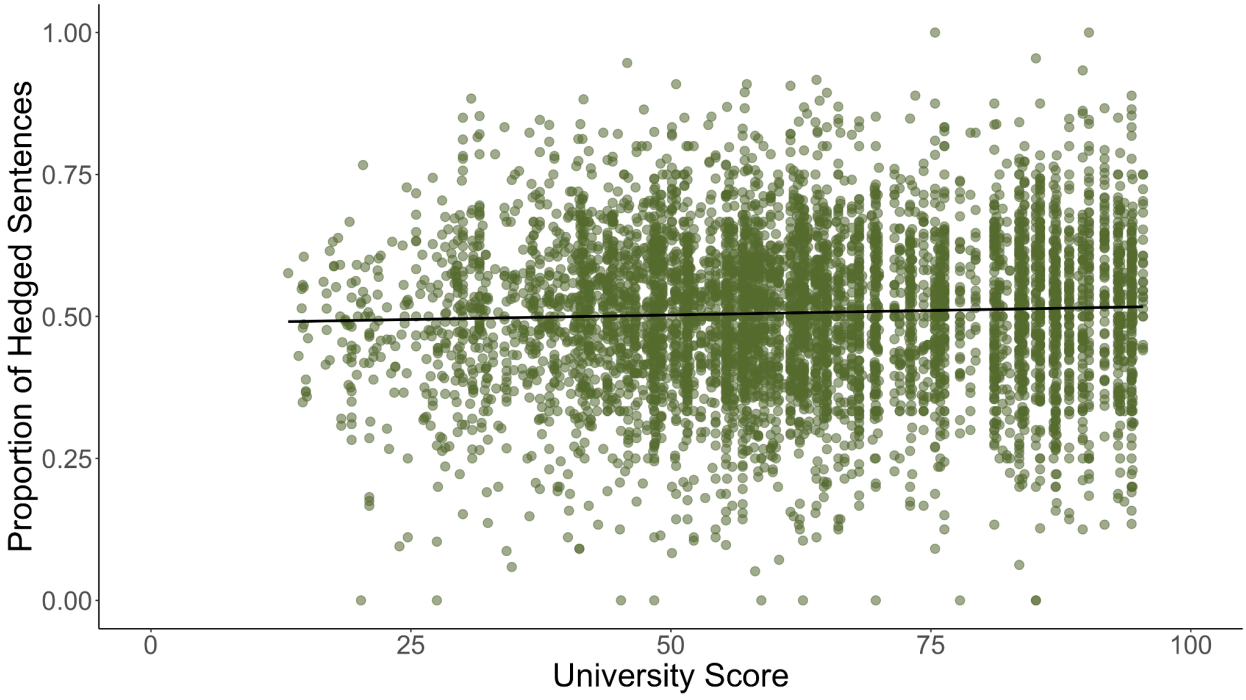

Figure 2.4: Association between institutional prestige and hedging in article discussion sections

We fit a similar model to examine the association between institutional prestige and boosting, with only random intercepts. We did not observe an association between institutional prestige and the proportion of boosted sentences in article discussion sections, B = 0.0000080, SE = 0.000070, p = .91, 95% CI [–0.00012, 0.00016]. Once again, we did not preregister how we would interpret nonsignificant results, nor a smallest effect size of interest. As with the previous research question, we made the very post hoc decision to use the magnitude of the point estimate for the hedging result (|B| = .00034) as the smallest effect size of interest. Because the 95% confidence interval included only effect sizes that were less than half as large as the point estimate for the association between institutional prestige and hedging, we can rule out the possibility that there was a practically meaningful effect using this metric. Therefore, we believe this constitutes evidence of absence of an association between institutional prestige and boosting in article discussion sections.

RQ3: Majority spoken language

To examine the association between the majority spoken language (English vs. other) in the country where the first author's institution was located and hedging, we fit a model with both random slopes and random intercepts. We found evidence of a significant association between the majority spoken language and the proportion of hedged sentences in article discussion sections, such that authors from predominantly non-anglophone countries hedged less in their discussion sections than authors from predominantly anglophone countries, B = –0.04, SE = 0.0047, $p < .001$, 95% CI [–0.05, –0.03]. In other words, there was a small effect showing that authors from countries with majority spoken languages other than English on average hedged 48.6% of sentences, and authors from majority native English-speaking countries on average hedged 52.4% of sentences in their discussion sections (see Figure 4).

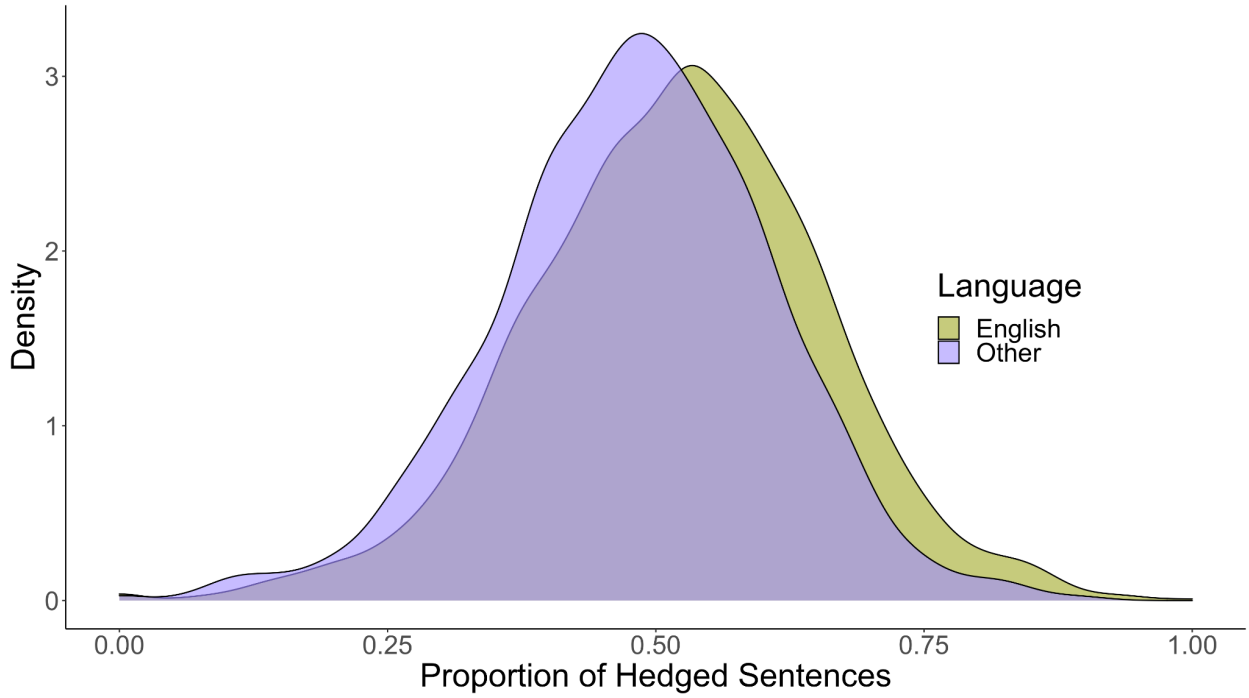

Figure 2.5: Proportion of hedged sentences in article discussion sections by majority spoken language in the author's country

To examine the association between majority spoken language and boosting, we fit a model with both random intercepts and random slopes for journals. We did not observe an association between the majority spoken language in the author's country and the proportion of boosted sentences in article discussion sections. Authors from countries with English as the majority spoken language and those from countries with non-English majority spoken languages did not significantly differ on the average proportion of boosted sentences in article discussion sections, B = 0.01, SE = 0.0049, p = .14, 95% CI [–0.00027, 0.02], with authors from anglophone countries boosting 17.2% of sentences on average while authors from non-anglophone countries boosted 17.8% of sentences on average. Once again, we did not preregister how we would interpret nonsignificant results, nor a smallest effect size of interest. As with the previous research question, we made a very post hoc decision about what should be considered the smallest effect size of interest. Here, we used the point estimate for the effect size of the gender

difference in hedging (|B| = .01) as the smallest effect size of interest, because we can compare the two effects (they both involve categorical predictors), and we considered this effect size meaningful in interpreting our results for gender differences in hedging. Because the 95% confidence interval for the current result included effect sizes twice as large the point estimate of the effect size for the gender difference in hedging, we cannot rule out the possibility that there was a practically meaningful effect which we failed to detect. The upper limit of the 95% confidence interval (effect size of B = 0.02) would represent authors from countries where English is not the majority spoken language boosting 2% more sentences (19%) compared to authors from countries where English is the majority spoken language (17%).

Discussion

In this study, we used a dictionary approach to measure hedging and boosting in the discussion sections of psychology articles. We were not sure of the validity of these measures, so our results should be interpreted with caution, and may help shed light on what, if anything, these measures are picking up on. We examined associations between hedging and boosting and three first author-related characteristics: first-author presumed gender, the institutional prestige of the author's university, and whether the country where that institution is located is majority native English speaking or not. We intended hedging and boosting to be a proxy for the strength of claims being made in an article's discussion section, and see this as a first step towards examining whether these measures are useful, and if so, how they might help us understand the factors underlying authors' willingness to make strong claims.

The ultimate goal of this research program is to understand the causal factors that lead authors to make strong claims (or not). However, our study can only speak to associations between the examined predictor variables, as operationalized here, with outcome measures (hedging and boosting) with

unknown validity. We found that hedging in discussion sections is significantly associated with all three author characteristics we investigated, with the association between the proportion of hedged sentences and majority spoken language in the author's country being the biggest effect. Authors in majority native English-speaking countries hedged 3-4% more on average than authors from countries with a different predominant language. This means that, if a discussion section had 100 sentences, an author from an institution in an anglophone country would hedge about 52 sentences, while an author from an institution in a non-anglophone country would hedge about 48-49 sentences.

The associations between gender or institutional prestige and hedging are smaller. Both effects had a similar magnitude of about 1% difference. Women first authors tended to hedge 1% more than men first authors; if a discussion section had 100 sentences, an average male author would hedge 50 sentences while an average female author would hedge 51 sentences. Meanwhile, authors from high prestige institutions tended to hedge 1% more than authors from a moderate prestige institution, meaning that, in a discussion section with 100 sentences, an author from a moderate prestige institution would hedge about 51 sentences on average, while an author from a high prestige institution would hedge about 52 sentences. Interestingly, while the gender difference in hedging was such that the group with less status (women) hedged more, the other two effects went in the opposite direction: the groups with more status (authors at high prestige institutions and in anglophone countries) hedged more.

