UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Conry-Murray, C., Mcconnon, A., & Bower, M. (2022). The Effect of Preregistration and P-Value Patterns on Trust in Psychology and Biology Research. *Collabra: Psychology*, *8*(1). https://doi.org/10.1525/collabra.36306

Methodology and Research Practice

The Effect of Preregistration and P-Value Patterns on Trust in Psychology and Biology Research

Clare Conry-Murray[1] [a], Ann Mcconnon[1], Morgan Bower[2]

[1] Psychology, Saint Joseph's University, Philadelphia, PA , US, [2] Biology, Saint Joseph's University, Philadelphia, PA , US

Keywords: practitioners, replication crisis, metascience

https://doi.org/10.1525/collabra.36306

Collabra: Psychology

Vol. 8, Issue 1, 2022

The replication crisis has shown that research in psychology and other fields including biology is not as robust as previously thought. In response, methods have been introduced to address the problem and increase reproducibility, including two methods that are the focus here: (1) preregistration of study hypotheses and methods, and (2) analysis of whether p-hacking may have occurred through patterns of p-values. Each is easy to find, even in short summaries of research, but do consumers of research recognize these indicators as evidence of trustworthiness? In the current study, we examined how professionals (n = 111), researchers (n = 74) and undergraduate students (n = 78) judged the trustworthiness of short descriptions of research in their field, which varied in terms of whether there was a reference to a preregistration or evidence of potential p-hacking. Overall, participants trusted studies less when they were not preregistered. Researchers and professionals, but not students were sensitive to evidence of p-hacking. We suggest that education about questionable research practices like p-hacking and hypothesizing after the results are known needs to be improved.

Across several fields, the replication crisis indicates that peer-reviewed research is not always robust enough to replicate (e.g., Baker, 2016; Open Science Collaboration, 2015; Simmons et al., 2011; Wiggins & Christopherson, 2019). Although the problem of unreliable research has been highlighted in psychology, it affects many fields, including biological and medical research (Carroll, 2017; Engber, 2016; Prinz et al., 2011). Several solutions have been proposed to make research more robust. One suggestion is preregistration, using a link to a time-stamped registration of methods and hypothesis. Another suggestion is the analysis of p-curves to determine whether p-values tend to fall just under .05, which can be an indication of inappropriate statistical methods (Rohrer, 2018). These practices provide easy-to-identify markers that could increase trust in research. The current study examines whether practitioners, researchers, and students trust research more when it has been preregistered and when it shows evidence of a series of p-values well under .05.

Research has shown that questionable techniques in data analysis and data collection are quite common (John et al., 2012). John et al. found that over a third of researchers surveyed indicated that they had engaged in questionable research practices that distort the meaning of a statistically significant result. Although the exact prevalence of questionable research practices is debated (Fiedler & Schwarz, 2016), most agree that questionable practices have resulted in too many failures to replicate studies. For example, the Open Science Collaboration (Open Science Collaboration, 2015) attempted to replicate 100 studies in psychology and found that though 97% of original studies had statistically significant results, only 36% of replications of those studies had statistically significant results.

The replication crisis is not limited to the field of psychology; other fields are also dealing with problematic research practices (e.g., in biology, Head et al., 2015). We focus on biology and psychology because both fields report concerns with questionable research practices and they encompass some of the most popular majors (*The NCES Fast Facts Tool Provides Quick Answers to Many Education Questions (National Center for Education Statistics)*, n.d.).

Common but questionable research practices include the following: only reporting results that fit researcher expectations, excluding data to make the results statistically significant, stopping data collection once results become statistically significant or adding covariates to push data into statistical significance. These are examples of "p-hacking," or manipulating the data until statistical significance is achieved. P-hacking increases the chances of false positives, from the standard alpha level of .05 to much higher levels (Chambers, 2017). When questionable practices like these are used to attain a p-value under .05, the rate of false

[a] **Corresponding author:**
cconrymu@sju.edu

positives can be very high. In fact, Simmons et al. (2011) found that almost any hypothesis could be found to be statistically significant through p-hacking. In particular, they showed that listening to the song "When I'm 64" by the Beatles, makes people younger, an impossible conclusion, but one that was statistically significant with the use of p-hacking. Since p-hacking is usually used to get p-values just under .05—the threshold for many publications, one signal of p-hacking is several p-values that are just under .05.

