One Decade Into the Replication Crisis,
How Have Psychological Results Changed?

Paul C. Bogdan[1]

[1] Psychology & Neuroscience, Duke University, Durham, 27708, NC, USA.

Corresponding Author:

Paul Bogdan
Levine Science Research Center
Duke University
308 Research Dr.
Durham, NC, 27708. USA
Phone: 630-935-5727
Email: paul.bogdan@duke.edu

Author note

Ethical approval: The present work did not require review by the local Institutional Review Board because it only uses publicly available data.

Acknowledgments: The author thanks Daniel Simons, Gregory Francis, Simon W. Davis, and Ulrich Schimmack for their feedback on an earlier version of this manuscript.

Funding: During data collection, P.C.B. was supported by a Dissertation Completion Fellowship provided by the University of Illinois at Urbana-Champaign.

Conflicts of Interest: The author declares that there were no conflicts of interest with respect to the authorship or the publication of this article.

Data availability: The data needed to reproduce the results has been deposited in a public repository (https://osf.io/mxs47/). The dataset contains every variable calculated for each paper along with article metadata (e.g., DOIs).

Code availability: The code used to collect the data and to produce the results has been uploaded to a public repository (https://github.com/paulcbogdan/PsychChange). Notably, this code can be readily adapted for fields beyond psychology.

Abstract

A psychology paper's p-values say a lot about how its studies were conducted and whether its results are likely to replicate. Examining p-values across the entire literature can, in turn, shed light on the state of psychology overall and how it has changed since the start of the replication crisis. The present research investigates strong ($p < .01$) and weak ($.01 \leq p < .05$) p-values reported across 240,398 empirical psychology articles from 2004-2024. Over this period and across every subdiscipline, the typical study has begun reporting markedly stronger p-values. Nowadays, papers reporting strong p-values are also more often published in top journals and receive more citations. Yet, it also appears that robust research is still not correspondingly linked to career success, as researchers at the highest-ranked universities tend to publish papers with the weakest p-values. Investigating language usage suggests that two-thirds of this association can be explained by highly ranked universities preferring laborious, expensive, and subtle research topics, even though these generally produce weaker results. Altogether, these findings point to the strength of most contemporary psychological research and suggest academic incentives have begun to promote such research. However, there remain key questions about the extent to which robustness is truly valued compared to other research aspects.

Keywords: Meta-science, replicability, text mining, open science

1. One Decade Into the Replication Crisis, How Have Psychological Results Changed?

The replication crisis revealed that many of psychology's seminal studies do not replicate (Doyen et al., 2012; Open Science Collaboration, 2015). Alongside these replication failures, countless methodological papers discussed the prevalence of questionable research practices and how such practices led to non-replicable findings (Francis, 2012; John et al., 2012; Simmons et al., 2011; Wagenmakers et al., 2011). In turn, new scientific standards were proposed (Hales et al., 2019; Nosek & Lakens, 2014; Schimmack, 2012; Simmons et al., 2012; Van't Veer & Giner-Sorolla, 2016). It has been a decade since the replication crisis reached widespread awareness. Has the subsequent push for replicability produced meaningful changes? The present research investigates this question and current state of psychological science from different angles.

Some aspects of replicability can be studied by examining the strength of a paper's p-values (Krawczyk, 2015; Lakens, 2015; Van Assen et al., 2015). For instance, it is problematic if a paper frequently reports p-values that barely fall under significance thresholds ($.01 \leq p \leq .05$). Even in a study with merely 50% power, most p-values should fall under .01, and in studies with 80% power, just 26% of significant results should land in this $.01 \leq p \leq .05$ interval (per simulations). If such p-values are instead commonplace, this points to questionable research practices (Simonsohn et al., 2014). One prior study of 103 replication attempts indeed found a 74% replication rate for findings reported at $p \leq .005$ and a 28% replication rate for findings at $.005 < p < .05$ (Gordon et al., 2021). P-values can be extremely informative.

The present research tracks p-values across the whole of psychology and how reported p-values may have shifted since the replication crisis began to percolate. Prior meta-analyses have operated on smaller scales, focusing on just narrower topics, restricted pools of journals, or limited time ranges (Boggero et al., 2017; Stuart et al., 2019; Vadillo et al., 2016; Olsson-Collentine et al., 2019; Pritschet et al., 2016; Schimmack, 2020; Youyou et al., 2023). The present research, instead, uses an original dataset that extends from 2004 to 2024 and is the largest of its kind, covering most of the non-predatory psychology literature. This expansive size opens the door to new and more comprehensive inquiries.

Along with investigating how p-values have changed over time, the present study leverages p-values to inspect the value structure of academic psychology. Discussions of incentives are part and parcel of replication crisis commentary (Asendorpf et al., 2013; Nosek et al., 2012, 2022). At the most basic level, loosening standards for replicability may increase the quantity of research output and its perceived innovativeness. In turn, looser standards may produce papers that are published in higher-ranked journals and accumulate more citations, which may ultimately allow authors to achieve more prestigious university positions. Alternatively, if academic psychology emphasizes robust results, the opposite may be the case, where even a researcher acting purely in self-interest would benefit from practicing replicable science. Creating this type of incentive structure has been described as an important end goal of the replicability movement (Nosek et al., 2022).

To better understand these issues, the present research probes p-values. Analyses focused on the percentage of papers' significant p-values (*ps* < .05) that that are "fragile" (.01 ≤ *p* < .05), narrowly crossing the typical threshold for significance. After validating that this percentage predicts replicability, the measure was used to investigate three sets of questions: First, since the replication crisis began, has psychology begun to publish statistically stronger results? Second, does contemporary psychology incentive strong results – i.e., do papers reporting strong p-values find publication in higher-impact journals, accrue more citations, and are its authors affiliated with top-ranked universities? Third, to contextualize findings on the first two questions, how may p-value strengths and these incentives relate to the research topics, hypotheses, and methodology used by different papers?

2. Materials and Methods

2.1. Disclosures

The present research was not preregistered. The number of papers acquired was designed to be as large as possible while respecting publishers' terms and conditions. Before collecting the full dataset, some preliminary analyses were done using a subset of the data, which are not reported – e.g., attempting to study researchers moving between universities or investigating national gross domestic product as a predictor of p-values. Given the exploratory nature of the present research, a threshold of $\alpha = .001$ was employed for analyses by default; this did not apply to tests that formally implemented multiple comparison correction, which still used $p_{corrected} < .05$. This research was exempt from approval by the local institutional review board.

2.2. Data collection

The original dataset generation procedure is shown in Figure 1 along with descriptive stats of the final dataset. To generate an initial list of possibly usable articles, metadata were downloaded from Lens.org, which is a free online platform that attempts to compile information on all scholarly records. The database contained 643,571 records from psychology journals between 2004 and 2024 among publishers amenable to downloading full-text articles (Elsevier, SAGE Publications, Springer-Nature, Wiley, and Frontiers). The lower bound was 2004 because before then, few journals published web versions of full-text articles (PDFs were not included in this dataset). This pool was pruned to 372,633 empirical papers by searching only for records containing a Results section published in journals that regularly publish empirical papers. Of articles in this pool, at least one p-value was extracted for 269,018 papers (72.2%). For the analysis, this pool was further pruned to 240,355 articles containing at least two significant p-values, as the focus was on significant results and requiring two p-values avoided papers simply mentioning a threshold off-hand (e.g., "$p < .05$"). Among these papers, journal scores (see below) were not available for 16,796 papers, and no university affiliation was found for 79,298 papers. For analyses engaging with all these variables simultaneously, the fully usable dataset consisted of 150,344 papers, spanning 384 journals and containing 16.8 p-values on average (SD = 14.8, median = 13 p-values). Supplemental Materials 1 provides additional details on dataset collection and organization.

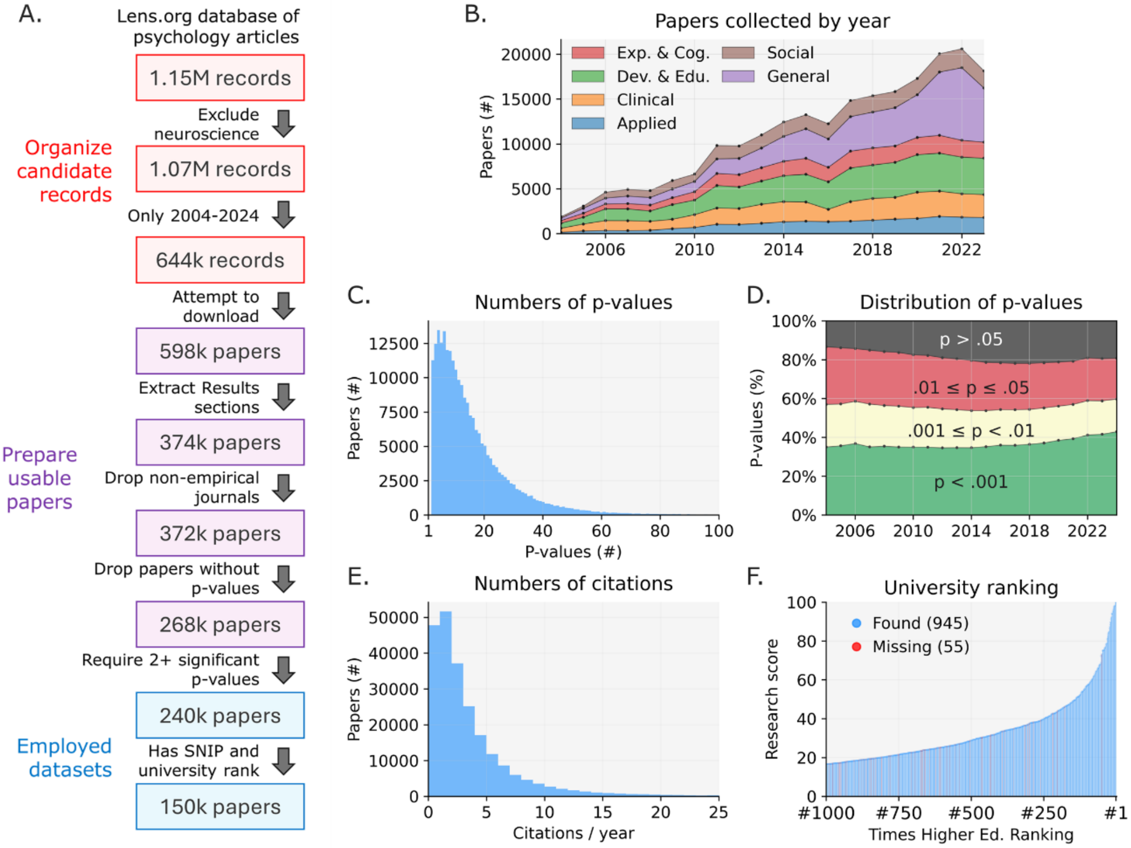


Figure 1. Summary of the organized dataset. A. Flow diagram showing how the final dataset was built over several stages. **B.** A stacked plot showing the number of articles collected each year from each psychology subject area; papers from journals with two or three areas contributed to both counts but were weighted 0.5 or 0.33, as primary aim is to report the overall number of papers; papers from 2024 were excluded because this year was only collected up to August; **C.** a histogram of how many p-values each paper contained; **D.** a breakdown of the percentage of p-values each year falling in different significance ranges; **E.** a histogram of how many citations each paper received per year since publication; **F.** a plot of each university's Times Higher Education ranking and its associated research score; universities were colored to denote ones that were found in the dataset. Dev. & Edu., Developmental and educational psychology; Exp. & Cog., Experimental and cognitive psychology.

2.3. Analysis overview

Retrieved papers were downloaded as HTML/XTML files. The papers were parsed, and p-values were extracted. Variables related to incentives were gathered (journal reputation, citation counts, university ranking). Then, after several validation tests were conducted, three main branches of analysis were done: (a) examining changes in psychological results over time by plotting temporal trends in p-values and other variables, (b) examining the link between p-values and SNIP, citations, and university ranking, and (c) analyzing how these variables relate to research topics and methodology by tracking papers' word usages. The organized dataset (https://osf.io/mxs47/) and code (https://github.com/paulcbogdan/PsychChange) have been uploaded to public repositories.