In contrast, boosting did not show any significant associations with the three examined author characteristics, and the confidence intervals obtained suggest that, if there are any boosting differences to be found using these particular measures, the effects would be minuscule. However, we take the liberty of speculating below on why these minuscule effects might still be indicative of relevant differences. Because of this, we note here that all these non-significant boosting differences are in the

direction one would expect if boosting and hedging were opposites; that is, in the opposite direction of all the hedging differences outlined above.

What evidence do these results provide regarding the potential validity of the hedging and boosting measures? First, the descriptive results show that there is more variance in our hedging measure than our boosting measure, suggesting that the hedging measure has more potential to detect differences across articles. Second, the fact that the hedging measure was associated with all three predictor variables, whereas we did not detect any associations with the boosting measure, is further evidence that the hedging measure is likely picking up on a real signal. Of course, it is too soon to say whether the hedging dictionary measure is a valid measure of the strength of claims (reverse-scored), but it is more promising than the boosting dictionary measure.

Next, we interpret our results on the assumption that at least the hedging measure is valid, but we urge readers to keep in mind that this assumption is not necessarily warranted. If we take these results at face value, we can think of some potential explanations for the difference in hedging between authors from institutions in anglophone vs. non-anglophone countries. First, cultural differences between how different authors express themselves might be at play. This might not be particularly likely given that the non-anglophone countries are quite heterogeneous. However, the top two non-anglophone countries in this dataset, Germany and the Netherlands, might be driving the effect, as they account for about 37% of all the articles from authors in non-anglophone countries. Perhaps there are cultural characteristics specific to German and Dutch culture that could be driving these effects.

Another possible interpretation is that, while authors from anglophone countries feel comfortable hedging their claims, authors from non-anglophone countries might feel like their claims could receive more scrutiny so they cannot afford to hedge as much if their conclusions are to be taken seriously. A similar explanation could apply to the (smaller, but present) difference in hedging between moderate

and high prestige institutions. It is possible that authors at high prestige institutions feel more comfortable hedging their claims while their counterparts at moderate prestige institutions feel the need to be more assertive to be believed. If some of these differences could be explained by authors trying to assert themselves and their work in the face of biases against them, one has to wonder whether the magnitude of the effects might indicate more perceived bias against non-anglophone scholars than those at less prestigious institutions.

Finally, our results for gender differences in hedging are consistent with those found by Riddle (2017), in which women tended to hedge more than men. Other works with a better theoretical foundation have posited much better explanations for this effect than we could come up with given our limited data and knowledge of those theories, so we will spare the reader from our speculation.

Although the effects described above are quite small, we would like to briefly argue that they may be indicative of actually meaningful differences in the strength of the claims being made in the discussion section of these articles. It is difficult to imagine that reading one extra sentence with a hedge word in a one-hundred-sentence long text could lead a reader to perceive that author as less assertive. Because of this, it is not clear how to define what is (or is not) an effect size that is "practically meaningful". However, the tool we used to measure the strength of the claims is so crude that it will undoubtedly produce an extremely noisy measure of the strength of the claims. Assuming that it contains any amount of signal at all (i.e., a real effect), the signal to noise ratio for this measure could lead to extremely small effects. If we had a better, less noisy measure of the strength of the claims, these effects might be revealed to be much larger. Therefore, we think it is reasonable to entertain the possibility that these small effects might be indicative of larger, actually practically meaningful effects that could be detected with improved measures. However, we are wary of using our measure's

unknown (and likely poor) psychometric properties as a point in favor of the importance of our effects, so we present this argument only half-heartedly.

Limitations

Of course, the biggest limitation of this study is that we do not know whether the hedging and boosting dictionary scores are valid measures of the strength of authors' claims. We believe they can be taken as valid measures of the frequency with which authors use the words and word stems in the respective dictionaries, but any inferences beyond that (including our tentative interpretations of our results above), rest on the assumption that this is a valid way to measure the strength of authors' claims. This assumption is not yet warranted.

One potential limitation of this study relates to the computational techniques that were used in the extraction and coding of the data. First, we relied on the XML structure of the articles to select the appropriate sections, as described in the method, but issues with the XML formatting could result in the extraction of undesirable text or the failure to extract desirable text. For example, some discussion sections might end up including figure captions if a figure was included in the discussion, or we may have missed discussion sections with labels we did not anticipate. Second, we based our best guess of the first author's gender on their automatically extracted first name as reported in the article's metadata. This is clearly not ideal for several reasons, including the fact that names may not always align with gender, that we are limited to examining only two genders, and that the most likely gender for a first name might change depending on country or culture (e.g., Simone). We have tried our best to address these issues by performing random checks on the data at different points in the process, and we are reasonably confident that, if any of these problems occurred, they do not substantially impact the results. We thought they were worth mentioning nonetheless.

Another limitation of our study is that we assumed author characteristics might be related to the articles those authors produced, but we only used the first author's characteristics when investigating these associations. This is unrealistic because multi-author articles, as the articles in this sample almost always were, are a group effort; if characteristics of the authors affect how the article is written, that effect would be diluted depending on how much of the article each author wrote.

Finally, our sample only included articles published in the last decade, and only in 7 specific psychology journals. Therefore, we can generalize our results to similar journals in this decade, but we do not think our results would generalize beyond this time period, or to psychology journals that are very different from the ones analyzed.

Conclusion

If we are in fact measuring hedging in these articles' discussion sections, then the patterns we found offer some interesting insight into which groups might be more or less prone to hedging their conclusions. However, we are ultimately interested in causes of strong claims, including why some authors might be more likely to make strong claims than others. Any causal interpretation requires assumptions, and our study did not provide any data that can speak to those assumptions.

Strong claims can, of course, be justified, if the quality of the studies and the strength of the evidence is high. Ultimately, understanding whether the claims being made are calibrated to the strength of the evidence being presented requires not only measuring the strength of the claims but also the strength of the evidence. We look forward to continued work developing better and better measures of both halves of this equation.

Chapter 3

How Do Science Journalists Evaluate Psychology Research?

The content of this chapter is currently under review at *Advances in Methods and Practices in Psychological Science*. Below is the citation for the corresponding manuscript.

Cite: Bottesini, J. G., Aschwanden, C., Rhemtulla, M., &Vazire, S. (invited revision). How do science journalists evaluate psychology research?

Abstract

What information do science journalists use when evaluating psychology findings? We examined this in a preregistered, controlled experiment by manipulating four factors in descriptions of fictitious behavioral psychology studies: (1) the study's sample size, (2) the representativeness of the study's sample, (3) the *p*-value associated with the finding, and (4) institutional prestige of the researcher who conducted the study. We investigated the effects of these manipulations on real journalists' perceptions of each study's trustworthiness and newsworthiness. Sample size was the only factor that had a robust influence on journalists' ratings of how trustworthy and newsworthy a finding was, with larger sample sizes leading to an increase of about two thirds of one point on a 7-point scale. Due to high precision, we can confidently rule out any practically meaningful effect in this controlled setting of sample representativeness, the p-value, and most surprisingly, university prestige. Exploratory analyses suggest that other types of prestige might be more important (i.e., journal prestige), and that study design (experimental vs. correlational) may also impact trustworthiness and newsworthiness.

Keywords: science communication, science reporting, research practices, metascience

Introduction

Science journalists play an important role in the scientific ecosystem. They are the primary external watchdogs that can monitor scientists and scientific institutions for problematic practices and call out dubious claims without much fear of harming their career prospects. In fact, the code of ethics from the Society for Professional Journalists instructs journalists to "Be vigilant and courageous about holding those with power accountable" (Society for Professional Journalists, 2014), a responsibility which extends to science journalists, who are charged with monitoring scientists and scientific institutions and keeping them accountable. However, for science journalists to play this important role, they need to have access to, and know how to use, relevant information when deciding whether to trust a research finding, and whether and how to report on it.

In this paper, we examine what information science journalists use when evaluating psychology findings in a controlled experimental setting. In particular, our study aims to better understand the influence of factors such as the research design (the study's sample size and the representativeness of the study's sample), the statistical evidence (e.g., the *p*-value associated with the finding), and reputational factors (institutional prestige) on science journalists' judgments of a finding's trustworthiness and newsworthiness. Understanding which, if any, of these factors influence science journalists' perceptions of research findings may help us understand which findings are more likely to be communicated to the public.

The Importance of Science Journalists

Criticism, scrutiny, and oversight are accepted, even valued, in many domains. From professional restaurant reviewers to corporate governance boards and environment-monitoring organizations, this is seen as an important, often remunerated, activity. Ideal candidates for these jobs know enough about the subject upon which they are reflecting to be able to scrutinize it, but have enough distance to

minimize conflicts of interest. Science is different. It is easy to imagine that science might not need external auditors: scientific scrutiny is meant to be a built-in feature of science, an integral part of the scientific process. Before any discoveries can be published, they must undergo evaluation by other scientists in the field — a process known as *peer review*. However, we believe that external science watchdogs are in fact needed, and play a crucial role in vetting the findings that get transmitted to the public. In many cases, they also determine which findings get shared with the public.

Although some science communication work can be done by university PR offices or the scientists themselves, science needs scrutiny from "inside-outsiders", more impartial critics (Ihde, 1997), a role which is often played by science journalists. Blum (2021) argues that the role of the modern science journalist is to "portray research accurately in both its rights and its wrongs", while Carr (2019) suggests science journalists should scrutinize research, "[a]nd [...] do so on behalf of readers, not scientists." Carr's defense of the role of science journalism as something done for readers, not scientists, reflects the deep changes the profession has undergone in the last century. Arising from a desire to popularize science in the early 20th century, what would eventually become science journalism was originally focused on publicizing "smart and positive science stories" (Blum, 2021). Although framed as a way to inform American citizens so they could better participate in democracy, science popularization also aimed to secure funding for future scientific endeavors by creating a public that cared enough about science to demand more of it (Katz, 2016).