P-hacking may be difficult to avoid in all fields with hypothesis testing. This is because researchers across fields have many methodological and statistical decisions to make, and it can be difficult to make each decision without consideration of hypotheses and the data when it is in front of the researchers (Gelman & Loken, 2013), unless the plan is specified beforehand (Nosek et al., 2018), as in a preregistration of study methods.

A related questionable research practice is hypothesizing after the results are known (known as "HARKing," Murphy & Aguinis, 2019). In this case, researchers change their hypotheses after looking at the data. The problem with HARKing is that, depending on the number of variables, it is quite common for at least one effect to be statistically significant in every analysis, just by chance alone (Bishop, 2013). Hypothesizing after seeing the data can raise the likelihood of false positives to over 50% in some cases, much higher than the reported level of 5%. Preregistering hypotheses can help researchers avoid the temptation to HARK because hypotheses are planned before seeing the data.

Preregistration—specifying both methods and hypotheses before data analysis begins in a time-stamped and publicly available repository, makes it more difficult to choose statistical tests based on a desire to confirm a hypothesis. It also makes it more difficult to develop a "hypothesis" after seeing the data. In the current study, we examine whether readers of psychology and biology research view preregistration as a sign of trustworthy research. We examine this question in a situation where the content of the preregistration is not available, instead there is only the indication that the study has been preregistered.

A second solution to the problem of p-hacking proposes to focus on determining whether questionable research practices have occurred by examining the evidence *after* the data has been collected and reported. P-curve analysis (Simonsohn et al., 2014) examines whether p-values are suspiciously close to .05, the typical alpha used in many fields, including psychology and biology. A series of p-values that are all very close to .05 is more likely to be p-hacked than very low p-values because researchers who p-hack are only motivated to get the p-value under .05, and once that level is achieved, they may stop p-hacking. True effects are unlikely to generate p-values that are consistently close to .05 over a series of many tests (Rohrer, 2018; Simonsohn et al., 2014). Instead, p-values that are not p-hacked typically have a much wider range. However, unlike preregistration, p-curves require some knowledge of statistics, and therefore many readers may not notice when p-values are clustered near .05, and if they do notice, it is an open question whether readers of research judge that p-value patterns provide evidence of trustworthiness. In the current study, we examine whether readers of research are sensitive to patterns of p-values that may indicate the possibility of p-hacking because all values are close to .05.

Both preregistration and p-curve analysis have been shown to be helpful under some circumstances. Clinical trials have been preregistered for many years (*History, Policies, and Laws—ClinicalTrials.Gov*, n.d.), but preregistration in psychology is more recent, although it is becoming more common (Lindsay & Nosek, 2018). Preregistering methodological plans, including a stopping rule, exclusion criteria etc., can help prevent p-hacking, and preregistering hypotheses can help prevent HARKing because hypotheses are specified and documented before data collection. Therefore, preregistration can help researchers avoid both p-hacking and HARKing, if it is used correctly (Rubin, 2017).

However, both preregistration and p-curve analysis are imperfect signs of research quality. Preregistration is not required for publication in APA journals (American Psychological Association, 2021). Even when it is done, many reviewers do not examine the preregistration to ensure that it has been followed. (See Registered Reports for a more robust form of preregistration, where peer reviewers evaluate proposed methods and accept or reject the paper before data is collected, Chambers, 2019). Additionally, it is not a perfect process (Claesen et al., 2019); a preregistration can be completed, fraudulently, after data has been analyzed. In addition, exploratory research that does not have planned tests and is not preregistered can also very valuable for theory building (Devezer et al., 2020). Furthermore, although preregistration enhances the credibility of research findings (Nosek et al., 2019), the lack of a preregistration does not mean a researcher used questionable research practices.

Patterns of p-values are also an imperfect measure of potential p-hacking since they only indicate the *possibility* of p-hacking. Researchers may be more aware of the limitations of these indicators than professionals or students, but they may also understand its potential value more. Thus, it is important to explore whether readers with different levels of expertise judge preregistration and patterns in p-values differently.

In the current study, we examine how preregistration and patterns in p-values affect trust in research across three groups: undergraduate students, researchers, and professionals. Current students may be more aware of the issues related to questionable research practices than in the past because discussion of these practices has increased in recent years (Motyl et al., 2017). There has been an increase in teaching about the issues and tools to address the issues in both biology (Toelch & Ostwald, 2018) and psychology (Chopik et al., 2018; Sarafoglou et al., 2020). These classes address issues such as preregistration, power analysis, and reporting effect sizes (Sarafoglou et al., 2020), which may help students to be more sensitive to evidence of questionable research practices. In addition, students have initiated clubs such as ReproducibiliTea, and some are pushing faculty for more open and transparent practices (Orben, 2019). Therefore, students may have a variety of different responses to evidence of questionable research practices. Still, most undergraduate students do not conduct original research (Krishna & Peter, 2018) and most undergraduates are unlikely to be at the forefront of open science, and therefore they may not be as knowledgeable as active researchers. In fact, research has shown that experience with preregistration and expertise in statistics are associated with more accurate judgments of the replicability of research in the RepliCATS project (Wintle et al., 2021).