2.4. P-value and statistic extraction

2.4.1. P-value extraction

After stripping papers down to their Results section (Supplemental Materials 2), omitting captions, and removing formatting, p-values were extracted with the following regular expression:

```
[whitespace/parentheses/bracket][p][whitespace/null][sign]
 [whitespace/null][leading zero][whitespace/null][number]
```

Details on the expression and its components are provided in Supplemental Materials 3. After p-value extraction, the proportion of significant p-values falling in the fragile range was calculated for each paper, as the number of p-values between .01 and .05 inclusive ($.01 \leq p < .05$) divided by the number of p-values of .05 or lower ($p < .05$). The analysis focused only on p-values reported with equal (“=”) or less-than (“<” or “≤”) signs. “≤” was treated as “<”. The main analyses did not distinguish one versus two-tailed analyses, but this matter is explored in

Supplemental Materials 5.2, and although the main text focuses on significant results, trends related to insignificant results are described in Supplemental Materials 6.

2.4.2. Statistic extraction

In addition to p-values, measures describing other outputs of statistical tests were also extracted and assigned to a nearby p-value. The test-statistic extraction procedure is described in Supplemental Materials 4. Test statistics (t-values, F-values, chi-square scores, r-values, and z-scores) with complete degrees of freedom were extracted for 1,398,189 p-values (35.8%) along with many extracted without degrees of freedom (using the 240,398 paper pool).

2.4.3. P-value survey

Because papers differ in how they report p-values, a survey was performed, and a taxonomy of reporting styles was developed. For instance, some papers follow the American Psychological Association (APA) style (reporting exact p-values unless $p < .001$), some papers report a mix of inequalities (e.g., "$p < .05$", "$p < .01$", and "$p < .001$"), and some other papers only report significance as a binary (always "$p < .05$"). A survey of how many papers fall into these and other categories is described in Supplemental Materials 5. The only papers that are problematic for the present analyses are the 2.3% of papers that exclusively report "$p < .05$" for significant results. For these papers, their fragile p-value percentage was recomputed based on the p-values implied by nearby test statistics if available. For "$p < .05$" papers that did not report any test statistics, their fragile p-value percentage was set at 51%, which was the mean of among the "$p < .05$" papers that reported test statistics. Supplemental Materials 5 elaborates upon this and discusses the mild degree of underreporting p-value more generally. To ensure that the present conclusions do not hinge on reporting-style phenomena, tests were also done using p-values implied from test statistics (Supplemental Materials 13).

2.5. Validation

Dataset validation was performed in three ways (Supplemental Materials 7). First, a relatively small pool of 40 papers marked as having at least one p-value were manually inspected. In every case, the present approach identified every p-value, properly accounting for different reporting styles and omitting non-Results sections and captions. Second, extracted exact (“=”) p-values were cross-checked with the p-values implied from nearby test statistics, which produced a tight correlation (*r* = .97), verifying the accuracy of the p-values at a wide scale. Third, data on 113 replication attempts was downloaded, and analyses were performed demonstrating that a paper’s fragile p-value percentage strongly predicts its chance of replicating (63.7% cross-validated accuracy).

2.6. Incentives variables

Each paper was assigned variables representing the three incentives of interest: (i) its journal’s yearly Source Normalized Impact Performance (SNIP), which is a score similar to a traditional impact factor but normalized by the number of citations typically received by papers in the same discipline (Moed, 2010), (ii) a log-transformed and year-normalized citation score, and (iii) the Times Higher Education 2024 World University Ranking research score assigned to the most commonly listed university on the paper (frequency ties broken randomly). The collection and organization of these variables are described in Supplemental Materials 8. Additionally, for the university ranking measure, Supplemental Materials 9 discusses alternative ways of assigning one score to each paper (e.g., averaging across schools), although the main text conclusions do not change regardless of how this is done.

2.7. Multilevel regression analysis

Each paper was submitted to three multilevel regressions that (i) used fragile p-value percentages to predict SNIP, (ii) used fragile p-value percentages to predict citations while controlling for SNIP, or (iii) used the authors’ university ranking scores to predict fragile p-value percentages. The multilevel regressions’ structures are detailed in Supplemental Materials 10.

To investigate some of the factors that could bias the regression analysis, several further analyses were done probing other variables that may be relevant. These additional tests covered: (1) whether it is more meaningful to examine the total number of fragile p-values rather than the percentage, (2) whether authors’ ages could be a confound, (3) whether papers containing Results section(s) but no p-values could bias the analysis, and (4) whether papers using Bayesian or machine learning methods could bias the analysis. These tests are all reported in Supplemental Materials 11, and none showed patterns challenging the conclusions below.

2.8. Language analysis

Text analyses were performed on the sentences preceding each reported p-value (details on sentence extraction in Supplemental Materials 12.1). For each paper, the scripts attempted to extract one sentence for each significant p-value. Then, for the 2500 most common words, papers were assigned normalized word usage scores; computed by counting how many times a word appeared among the paper’s sentences and dividing by the total number of words across all the paper’s sentences. For each of the 2500 most common words, a separate linear regression was fit predicting the word’s usage score based on the paper’s fragile p-values percentage, e.g.,

$$\mathit{normed\_usage}[\text{“}\mathit{the}\text{”}] \sim 1 + p\ \mathit{fragile\ percentage}$$

The identification of 2500 words and the regressions were done four times, separately for sentences reporting t-values, F-values, chi-square values, or correlation coefficients/betas (see the quantity of each in Supplemental Materials 12.2). Using data associated with all p-values, irrespective of nearby test statistics, further regressions were also tested which added the year, SNIP, citation, or ranking score as predictors, e.g.,

$$\mathit{normed\_usage}[\text{“}\mathit{the}\text{”}] \sim 1 + \mathit{p\_fragile\_percentage} + \mathit{ranking\ score}$$

Although not formally tested, overlapping associations between word usage and two other variables would notably point to statistical mediation (e.g., ranking → word → p-values or ranking → p-values → word); see Bogdan, Cervantes, and Regenwetter (2023) for discussion and intuition on the close link between mediation, multivariate distributions, and overlapping in variable patterns.

3. Results

3.1. Fragile p-values have declined over time

From before the replication crisis (2004-2011) to today (2024), the overall percentage of significant p-values in the "fragile" range has dropped from 32% to nearly 26% (Figure 2A). This matches the percentage of fragile p-values expected from studies with 80% power (see the dashed lines in Figure 2). A similar trend emerges if the focus is instead placed on the p-values implied papers' test statistics (Figure 2B). Interestingly, the fragile percentage for implied p-values is usually 2-4% lower than the percentage calculated with reported p-values. This partly reflects a selection effect between papers that report versus do not report test statistics (furthered probed in Supplemental Materials 13.2). Regardless, the data overall shows fragile p-value rates markedly decreasing over time.

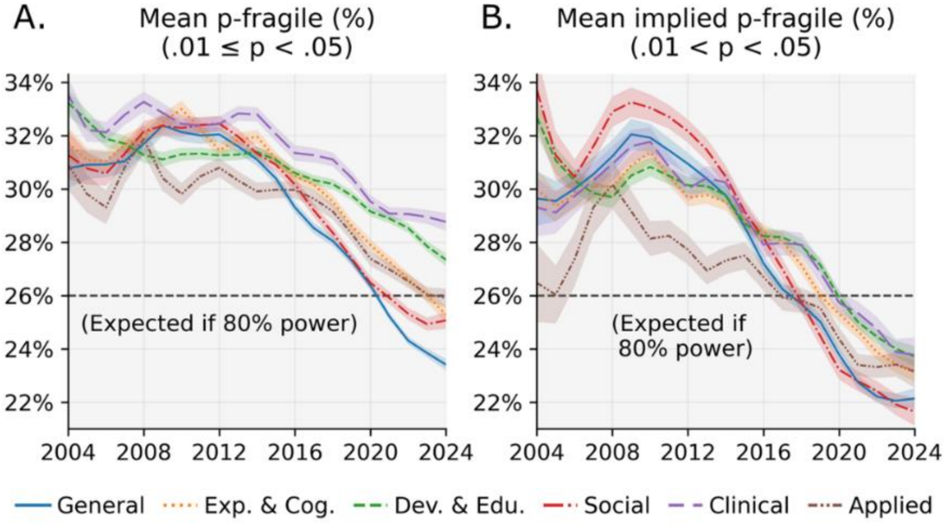


Figure 2. Changes in fragile p-values over time. A. The mean fragile p-value percentage for each subfield and year was calculated. Papers were assigned to specific subfields based on their journal's Scopus classification. Papers in journals associated with two or three subfields contributed to each subfield's plot (contributed in full, not weighed by a half or third). A correction, subtracting 2.3% from the fragile p-value percentage, has been applied to account for papers underreporting the strength of results (e.g., reporting "p < .05" unnecessarily); see Supplemental Materials 5.1. The dashed line at 26% is a reference showing the fragile p-value rates expected from studies with 80% power and α = .05. **B.** Mean fragile percentage, calculated using p-values implied from nearby test statistics. Shaded regions represent ± 1 standard error.

Examining the distribution of fragile p-values more precisely suggests that these aggregate shifts derive from the average study reporting fewer fragile p-values (Figure 3); see how the histogram averages steadily shift leftward over time. However, as the right tails of Figure 3 show, there remain many studies publishing weak p-values. This deserves consideration despite the overall trend toward fewer fragile findings.

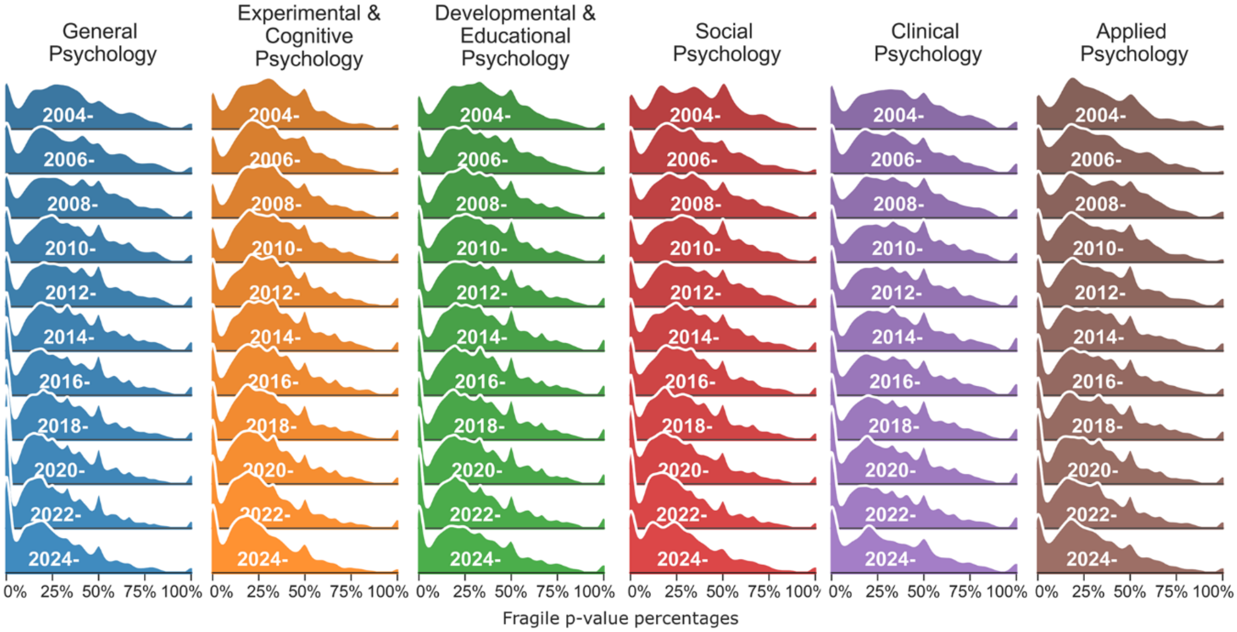


Figure 3. Year-wise ridgeplots for different subfields. Each column is a series of density plots generated independently for each of the six psychology subfields. Each density plot represents the data for two years (e.g., "2004-" corresponds to 2004 & 2015). These year labels are placed at the average of each histogram.