Science journalism's original penchant for positively-framed science stories changed over the last century. Around the 1960s and 70s, journalism saw the incorporation of more diverse voices and viewpoints into the news, and more skeptical, critical science journalism, as it became increasingly clear that scientific progress also had its downsides (Crewdson, 1993; Blum, 2021). Despite these changes, some journalists like Crewdson (1993) heavily criticized science journalism in the 1990s, arguing that

journalists are often too close to the scientists they report on and too keen to cheer science on — leading Crewdson to describe science journalists as “perky cheerleaders” for science.

Today, science journalism is in the process of “growing up” (Blum, 2021). More emphasis on training and investigative reporting, starting in the 1980s and 90s, has better positioned the profession to act as the science watchdog we need. New technological developments have helped science journalists by increasing the visibility of their work and allowing them to tackle the complexities of reporting on scientific findings with new communication tools (e.g., by creating data visualizations or recording podcasts). Science journalists now have more opportunity to become good science watchdogs who can help the public consume scientific research through a critical lens and draw the public’s attention to more rigorous research. Clearly, a more informed public with access to more nuanced scientific information is a social benefit of having more critical science journalists.

Critical science journalism also has many benefits for science itself. First, if uncertain science is communicated uncritically, without the proper context and caveats, science risks losing credibility. Although scientific theories can never be proven definitively, some research areas or findings are much more certain and justify bold claims more than others. But if these differences in certainty are not made explicit, the public could get the impression that all published findings are equally solid. When it becomes clear that this is not warranted — if coffee cures cancer one week but causes it the next — this can erode public trust in science, and at the extremes, facilitate the types of science denialism we see, for example, toward climate change research. A good science journalist can evaluate each research finding to see if it is solid enough to warrant reporting on, and if it is, help contextualize it and calibrate their claims to the evidence.

Second, scientists need science watchdogs so that incentives on scientists reward better research practices, and therefore better science. That is, good science journalism helps ensure that the best

research, however we define it, is receiving the most attention. Media attention is part of the incentive structure in science, and there is some indication that it is rewarding to scientists — getting media coverage for their work may boost scientists' citations, facilitate invitations to collaborate or give talks, and may help them attract more investors or donors to fund their research (Dance, 2018). Media attention is only one of many rewards that scientists may respond to, but it can be an important one. Therefore, it is important to know whether journalists' decisions about whether and how to report on research findings track the quality of the research.

Finally, having competent science watchdogs serve as intermediaries between researchers and the general public can help keep scientists honest. Scientists are only human, and like all humans, we are prone to biases about our own research. For example, public engagement with our work may help us get funding, and therefore we have an incentive to exaggerate our own accomplishments. But if we anticipate that our work will receive the appropriate amount of scrutiny as it is being transmitted to the general public, this provides an incentive to improve our research practices and better calibrate our claims to the evidence.

In light of the important role that science journalists play in communicating and critiquing science, it is worth understanding how science journalists form opinions about findings they could potentially report on. In this study, we ask what information science journalists use when deciding what findings are trustworthy and newsworthy, and we investigate four potential factors in a controlled experimental setting.

The present study

As a first step to understanding the factors that influence science journalists' decision-making, we investigated how science journalists evaluate research findings in psychology. Specifically, we invented findings similar to those that might arise in the field of social and personality psychology (our area of

expertise), and experimentally manipulated four features of these research studies to investigate how these features affect science journalists' perceptions of the research. We sampled primarily U.S.-based science journalists, and the study descriptions (locations, samples) were also U.S.-centric. We examined the effect of four variables. All four manipulated variables pertained to some aspect of the research studies that the science journalists read and evaluated: the sample size (number of participants in the study), sample representativeness (whether the participants in the study were from a convenience sample or a more representative sample), the statistical significance level of the result (just barely statistically significant or well below the significance threshold), and the prestige of the researchers' university. Our aim was to examine the causal effects of these factors on our dependent variables: journalists' perceptions of the trustworthiness and newsworthiness of the research presented.

We selected the four manipulated variables from a larger pool of potential factors that either do impact the strength of the evidence a study can provide, or are commonly thought to be related to study quality. Our final selection of these four variables was the result of balancing various considerations, including making sure that the variables could easily be manipulated, that science journalists could understand the information presented, and that we manipulated a diverse set of factors (e.g., some that would require some statistical literacy and others that would not). We describe the four variables we manipulated, along with an explanation of why we might expect each to have an effect on journalists' judgments of trustworthiness and/or newsworthiness, though we should emphasize that we did not make any predictions about these effects (though we did preregister our research questions and analysis plan, see next section).

First, we chose to manipulate sample size, a feature of studies that has recently (and historically) received a lot of attention in psychology. Quantitative studies in psychology generally aim to measure or estimate a parameter (e.g., a correlation or group difference), and the more data that is available (i.e.,

the larger the sample size or number of observations), the more precise that estimate will be. Thus, if science journalists' reasoning is consistent with scientific values, we would expect studies with larger sample sizes to be judged as more trustworthy and more newsworthy, all else being equal.

We also manipulated sample representativeness as it directly impacts how much a given finding can be generalized to the population as a whole. The mismatch between psychology's broad claims and the samples those claims are based on has been the subject of criticism for decades (Henrich et al., 2010; Sears, 1986). Findings based on unrepresentative samples (e.g., studies of human adults based on convenience samples of college students at one university) suffer from greater threats to their generalizability than do studies with more representative samples. In our study, every vignette presented to journalists included a quotation from the researcher making a general claim about people based on their finding. Thus, the representativeness of the sample (which we manipulated) should impact science journalists' evaluations of the trustworthiness (and potentially newsworthiness) of the claims in the vignettes.

The third variable we manipulated is the *p*-value associated with the finding. As a result of common pressures and practices in psychology, the *p*-value can be a useful clue for researchers to differentiate real effects from noise and guard against false positives (though, like many clues, it is far from perfect). Mathematically, well-designed studies examining real effects should most often produce very small *p*-values, well below the .05 cutoff that is commonly used. Although studies that present *p*-values closer to that threshold (e.g., between .03 and .05) are not necessarily suspect, *p*-values in that range are more common when researchers engage in questionable practices or overfit their data, and may be an indication that a finding is not as trustworthy. Thus, if science journalists are aware of and agree with this reasoning, findings with *p*-values closer to the .05 cutoff should lead them to judge the finding as less trustworthy (and presumably less newsworthy) than findings with *p*-values close to 0.

Finally, we manipulated the level of prestige of the researcher's university. While this may or may not have any bearing on the actual trustworthiness or newsworthiness of a finding, we expected that science journalists may judge research from more prestigious universities as more newsworthy, and potentially also more trustworthy, due to stereotypes journalists might hold, or might expect the public to hold, about how prestige is related to research quality and rigor. We reasoned that this might be especially likely in the context of our experiment, where the science journalists were given relatively little information.

We chose to maximize internal validity (the validity of our causal inferences) by using fictitious summaries of research to allow us to manipulate these variables independently from the content of the research presented. This design decision comes at the expense of external validity and realism. For example, our findings may have limited applicability in contexts where journalists have much more information about the research, or do not have the kind of information presented in our fictitious summaries. We discuss these and other limitations in our discussion. We believe that understanding the causal influence of the features examined here provides a foundation for future studies examining or intervening on science journalists' approach to evaluating scientific findings. Understanding whether and how science journalists' judgments are influenced by features of the research design, or by the prestige of the researchers' institutions, can point to avenues for further strengthening the role of science journalists as critics.

Method

Participants

We aimed to recruit United States-based journalists who self-identify as science journalists or journalists who sometimes report on scientific research. Participants were offered a $25 Amazon.com gift card as

compensation for the time spent taking the survey. Participants were recruited through professional networks of the 3rd author (CA), and through snowball sampling where members of CA's professional network were asked to help identify other potential participants.

Data collection started on December 22, 2020, and ended 3 months later, according to our preregistered stopping rule (see https://osf.io/kv9uw/ for the full preregistration), resulting in 186 complete observations. We excluded 5 participants who self-reported having provided inaccurate data, resulting in 181 participants. No participants were excluded for non-serious responding, and the pattern of responses did not indicate a need to worry about non-journalists having completed the survey. Although we asked participants to guess what variables we were manipulating, as per our preregistration, we did not exclude those who guessed correctly.

Stimuli

Participants were shown vignettes created by two of the authors (JGB and SV) to resemble real social and personality psychology study results. We aimed to create vignettes that varied in topic, methodology, and other characteristics, but that were similar in their format and average in plausibility and interestingness (to avoid floor and ceiling effects in ratings of trustworthiness and newsworthiness). To ensure that we achieved this goal, we first created a pool of 25 vignettes (https://osf.io/9xvfa/) that we pretested[5] to assure they were near the middle of the response scale in plausibility and interestingness. We also received feedback from one of the authors (CA) on which ones seemed to be too implausible. Based on this feedback, we eliminated 3 vignettes, for a final set of 22 vignettes, which can be found at https://osf.io/xej8k/.

[5] We tested both the full vignettes and the main claims being made in each of them. All the pretest data and code can be found at https://osf.io/tnmfu/

Design

We varied 4 characteristics in each vignette: the *sample size*, which could be small (N between 50 and 89) or large (N between 500 and 1000); the *sample type*, which could be a convenience sample (e.g., "local volunteers") or a more representative U.S. sample (e.g., "people from a nationwide sample"); the *p-value*, which could be high (between .05 and .03) or low (between .005 and .0001); and the *prestige of the university* where the research was done, which could be higher (e.g., "Yale University") or lower ("East Carolina University"). This created a 2(sample size: small vs. large) x 2(sample type: convenience vs. more representative) x 2(*p*-value: high vs. low) x 2(university prestige: higher vs. lower) design for a total of 16 conditions. We randomly assigned participants to see 8 of these 16 conditions, in a planned-missingness within-subjects design. Each of the 8 conditions was superimposed on a different, randomly selected vignette.