Practitioners rely on research but may not be active researchers themselves, and therefore they may be less aware of recent practices to increase robust research results. However, even active researchers within clinical psychology have been slow to engage in open science practices (Tackett et al., 2019). In any case, it is important for practitioners to be able to find articles and assess their trustworthiness, in order for a practitioner to properly diagnose and treat patients. However, this can be a challenge for busy practitioners. Research shows that for a practitioner to stay on top of new research, they need to spend approximately 300 hours a month reading and studying new journals (Alper et al., 2004). This amount of time can be cut down if the practitioner uses signs of questionable research practices.

The Current Study

In the current study, we examine how undergraduates, researchers and practitioners in biology and psychology judge indicators of actions taken to prevent questionable research practices and signals of robust effects. Thus, we examine judgments of the trustworthiness of research abstracts varying two key factors: 1) whether the study reports being preregistered, and 2) whether the findings in a series of p-values are all well under .05. Previous research on whether preregistration increases trust has found ambiguous results (Field et al., 2020). We know of no research that has examined perception of a series of p-values.

In a factorial design, participants were asked to judge short descriptions of research in their field, which varied whether there was reference to a preregistration or evidence of p-hacking. We targeted biology and psychology undergraduate majors, researchers in those fields who spend significant time doing research, and applied professionals working in the fields of biology and psychology.

We expected to find a main effect for preregistration because it can be understood as a marker of a higher quality study that is easy to find. However, we also expected that researchers, more than other groups, would indicate that preregistered studies were trustworthy. Finally, we also expected expertise to affect whether participants were sensitive to the negative implications of p-values that cluster just below .05. We expected that researchers would view studies with more variability in significant p-values as more trustworthy compared to participants with less research expertise (students and professionals).

In exploratory analyses, we also examined whether one field is more or less sensitive to p-hacking and whether our other variables interact with the field, but we have no specific hypothesis about the effect of field of study. We also explored whether evidence of p-hacking or preregistration has a bigger impact on trust in a study.

The preregistration of the current study is available here: https://osf.io/7grkq?view_only=None.

Method

Participants

Participants were recruited through social media, word-of-mouth, and Prolific. On Prolific, participants were paid $2.00 to complete the survey, which averaged 7 minutes. On Prolific, we were able to add filters for biology- and psychology-related fields, such as medicine and nursing in biology, and counseling and social work in psychology. Our initial sample size was 318 participants; however, we excluded participants who were not in the fields of psychology or biology or related fields (n= 30 people), or who skipped the question about primary field (n = 6). We also excluded participants who completed the survey in under 2.5 minutes (n = 11 participants excluded). Finally, a small number of participants who noted that they had never studied or worked in biology or psychology-related fields or who reported data that seemed very unlikely to be true (e.g., a 19-year-old with 8 years of research experience, n = 8) were excluded. Although we did not preregister these exclusions, we found them to be necessary to ensure the data was valid. The final dataset included 263 participants.

We divided participants into three groups depending on their educational level and their experience with research. The three levels included undergraduate students who were currently students in college; researchers, who were defined as having completed an undergraduate degree and who reported spending at least 20% of their work time (i.e., one work-day) on research each week; and professionals, who were defined as having completed an undergraduate degree and who reported spending less than 20% of their time on research. These groups differ from the preregistration in that Master's level students were included with researchers but we preregistered that they would be included with students. We made the decision to include Master's level students with researchers to make the expertise groups sizes more equal. The analysis with master's students in the "student" group is available in the supplementary materials.[1]

We did not collect data on the current professions of all participants because we added that question after data collection had commenced. Examples of biology-related professions included nurses, doctors, and athletic trainers. Psychology-related professions included education, recruitment, and mental health. Some "professionals" were graduate students, but they were in programs that did not require research. Table 1 shows the distribution of participants in these categories by field and gender.