The drop in fragile p-values may be driven by increases in statistical power. The median sample size has substantially increased over time (Figure 4A). Among studies reporting t-values

(55,342 papers), larger samples predict lower p-value percentages (paper-by-paper Spearman correlation: ρ = -.22, *p* < .0001). Effect sizes are also relevant to power, but their relationship to fragile p-values is more ambiguous. Reported effect sizes have generally dropped over time (Figure 4B) and this may reflect effect size estimates becoming more accurate as larger samples are used; endorsing this idea, the median reported Cohen's *d* is strongly negatively correlated with sample sizes (ρ = -.68, *p* < .0001). In turn, the correlation between Cohen's *d* and the fragile percentage is weak (ρ = -.13, *p* < .0001). Regardless, the sample size trends make a strong case that increases in statistical power may partially underlie the move away from fragile results.

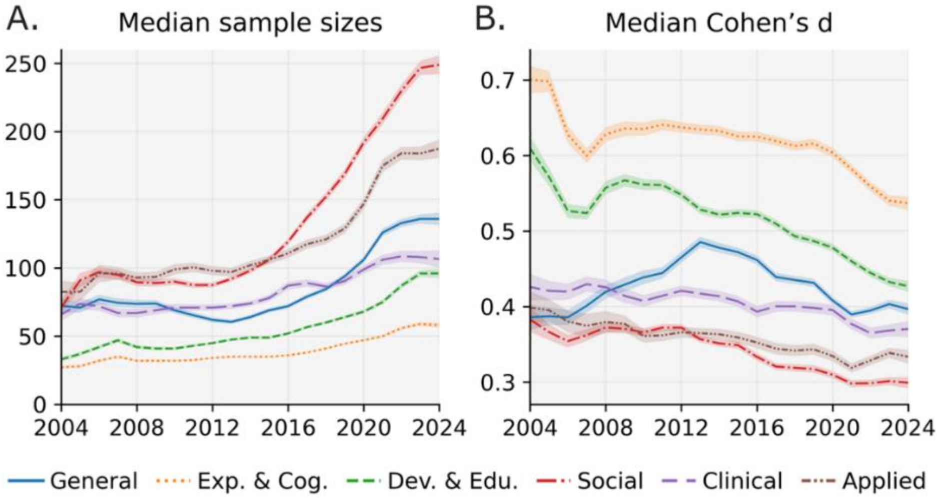


Figure 4. Changes in sample sizes and effect sizes over time. A. Median of the median sample size used for studies' t-tests, calculated from t-test degrees of freedom of significant results. **B.** Median of the median Cohen's d of studies' significant t-tests, calculated by dividing t-values with the square root of the sample size. Shaded regions represent ± 1 standard error.

3.2. Fragile p-values and incentives

To assess the relationship between fragile p-values and academic psychology's incentives, multilevel regressions were tested linking papers' fragile p-value percentages with (i) its journal's SNIP/normalized-impact-factor, (ii) the paper's log-transformed year-standardized citation count, and (iii) the university rankings of its authors. For each test, the regression included interactions with Year to capture whether any relationship changed over time.

All three variables representing academic psychology's incentives were linked to fragile p-values in some way. Fragile p-values did not significantly predict SNIP (*β* = -.004 [-.009, .001], *B* = -.009, *p* = .14; Figure 5A)[1] but yielded a significant fragile p-value x Year interaction effect on SNIP (*β* = -.009 [-.010, -.008], *B* = -.0039, *p* < .0001. In other words, more esteemed journals have historically published papers with weaker results, but nowadays top journals mostly publish strong findings. On top of this journal effect, papers with fewer fragile p-values also receive more citations (*β* = -.036 [-.044, -.028], *B* = -.13, *p* < .0001; Figure 5B). Further, an interaction with Year suggests the inverse link between fragile p-values and citations has grown since the replication crisis' start (*β* = -.011 [-.016, -.006], *B* = -.008, *p* < .0001). Overall, these results point to important growth in the standards used to evaluate research. Yet, bucking these optimistic trends, it is also the case that papers from higher-ranked universities tend to have more fragile p-values (*β* = .016 [.009, .025], *B* = .00018, *p* = .0005; Figure 5C), and a null interaction with Year shows that this link between fragile p-values and rankings has not significantly changed over time (*β* =-.004 [-.001, .010], *B* = -.00001, *p* = .09). Hence, the findings on these three incentive variables altogether paint a mixed picture.

[1] β and B refer to standardized and unstandardized coefficients, respectively.

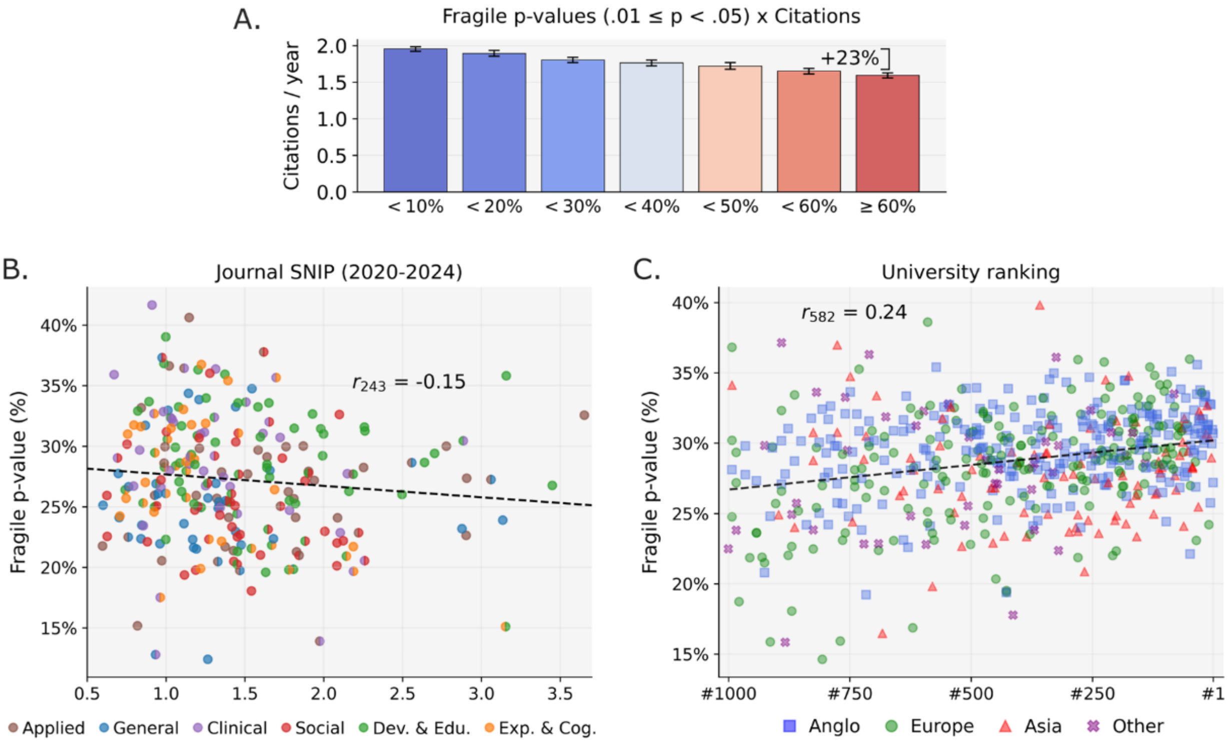


Figure 5. Links between p-values and academic values/incentives. A. Bars represent the citations per year received by recent papers (2020-2024), depending on their fragile p-value percentage. Although the main text analyses use a log-transformed and year-standardized measure of citations, these have been reverted here back to a more interpretable quantity (citations per year). This measure was averaged for each year and then averaged across years; SNIP was not controlled for. **B.** Scatterplot represents each journal's mean fragile p-value percentage and mean SNIP from 2020-2024. Journals with two areas are colored with both. Four journals had 3 or more areas, and for those, two areas were selected randomly. Journals with fewer than 10 papers with p-values and journals with an unreliable fragile p-value percentage measure (standard error .04) were excluded. A Spearman correlation is reported. **C.** Scatterplot represents each university's mean fragile p-value percentage and its Times Higher Education 2024 ranking. The whole 2004-2024 period was used to ensure an adaquete sample size with the effect of year on fragile p-values regressed out. Rankings are used for the x-axis rather than research score because rankings were expected to be more intuitive. Universities with fewer than 10 papers with p-values or with an unreliable fragile p-value measure (standard error over .04) were excluded. Dots are colored based on a university's region. "Anglo" refers the United States, United Kingdom, Canada, Australia, and New Zealand. This regional organization was designed to avoid overly small or large groups but may ignore heterogeneity within the divisions.

3.3. Topics and methodologies linked to fragile p-values

The final analysis examined word usage to enhance the interpretation of the p-value findings thus far. Figure 6 shows the results of 10,000 independent regressions, examining how papers' fragile p-value percentages predict 2500 different words' usages. For example, the red word "*completer*" in the bottom-left corner of Figure 6 indicates that papers reporting fragile p-values for t-tests are more likely to also use the word "*completer*" near those t-tests (this term refers to clinical intervention research). To cover a wide range of statistical approaches, this word analysis was done separately for results associated with different test statistics. This yielded many patterns. Words tied to often criticized psychological topics lay near the bottom of these lists, like "[social] *priming*" and "*genotype*". Topics where data recruitment is expensive also stand out, particularly topics related to clinical, developmental, and/or biological psychology, such as "*infant*", "*ASD* [autism spectrum disorder]", "*intervention*", "*pupil*", "*cortisol*", "*amplitude*", and "*gyrus*". In addition, many words linked to weak p-values reference analytic approaches offering low power or permitting many degrees of freedom: "*between*[-]*group* [analysis]", "*moderated*", "*ANCOVA*", "*left*" vs. "*right*" "*hemisphere*" differences, or "*sex*" effects. Interestingly, among the predictors of strong p-values, few words pertain to scientific topics but instead predominantly concern methods – e.g., "*multivariate*", "*hierarchical*", "*validity*" or "*repeated*[-]*measures*". Altogether, these results begin to illustrate the literature producing strong or weak p-values.