Due to the within-subjects design, participants were likely to see the same level of the manipulated variables multiple times (for example, small sample sizes or a high-prestige university). To make it more difficult for participants to guess what we were systematically varying across vignettes, we avoided presenting the exact same value by operationalizing each level of each variable in multiple ways. For the two numeric independent variables (sample size and *p*-value), numbers from the corresponding range were randomly sampled from a uniform distribution. For example, if a participant was assigned to see a small sample size on a given vignette, the actual sample size presented was a whole number between 50 and 89 inclusive, sampled with uniform probability. For the two independent variables that were non-numeric (sample type and university prestige), each level had a corresponding list of 8 options, and one of the options was randomly selected with equal probability. For example, if a participant was assigned to see a more representative sample type on a given vignette, the description of the sample was one of

the eight possible descriptions for more representative samples. All options can be found here: https://osf.io/xej8k/.

For each participant who started the study, all 16 possible conditions (i.e., all combinations of the 4 factors) were created, and a random operationalization of each level was selected without replacement. From these 16 conditions, a random set of 8 conditions was selected without replacement. These 8 conditions were superimposed on 8 randomly selected vignettes out of the 22, virtually guaranteeing that no participant saw the same stimuli (e.g., the same description of a convenience sample, or the same exact *p*-value) more than once. After providing their ratings of the 8 vignettes, all participants saw the rest of the questions in the same order.

Based on feedback from CA and a few other journalists who pre-tested the survey, we opted to present only half of the 16 conditions to each participant to keep the study relatively short and avoid participant fatigue. A simulation of this study design conducted a priori with 10,000 iterations revealed that a sample size between 150 to 250 participants, who each saw 8 vignettes, would give us 91 to 99% power to detect an unstandardized main effect of 0.2 scale points for each of the four variables, with alpha = .05, two-tailed. Based on the simulation results, we estimate that our obtained sample afforded us at least 95% power to detect these effects.

Procedure

After consenting to participate and confirming they met the inclusion criteria ("a journalist who (at least sometimes) covers science and/or health, medicine, psychology, social sciences, or wellbeing"), participants were presented with a series of eight one-paragraph vignettes describing fictitious findings in psychology. Participants were asked to evaluate the research described in each vignette on our two dependent variables: its trustworthiness (4 likert-type items; example item: "The methodology of this

study is rigorous.") and its newsworthiness (2 likert-type items; example item: "This study is worthy of being reported on."). See https://osf.io/xej8k/ for the full survey used.

After completing the vignette evaluations, participants were asked three open ended questions. First, they were asked to describe how they typically evaluate research findings. Then, participants were asked to describe how they evaluated the findings that were presented in the study. Finally, they were asked whether they had any guesses about what characteristics of the fictitious study vignettes we were systematically varying; if participants answered yes, they were then provided with a text box to describe their guesses.

Participants were then debriefed about each of the manipulated variables sequentially. First, we described the variable that we manipulated. Then, we asked whether they agreed that they were familiar with this variable as a factor that can be used to evaluate the quality of a study, and whether they agreed that a "better" level of this factor (e.g., larger sample size or lower *p*-value) increased the validity or newsworthiness of a scientific study. This process happened for each of the manipulated variables in turn, and all three questions were answered on a scale from -3 ("Strongly disagree") to +3 ("Strongly agree") for all four variables.

Finally, we asked participants for some demographic information (including gender, the topics they typically cover, and educational background), offered a space for any general comments, and then provided them with an opportunity to self-report that their data shouldn't be included in the analyses. At the end of the survey, participants were redirected to a separate survey where they entered identifiable information so that we could verify their identities as journalists. The two surveys were completely unlinked, and participants were told this repeatedly.

Results

Sample

The majority of the 181 participants in the final sample reported identifying as women (76.8%), with 19.3% of participants self-identifying as men, 2.8% as non-binary, and 1.1% preferred not to say. No participants chose to self-describe their gender.

Participants reported regularly covering a variety of scientific domains: life sciences (63.5%), health & medicine (58.0%), general science (46.4%), psychical sciences (37.6%), psychology (26.5%), social science (24.9%), lifestyle & wellbeing (16.6%) and other (16.6%).

In terms of the types of news organizations they primarily worked for, most participants reported working for online news organizations (92.3%), followed by print news organizations (58.6%), with a few working for audio (7.7%), video, or TV (3.9%), or another type (2.2%) of news organization. The audience of these news organizations was described as being primarily a general audience (36.5%), primarily science oriented (29.8%), or an even mix of both (33.7%).

We also asked participants about their educational background. 61.9% reported having studied physical or natural sciences at the undergraduate level, while 24.3% reported studying social sciences at the undergraduate level. Further, 35.4% reported having studied physical or natural sciences at the graduate level, while 13.3% reported studying social sciences at the graduate level; 48.1% reported having a journalism degree.

Descriptives

Our two dependent variables, trustworthiness and newsworthiness, were calculated by reverse-scoring the relevant items and averaging the items measuring each construct (4 items for Trustworthiness, 2

items for newsworthiness; Cronbach's alpha = 0.92, 0.95 respectively). Trustworthiness (*Mdn* = -0.5, *IQR* = 2.25, range = [-3, +3]) and newsworthiness (*Mdn* = -0.5, *IQR* = 3, range = [-3, +3]) did not show any evidence of floor or ceiling effects, and demonstrated reasonable variance. Judgments of trustworthiness and newsworthiness were highly correlated: participants who tended to rate vignettes as trustworthy also tended to rate vignettes as newsworthy ($r = 0.64$, $p < 2*10^{-16}$). In turn, vignettes that were typically rated as trustworthy were also more likely to be rated as newsworthy ($r = 0.85$, $p = 0.0000004$). Finally, across vignettes and participants, judgments of trustworthiness were highly associated with judgments of newsworthiness ($B = 0.73$, $p < 2*10^{-16}$).

Answers to the three open ended questions — descriptions of how participants typically evaluate research findings, how they evaluated the findings in the present study, and their guesses about which characteristics of the vignettes we manipulated — were manually coded by two coders for whether or not they contained mentions of the four manipulated variables. Initial agreement was good (Cohen's Kappa = 0.90). Disagreements were resolved by discussion between the two coders, deviating from the preregistered procedure, which stated we would use a third coder to resolve disagreements.

First, participants reported what information they typically use to evaluate findings. 66.9% of participants mentioned sample size, 27.1% mentioned the representativeness of the sample, 30.9% mentioned *p*-values, and only 16.0% mentioned the prestige of the institution where the research was conducted.

Then, participants reported what characteristics they used in forming their judgments of vignettes in the present study. 79.0% mentioned sample size, 34.3% mentioned the representativeness of the sample, 38.1% mentioned *p*-values, and only 9.4% mentioned institutional prestige.

Finally, when asked whether they had any guesses about what was being systematically varied in the vignettes, 65.7% of participants said they had a guess. Of those, 83.2% mentioned sample size, 38.7%

mentioned the type of sample, 64.7% mentioned *p*-values, and 30.3% mentioned the prestige of the university. We expected this number to be quite high, especially given that each participant read eight vignettes, so we preregistered our *a priori* decision *not* to exclude participants who correctly guessed our manipulated variables. However, this does present some concerns in terms of demand characteristics or self-presentation from the participants. We further examine some possible consequences of this for the interpretation of our results in the Limitations section of our discussion.

Table 3.1. Percent of science journalist participants who identified each of the four manipulated variables in their answers to each of three open-ended questions (answered after participants rated the eight fictitious vignettes).

Question	**Sample Size**	**Sample type**	***p*-Value**	**Uni Prestige**
What characteristics do you consider when evaluating the trustworthiness of a scientific article?	66.9%	27.1%	30.9%	16.0%
What characteristics did you weigh in judging the trustworthiness of the findings presented?	79.0%	34.3%	38.1%	9.4%
Before we tell you what [characteristics we varied], do you think you know any of them?	83.2%	38.7%	64.7%	30.3%

When asked about their familiarity with each of the independent variables as a factor that may be used when evaluating the quality of a study on a scale of -3 to +3, participants reported higher familiarity with sample size (*M* = 2.66, *Mdn* = 3, *IQR* = 1, range = [+1, +3]), followed by the representativeness of the sample (*M* = 2.31, *Mdn* = 3, *IQR* = 1, range = [-3, +3]), the *p*-value (*M* = 2.11, *Mdn* = 2, *IQR* = 1, range = [-3, +3]), and university prestige (*M* = 1.45, *Mdn* = 2, *IQR* = 2, range = [-3, +3]). The lower mean for university prestige suggests that participants may not have interpreted this question as being strictly

about familiarity with the variable, as we expect that close to 100% of our sample is likely familiar with the idea that university prestige is sometimes used as a factor when evaluating the quality of a study. We suspect participants may have treated this rating in part as an opportunity to express endorsement of the use of this factor in evaluating the quality of research studies (a construct we aimed to capture with the next item).

In terms of whether these factors could increase the validity of a study, participants generally agreed that larger sample sizes (*M* = 2.10, *Mdn* = 2, *IQR* = 1), more representative samples (M = 2.18, Mdn = 2, IQR = 1), and smaller *p*-values (*M* = 1.23, *Mdn* = 2, *IQR* = 2) can increase the validity of a study, but tended to disagree that higher university prestige could do the same (*M* = -0.57, *Mdn* = -1, *IQR* = 3).

Results were similar for newsworthiness: participants generally agreed that larger sample sizes (*M* = 1.03, *Mdn* = 1, *IQR* = 2), more representative samples (*M* = 1.33, *Mdn* = 2, *IQR* = 1), and smaller *p*-values (*M* = 0.62, *Mdn* = 1, *IQR* = 2) can increase the newsworthiness of a study, but tended to disagree that higher university prestige could do the same (*M* = -0.41, *Mdn* = 0, *IQR* = 3).

Main research questions

To examine our first research question, we fit a linear mixed-effects model in which trustworthiness ratings were predicted by each of the four independent variables: the sample size, the sample type, the *p*-value, and the prestige of the university. In addition to the main effect of each variable, our model also included random intercepts for participant and vignette, allowing us to account for variability due to their particular characteristics (e.g., a participant's general tendency to rate vignettes as more trustworthy).