The exclusions and group criteria include some unanticipated criteria that was added after preregistration to make

[1] The majority of effects were the same when Master's students were included as "students" rather than "researchers." However, there was one difference. We found that distinction between abstracts with evidence of p-hacking or not were also made by students when Master's students were included with undergraduates, *t*(794.4)= -3.73, *p* < .001, *d* = .26. More details are in the supplementary material.

Table 1. Participant Characteristics

	N	Gender: Women	Field: Biology	Age (M)	Percent Time on research (M)	Years of work experience (M)	Years of research experience (M)	Typicality of studies (M)
Undergraduates	78	61.5%	60.3%	22.5	5.80%	1.46	.08	4.99
Researchers	74	61.2%	46.3%	28.6	36.3%	4.00	4.53	5.08
Professionals	111	74.8%	54.0%	30.7	2.49%	5.53	2.36	4.38
Total	263	68.6%	54.0%	27.7	13.0%	3.90	2.52	4.77

Note. Typicality of studies was coded from 0, very unusual, I don't usually see studies like this to 10, very typical of my field.

Table 2. Number of Participants in Each Group

	Students	Researchers	Professionals	Total
Psychology	47	35	51	133
Biology	31	39	60	130
Total	78	74	111	263

groups approximately equal or to address unusual data. The criteria implemented after the preregistration were completed before data analysis.

Power analysis

We completed an a priori power analyses for each hypothesis using WebPower, reported in our preregistration on OSF, available here: https://osf.io/7grkq?view_only=None. In each, we used an alpha of .05, with an effect size f = .25 and a power of .90. We chose to use a small-medium effect size because it was the smallest effect size of interest, given practical constraints on the amount of data we could collect. For the main effect of preregistration on judgments of abstracts, we needed 170 participants. For the expertise group x preregistration and expertise group x p-hacking evidence in abstracts, we needed 206 participants. We were able to exceed these thresholds. Table 2 shows the N per cell.

Procedure

After consenting to participation in accordance with Saint Joseph's University's IRB (approval #1611771), participants reported on their demographics. Next, participants were directed, based on their choice of major or field (psychology or biology) to read and rate how much they trusted a series of eight short abstracts in their field, on a scale from 0 = not at all to 10 = very much. In both psychology and biology abstracts, there were two trials. One series of abstracts described an experimental manipulation addressing depression. In biology it was a drug, and in psychology it was different types of therapy. In each case, the treatment was identified by various acronyms. The second trial included abstracts describing an association between sleep deprivation and motor activity being moderated by either a physiological trait in mice for biology, or a personality trait in people for psychology, again described using an acronym. An example of a Biology abstract is below:

> This study tested whether the effect of sleep deprivation on motor activity in mice is smaller for mice who have high levels of TSDM. In three studies (n = 77, 78 and 70) with different measures of motor activity, CD-1 mice with higher TSDM levels showed higher activity levels after of sleep deprivation than those with lower levels at statistically significant levels, at p = .048, p = .045 and p = .046, respectively. Preregistration is available at osf.io/jfiu8nm

An example of a Psychology abstract is below:

> This study tested whether the effect of sleep deprivation on motor activity is smaller in participants who scored high on the TSDM personality scale. In three studies (n = 77, 78 and 70) with different measures of motor activity, participants with higher TSDM scores showed higher activity levels after of sleep deprivation than those with lower scores at statistically significant levels, at p =.048, p = .045 and p = .046, respectively. Preregistration is available at osf.io/jkiu8nm

Each abstract described three studies with different p-values. The p-values were written in groups of three with values as such: .04x, .04x and .04x for p-values clustered near .05 or .04x, .01x and .00x for p-values with more variability. The value "x" ranged from 0-9 and was alternated for each abstract. Each abstract described a total sample size of 280-290, and each study within the abstract was described as having between 70-80 participants. The drug names used were all created from a fake drug generator: https://perchance.org/fake-drug-name. All "preregistered" studies included a fabricated link to websites, and participants were told the link was real, but they did not need to follow it. Anyone who did try to follow a link got an error message. At the end of the survey, participants were asked whether the abstracts were typical for studies within their field, and they saw them as somewhat typical (see Table 1). Data, code and study materials are available here: https://osf.io/wdptv/?view_only=d4415cacb48345c8b668f99aeacbf6a0.