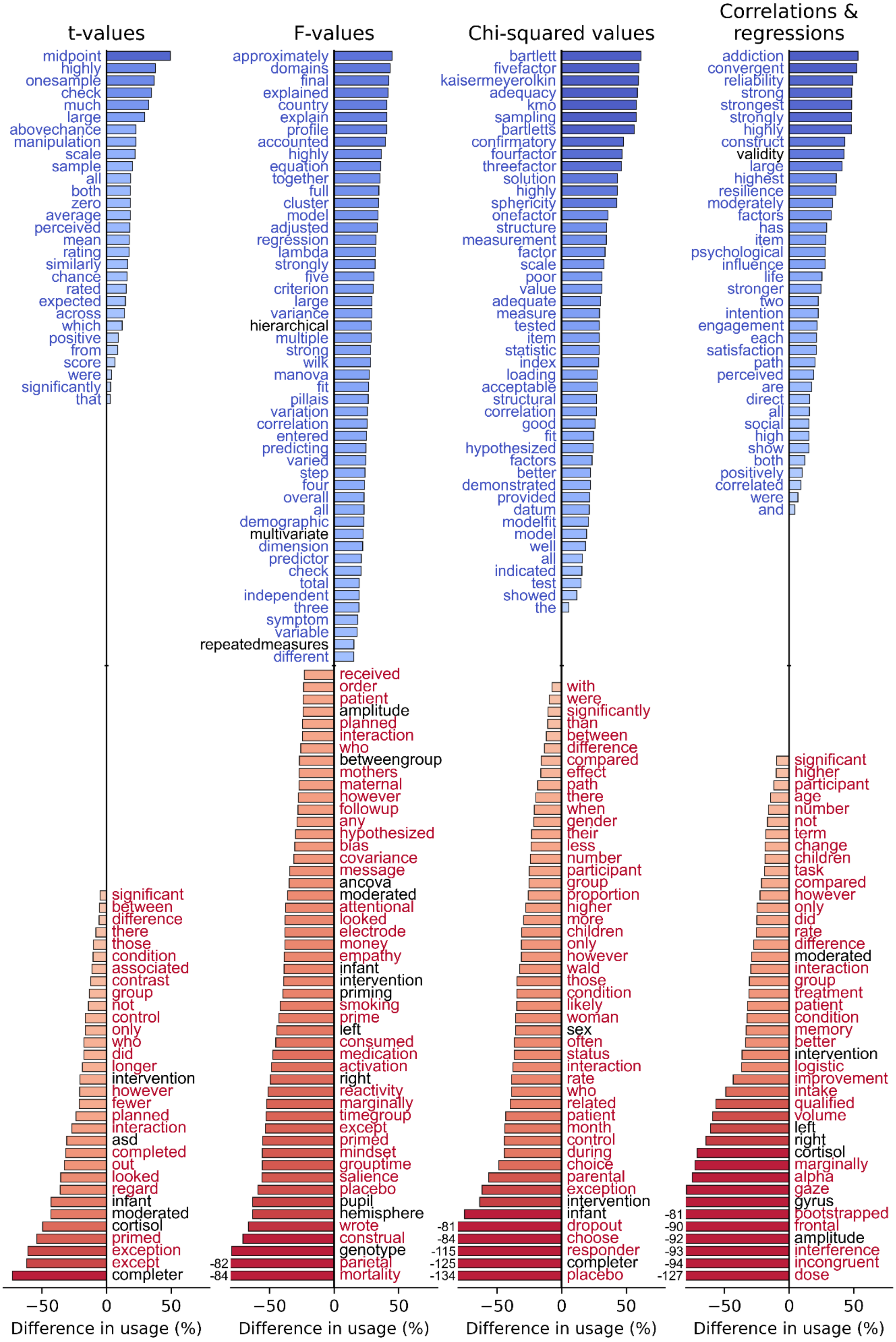

t-values
midpoint
highly
onesample
check
much
large
abovechance
manipulation
scale
sample
all
both
zero
average
perceived
mean
rating
similarly
chance
rated
expected
across
which
positive
from
score
were
significantly
that
significant
between
difference
there
those
condition
associated
contrast
group
not
control
only
who
did
longer
intervention
however
fewer
planned
interaction
asd
completed
out
looked
regard
infant
moderated
cortisol
primed
exception
except
completer
−50
0
50
Difference in usage (%)
F-values
approximately
domains
final
explained
country
explain
profile
accounted
highly
equation
together
full
cluster
model
adjusted
regression
lambda
strongly
five
criterion
large
variance
hierarchical
multiple
strong
wilk
manova
fit
pillais
variation
correlation
entered
predicting
varied
step
four
overall
all
demographic
multivariate
dimension
predictor
check
total
independent
three
symptom
variable
repeatedmeasures
different
received
order
patient
amplitude
planned
interaction
who
betweengroup
mothers
maternal
however
followup
any
hypothesized
bias
covariance
message
ancova
moderated
attentional
looked
electrode
money
empathy
infant
intervention
priming
smoking
prime
left
consumed
medication
activation
right
reactivity
marginally
timegroup
except
primed
mindset
grouptime
salience
placebo
pupil
hemisphere
wrote
construal
genotype
-82 parietal
-84 mortality
−50
0
50
Difference in usage (%)
Chi-squared values
bartlett
fivefactor
kaisermeyerolkin
adequacy
kmo
sampling
bartletts
confirmatory
fourfactor
threefactor
solution
highly
sphericity
onefactor
structure
measurement
factor
scale
poor
value
adequate
measure
tested
item
statistic
index
loading
acceptable
structural
correlation
good
fit
hypothesized
factors
better
demonstrated
provided
datum
modelfit
model
well
all
indicated
test
showed
the
with
were
significantly
than
between
difference
compared
effect
path
there
when
gender
their
less
number
participant
group
proportion
higher
more
children
only
however
wald
those
condition
likely
woman
sex
often
status
interaction
rate
who
related
patient
month
control
during
choice
parental
exception
intervention
infant
-81 dropout
-84 choose
-115 responder
-125 completer
-134 placebo
−50
0
50
Difference in usage (%)
Correlations &
regressions
addiction
convergent
reliability
strong
strongest
strongly
highly
construct
validity
large
highest
resilience
moderately
factors
has
item
psychological
influence
life
stronger
two
intention
engagement
each
satisfaction
path
perceived
are
direct
all
social
high
show
both
positively
correlated
were
and
significant
higher
participant
age
number
not
term
change
children
task
compared
however
only
did
rate
difference
moderated
interaction
group
treatment
patient
condition
memory
better
intervention
logistic
improvement
intake
qualified
volume
left
right
cortisol
marginally
alpha
gaze
gyrus
-81 bootstrapped
-90 frontal
-92 amplitude
-93 interference
-94 incongruent
-127 dose
−50
0
50
Difference in usage (%)

Figure 6. Words linked with strong (blue) or weak (red) p-values. Across 2500 words and 4 statistic types, 10,000 regressions were fit, attempting to predict each word's usage based on papers' fragile p-value percentages. Bars represent the standardized coefficient of the fragile-p-value predictor divided by a word's baseline frequency. Every listed word was significant following correction at a family-wise error level ($p_{\text{FWE}} < .05$) across families of 2500 tests (correction applied separately for each test statistic). Correction used the Holm-Sidak method, which is similar to Bonferroni correction. Words mentioned in the main text (e.g., "*completer*") are colored black in the figure for emphasis.

To shed light on the association between papers' fragile p-values and their authors' universities' rankings, further regressions were tested. Among the links to academic incentives shown earlier (Figure 5), the link to university rankings deserves this additional special focus because it implies a disconnect where academic incentives do not promote strong results. To help unpack this university effect, the present analyses regressed each word's usage on a paper's fragile p-value percentage along with its university ranking score (examining p-values across t-tests, F-tests, etc.). Then, the conjunction was taken between words yielding positive associations with both predictors. These words are listed in Table 1 – e.g., the word "*abstinence*" here indicates that (i) this word is often used to describe results with fragile p-values, and (ii) this word is also often used by authors from highly ranked universities. This analysis can also be performed with respect to negative associations and for other variables (Year, citations, SNIP), and those overlaps are reported in Supplemental Tables S2-12. In the present section, the focus is solely on words linked to higher rates of fragile p-values and higher-ranking university papers.

Words linked *more* fragile p-values and to *higher-ranked* universities						
abstinence	twotailed	looked	HIV	unadjusted	verb	choose
chose	successfully	saw	arm	infant	drug	HAMD
driven	looking	bar	binomial	provider	motivated	multivariate
completer	look	remission	window	responder	experiment.	MDD
reliable	money	novel	connectivity	speech	adjusting	trial
late	tau	joint	toddler	days	visit	primed
exact	label	unexpected	antidepressant	brain	give	implicit
race	her	attended	lifetime	activation	cause	cortisol
relative	pupil	likely	week	bilingual	stories	placebo
priming	bipolar	continued	earlier	versus	object	volume
familiar	medication	minority	onset	caregivers	baseline	learned
prior	BDI	caregiver	depressed	children	offer	attendance
subject	longer	day	assigned	attempt	whose	reduction
took	odds	received	spatial	month	condition	they
less	choice	prime	vocabulary	participant	smoking	expressed
history	later	memory	treatment	temporal	disorder	making
receiving	reduced	interacted	pair	exposure	either	fewer
report	course	end	region	preference	read	episode
greater	remained	planned	more	ethnicity	consistent	faster
who	care	bias	did	early	outcome	event
without	completed	left	larger	any	cue	such
slower	during	within	those	diagnosis	decreased	logistic
patient	than	rate	times	younger	increased	neutral
sensitivity	experienced	individual	response	contrast	compared	made
session	mothers	qualified	partner	problem	their	when
face	child	having	but	use	associated	reported
number	symptom	controlling	age	lower	after	change
among	not	main	significant	time	with	interaction

Table 1. Words positively associated with fragile p-values and with higher-ranked universities. For 2500 words, 2500 regressions were fit. Each regression attempted to predict one word's usage based on the paper's fragile p-value percentage and the paper's university ranking score (*word usage ~ 1 + fragile p-value percentage + ranking*). Per the regression coefficients, every word listed here is both used significantly more by authors from highly ranked universities and is positively associated with fragile p-values. Because the analysis here requires overlaps in significance across the two predictors, the requirements for significance were loosened to use false-discovery rate correction ($p_{\text{FDR}} < .05$), unlike the family-wise correction used for Figure 6 (Benjamini & Hochberg, 1995).

Top institutions often study special populations, such as "*HIV*" patients, major depressive disorder ("*MDD*") patients along with "*caregivers*" and smokers ("*abstinence*" & "*smoking*"). Along with being clinically oriented, top-ranked institutions emphasize biology, focusing on medication ("*antidepressant*"), hormones ("*cortisol*"), and neural "*activation*" or "*tau*". In this type of research, achieving high statistical power can be challenging. Beyond biological and clinical psychology, Table 1 more generally suggests that top-ranked institutions are interested in expensive behavioral work and subtle mechanisms. For instance, the terms "*day*" and "*week*" emerge, which refer to multi-session studies. Such studies are difficult to run via online platforms and can be much more labor-intensive than single-session research. "*Looked*" is tied to eye-tracking research, which requires costly equipment. "*Choose*" and "*money*" point specifically to behavioral economics and decision-making research, which can be difficult to run online in a convincing manner. "*Memory*" research can be time-consuming because it often requires from lengthy retention intervals. Memory effects may also be subtle, benefiting from the collection of many trials because hits/miss responses are effectively drawn from a probability distribution. Related to the interest in subtlety, highly ranked universities also show a preference for "*priming*" research and "*implicit*" mechanisms. By contrast, many of the words associated with low-ranked universities and few fragile p-values stem from survey and correlational research (Supplemental Table S2). Such surveys come with their own limitations, but they can presumably be collected more widely and cheaply than the type of experimental work preferred by top institutions.

Altogether, these patterns suggest that the link between university rankings and weaker results can be explained as the pursuit of topics and methods where statistical power is likely more limited. However, this conclusion alone may not be complete. For the final analysis, a non-multilevel linear regression was performed regressing a paper's fragile p-value percentage its university ranking while including all 2500 words' usage levels as covariates. The original link between fragile p-values and rankings remains significant ($\beta = .014$, $p < .0001$). Compared to another non-multilevel regression without the covariates, the link to rankings has dropped by two-thirds (reference: $\beta = .039$, $p < .0001$). Yet, the continued existence of the pattern suggests the association is only partly explained by the research topic and method employed.

4. Discussion

The present research investigated reported p-values in psychology papers from 2004 to 2024, putting forth three main conclusions: First, psychological research has begun to publish considerably stronger p-values in recent years, pointing to the success of many replication crisis efforts. Second, analyses linking p-values to academic incentives show that contemporary papers reporting strong p-values tend to find publication in more esteemed journals and receive more citations. However, there are also signs that robust research is still not linked to success, as top-ranked universities today tend to publish papers with weaker p-values. Third, dissecting these patterns by analyzing language usage shows how some methods and topics consistently produce findings with fragile p-values. The link between high-end universities publishing weak p-values can be partially explained by top universities emphasizing studies that are resource-intensive, laborious, and linked to subtle effects. Along with these main results, readers are encouraged to see the extensive supplemental analyses, many putting forth original findings (e.g., on p-value reporting styles [section 5], insignificant p-values [section 6], and Bayesian or machine learning analysis [section 11]). Possible interpretations and implications of the primary findings are discussed below.