University prestige did not affect trustworthiness ratings (*B* = -0.01, bootstrapped 95% CI = [-0.13, 0.11], *t* = -0.09, *p* = .926), and neither did having a more representative sample instead of a convenience

sample (*B* = 0.12 [-0.00 0.23], *t* = 1.95, *p* = .051). Having a lower *p*-value did have a very small but statistically significant effect on trustworthiness ratings (*B* = 0.15 [0.05, 0.27], *t* = 2.61, p = .009). The most robust effect was that larger sample sizes led to higher ratings of trustworthiness (*B* = 0.73 [0.61, 0.85], *t* = 12.44, $p < 2*10^{-16}$), such that having a small sample (*N* between 50 and 89) led to studies being rated 0.73 points lower on a -3 to +3 scale than having a larger sample (N between 500 and 1000).

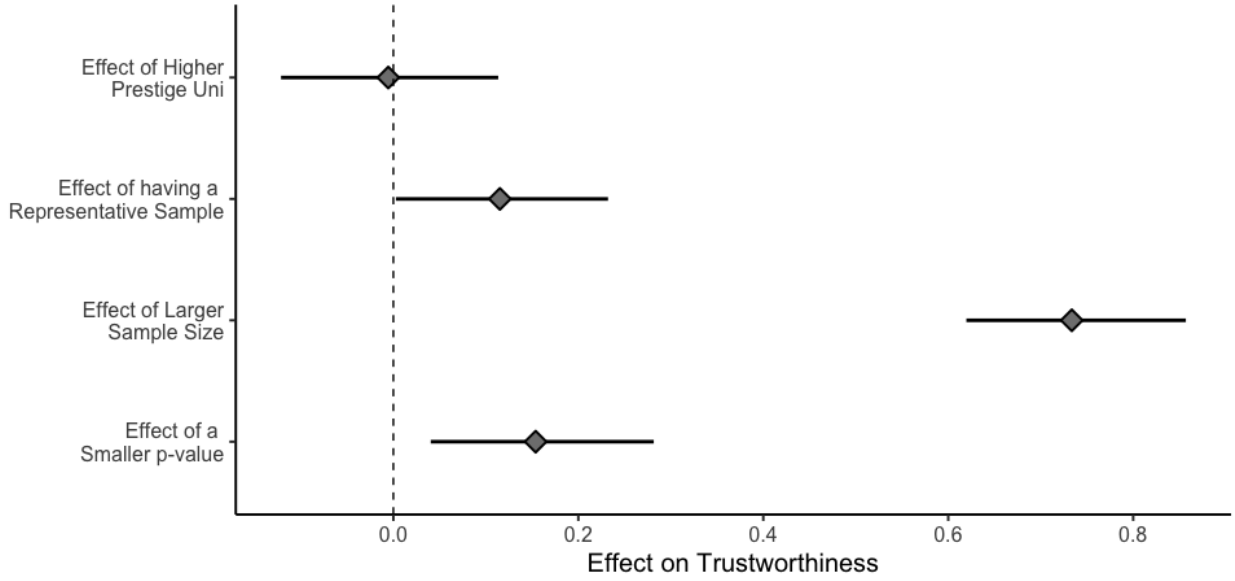

Figure 3.1. What factors influence journalists' ratings of trustworthiness? Trustworthiness was rated on a -3 to +3 scale; bootstrapped 95% CIs. Effects are presented in raw units (difference between two conditions, e.g., smaller vs. larger *p*-value)

We fit a similar mixed-effects model to examine our second research question: whether newsworthiness ratings were predicted by each of the four independent variables. For the same reasons as before, we also included random intercepts for participant and vignette.

Results were quite similar: university prestige did not affect newsworthiness ratings (*B* = 0.03, bootstrapped 95% CI = [-0.12, 0.16], *t* = 0.39, *p* = .696), and neither did having a more representative sample instead of a convenience sample (*B* = 0.09 [-0.04, 0.22], *t* = 1.35, *p* = .179). Having a lower *p*-

value had a very small but significant effect on newsworthiness ratings (B = 0.15 [0.03, 0.28], t = 2.31, p = .021). Once more, the only robust effect was that vignettes with larger sample sizes were perceived as more newsworthy (B = 0.59 [0.45, 0.72], t = 9.01, p <$2*10^{-16}$), such that having a small sample led to studies being rated 0.59 points lower on newsworthiness on our -3 to +3 scale compared to having a larger sample.

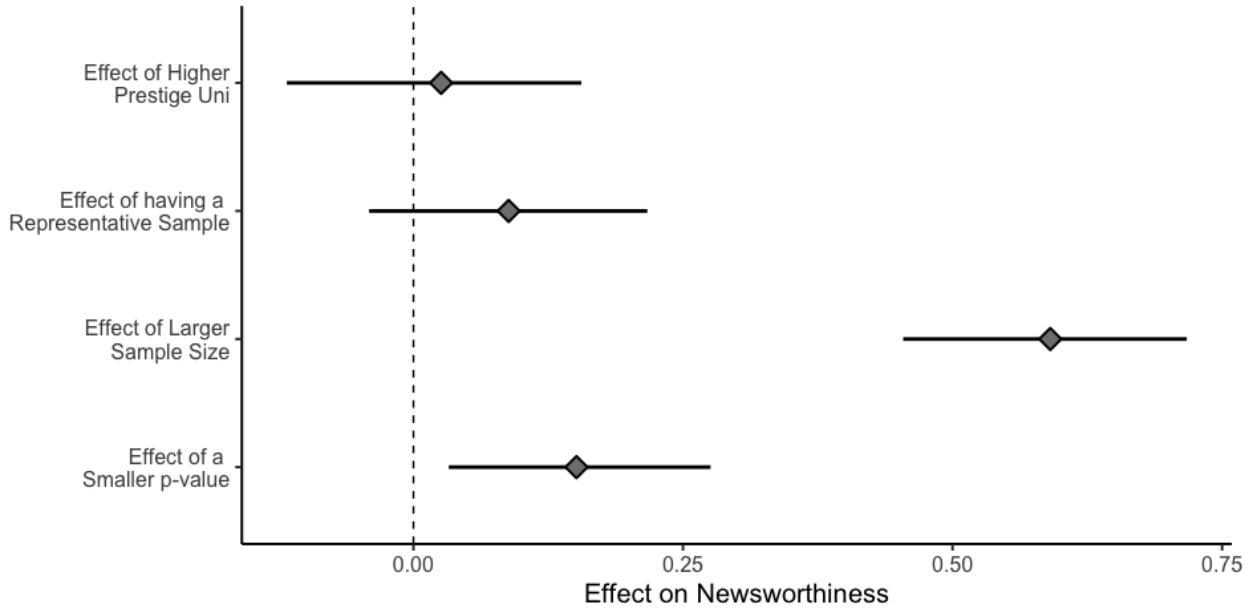

Figure 3.2. What factors influence journalists' ratings of newsworthiness? Newsworthiness was rated on a -3 to +3 scale; bootstrapped 95% CIs. Effects are presented in raw units (difference between two conditions, e.g., smaller vs. larger *p*-value).

Exploratory analyses

All analyses presented below were not preregistered and were run after we had already seen the data and results.

Variance partitioning

How much of the variance in the dependent variables (trustworthiness and newsworthiness) can our model explain, and how much of that is due to the vignettes, the participants, or the manipulated variables? For the trustworthiness model, 38% of the variance in this dependent variable can be explained by different parts of the model. First, about 20% of the variance is due to between-participant differences, while 11% is due to between-vignette differences. Finally, around 7% of the variance is explained by our manipulated variables, most of which is due to the sample size variable.

For the newsworthiness model, 39% of the variance in this dependent variable can be explained by different parts of the model. First, 26% of the variance is due to between-participant differences, while 9% is due to between-vignette differences. Finally, 4% of the variance is explained by our manipulated variables, most of which is due to the sample size variable.

Associations with Vignette Characteristics

The vignettes naturally varied in several ways other than the manipulated characteristics. These differences were accidental — in coming up with hypothetical studies and findings, we naturally came up with vignettes that varied in their design, and in the characteristics of the researchers. Importantly, unlike our manipulated variables, these variables were not randomly assigned to vignettes — they were confounded with the content of the vignettes. While we had no plans to examine any of the variables on which the vignettes naturally varied, after reading participants' open ended answers about what factors they used to evaluate the vignettes (see below for a summary of some of the most common themes), two characteristics stood out to the coders as potentially explaining some of the variance in participants' ratings. The first was the design of the study — whether it was experimental or correlational. As it happened, 11 out of the 22 vignettes we used presented a study that was described as (or strongly implied to be) an experiment, and 11 presented a study whose design was correlational. The second

characteristic that stood out to us was the implied ethnicity of the researcher. Every vignette included a quote from the fictitious researcher, along with the researcher's last name. As it happened, 11 out of the 22 vignettes we used presented a last name that was likely to be perceived as non-White or as Hispanic (e.g., Zheng, Mustafa, Rivera), and 11 presented a last name that was unlikely to be perceived as non-White or as Hispanic (e.g., Quinn, Cabot, Carter). Although these two characteristics were not experimentally manipulated, we could still examine the association between judgments of trustworthiness/newsworthiness and these variables as they naturally varied across vignettes.

To examine the association between the study design (experimental vs. correlational) and trustworthiness ratings, we fit a mixed-effects model with the study design variable as the only predictor, as well as random intercepts for participant and vignette, similarly to our models for the main analyses. We see that having an experimental design was a good predictor of trust in a vignette's findings (B = 0.56, bootstrapped 95% CI = [0.26, 0.92], t = 3.31, p = .003), such that people rated vignettes with an experimental design as being half a point higher on trustworthiness on average. This effect was similar for newsworthiness: vignettes with experimental designs received higher newsworthiness ratings than vignettes with correlational designs (B = 0.47 [0.11, 0.82], t = 2.59, p = .018).

There was no association between the implied race/ethnicity of the researcher and either trustworthiness or newsworthiness ratings (trustworthiness: B = -0.12 [-0.58, 0.30], t = -0.59, p = .563; newsworthiness: B = -0.01 [-0.40, 0.38], t = -0.03, p = .974).

Themes in the open-ended responses

After participants had read all eight vignettes, we asked them three open-ended questions: "What characteristics do you [typically] consider when evaluating the trustworthiness of a scientific finding?", "What characteristics did you weigh in judging the trustworthiness of the findings presented?" and "[...]

we varied some characteristics of the fictional studies you read about. [...] do you think you know what any of them were?" As reported above (Table 1), participants often mentioned the four manipulated variables — the sample size and its representativeness, the *p*-value, and the prestige of the university where the research was done — across all three questions. However, there were also a few other recurring themes. Table 2 shows the most interesting themes we noticed in responses to each of the three questions, with a few selected examples for each. This is a subjective, non-systematic exploration of the topics brought up by the participants in their free responses, but it serves to illustrate the content of the answers given by the participants and potentially generate new research questions for future research.