Results

We preregistered an analytic plan that includes both repeated measures ANOVA and linear-mixed effects models. However, because model comparisons indicated the need for random effects, we report the mixed-effects analysis here.[2]

We used the *lme4* package (Bates et al., 2015) in R (R Core Team, 2021) to fit a linear mixed-effects model with judgments of trustworthiness as the response term. A χ^2 test indicated that the mixed-effects model with random effects included for trial and subject was an improvement on the intercept-only model, $\chi^2(2) = 797.4$, $p < .001$. We then tested a model that included preregistration, preregistration x expertise group and p-hacking x expertise group as fixed effects, with participant identity and trial as random intercepts to account for by-subject and by-trial variability. This model showed a marginal R^2 of .03, and a conditional R^2 of .52. A significant main effect for preregistration, $F(1, 1798.69) = 34.98$, $p < .001$, $d = .17$, indicated that preregistered abstracts were judged to be more trustworthy ($M = 6.43$, $SD = 2.22$) than abstracts without preregistration ($M = 6.06$, $SD = 2.19$).

There was also an interaction between expertise group x p-hacked, $F(5, 570.31) = 18.67$, $p < .001$. Bonferroni-corrected planned contrasts showed that the researchers, $t(578.45)= -5.31$, $p < .001$, $d = .44$ and the professionals, $t(866.9)= -2.80$, $p = .002$, $d = .19$ both distinguished between abstracts with evidence of p-hacking or not, but students did not distinguish the two, given reduced alpha level with the Bonferroni correction, $t(609.62)= -2.03$, $p = .021$, $d = .16$. The expected interaction between preregistration and the expertise group was not significant. See Table 3 for means and SDs.

To explore the field of study effects, we also examined a model that included preregistration, p-hacking, and field (biology or psychology), as well as the following interactions: field x preregistration and field ax p-hacking as fixed effects. Participant identity and trial were random intercepts. In addition to the hypothesized effect of preregistration discussed above, there was an unexpected main effect for p-hacked abstracts, $F(1, 1801.37) = 79.58$, $p < .001$, $d = .26$, such that abstracts with evidence of p-hacking were judged to be less trustworthy ($M = 5.96$, $SD = 2.18$) than abstracts without evidence of p-hacking ($M = 6.53$, $SD = 2.21$).

There was also a main effect for field, $F(1, 257.19) = 7.98$, $p = .005$, $d = .26$, showing that people in biology rated the biology abstracts as more trustworthy (M = 6.50, $SD = 2.08$) than people in psychology judged the psychology abstracts (M = 5.93, $SD = 2.33$). Finally, there were also interactions between field x p-hacked, $F(1, 1801.38) = 14.65$, $p < .001$, and field x preregistration, $F(1, 1800.70) = 6.38$, $p = .011$. These indicated that participants in the field of psychology judged preregistered abstracts as more trustworthy than those lacking preregistration, but in biology this distinction was not made. Both fields distinguished differences in patterns of p-values, but psychology made a larger distinction. See Table 4 for means by field.

Discussion

The most important finding in this study is that across different educational levels and levels of research experience, people trust studies that have been preregistered slightly more than ones that do not have preregistered findings. On the other hand, studies that show statistical indicators of possible p-hacking are less likely to be detected for those with less training or experience.

We expected and found a main effect for preregistration, perhaps because preregistration was easy to notice in our abstracts. Since participants saw both preregistered and not preregistered abstracts, they were likely aware that this factor was being tested. Preregistration has the goal of decreasing p-hacking (Gonzales & Cunningham, 2015). In contrast to Field et al. (2020), whose results were ambiguous about whether preregistration affects researchers' trust in published results, our findings show that participants recognized preregistration as a measure of quality and they trusted preregistered research more; however, our findings showed only a small effect. The difference between our finding and previous research by Field et al. may be because Field et al. only examined judgements of trustworthiness from researchers, whereas we also examined judgments from students and professionals, who may assume that preregistration is rigorous. It is also possible that some people did not know what preregistration was, but they may have rated preregistered studies as more trustworthy because those abstracts were assumed to have included additional information in the link. Researchers may recognize that preregistration can be used inappropriately, and reviewers rarely check the preregistration to ensure that it was fol-

[2] We present the repeated measure ANOVA results here. We conducted a 2 preregistration status (preregistered or not) x 2 p-hacking status (p-hacking evidence in p-values or not) x 3 expertise type (undergraduate, professional or researcher) repeated measures ANOVA with preregistration and p-hacking evidence as repeated measures. We expected and found a main effect for preregistration, $F(1, 268) = 8.07$, $p = .005$, $\eta_p^2 = .03$, which indicated that preregistered studies were rated as more trustworthy, $M = 5.49$, $SE = .13$, than studies that were not preregistered, $M = 5.32$, $SE = .13$.