The percentage of significant p-values that are fragile ($.01 < p < .05$) has dropped from 32% before the replication crisis to just over 26% today (Figure 2). This percentage nearly matches the level of fragile p-values expected from studies with 80% power. Further, as Supplemental Materials 7.3 and 13.2 show, lower rates of fragile p-values significantly predict replicability. These decreases in the fragile p-value rates are evident across every psychological discipline. Although there remain *very many papers* that continue to report weak evidence (see Figure 3 ridgeplots), there overall appears to have been considerable progress in the strength of psychology’s findings since the replication crisis began.

Drops in fragile p-values may be partially explained by many studies increasing their statistical power. Power is closely linked to sample size, and sample sizes began to rapidly rise around 2015 (Figure 4A), which coincides with fragile p-values precipitously. The expansion of sample sizes is likely intertwined with the rise of online recruitment platforms, such as Amazon Mechanical Turk or Prolific, which have made large sample sizes more widely accessible (Buhrmester, Talaifar, Gosling, 2018). Effect sizes are another piece of statistical power, but the association here is more ambiguous. All else kept equal, smaller effect sizes will lead to lower statistical power. However, in practice, published studies with low power will report inflated effect sizes (Kühberger et al., 2014), so the link between these variables becomes muddled; see the strong negative paper-by-paper correlation between sample sizes and effect sizes (Spearman ρ = -.68). Thus, decreases in effect sizes over time (Figure 4B) may actually further endorse that statistical power is rising in psychological research. In contrast, prior studies have put forth that statistical power has remained low in the social/behavioral sciences from 1955 to 2015 (Smaldino & McElreath, 2016) and the incentives for fast scientific discovery dissuade well-powered research (Tiokhin et al., 2021). The present patterns are instead more consistent with an emerging upward trajectory in statistical power, and statistical power is foundational to replicable science (Stanley et al., 2018).

These demonstrated improvements in psychological research will hopefully push back against the public distrust in science that has grown in recent years. Surveys show that 18% of laypeople report having heard of recent failures to replicate psychology studies and up to 29% report awareness of such failures in other fields (Anvari & Lakens, 2018). A considerable minority of the public uses replication failures to justify distrust in scientific research (Anvari & Lakens, 2018), and experimental research concurs that informing people of replication failures dampens scientific trust (Hendriks et al., 2020). Hopefully, the results here can serve as a springboard to communicate the rigor in much of contemporary psychology.

The present analyses also demonstrate that papers reporting stronger p-values tend to be published in more esteemed journals. The nominal effects between fragile p-values and journal SNIP may appear to be minor, as an upper echelon journal (2.5 SNIP, 95th percentile) shows just 2% lower rates of fragile p-values than less esteemed journals (1.0 SNIP, 20th percentile) (Figure 5B). However, visual inspection of the scatterplot suggests that there is considerably more variability among lower journals. Whereas most high SNIP journals predominantly publish strong results, at the lower end, there are journals with results of all sorts. Hence, although a line of best fit may not show a steep slope, top journals consistently appear to hold papers to a high standard. On top of this journal effect, papers reporting fewer fragile p-values also tend to receive more citations. Papers reporting strong p-values (less than 10% fragile) can expect to receive 22% more citations than papers reporting mostly weak p-values (over 60% fragile) (Figure 5A). Although not illustrated, comparing papers with 10% vs. 90% fragile p-values further reveals a 30% gain. These are substantial boosts that exist for conducting seemingly more robust research.

The present findings on journal destinations and citations push back on some pessimistic conclusions put forth by earlier work. Examining replication outcomes and reported statistics, Dougherty and Horne (2022) along with Gupta and Bosco (2023) suggest that higher impact factor journals tend to publish less robust findings. Investigating links to replication outcomes, Schafmeister (2021) and Serra-Garcia and Gneezy (2021) argue that successful replication and replicability do not impart any benefits to a paper’s citations. However, unlike the present research, these four prior studies all focused on older papers (overwhelmingly pre-2017). Additionally, these prior studies examined smaller portions of the literature, whereas the preset research better covers the entirety of psychological science and is thus most resilient to selection biases. With these changes, a brighter picture unfolds wherein robust results are nowadays published in higher SNIP journals and receive more citations.

All this being said, the final result linking fragile p-values and university ranking adds nuance to the otherwise positive trends. Specifically, the #1 globally ranked university will tend to publish papers with 3.5% more fragile p-values than the #1000 university (Figure 5C). This is a considerable fraction of the temporal trend from before the replication crisis to today. Furthermore, examining the confidence intervals of the null Ranking x Year interaction suggests that this gap between high/low ranking universities has minimally shifted since the replication crisis began, if at all (for a discussion of interpreting the absence of effects, see Lakens et al., 2018). The language analyses suggest that these patterns can be partly explained by high-ranking universities' preferences for difficult research. This preference manifests as a focus on clinical and biological psychology along with tendencies to conduct behavioral studies involving costly equipment, multiple days of labor, and in-person data collection. These factors presumably limit sample sizes, and the large investments required may encourage questionable research practices.

The apparent link between university rankings and weaker findings begs questions about what exactly the psychology community should aspire for. Some of this difficult research may have great practical importance (e.g., medical value). Moreover, this type of 'prestigious' experimental work may also have higher validity and causal power than despite correlational work producing stronger p-values (Table S2). If the type of research pursued by high-ranking institutions must be done but must also be conducted in a robust fashion, then what systematic changes are necessary? The present research will hopefully inform these types of policy and institutional questions.

Limitations

The principal assumption is that the present approach does not introduce selection bias in a way that meaningfully confounds the link between fragile p-values and other variables. Selection bias could operate in terms of which papers were included and which results within a paper are extracted. Regarding paper inclusion, p-values were extracted for 72% of papers containing a Results section. This is a clear majority but leaves a meaningful minority, including qualitative papers, methodological papers, and papers using Bayesian statistics or machine learning. The results in Supplemental Materials 11.4 and 11.5 investigate such papers, showing no evidence of biases against the main conclusions. Nonetheless, these papers create some ambiguity. Selection bias may also occur within papers because the analyses ignore figures and tables. It is unclear whether this causes p-value fragility to be underestimated or overestimated. However, for most research areas, papers' most central findings will presumably still be mentioned in the text. This would suggest that ignoring figures and tables may actually yield a more refined measure, although this cannot be said with certainty.

A final source of selection bias may stem from some journals not being included in the dataset. In particular, the dataset omitted journals of miscellaneous research (e.g., *Science* or the *Proceedings of the National Academy of Sciences*) to ensure that all of the papers covered here were specifically on psychological research. As these journals have among the highest impact factors, their omission may have caused some selection bias. However, this is expected to be minor, as these journals' papers are a fairly small fraction of the psychology literature.

Author Contributions: P.C.B. is the sole author of this article.

Conflicts of Interest: The author declares that there were no conflicts of interest with respect to the authorship or the publication of this article.

References

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666-678.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bogdan, P. C., Cervantes, V. H., & Regenwetter, M. (2024). What does a population-level mediation reveal about individual people?. *Behavior Research Methods*, *56*(6), 5667-5692.

Boggero, I. A., Hostinar, C. E., Haak, E. A., Murphy, M. L., & Segerstrom, S. C. (2017). Psychosocial functioning and the cortisol awakening response: Meta-analysis, P-curve analysis, and evaluation of the evidential value in existing studies. *Biological Psychology*, *129*, 207-230.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149-154.

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

Dougherty, M. R., & Horne, Z. (2022). Citation counts and journal impact factors do not capture some indicators of research quality in the behavioural and brain sciences. *Royal Society Open Science*, *9*(8), 220334.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS ONE*, *7*(1), e29081.

Fox, C. W., Meyer, J., & Aimé, E. (2023). Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal. *Functional Ecology*.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 585-594.

Gordon M, Viganola D, Dreber A, Johannesson M, Pfeiffer T (2021) Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLoS ONE* 16(4): e0248780.

Gupta, A., & Bosco, F. (2023). Tempest in a teacup: An analysis of p-Hacking in organizational research. *PloS ONE*, *18*(2), e0281938.

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, *42*, 13-31.

Hendriks, F., Kienhues, D., & Bromme, R. (2020). Replication crisis= trust crisis? The effect of successful vs failed replications on laypeople's trust in researchers and research. *Public Understanding of Science*, *29*(3), 270-288.

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.

Kitayama, S. (2017). Journal of Personality and Social Psychology: Attitudes and social cognition.

Krawczyk, M. (2015). The search for significance: a few peculiarities in the distribution of P values in experimental psychology literature. *PloS ONE*, *10*(6), e0127872.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PloS ONE*, *9*(9), e105825.

Lakens, D. (2015). On the challenges of drawing conclusions from p-values just below 0.05. *PeerJ*, *3*, e1142.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in methods and practices in psychological science*, *1*(2), 259-269.

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265-277.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., & Nuijten, M. B. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719-748.

Nosek, B. A., & Lakens, D. (2014). Registered reports. In: Hogrefe Publishing.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615-631.

Olsson-Collentine, A., Van Assen, M. A., & Hartgerink, C. H. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, *30*(4), 576-586.

Pritschet, L., Powell, D., & Horne, Z. (2016). Marginally significant effects as evidence for hypotheses: Changing attitudes over four decades. *Psychological Science*, *27*(7), 1036-1042.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551.

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 364.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *Available at SSRN 2160588*.

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.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological bulletin*, 144(12), 1325.

Stuart, M. T., Colaço, D., & Machery, E. (2019). P-curving x-phi: Does experimental philosophy have evidential value? *Analysis*, *79*(4), 669-684.

Tiokhin, L., Yan, M., & Morgan, T. J. (2021). Competition for priority harms the reliability of science, but reforms can help. *Nature Human Behaviour*, *5*(7), 857-867.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, *114*(48), 12708-12713.

Vadillo, M. A., Gold, N., & Osman, M. (2016). The bitter truth about sugar and willpower: The limited evidential value of the glucose model of ego depletion. *Psychological Science*, *27*(9), 1207-1214.

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12.

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).

Youyou, W., Yang, Y., & Uzzi, B. (2023). A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proceedings of the National Academy of Sciences*, *120*(6), e2208863120.

1. Additional data collection details

Data collection began with the retrieval of metadata from Lens.org. The Lens.org database was filtered to solely records from psychology journals; that is, journals classified by Scopus as: General Psychology, Social Psychology, Experimental and Cognitive Psychology, Applied Psychology, Developmental and Educational Psychology, or Psychology (Miscellaneous). Additionally, the database was filtered to only records from large non-predatory publishers amenable to data mining research: Elsevier, SAGE Publications, Springer-Nature, Wiley, and Frontiers. This yielded 1,156,855 scholarly records from 1900 to August 2024. Notably, these are not all articles but instead records (i.e., any item with a DOI) and include entries such as editorials, cover artwork, and seemingly erroneous empty items. Publishers American Psychological Association and Taylor and Francis could not be used due to restrictions in their terms and conditions. These two publishers produced 549,212 records from 1900 to August 2024 that were not analyzed. Nonetheless, these counts suggest that the five analyzed publishers make up a large majority of the psychological literature.