When asked what characteristics of a study they considered when evaluating its trustworthiness, participants often mentioned the prestige of the journal where it was published or the fact that it had been peer reviewed. Many participants also seemed to value experimental methodology, or methodology that allows researchers to make causal claims. Some answers suggested that journalists do take statistical significance into account, but only very few included explanations that suggested they made any distinction between higher or lower *p*-values; instead, most mentions of *p*-values suggests journalists focused on whether or not the key result was statistically significant

Many participants mentioned that it was very important to talk to outside experts or researchers in the same field to get a better understanding of the finding and whether it could be trusted. Journalists also expressed that it was important to understand who funded the study, and whether the researchers or funders had any conflicts of interest. Finally, they indicated that making claims that were calibrated to the evidence was also important and expressed misgivings about studies for which the conclusions don't follow from the evidence.

When explaining what details in the vignettes they had relied on to judge the trustworthiness of the presented studies, many of these same themes resurfaced, including mentions of overclaiming or claims that were not calibrated to the evidence or design of the study and concerns about the methods that were used, including making causal claims based on correlational evidence. Participants also mentioned the general plausibility of the findings as an important factor in helping them evaluate the trustworthiness of each study. Importantly, many participants said they did not have enough information to evaluate the presented study vignettes, and implied that their typical evaluation when deciding whether to pursue a story is much more thorough, sometimes relying on outside experts, a full press release, or the entire paper.

Finally, participants correctly guessed our manipulated variables quite often, but many other naturally varying characteristics of the vignettes were also brought up. Many participants guessed we were manipulating the study design, or the ethnicity of the researcher through last names, as well as the level of overclaiming or the calibration of the claim being made at the end to the evidence presented. Interestingly, a few participants guessed that we manipulated the gender of the researcher (e.g., "I also noticed that the gender of the quoted researcher seemed to be male in most of the paragraphs."), even though no information about researcher gender was provided in the vignettes.

Table 3.2. Selected examples of themes in open-ended responses beyond the four manipulated variables.

Question: "What characteristics do you [typically] consider when evaluating the trustworthiness of a scientific finding?"	
Theme	**Examples**
Using journal prestige as an indicator of research quality	"The journal the work is reported in." "I want to know that it has been published in a reputable journal" "the journal itself where the findings were published and its impact factor" "The name of the academic journal the study is in matters more to me than the name of

	the university the research came from." "Was it published in a legit journal?" "Quality of journal is the first key barometer."
Published in peer reviewed journal	"Also has it been published in a peer-reviewed journal or other publication." "Whether the study was rigorously peer reviewed or not." "publication status (preferably with peer review but that's no guarantee of good quality)" "I don't have the training to do forensics on papers, so to stay out of trouble I make sure the finding has passed through peer review for a reputable journal and assume the reviewers would have picked up anything egregious."
Experimental vs. correlational	"Does the methodology allow for causal interpretations or just associations?" "I'm also always looking for whether there is a baseline assessment of whether findings are actually causal or whether there is acknowledgement that the study only establishes a correlation." "I would want to see that they made an effort to distinguish between correlation and causation." "Are they claiming causation or simply showing correlation? If causation, is it a randomized control trial?" "observational vs randomized design"
Significant *p*-value vs. non-significant	"A p-value of 0.01 or less is preferred, but must be less than 0.05" "whether the finding was statistically significant" "I also look at p-values for trustworthiness, but I admit that I'm not great at statistics" "And p-value does play a role, but that's just the lowest bar to clear" "How significant was the outcome?" "[...] very small p value." "I do look at P-values even though I know those can be massaged."
Comments from other researchers	"And perhaps most importantly, I interview external sources who are experts in the field at hand and gauge what their perspectives are on the new finding. Are they enthusiastic? Skeptical? Unsupportive?" "whether outside experts trust the findings" "[...]I also rely on outside experts to provide crucial comments. I have *never* reported a story without including at least one outside comment, in order to place the study within the proper context." "I lean heavily on interviews with outside sources when evaluating the trustworthiness

	of a study." "the opinions of outside experts who are in the field but who weren't involved with the study"
Funding and other conflicts of interest	"Researcher conflicts of interest or funding sources." "Who paid for the research? Who conducted the research and would influencing the result consciously or unconsciously benefit them?" "A primary factor is the potential presence of conflicts of interest" " I also like to [...] investigate whether the researcher or their institution has any stake in the results." "Have they disclosed potential conflicts of interest, and if so, what are they?"
Overstating conclusions / claims vs. claims calibrated to evidence	"Is the researcher adamant that this study of 40 college kids is representative? If so, that's a red flag." "whether authors make sweeping generalizations based on the study or take a more measured approach to sharing and promoting it" "I also consider how surprising the claim is. If very unexpected, the evidence must be very strong." "Another major point for me is how 'certain' the scientists appear to be when commenting on their findings. If a researcher makes claims which I consider to be over-the-top about the validity or impact of their findings, I often won't cover." "I also look at the difference between what an experiment actually shows versus the conclusion researchers draw from it--if there's a big gap, that's a huge red flag"
Question: "What characteristics did you weigh in judging the trustworthiness of the findings presented?"	
Theme	**Examples**
Overclaiming	"I paid attention to whether the interpretations of the findings were reasonable, whether the quotes from researchers made statements implying causation or overgeneralization" "Often, my reaction was based on the quote from the scientist. If they're applying a complex human behavior based on a correlation, whatever its strength, I would want to see more about their study design." "I looked for whether the scientist's quote seemed to overplay the findings." "I was forced to base my decision on if it seemed like the results were being overstated or there were other plausible explanations for the result they got that they failed to mention." "The strength/confidence of the concluding statement in relation to the statistical significance (overconfident=less trustworthy)"

General plausibility and relevance of finding/claims	"some of the studies were frivolous--like who would even care" "Also, I evaluated how much the given task was related to the real world. For example, a driving game might be a useful paradigm, but in the end it really only tells you about people's behavior inside a driving game, not about how people behave when driving an actual car on a real road." "some statements didn't seem plausible (e.g., news release that said sleeping entirely explains memory lapses) but I assumed that was the fault of the person writing the news release and not reflective of the trustworthiness of the actual study" "whether the research question seemed kind of absurd or silly to begin with" "most findings barely more astute than obvious"
Design/methods	"Whether they were randomized; whether good controls were in place" "Or if the method seemed wacky (e.g. watching an "anger-inducing video" and then playing a video game, to judge people's driving)." "whether the experiment was measuring association or causality" "I found that small sample sizes and experimental design that didn't seem to adequately address the question that was supposedly being answered were red flags" "3rd most important--what was being measured and how" "whether the experiment was observational versus interventional." "In addition, experimental studies vs online surveys and self-reports are more trustworthy."
Not having enough information when evaluating results / implying evaluation in realistic context is more thorough	"This was a difficult exercise as there wasn't much information to go on." "Since I couldn't vet the researchers described, I weighed the strength of the study design as best I could from details provided." "My gut reactions here on what was more trustworthy would just guide which studies I might go on to investigate further, not what I would pitch to an editor or report and write." "Most of these paragraph lacked sufficient detail I would need to make a decision." "For any of these cases, I would have wanted more information before deciding to cover or not cover any of them in real life." "I didn't have enough information to adequately evaluate any of these samples."
Question: "[...] we varied some characteristics of the fictional studies you read about. [...] do you think you know what any of them were?"	
Theme	**Examples**

Study design	"whether the studies were survey or experiments" "observational studies vs. randomized controlled trials" "whether randomized or not" "one pattern I noticed was that some studies were observational and others were experimental" "intervention vs observational"
Researcher ethnicity / names	"names of the researchers which might indicate different races/ethnicities" "names/ ethnicity of researcher" "And the perceived ethnicity/gender of the researcher quoted." Race/ethnicity of the authors"
Quotes or claims	"author claims/conclusions and their level of speculation" "presence of a quote from a researcher" "strength of what the researcher said about the study in their quote" "To what extent researchers made conclusions that the study could not support--people driving less carefully while irritated seemed a more reasonable interpretation of a study than some of the others, like the case of the video and gift expectations." "how strongly the results were stated" "whether the findings were presented in a way that went beyond the actual findings"

Note. See https://osf.io/drzga/ for an extended version of this table of examples.

Discussion

In this study, we examined the causal influence of four factors on science journalists' evaluations of the trustworthiness and newsworthiness of fictitious psychology findings. This study is a first step towards understanding the process that science journalists go through when reading, evaluating, and reporting on science.

Sample size was the only one of the four manipulated factors that had a robust influence on participants' ratings of trustworthiness and newsworthiness across vignettes. This effect was consistent

with scientific reasoning; larger samples provide more evidence and precision, and so are generally considered more trustworthy, all else being equal. The magnitude of the effect of sample size on science journalists' evaluations was modest; vignettes describing studies with larger samples were perceived as more trustworthy and newsworthy by about two thirds of one point on a 7-point scale compared to vignettes describing studies with smaller samples. The finding that this was the only manipulated variable that influenced participants' evaluations is consistent with participants' self-reports on the open-ended question: when asked what factors they used to evaluate the findings in this study, 79% mentioned using sample size, the highest of any of the four manipulated factors.

In contrast, the other three factors did not have appreciable effects on science journalists' perceptions of the trustworthiness or newsworthiness of studies. Studies with samples that were more versus less representative, results with small (close to zero) versus large (close to .05, but still significant) *p*-values, and research from more versus less prestigious institutions were all perceived similarly by our participants. Our study was designed to provide quite precise estimates, and indeed the 95% confidence intervals around these results exclude all effects larger than about one third of a point on a seven-point scale. Thus, we can be fairly confident that these factors do not have a practically significant impact on these journalists' perceptions in this kind of context. In short, a study's sample size seemed to matter quite a bit more for journalists' evaluations of research than did the representativeness of the study's sample, the *p*-value of the (significant) result, or the prestige of the researcher's institution.