We also expected an expertise x preregistration interaction such that those with more research experience (i.e. researchers and to a lesser degree students) would view preregistered studies as more trustworthy, but this effect was not found, $F(2, 268) = 1.38$, $p = .253$, $\eta_p^2 = .01$. See Table 3 for means.

Finally, we expected an expertise x evidence of p-hacking interaction such that researchers would view studies with consistently lower p-values as more trustworthy than the other groups. This effect was found, $F(2, 268) = 3.19$, $p = .043$, $\eta_p^2 = .02$, and it indicated that only researchers rated studies with evidence of p-hacking as significantly less trustworthy, ($p = .006$), with Tukey correction for multiple comparisons. See Table 3 for means.

Table 3. Mean (and *SDs*) judgments of trustworthiness by condition and expertise level

	Preregistration		Evidence of p-hacking	
	Yes	No	Yes	No
Undergraduates	6.47 (2.02)	6.16 (1.95)	6.15_a (1.88)	6.48_a (2.09)
Researchers	6.50 (2.37)	6.05 (2.36)	5.76_a (2.39)	6.78_b (2.26)
Professionals	6.36 (2.26)	5.99 (2.23)	5.96* (2.22)	6.39* (2.26)
Total	6.43_a (2.22)	6.06_b (2.19)	5.96* (2.18)	6.52* (2.21)

Note. Means are based on judgments of trustworthiness on a scale from 0 = not at all to 10 = very much. Mean comparisons were tested within conditions (Preregistration or Evidence of p-hacking conditions) for each level of expertise. Means with different subscripts in each row within each condition indicate that means differ at $p < .006$, the Bonferroni corrected significance level, as predicted. Means with asterisks also differed at $p < .006$ in the exploratory analysis and no asterisks indicates exploratory differences were not found.

Table 4. Exploratory Findings by Field of Study

	Preregistration			Evidence of p-hacking			Total
	Yes	No	Cohen's d	Yes	No	Cohen's d	
Psychology	6.22* (2.35)	5.66* (2.28)	.24	5.51* (2.26)	6.36* (2.31)	.37	5.94 (2.33)
Biology	6.62 (2.10)	6.39 (2.06)	.11	6.34* (2.02)	6.67* (2.12)	.16	6.50 2.08)

Note. Means are based on judgments of trustworthiness on a scale from 0 = not at all to 10 = very much. * indicates means are different within the conditions (Pre registration or the Evidence of p-hacking) at $p < .013$, the Bonferroni corrected level, within each field.

lowed. Other research has found that the majority of preregistrations included deviations that were not disclosed (Claesen et al., 2019). In addition, not all preregistrations are made public (Field et al., 2020), making it impossible to check whether there were deviations. The fact that the expertise x preregistration hypothesized interaction was not found could be the result of different groups making different assumptions about what the preregistration meant, but unfortunately, we cannot be sure of this because we did not ask for reasons for judgments of trustworthiness. Still, given that there is some effect of increased trust in preregistered studies across our groups, even if small, it is important to ensure that preregistrations truly reflect a more trustworthy study.

We expected and found an interaction between expertise and evidence of p-hacking, such that researchers and professionals were sensitive to p-hacking but students were not. Researchers may be especially aware of issues of p-hacking because they are likely to use p-values in their own work. Indeed, the effect size for this difference was medium for researchers. Professionals also distinguished between abstracts with evidence of p-hacking or not, though the effect for this difference size was small for professionals. Professionals who have been out of school for a long time are consumers of research, and they may need to use research findings in their day-to-day work life (*How Does Research Impact Your Everyday Life?*, 2016).

However, the researchers and professionals who completed our survey rated p-hacked studies as only slightly less trustworthy despite that there is evidence that suggests it is very rare to find three p-values just under .05 (Rohrer, 2018; Simonsohn et al., 2014). It may be that they judged lower p-values as an indication of a stronger effect, and not as p-hacking, which is a more serious violation of research ethics that can lead to many false positives. In addition, there has been criticism of the use of p-curve to detect p-hacking (Erdfelder & Heck, 2019), which argues that researchers may choose not to publish even statistically significant results when results are unfavorable to a researcher's expectation, affecting the distribution of p-values. This criticism rightly suggests caution. However, journal-level publication bias tends to skew towards positive and significant results (Hopewell et al., 2009), which supports the use of p-curves. Thus, we believe that it is appropriate that researchers judge that a series of p-values just under .05 may be less trustworthy.