Of the 1.15M records, some were of journals also considered to be Cognitive Neuroscience based on their Scopus categories. Yet, some of the journals labeled Cognitive Neuroscience only publish a relatively small number of actual cognitive neuroscience papers. For example, both *Cortex* and *Cognition* have cognitive neuroscience labels, but the former largely publishes actual cognitive neuroscience papers whereas the latter mostly publishes cognitive psychology papers (per the Author's assessment). Hence, every journal with the Cognitive Neuroscience category was manually inspected, and the following 30 journals were selected for inclusion despite their label (possible near-duplicates related to inconsistent naming in the Lens.org database): *Adaptive Behavior, Asian Journal Of Sport And Exercise Psychology, British Journal Of Developmental Psychology, Child Development Research, Chronic Stress, Cognition, Cognitive Processing, Cognitive Psychology, Cognitive Science, Cognitive Systems Research, Developmental Science, Evolutionary Psychology, Human Factors, Human Factors The Journal Of The Human Factors And Ergonomics Society, International Journal Of Behavioral Development, Journal Of Behavioral And Cognitive Therapy, Journal Of Communication Disorders, Journal Of Contextual Behavioral Science, Journal Of Fluency Disorders, Journal Of Memory And Language, Journal Of Research On Adolescence, Journal Of The Experimental Analysis Of Behavior, Learning Amp Behavior, Learning And Motivation, Memory Amp Cognition, Nature Human Behaviour, Quarterly Journal Of Experimental Psychology, Sleep Medicine Clinics, The Quarterly Journal Of Experimental Psychology, Topics In Cognitive Science*. After excluding every Cognitive Neuroscience journal except for those 30, the pool of records was reduced to 1,074,465. Of these, 643,571 were from 2004-2024.

Attempts were made to download web versions (HTML or XML files) of these 644k records. PDFs were not considered, as the HTML/XML versions permit more precise text mining, allowing the removal of tables/figures/captions from the text and the segmentation of articles into individual sections. The data collection scripts successfully downloaded content for 598,364 records. Of those, 374,198 contained a Results section, which was cut down to 372,633 that were from empirical journals (i.e., journals that regularly published papers containing Results sections; see Supplemental Materials 2).

Note that psychology papers from multidisciplinary journals (e.g., *Science* or the *Proceedings of the National Academy of Sciences*) were not included in the initial pool of records or the final dataset. This is because a paper's categorization, given by Scopus, depends entirely on its journal, and every paper from journals like these is labeled "multidisciplinary."

2. Results section extraction

The papers were pruned to just their Results sections to avoid mentions of p-values in the Methods or other sections, which are assumed to be less relevant to the strength of a paper's findings. Pruning involved identifying section titles based on HTML/XML elements. For example, Elsevier encodes "Results" section titles using <ce:section-title, id="…"> Results</ce:section-title>. Processing specifically searched for titles containing the words "Result" or "Finding" (e.g., "Results" or "Results and Discussion"). In the case of multi-study reports containing multiple Results sections, each one's text was concatenated. HTML/XML elements were likewise used to remove figures and table captions. The manual validation confirmed the efficacy of these scripts in preventing Methods or caption p-values from entering the final datasets.

Review papers occasionally contained sections with the term "Results". Although such papers are harmless to the analysis, as review papers are unlikely to contain p-values, an effort was made to eliminate them. Hence, for a given year, all papers from journals that published fewer than five papers with a Results section were deemed to not be empirical journals, and their papers were considered to not have Results sections (changes the status of 1730 papers; the final count in Section 2.2 accounted for this).

3. P-value regular expression

After omitting captions and removing formatting, p-values were extracted with the following regular expression:

```
[whitespace/parentheses/bracket][p][whitespace/null][sign]
 [whitespace/null][leading zero][whitespace/null][number]
```

[whitespace/parentheses/bracket] specifies a whitespace, opening parentheses, or opening bracket. The whitespace aspect covered the different Unicode characters used to represent spaces (e.g., a standard space " ", a non-breaking space U+00A0, etc.). [p] specifies "p" or "P". [whitespace/null] specifies whitespace or the lack of any character. [sign] corresponds to "=", ">", "≥", "<", or "≤". [leading-zero] corresponds to "0" or the lack of any character (e.g., "*p* = .04" and "*p* = 0.04"). Finally, [number] corresponds to decimal numbers and scientific notation (e.g., "p = 2.7 × 10-5"; given the diversity in how scientific notation is written, some cases were presumably not covered). This elaborate expression was developed to satisfy many different journals' reporting styles.

4. Test statistic extraction and processing

In addition to p-values, measures describing other outputs of statistical tests were also extracted and assigned to a nearby p-value. Foremost, extraction attempted to extract complete test statistics (t-values, F-values, chi-square scores, r-values, and z-scores) with complete degrees of freedom. To do this, the text up to 32 characters prior to a given extracted p-value was scanned using a regular expression designed for the extraction of test statistics (see released code). If, for example, a different p-value was reported 20 characters prior, then a search would only cover the prior 19 characters to avoid ambiguity. Overall, 947,117 extracted p-values (37.4%) were paired with a test statistic with complete degrees of freedom.

In some cases, F-values and chi-square values were reported inversely. That is, an F-value is typically used to indicate that one variance is significantly greater than another variance (e.g., F[1, 100] = 4 indicates $p = .05$). However, F-values can also be used to report that one variance is significantly smaller than another (F[1, 100] = 0.25 may also indicate $p = .05$). These cases were identified by assessing whether one minus the p-value implied by a test statistic is within 0.1 of the nearby reported p-value. For example, if a paper reports "F[1, 100] = 0.25, $p = .05$", then the F-value is treated as 4. As so, 0.1% of F-values were flipped. The same was done for chi-square statistics, but these were more common (1.7% of chi-square values flipped).

5. Underreporting and overreporting p-value robustness

5.1. Reporting styles

Papers differ in how they report p-values. The APA style manual 6th and 7th editions state that p-values should be reported exactly (e.g., "$p = .04$") unless they are less than .001 (e.g., "$p < .001$"). However, many papers deviate from this. Papers may use a weaker lower-bound cutoff, such as using "$p < .01$" or possibly have a $p < .001$ resolution but use less-than signs throughout (e.g., reporting always reporting "$p < .05$", "$p < .01$", or "$p < .001$"). One reporting style can be problematic – namely, papers that write "p < .05" for every significant p-value even if the p-value is far below the significance threshold (e.g., t[50] = 3.0 corresponds to $p = .004$ but may be reported as "$p < .05$"). These papers may present strong results with fragile p-values, which is an issue for the present focus on examining percentages of significant p-values that are fragile. To investigate this issue, a taxonomy of p-value reporting styles was created, and the frequencies of different styles are listed in Table S1.

Style		**Quantity**		**Percentage fragile**		**Number with p-implied**	
Equal/less	Cutoff	Count	(%)	Raw p-values	P-implied	Count	(%)
All less	0.05	5549	2.3%	100.0%	50.9%	1614	29.1%
	0.01	4062	1.7%	0.0%	3.5%	1110	27.3%
	0.001	14591	6.1%	0.0%	0.8%	5819	39.9%
	0.0001	1130	0.5%	0.0%	1.0%	351	31.1%
All less mixed	0.05	260	0.1%	100.0%	82.8%	113	43.5%
	0.01	12942	5.4%	44.5%	38.2%	5340	41.3%
	0.001	44515	18.5%	26.6%	22.1%	22718	51.0%
	0.0001	6833	2.8%	22.6%	19.7%	3970	58.1%
Equal less	0.05	44	0.0%	100.0%	87.7%	23	52.3%
	0.01	3374	1.4%	44.3%	39.7%	1639	48.6%
	0.001	64420	26.8%	29.0%	25.3%	41507	64.4%
	0.0001	4484	1.9%	29.2%	27.1%	2071	46.2%
All equal	0.05	3447	1.4%	100.0%	96.4%	1236	35.8%
	0.01	9197	3.8%	63.8%	59.8%	3917	42.6%
	0.001	5613	2.3%	43.8%	42.9%	2601	46.3%
	0.0001	5658	2.4%	29.5%	28.0%	2510	44.4%
Eclectic		54247	22.6%	35.6%	28.8%	34595	63.8%

Table S1. Summary of p-value reporting styles. Five categories of significant p-value reporting styles were defined: *All less* represents papers that report every p-value as "p < x" where x is identical for every reported result. The Cutoff column represents the value of "x". Papers in between the listed levels (e.g., "p < .005" or "p < .02") were uncommon, and are listed here by rounding up to .0001, .001, .01, or .05. *All less mixed* papers always report "p < x" but the "x" may vary across results. The cutoff represents the lowest possible x; note that *All less mixed* .05 represents cases with, e.g., "p < .02" and "p < .05", where x = .02 was rounded up here in the table categorization per above. *Equal less* represents papers that report "p = y" for varying y unless "p < x", and the cutoff represents this x; APA style is generally described as this with a cutoff of .001. *All equal* papers always report "p = y" for varying y. *Eclectic* papers do not fall into any other category (e.g., reporting "p = .008" and "p < .01"). The Percentage Fragile columns represent the mean percentage of fragile p-values across the papers in a row, calculated with respect to the reported p-values themselves or in terms of p-values implied from neighboring test statistics. The rightmost columns indicate the number of papers in a row reported with any test statistics at all. The counts reflect the dataset pre-no-SNIP and pre-no-affiliation exclusion (total 240,355 papers). Overall, 67003 papers (45.1%) adhere to APA style ("p = …" unless "p < .001") or a stricter form of APA style where the lower bound is below .001; APA style corresponds to *All less* .001 or .0001, *Equal less* .001 or .0001, *All equal* .05, .01, .001, .0001. Categorization omitted p-values reported "p = .05" as some papers treat "p = .05" as significant whereas some treat it as insignificant (overall, just 0.7% of p-values in the dataset are "p = .05").

Fortunately, papers that exclusively report "p < .05" make up only 2.3% of papers in the dataset. These could be excluded entirely from the analysis. However, calculating the fragile percentage using p-values implied from test statistics reveals that papers that always report "p < .05" mostly put forth weak results (implied p-value fragile percentage = 51% in Table S1). Hence, simply excluding those 2.3% of papers would downplay fragile p-value levels. Instead, the present analyses defined papers that report every p-value as "p < .05" Hence, for these papers, their fragile p-value percentage was recomputed based on the p-values implied by nearby test statistics. For "p < .05" papers that did not report any statistics, their fragile p-value percentages were all set to 51%.

Aside from papers that report "p < .05" for every result, papers in general regularly report "p < .05" even when their reported test statistic suggests that the true p-value is p < .01. Specifically, going beyond "always p < .05" papers, examining all p-values with valid test statistics reveals that 9.2% of p-values recorded as fragile ($.01 \leq p < .05$) are actually robust per the implied p-value ($p_{implied} < .01$). These authors understated their findings' robustness. This type of underreporting is more common than erroneous overreporting: Among significant p-values reported as robust ("p < .01"), 1.3% of p-values reported are actually fragile per the implied p-value.[2] Using the overreporting and underreporting measures, a correction can be applied. Specifically, if the p-fragile-percentage rate is recorded as 35% (see main-text Figure 2A), when underreporting and overreporting are accounted for, this suggests that the true fragile p-value percentage is actually 32.7% (a 2.3% difference); note $.023 = .35 \times .092 + .65 \times .013$.[3]

The plot of fragile p-value percentages over time in the main text Figure 2A has 2.3% subtracted from each point. Yet, even with this correction, plotting the implied fragile p-values themselves shows lower levels, particularly in 2024 where there appears to be a gap as large as

[2] Note that these discrepancies are largely not due to rounding issues related to "p = .01". If such p-values are ignored, there remains a 6.6% (from 9.2%) rate of underreporting robustness.

[3] It is only by coincidence that the number 2.3% appears twice here as both the percentage of "p < .05" papers and the amount of correction necessary. Those calculations are independent.

4%. This difference may be explainable as a selection effect, in terms of which papers properly report test statistics and which results are reported alongside test statistics (e.g., usually no test statistics are reported with regressions).

Overall, the effects of papers misreporting robust results as "$p < .05$" appear to be fairly small, demonstrating the validity of the present focus on p-values. Nonetheless, given the minor ambiguity, the main text multilevel regression was also done while using the fragile implied p-value percentage. (Supplemental Materials 13).

5.2. One-tailed testing and misreported p-values

Along with papers varying in their overall reporting style, any papers may simply misreport individual p-values. The present subsection investigates this topic to clarify whether misreports could induce biases or shift interpretations. This issue can be examined by analyzing the relationship between reported p-values and the p-values implied by nearby test statistics. In general, severe misreporting is rare (see the $\rho = .972$ correlation of p-values x implied p-values in Supplemental Materials Section 7.2). Nonetheless, the focus here is on those uncommon instances of misreport.