While we did not preregister any predictions, we admit to being surprised by the finding that journalists' evaluations of research were not affected by the prestige of the authors' institutions. We expected journalists to perceive research conducted at more prestigious institutions to be both more trustworthy and more newsworthy. Our results suggest this stereotype — that science journalists are influenced by

flashy university names — is wrong, at least in contexts similar to this experiment and with journalists similar to those in our sample. We discuss some potential alternative explanations for this finding below.

We also found that manipulating the representativeness of the samples did not affect journalists' perceptions of trustworthiness or newsworthiness. Again, while we did not preregister any predictions, we found this surprising, given how often social science research is critiqued for relying too heavily on student samples (e.g., Rad et al., 2018). Moreover, there has been a big push for more diverse samples in the social sciences (Henrich et al., 2010), though it is not clear that this has translated into better research practices on this front. Nevertheless, we suspect that most social scientists assume that their research is judged, in part, based on the effort they put into recruiting samples that match the population they wish to understand. If it is in fact the case that the representativeness of the sample does not affect how science journalists evaluate research, this would suggest a mismatch between what social scientists believe that the media or the public care about, and what science journalists care about. Below we discuss some alternative explanations for this finding.

We also did not detect much effect of the research finding's exact *p*-value on journalists' evaluations. Both levels of this manipulation presented vignettes whose results had statistically significant *p*-values, but in one condition *p*-values were far below the significance threshold of .05 (e.g., $p = .003$) whereas in the other condition *p*-values were only just below .05 (e.g., $p = .041$). Effects that are non-zero in the population are much more likely to produce *p*-values close to zero than close to .05 in sample data, and that *p*-values close to .05 indicate results that are more likely to have been influenced (likely inadvertently) by *p*-hacking or overfitting (compared to results with *p*-values close to zero; Simmons et al., 2011; Simonsohn et al., 2014). We did not necessarily expect science journalists to be familiar with, or to apply, this reasoning (and, again, we did not make any predictions regarding the effect of this manipulation). Our results, in conjunction with participants' open ended responses (see Table 2),

suggest that there may be some journalists who were influenced by this manipulation of *p*-values, but most were not. This suggests a potential area for greater communication between methodologists and science journalists, to ensure that information about how to evaluate statistical results is shared with those who can use it.

Exploratory analyses of participants' open-ended comments (see Table 2) suggest that other factors we did not manipulate may be influential. For example, while manipulating *p*-values within the statistically-significant range (0 to .05) did not have much of an effect on journalists' evaluations, their comments suggest that whether or not the *p*-value is statistically significant (i.e., $p < .05$ vs. $p > .05$) may matter more. Similarly, while manipulating the prestige of the researchers' universities did not have an effect on journalists' evaluations, their comments suggest that the prestige of the journal in which the research is published may matter more. Finally, whether the study was experimental or observational (which we varied incidentally across vignettes but did not manipulate in a controlled fashion) also seemed to potentially play a role in journalists' evaluations, according to their comments. We conducted exploratory analyses to test this and found that, indeed, vignettes presenting experimental research were perceived as more trustworthy and newsworthy than those presenting observational research (though we cannot rule out confounds as we did not systematically manipulate this variable).

Limitations

Although we believe our results are an important first step in understanding the process through which scientific findings get communicated to the public, our study has several limitations. First, having journalists evaluate scientific findings by presenting them only with short, fictitious study vignettes enhances experimental control but is highly artificial when compared to science journalists' usual process. This was reflected in participants' answers to the open-ended question asking them what information they used when evaluating the study vignettes; many participants mentioned not having

enough information, or implied that their real evaluation process is much more thorough than this context allowed (e.g., “This was a difficult exercise as there wasn't much information to go on.”, see Table 2 for more examples) This problem could be mitigated by more accurately mirroring journalists’ process as they evaluate research findings. One solution might be to use press releases for real findings, which would include more information than our vignettes, and are closer in format to what journalists first see when they encounter a potential topic to write about.

We feel confident that we can generalize from the results of our 22 specific vignettes to other possible vignettes containing similar types and amounts of information because we modeled vignette as a random factor, and because journalists’ ratings of the vignettes’ trustworthiness and newsworthiness suggest there was a good amount of variance among these 22. However, we cannot generalize beyond this type of vignette, which all presented fictitious behavioral psychology findings. Thus, we have no reason to believe these results would generalize even to studies of how journalists evaluate findings in other subfields of psychology, much less findings in other fields.

Another limitation of this study is how the sample of science journalists was obtained. We used a snowball-like sampling approach: one of the authors (Christie Aschwanden) sent a recruiting message to her professional network, and the message included a request to forward it to other science journalists they might know. Although the size of the sample we obtained (N = 181) suggests we were able to collect a range of perspectives, we suspect this sample is biased by an “Aschwanden effect”: that science journalists in the same professional network as Aschwanden will be more familiar with issues related to the replication crisis in psychology and subsequent methodological reform, a topic Aschwanden has covered extensively in her work. Therefore, we are not confident that the results obtained here can be generalized to all U.S.-based science journalists. Instead, our conclusions should be

circumscribed to U.S.-based science journalists who are at least somewhat familiar with the statistical and replication challenges facing science.

Another limitation is that we used ad-hoc items and scales for all of our measures. We do not think this was a problem for the key measures: both the trustworthiness and newsworthiness dependent variables had multiple items with very good reliability, and the concepts were straightforward. However, some of the measures used for secondary analyses clearly had problems. For example, when asked to describe their familiarity with all four manipulated variables as a factor that may be used when evaluating the quality of a study, participants reported being much less familiar with university prestige than any of the other three variables. This suggests participants did not interpret this question as intended (i.e., as being about familiarity), as we would expect all participants to be extremely familiar with university prestige as a factor that could be used as an indicator of research quality. Rather, the result suggests at least some participants were expressing their disagreement that university prestige is a *valid* indicator that a particular finding can be trusted. These items were not used in our primary analyses.

Another concern is that the journalists in our study were aware of being studied, which may give rise to demand characteristics or concerns about self-presentation. The participants were often highly-trained science journalists, and many accurately guessed what variables were being manipulated across vignettes: nearly two thirds hazarded a guess about what variables were being manipulated, with over 95% of those correctly guessing at least one manipulated variable. We anticipated this, and decided a priori not to treat data from those participants differently, because we did not think that being aware of the research hypotheses would alter participants' responses or present a threat to the validity of our conclusions. Nevertheless, whether or not participants could guess what was being manipulated, we are concerned about the possibility that participants' responses could have been influenced by feeling that their answers might be judged. Specifically, it is also possible that participants believed there was a

"right" answer (or believed that the researchers were looking for a specific answer) and, instead of answering honestly, gave the answer that they believed would be judged as correct. This poses a problem for interpreting our results. For example, it may not be that participants truly trust research with larger samples more, but instead, believe that to be the "correct" answer, or the answer the researchers were looking for. This is especially plausible because, indeed, we (and likely other researchers) do in fact believe that some of the factors we manipulated should influence science journalists' judgments of trustworthiness and newsworthiness.

Another challenge to interpreting our results is that it is difficult to compare the results across the four manipulated factors. Some manipulations were probably more salient than others. For example, while the representativeness of the sample and the university where the research was done were presented in text and blended in with the rest of the vignette, the *p*-value and sample size were presented numerically, and likely stood out more, visually. In particular, the sample size information may have stood out because in 10 of the 22 vignettes, the sample size number was at the start of a sentence, which is unusual — and incorrect, according to Associated Press style guidelines, which journalists will certainly be familiar with. Such differences in how salient the manipulated factors were likely to be can make it difficult to compare effects across these factors.

Some of these problems arise from the fact that we had to dichotomize our variables to be able to manipulate them (e.g., decide on a "high" vs. "low" level) when all four are, in fact, continuous. We attempted to select two slices of these continuous variables that we thought would show the largest difference while still being realistic and appropriate for our study design. However, our findings may not generalize to other levels of these variables. For example, sample size may matter a lot more than the other factors for the specific range we used, but there is almost certainly a range after which it will stop mattering as much. Is a study with 3,000 people noticeably more trustworthy than one with 1,000?

Further, another characteristic that journalists reported using when they evaluate findings is whether or not a study was significant (p-values above vs. below 0.05), suggesting that this "slice" of the *p*-value variable might also be worth investigating. In the end, we cannot be sure that other slices of the same variables would not show stronger or weaker effects.

More broadly, this raises the question of whether we selected the most relevant factors to manipulate. Our exploratory analyses suggest that vignettes presenting an experimental study received higher ratings of trustworthiness and newsworthiness when compared to those that included an observational study. This correlational evidence, in conjunction with the comments from journalists which often mentioned experimental design as a factor they use when evaluating findings, suggests that this and other variables we did not manipulate might have an important impact on journalists' evaluations.

Implications

Despite these issues, our results do suggest some practical implications. In particular, the result for sample size might indicate that, for some science journalists, samples smaller than a hundred participants (in a typical, between-subjects social psychology study) are particularly untrustworthy — enough to noticeably decrease their evaluations. This might suggest that the perils of small sample sizes are a more talked-about issue, making it an easy basis for a "rule of thumb". Although a small sample can be indicative of methodological issues, if attention to sample size comes at the expense of attention to more diagnostic information presented, this could be a problem. In our study, sample size had a substantially larger effect on journalists' evaluations of research than did other factors that are arguably equally — or more — important, such as the extremity of the statistical result and the representativeness of the sample.

The lack of any effect of sample representativeness is interesting, especially for psychologists, who often do research on convenience samples. Within psychology, there seems to be much talk and little action

about this problem. The results of the current study suggest that perhaps the lack of action could be due in part to a lack of consequences — findings based on samples that are very unrepresentative of the population that researchers claim to be studying (e.g., "undergraduates at the university") were rated as just as trustworthy and newsworthy as findings from studies where the sample and population are more similar (e.g., "people from a nationwide sample"). Given the extra effort often required to recruit more representative samples, if the consequences are trivial, at least as far as media exposure and criticism from journalists go, then this could help perpetuate the status quo.