Undergraduate students rated studies similarly whether there was evidence of p-hacking or not. Undergraduate students may still be learning about good research practices and statistics, and thus, improved teaching related to research practices is likely to be beneficial. Research shows that the attitudes of mentors and professors in psychology about questionable research practices are often reflected in students' attitudes (Krishna & Peter, 2018). Though students who learn about open science and the reproducibility crisis do tend to trust psychology research less, they also see psychology as more scientific, similar to natural science fields (Chopik et al., 2018).

We had two exploratory findings of note. First, we found a main effect for the p-hacking condition. However, given that the main effect was not hypothesized and it was also qualified by a hypothesized interaction with an expertise group, it may be that the researchers and the professionals are driving the main effect for p-hacking. Overall, it appears that preregistration is a signal of research quality across groups, but p-value patterns were considered primarily by researchers and professionals.

We also explored whether one field is more or less sensitive to p-hacking. We found that people in the field of biology generally trusted biology abstracts more than people in the field of psychology. It may be that people in psychology are more aware of the limitations of research.

There were several limitations to our study. We did not include a measure of what participants already knew about preregistration or p-hacking. Future research should also ask participants to justify their judgments in order to examine what aspects of preregistration and p-values affected their judgments.

Future research should also examine whether specific educational interventions can aid the detection of robust research. In fact, there is a DARPA initiative to build artificial intelligence to systematically examine the literature to evaluate the credibility of research (Alipourfard et al., 2021). We advocate for more education and resources in this area, but our results, on their own, do not indicate whether these practices are good or bad in terms of validity or reproducibility of the research. Instead, our findings indicate whether researchers, practitioners, and students believe these factors indicate quality. Future research should examine how researchers and others who use research reason about open science practices and statistical knowledge that can combat p-hacking. Awareness and beliefs in these areas should be more explicitly measured to determine how these beliefs affect judgments of research results.

Overall, we found that undergraduate students, professionals, and researchers trust studies less when they were not preregistered. In addition, researchers and professionals were sensitive to evidence of p-hacking. We suggest that education about research practices like p-hacking, HARKing, and preregistration needs to be improved in order for both consumers and producers of research to be able to better recognize when findings can be expected to be robust or not.

Author Contributions

Contributed to conception and design: CCM, AM, MB
Contributed to acquisition of data: CCM, AM, MB
Contributed to analysis and interpretation of data: CCM
Drafted and/or revised the article: CCM, AM, MB
Approved the submitted version for publication: CCM, AM, MB

Competing Interests

No competing interests exist.

CCM is an action editor at Collabra: Psychology. She was not involved in the review process of this article.

Data Accessibility Statement

Data and the survey are available here: [OSF, https://osf.io/wdptv/?view_only=d4415cacb48345c8b668f99aeacbf6a0]

Submitted: March 02, 2021 PDT, Accepted: April 01, 2022 PDT

References

Alipourfard, N., Arendt, B., Benjamin, D. M., Benkler, N., Bishop, M. M., Burstein, M., Bush, M., Caverlee, J., Chen, Y., Clark, C., Dreber, A., Errington, T. M., Fidler, F., Fox, N. W., Frank, A., Fraser, H., Friedman, S., Gelman, B., Gentile, J., … Wu, J. (2021). *Systematizing Confidence in Open Research and Evidence (SCORE)*. https://doi.org/10.31235/osf.io/46mnb

Alper, B. S., Hand, J. A., Elliott, S. G., Kinkade, S., Hauan, M. J., Onion, D. K., & Sklar, B. M. (2004). How much effort is needed to keep up with the literature relevant for primary care? *Journal of the Medical Library Association: JMLA*, *92*(4), 429–437.

American Psychological Association. (2021, January). *Preregistration*. American Psychological Association. https://www.apa.org/pubs/journals/resources/preregistration

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bishop, D. (2013, June 7). BishopBlog: Interpreting unexpected significant results. *BishopBlog*. http://deevybee.blogspot.com/2013/06/interpreting-unexpected-significant.html

Carroll, A. E. (2017). Science needs a solution for the temptation of positive results. *New York Times*. https://www.nytimes.com/2017/05/29/upshot/science-needs-a-solution-for-the-temptation-of-positive-results.html

Chambers, C. (2017). *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice* (1st Edition). Princeton University Press.