The present analyses use a technique akin to the *statcheck* method, evaluating test-statistic/ p-value consistency while accounting for possible rounding errors. For example, $t[19] = 2.3$ would be consistent with p-values anywhere in the range of $t[19] = 2.25$ ($p = .036$) to $t[19] = 2.35$ ($p = .030$). Consistency in this way was tested for every p-value in the dataset that was reported alongside a test statistic. Only significant p-values with an equal sign were considered.

Of the results, 12.7% showed inconsistency between their reported p-value and implied p-value. For inconsistent results, Figure S1A displays the differences in the z-scores corresponding to the reported p-value and the implied p-value; negative values indicate that the reported p-value was underreported (e.g., "$p = .02$" while $p_{implied} = .01$) whereas positive values indicate that the p-value was overreported. One emerging pattern is the spike around 0.3. This spike reflects one-tailed testing (e.g., $p = .025$ corresponds to $z = 1.96$ while $p = .05$ corresponds to $z = 1.64$); calculation of implied p-values by contrast always assumed two-tailed testing. To identify cases of one-tailed testing, every inconsistent result was assessed for whether its implied p-value was 2x the reported procedure (i.e., re-doing the *statcheck* strategy but for one-tailed implied p-values). The inconsistent results identified used this strategy are shown in red in Figure S1B. One-tailed testing can also be found by examining language usage. Nuijten et al. (2016; *BRM*) attempted to delineate one-tailed tests by searching entire papers for "one-tailed", "one-sided", or "directional". Here, a similar strategy was used (searching more broadly for the terms "-tailed", " tailed", "-sided", " sided", or "directional"). This identified the results shown in blue in Figure S1C. Examining trends over time shows that rates of one-tailed testing have gradually decreased (Figure S1D). Note that one-tailed testing does not necessarily explain the gap between fragile p-value percentages and fragile implied p-value percentages reported in main text Figure 2. Many of those one-tailed results would not even be considered per their implied p-values.

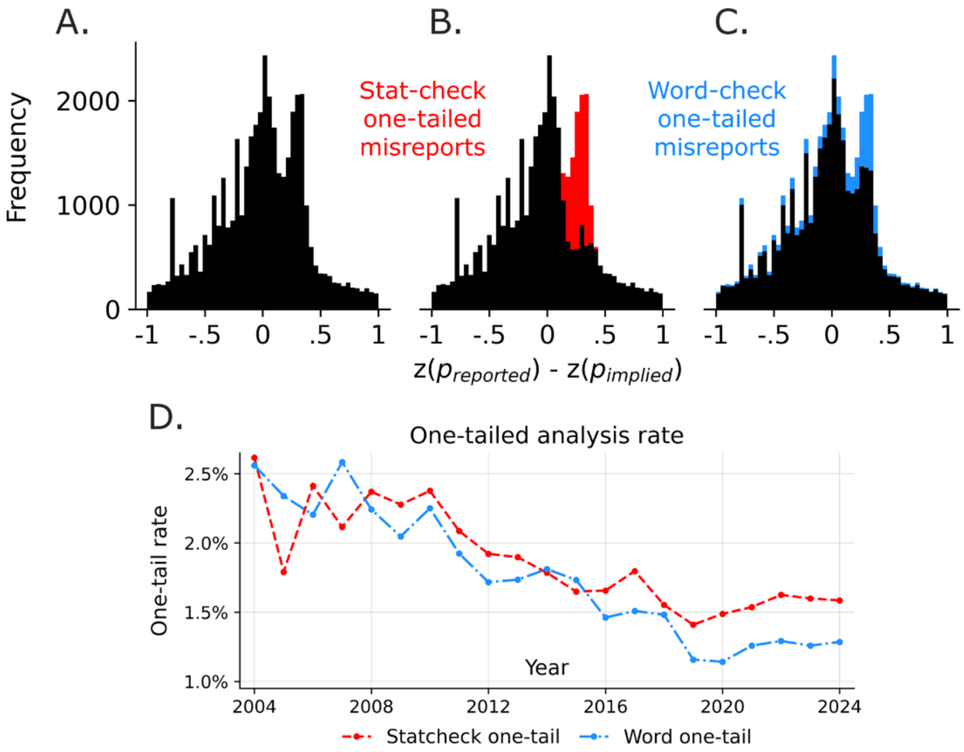


Figure S1. Frequencies of misreporting p-values and one-tailed statistical tests. A. The black area represents a histogram, wherein each entry is one case where a result's reported p-value deviates from the p-value implied by a nearby test statistic. This histogram notably shows a clear secondary peak around z = 0.35. **B.** The secondary peak is driven by the p-values from one-tailed significance tests. Here, the red indicates entries where the "misreported" p-value entry aligns perfectly with twice the statistic-implied p-value, which would be produced by a one-tailed test. **C.** Another way to identify one-tailed tests is by searching papers for the terms "-tailed", " tailed", "-sided", " sided", or "directional". Inconsistent results that are in papers containing one of these keywords are shown here in blue.

Moving beyond just one-tailed tests, *Figure* S2 quantifies more general temporal trends in misreporting (with the colored lines representing the trends while one-tailed inconsistencies are omitted). Interestingly, the rate and magnitude of misreporting seem to have gone down until 2015 and then increased in recent years (Figures 2A & 2B). Yet, authors do not appear to be misreporting statistics nowadays in a way that specifically favors them, as the rate of overreporting and underreporting occurs at roughly 50%/50% rates (Figure 2C). The reasons for this are unclear. One possibility is that this is that papers are now reporting statistics with more precision, creating more room for error; test statistics were reported with an average of ~1.9 decimal points around 2004 and this increased to ~2.05 decimal points today; p-values were reported with ~2.65 decimal points in 2004 to ~2.8 decimal points today. Potentially, there are selection effects emerging related to the types of papers that do or do not report test statistics. Fully unpacking this question is beyond the present scope. Most relevant for the present work, there do not appear to be stark differences across sub-fields in misreporting or one-tailed testing (Figures 2D & 2E), albeit with social psychology possibly showing ~2% lower misreporting in most years. The reason for this trend is unclear. Overall, examining misreports did not yield any evidence that would challenge the validity of the main text analyses or interpretations.

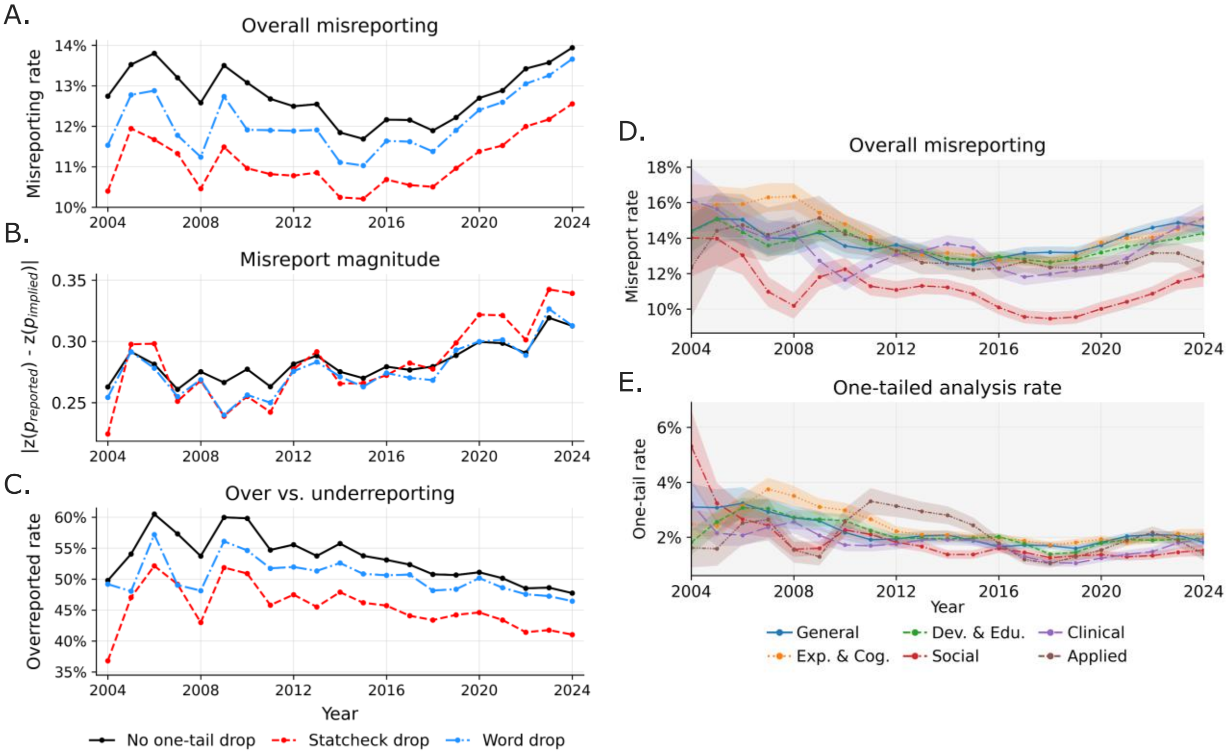


Figure S2. Temporal trends in misreporting p-values and one-tailed statistical tests. A. The lines represent the p-value/implied-p-value inconsistency rate each year. The red and blue lines correspond to the rates when one-tailed results are omitted (detected via one of the two ways noted in the text and the Figure S1 caption). One-tailed results would otherwise be considered misreported. **B.** This line represents the absolute difference in z-scores between p-values and implied p-values, averaged across all misreported cases. **C.** The percentage represents the rate at which a misreport produces a p-value that is stronger than the implied p-value. **D.** This corresponds to the black line from plot A but is now computed separately for each sub-field; shaded regions were computed by averaging the misreporting rate by paper and then taking the standard error (shaded regions represent ±1 standard error). **E.** This corresponds to the one-tailed rate shown in Figure S1D but computed separately for each field.

6. Insignificant results

The main text analyses focus on statistically significant results, but it is also interesting to consider insignificant results and how the percentage of insignificant findings may have shifted over time. Potentially, rises in registered reports and preregistrations have encouraged more insignificant findings to be reported. This topic is investigated here.

Among the main dataset – i.e., papers reporting at least two significant p-values – the average paper now reported a higher percentage of its results as insignificant, compared to two decades ago (Figure S3). This may reflect papers now providing more elaborate statistical descriptions of their data, including insignificant associations. However, there was an interesting downward trend from 2019 to 2022 with a major downtick between 2021 and 2022. The reasons for this are unclear, although potentially the large downtick is related to the COVID-19 pandemic in some way.

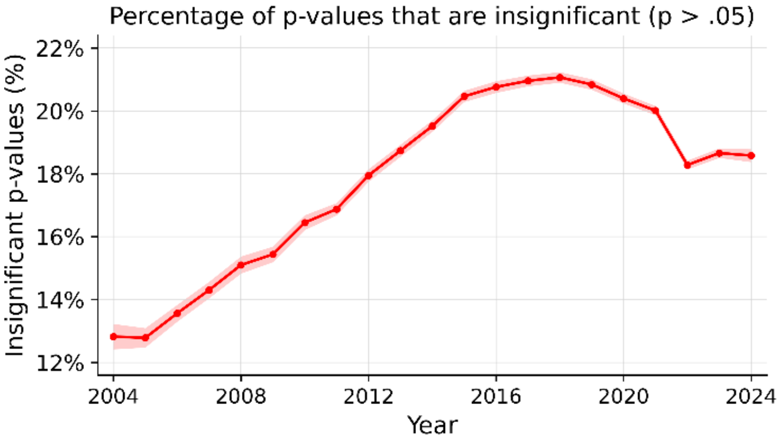


Figure S3. Rates of insignificant p-values. This line reflects the percentage of reported p-values that are insignificant (*p* > .05) within the main dataset representing papers otherwise reporting significant findings. The shaded region represents ±1 standard error.