In addition, we feel confident concluding that even for science journalists who are familiar with the replication crisis, how close a significant *p*-value is to the threshold for significance does not affect journalists' evaluations, on average. We suspect that if this sample of journalists was not influenced by variation in the extremity of statistically significant *p*-values, then it is unlikely other science journalists would be. Once again, this suggests that there may not be negative consequences (in terms of media attention and credibility in journalists' eyes) for engaging in questionable research practices (e.g., *p*-hacking), nor much reward for engaging in practices that reduce the risk of *p*-hacking.

Finally, while we are heartened to see that university prestige did not seem to influence journalists' evaluations of research findings in our study, we do not think that this means that journalists do not use prestige-related clues. Given that many prestige-related variables are problematic (i.e., perpetuate inequalities and are poor proxies for quality), we encourage more research into other possible prestige-related factors that influence journalists' evaluations. For example, several of the journalists in our study mentioned using journal prestige when asked about how they normally evaluate findings ("[I pay attention to] the journal itself where the findings were published and its impact factor", see Table 2 for more examples). If journalists are influenced by journal prestige more than by university prestige, this may suggest that journalists' evaluation process is similar to researchers' (who notoriously use journal

prestige as a shorthand for research quality (Harney et al., 2021). This would be concerning because of the growing evidence that using journal prestige as an indicator of research quality is invalid and harmful (Vazire, 2017; Brembs et al., 2013).

Future Directions

How can we gain more insight into what factors science journalists use to evaluate research? First, qualitative studies would be helpful for generating hypotheses. For example, our selection of the four manipulated variables was driven by our interests in these factors and our perception that they may or should play a role in journalists' evaluations. However, if we had asked science journalists to describe their process, we might have selected a different set of variables (e.g., journal prestige instead of university prestige). The open ended questions in this study provide a glimpse into what we could learn from a qualitative study, and are a good starting point when generating ideas for follow-up studies.

One avenue for future research is to observe journalists in their usual decision-making process. For example, by collaborating with practicing science journalists, observational studies could compare what findings journalists report on versus what they hear about but decide not to report on. This would allow us to examine what factors are associated with these decisions. Observational studies could also examine what questions journalists ask when they contact independent researchers to inform their evaluation of the research they are considering reporting on. This may offer insights into what aspects of the study are most important for journalists' judgments of trustworthiness and newsworthiness. Finally, observational studies could examine what is taught in journalism schools and training programs. What tools are these training programs providing prospective journalists for evaluating research?

Finally, if the goal is to improve the quality of science communication, conducting intervention studies might clarify what actions would get us there faster. For example, would educating journalists about methodological issues have a substantial impact on how they decide which findings to report on? This

might be especially valuable if done at journalism schools, by providing journalists with the tools they need to evaluate the soundness of research early in their careers. Another possible avenue for interventions is training journalists on how to ask better questions of scientists. Several participants mentioned that getting the perspective of independent researchers was part of their process when deciding whether to report on a study; doing so more effectively might improve their ability to evaluate the quality of the evidence in a scientific article.

Conclusion

In this contrived setting, with this specific sample, the sample size of the study presented to science journalists seemed to have an important influence on their evaluations of the trustworthiness and newsworthiness of the research. In contrast, the representativeness of the sample in the research presented, the level of statistical significance of the research finding's *p*-value, and the prestige of the researchers' affiliation did not seem to influence science journalists' evaluations. Our exploratory results suggest that experimental research is trusted more than observational research, while the open-ended responses suggest that the prestige of the journal where the research is published is also worth looking into. Quantitative and experimental studies like ours help us understand what implicit and/or explicit rules journalists might use when evaluating scientific studies.

References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PloS one*, *12*(3).

AllTrials. (n.d.). About AllTrials. Retrieved June 13, 2019, from AllTrials website: https://www.alltrials.net/find-out-more/all-trials/

Bastian, H. (2017, August 29). Bias in Open Science Advocacy: The Case of Article Badges for Data Sharing. Retrieved November 24, 2019, from Absolutely Maybe website: https://blogs.plos.org/absolutely-maybe/2017/08/29/bias-in-open-science-advocacy-the-case-of-article-badges-for-data-sharing/

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1-48. doi:10.18637/jss.v067.i01.

Bendels, M. H., Müller, R., Brueggmann, D., & Groneberg, D. A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. PloS one, 13(1), e0189136.

Blum, D. (2021). Science journalism grows up. *Science, 372*(6540), 323-323.

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: unintended consequences of journal rank. *Frontiers in human Neuroscience*, 291.

Carr, T. (2019, July 15). *Revisiting the Role of the Science Journalist.* Undark Magazine. https://undark.org/2019/07/15/science-journalism-communications/

Crewdson, J. (1993). Perky cheerleaders. *Nieman Reports*, *47*(4), 11-16.

Cummings, J. A., Zagrodney, J. M., & Day, T. E. (2015). Impact of Open Data Policies on Consent to Participate in Human Subjects Research: Discrepancies between Participant Action and Reported Concerns. *PLoS ONE*, *10*(5). https://doi.org/10.1371/journal.pone.0125208

Dance, A. (2018). On the record. *Nature*, *562*(7725), 153-156.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378.

Fox, N. W., Honeycutt, N., & Jussim, L. (2018, August 14). How Many Psychologists Use Questionable Research Practices? Estimating the Population Size of Current QRP Users. https://doi.org/10.31234/osf.io/3v7hx

Frequently Asked Questions about the NIH Public Access Policy | publicaccess.nih.gov. (n.d.). Retrieved June 13, 2019, from https://publicaccess.nih.gov/faq.htm#753

Harney, J., Mayville, L., Hrynaszkiewicz, I., & Kiermer, V. (2021, July 29). Researchers' Goals When Assessing Credibility and Impact in Committees and in Their Own Work. https://doi.org/10.31219/osf.io/ryds4

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature human behaviour*, *5*(12), 1602-1607.

Holman, L., Stuart-Fox, D., & Hauser, C. E. (2018). The gender gap in science: How long until women are equally represented?. *PLoS biology*, *16*(4), e2004956.

Ihde, D. (1997). Why not science critics?. I*nternational studies in philosophy, 29*(1), 45-54.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Katz, Y. (2016, October 25). *Cheerleading with an agenda: how the press covers science.* 3:AM Magazine. https://www.3ammagazine.com/3am/cheerleading-agenda-press-covers-science/

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Knobloch-Westerwick, S., Glynn, C. J., & Huge, M. (2013). The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. *Science communication*, *35*(5), 603-625.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, τ. 82, pp. 1–26. doi:10.18637/jss.v082.i13

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M. J. (2018). The accuracy of measurements with probability and nonprobability survey samples: Replication and extension. *Public Opinion Quarterly*, *82*(4), 707-744. https://doi.org/10.1093/poq/nfy038

Makel, M. C., Hodges, J., Cook, B. G., & Plucker, J. (2019, October 31). Questionable and Open Research Practices in Education Research. https://doi.org/10.35542/osf.io/f7srb

McSweeney, B., Allegretti, J. R., Fischer, M., Monaghan, T., Mullish, B. H., Petrof, E. O., … Kao, D. H. (n.d.). Potential Motivators and Deterrents for Stool Donors: A Multicenter Study. Retrieved February 14, 2019, from https://ep70.eventpilot.us/web/page.php?page=IntHtml&project=DDW18&id=2907807

Mello, M. M., Lieou, V., & Goodman, S. N. (2018). Clinical Trial Participants' Views of the Risks and Benefits of Data Sharing. *New England Journal of Medicine*, *378*(23), 2202–2211. https://doi.org/10.1056/NEJMsa1713258

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*(1), 34–58. https://doi.org/10.1037/pspa0000084

Open Science Badges. (n.d.). Retrieved June 6, 2019, from https://cos.io/our-services/open-science-badges/

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

Pashler, H., & De Ruiter, J. P. (2017). Taking responsibility for our field's reputation. *APS Observer*, *30*.

Patient Groups, Industry Seek Changes to Rare Disease Drug Guidance. (n.d.). Retrieved June 13, 2019, from https://www.raps.org/news-and-articles/news-articles/2019/4/patient-groups-industry-seek-changes-to-rare-dise

Pennebaker, J.W., Booth, R.J., Boyd, R.L., & Francis, M.E. (2015). Linguistic Inquiry and Word Count: LIWC2015. Austin, TX: Pennebaker Conglomerates (www.LIWC.net)

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

Protection of Human Subjects. , Pub. L. No. 45, § 46, C.F.R. (2009).

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401-11405.

Rowhani-Farid, A., Allen, M., & Barnett, A. G. (2017). What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review*, *2*(1), 4. https://doi.org/10.1186/s41073-017-0028-9

Sanderson, S. C., Linderman, M. D., Suckiel, S. A., Diaz, G. A., Zinberg, R. E., Ferryman, K., … Schadt, E. E. (2016). Motivations, concerns and preferences of personal genome sequencing research participants: Baseline findings from the HealthSeq project. *European Journal of Human Genetics*, *24*(1), 14–20. https://doi.org/10.1038/ejhg.2015.118

Schiavone, S. R., & Vazire, S. (2022). Reckoning With Our Crisis: An Agenda for the Field of Social and Personality Psychology. *Perspectives on Psychological Science*, 17456916221101060.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, *51*(3), 515. https://doi.org/10.1037/0022-3514.51.3.515

Silge, J. & Robinson, D. (2016). tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS*, *1*(3). doi: 10.21105/joss.00037

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.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of experimental psychology: General*, *143*(2), 534.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Society for Professional Journalists (2014). *SPJ Code of Ethics.* Retrieved June 9, 2022. https://www.spj.org/ethicscode.asp

Trinidad, S. B., Fullerton, S. M., Ludman, E. J., Jarvik, G. P., Larson, E. B., & Burke, W. (2011). Research Practice and Participant Preferences: The Growing Gulf. *Science*, *331*(6015), 287–288. https://doi.org/10.1126/science.1199000

Vazire, S. (2017, April 12). Against Eminence. https://doi.org/10.31234/osf.io/djbcw

Washburn, A. N., Hanson, B. E., Motyl, M., Skitka, L. J., Yantis, C., Wong, K. M., … Carsel, T. S. (2018). Why Do Some Psychology Researchers Resist Adopting Proposed Reforms to Research Practices? A Description of Researchers' Rationales. *Advances in Methods and Practices in Psychological Science*, *1*(2), 166–173. https://doi.org/10.1177/2515245918757427