Chambers, C. (2019). What's next for Registered Reports? *Nature*, *573*(7773), 187–189. https://doi.org/10.1038/d41586-019-02674-6

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology*, *45*(2), 158–163. https://doi.org/10.1177/0098628318762900

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F., & Vanpaemel, W. (2019). *Preregistration: Comparing Dream to Reality* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/d8wex

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2020). The case for formal methodology in scientific reform. *Royal Society Open Science*, *8*(3), 200805. https://doi.org/10.1098/rsos.200805

Engber, D. (2016, April 19). Think psychology's replication crisis is bad? Welcome to the one in medicine. *Slate Magazine*. https://slate.com/technology/2016/04/biomedicine-facing-a-worse-replication-crisis-than-the-one-plaguing-psychology.html

Erdfelder, E., & Heck, D. W. (2019). Detecting evidential value and p-hacking with the p-curve tool. *Zeitschrift Für Psychologie*, *227*(4), 249–260. https://doi.org/10.1027/2151-2604/a000383

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Field, S. M., Wagenmakers, E.-J., Kiers, H. A. L., Hoekstra, R., Ernst, A. F., & van Ravenzwaaij, D. (2020). The effect of preregistration on trust in empirical research findings: Results of a registered report. *Royal Society Open Science*, *7*(4), 181351. https://doi.org/10.1098/rsos.181351

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Department of Statistics, Columbia University.

Gonzales, J., & Cunningham, C. (2015). The promise of pre-registration in psychological research. *Psychological Science Agenda*. https://www.apa.org/science/about/psa/2015/08/pre-registration

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

History, Policies, and Laws—ClinicalTrials.gov. (n.d.). Retrieved August 24, 2020, from https://clinicaltrials.gov/ct2/about-site/history

Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*, *1*. https://doi.org/10.1002/14651858.mr000006.pub3

How does research impact your everyday life? (2016). Study International. https://www.studyinternational.com/news/how-does-research-impact-your-everyday-life/

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

Krishna, A., & Peter, S. M. (2018). Questionable research practices in student final theses – Prevalence, attitudes, and the role of the supervisor's perceived attitudes. *PLoS ONE*, *13*(8). https://doi.org/10.1371/journal.pone.0203470

Lindsay, S., & Nosek, B. A. (2018). Preregistration Becoming the Norm in Psychological Science. *APS Observer*, *31*(3). https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & 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

Murphy, K. R., & Aguinis, H. (2019). HARKing: How Badly Can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, *34*(1), 1–17. https://doi.org/10.1007/s10869-017-9524-7

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration Is Hard, And Worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orben, A. (2019). A journal club to fix science. *Nature*, *573*(7775), 465–466. https://www.nature.com/articles/d41586-019-02842-8

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rohrer, J. (2018, August 15). *The uncanny mountain: P-values between .01 and .10 are still a problem*. The 100% CI. http://www.the100.ci/2018/02/15/the-uncanny-mountain-p-values-between-01-and-10-are-still-a-problem/

Rubin, M. (2017). An Evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology*, *21*(4), 321–329. https://doi.org/10.1037/gpr0000135

Sarafoglou, A., Hoogeveen, S., Matzke, D., & Wagenmakers, E.-J. (2020). Teaching good research practices: Protocol of a research master course. *Psychology Learning & Teaching*, *19*(1), 46–59. https://doi.org/10.1177/1475725719858807

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*, *15*(1), 579–604. https://doi.org/10.1146/annurev-clinpsy-050718-095710

The NCES Fast Facts Tool provides quick answers to many education questions (National Center for Education Statistics). (n.d.). National Center for Education Statistics. Retrieved January 5, 2021, from https://nces.ed.gov/fastfacts/display.asp?id=37

Toelch, U., & Ostwald, D. (2018). Digital open science—Teaching digital tools for reproducible and transparent research. *PLOS Biology*, *16*(7), e2006022. https://doi.org/10.1371/journal.pbio.2006022

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Wintle, B., Mody, F., Smith, E. T., Hanea, A., Wilkinson, D. P., Hemming, V., Bush, M., Fraser, H., Singleton Thorn, F., McBride, M., Gould, E., Head, A., Hamilton, D. G., Rumpff, L., Hoekstra, R., & Fidler, F. (2021). Predicting and reasoning about replicability using structured groups. *MetaArXiv*. https://doi.org/10.31222/osf.io/vtpmb

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/36306-the-effect-of-preregistration-and-p-value-patterns-on-trust-in-psychology-and-biology-research/attachment/91807.docx?auth_token=YlbIDwvhWCBboxnQFSZj