Further descriptive statistics were calculated, now examining the percentage of papers that entirely report insignificant results. These analyses begin with a version of the dataset that contains all papers with at least one p-value (269k papers). Of these, 8,101 papers (3.0%) only reported insignificant results, although roughly half of those corresponded to papers reporting just a single finding. On average papers with entirely insignificant results report a mean of 2.54 p-values (SD = 2.75). Among papers reporting at least two p-values, 4,212 (1.6%) reported only insignificant results. Of note, the former pool of 8,101 papers may include non-empirical reports that mention "*p* > .05" as part of their prose without reporting specific results. To avoid these, the analyses proceed by filtering the dataset to only papers with at least two p-values (253k papers), which also parallels the design of the main text dataset.

There seems to be an increasing trend in the percentage of papers reporting entirely insignificant results. Figure S4A first illustrates this, examining the percentage of the 253k papers that exclusively report insignificant results. From 2007-2024 there is a gradual slow upward trajectory in the percentage of papers reporting entirely insignificant results. To be clear, the earliest dates (2004-2006) run counter to this idea. The high points in those early years may be noise related to the lower numbers of papers, although it is also possible that this is a selection effect, as in those years, papers were only available from journals that shifted to web versions of articles at that time. Regardless, when more complete sets of journals are investigated, the trajectory is upwards.

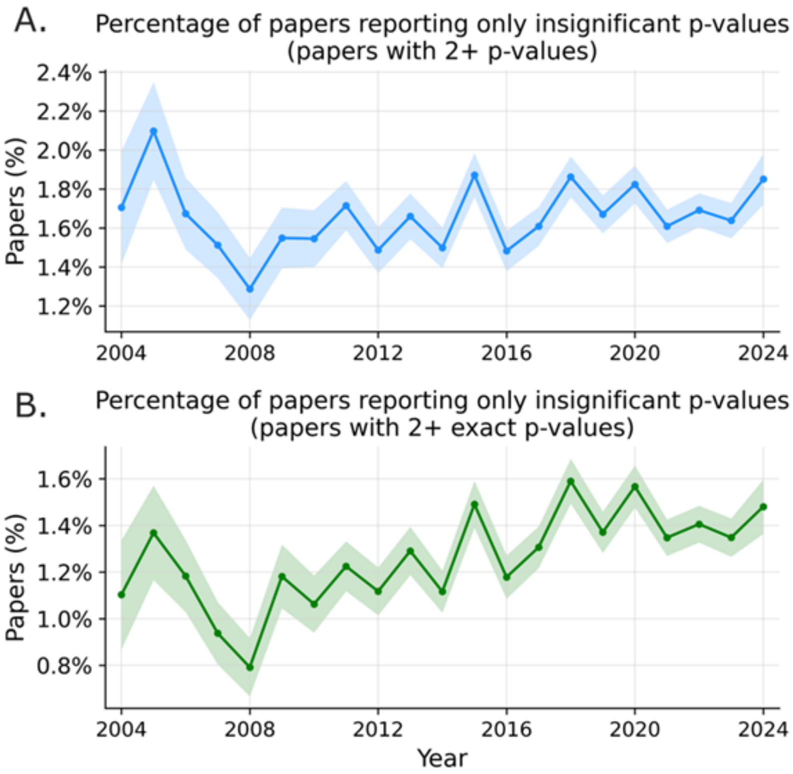


Figure S4. Percentage of papers reporting entirely insignificant results. **A.** This line shows the percentage of papers for which all of the reported p-values are insignificant (*p* > .05) (among all papers reporting at least two p-values in general). **B.** This second line is similar to A but now only tracks papers that report at least two exact ("=") p-values.

To further investigate papers reporting entirely insignificant findings, and to further ward off papers writing "*p* > .05" as prose without focusing on specific findings, an additional plot was made: Figure S4B shows the percentage of papers reporting entirely insignificant findings that reported at least two exact ("=") insignificant p-values (e.g., reported "*p* = .24"). This plot shows a clearer and larger upward trend from 2007-2024.

7. Validation

7.1. Manual validation

Manual inspection assessed whether the number of p-values extracted using the regular expression matches the actual number of p-values reported in a paper while properly ignoring non-Results sections, tables, and captions. Forty papers spanning every Publisher were inspected (DOIs listed in the linked OSF repository). Every paper's number of p-values aligned perfectly with the extracted p-values.

7.2 Cross-checking p-values and test statistics

Second, a wider validation was performed on the accuracy of the exact p-values extracted. As described above, for each p-value, an attempt was made to extract its associated test statistic. This was successful for 1,398,077 p-values (35.8%) and of those 671,119 are exact ("=") p-values (e.g., ignoring "*p* < .001"). Based on each test statistic, an implied p-value can be computed. Measuring the Spearman correlation between the extracted p-values and the implied p-values yields a tight correlation (ρ = .972). Measuring the Pearson correlation between the z-scores implied by the p-values and those implied by the test statistics also shows a tight correlation (*r* = .974). Note that the small bit of unexplained variance also does not necessarily imply even a small error in the present parsing, as misreporting and rounding will lower the correlation from 1.0. Overall, this demonstrates the accuracy of the present regular expressions.

7.3. Predicting replicability

The third form of validation assessed the premise that p-values shed major light on replicability. For this, two additional datasets detailing successful and unsuccessful replications were acquired: (1) the data by Youyou, Yang, Uzzi (2023) (*PNAS*) (https://osf.io/f5sxn/) and (2) the replication database by the Framework for Open and Reproducible Research Training Replication Database (https://osf.io/9r62x/); these two datasets were included in the present research's repository. Together, the downloaded data describe 498 replication attempts that yielded successful replications or decisive unsuccessful replications. Of these, 113 papers overlapped with the present dataset (49 successful replications and 64 conclusive failures); the considerable drop is partly because many of the replication attempts focused on older papers not gathered here. Nonetheless, as will be shown, the differences in the p-values between replicable and non-replicable papers were large, and 113 papers proved sufficient for a decisive demonstration.

Using the replication data, two analyses were done. First, a two-sample t-test was used to compare the proportion of fragile p-values among the original studies that replicated to the proportion from studies that did not replicate. Second, a basic machine learning analysis was also performed, attempting to predict replication outcomes based on a paper's p-values. For this, a logistic regression classifier was tested attempting to predict replicability from the proportion of a paper's p-values that were fragile. For classification, chance-level accuracy was set to be 50% by randomly selecting the dataset to 49 successful and 49 unsuccessful cases. Then, leave-two-out cross-validation was performed, while always leaving out one successful and one unsuccessful case to maintain the stratification during training. Cross-validation accuracy was averaged across 1,000 runs using random 49/49 subsets and leaving out random pairs.

As expected, studies that were later replicated showed distinctly fewer fragile p-values (M = .32 [SE = .03]) than studies that did not replicate (M = .51 [SE = .03]) (*t*[112] = -3.89, *p* = .0002, *d* = -0.74; Figure S5). Additionally, a logistic regression classifier was tested, which attempted to predict replication outcomes based on a paper's fragile p-value percentage. It achieved 63.7% cross-validated accuracy, relative to chance-level accuracy set to 50% via stratifying the data. Similar results emerged when the classifier instead used the fragile percentage calculated using p-values implied from test statistics, which yielded a classifier with 63.4% accuracy (Supplemental Materials 13.1). These accuracy gains relative to chance are notably higher than what a prior study achieved while attempting to predict replicability using an article Abstract's language usage (Yang et al., 2020) (*PNAS*), which speaks to the power of this p-value approach.

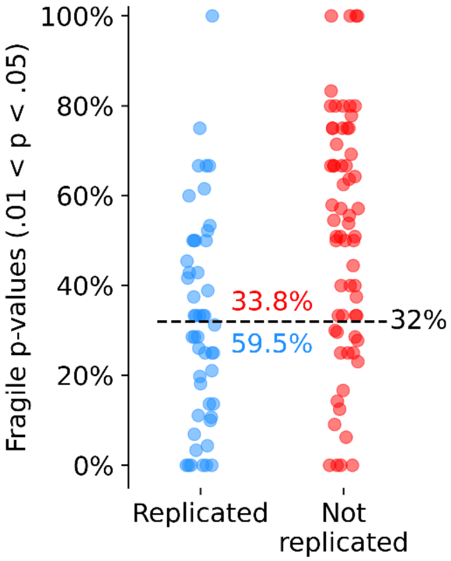


Figure S5. Replicable papers generally have fewer fragile p-values. Each dot represents the fragile p-value percentage for one paper and whether the paper's results were replicated or not. The dashed line at 32% represents the cutoff for whether a paper has over or under a 50% chance to replicate given its fragile p-value percentage, as identified using a logistic regression. The red "33.8%" indicates that 33.8% of papers above the dashed line replicated. The green "59.5%" indicates that 59.5% of papers below the dashed line replicated.

8. Incentives variables

8.1. Journal reputation

Journal reputation was defined as SNIP scores, which were gathered using the Scopus API. SNIP scores represent a journal's impact factor (citations over the last three years divided by the number of papers over the last three years) normalized with respect to the typical number of citations that papers in the journal's field tend to receive. SNIP varies for each journal by year. In years when no SNIP score was available for a given paper (e.g., when new journals were founded), the soonest subsequent score was assigned. For instance, if a journal's 2011 and 2012 scores were not available, but a 2013 score was, then papers that the journal published in 2011 and 2012 were assigned the 2013 score. This was done for just 5.0% of cases and hence was unlikely to introduce considerable bias.

8.2. Citations

Citations counts were provided for each paper by Lens.org as of August 3rd, 2024. Citations per year were calculated by dividing the count by 2024 minus publication year. Because citations per year are strongly right-skewed (main text Figure 1E), the values were log-transformed, log(x + 1). Then, to further normalize the citations per year data, the data were z-standardized on a year-by-year basis. For instance, to normalize papers from 2013, the 2013 log-citation-count mean was subtracted from each paper, and the resulting quantity was divided by the standard deviation of the 2013 papers' log-citation-counts. This procedure precisely accounts for the (strong) non-linear relationship between citations and publication year.

8.3. University ranking

University affiliations were assigned to papers by searching each paper's text for the names of universities listed in the Times Higher Education 2024 World University Rankings (e.g., searching for the string "Duke University"). Only the top 1000 ranked schools, sorted based on their 2024 research score, were considered. The search only covered the first half of each paper's text to avoid university names that may appear in its reference section (e.g., "Harvard University Press"). Paper text and university names were both stripped to only letters in the Modern English alphabet (e.g., diacritics were removed). Overall, for the final dataset, 945 universities from 69 countries were successfully linked to at least one paper (942 among papers with a SNIP score). Preliminary tests, which also covered documents without a Results section, found 978 universities, meaning that many universities not identified simply reflect their absence from the scholarly record and not a parsing limitation. Most papers were linked to multiple universities (M = 3.74 universities, SD = 3.99, median = 3). Among the potentially multiple universities, the one that appeared most often in the text was selected and assigned as the paper's school; ties were broken randomly. Selecting a single university for each paper was useful for the analyses below, which could then use multilevel regressions that permit modeling the one university as a grouping factor. Confirmatory analyses using different approaches to defining one score – either the maximum, minimum, median, or mean score of a paper's universities – yielded similar multilevel regression findings (Supplemental Materials 9.1).

Based on a paper's assigned university, a ranking score variable was assigned to it based on the university's research score (a continuous measure). Times Higher Education states that this score was calculated based on surveys of each school's reputation (60% weight) along with measures of research spending per staff (20% weight) and publications per staff (20% weight). All papers were assigned their school's 2024 research score regardless of the paper's publication year. This was done because, rather than a year-specific score, older rankings (e.g., 2011) included as few as just 200 schools. Additionally, the correlation of a school's rank across years is extremely strong. Supplemental Materials 9.2 discusses this and shows that the multilevel regression below produces similar results regardless of the strategy used.

9. Alternative definitions for papers' university ranking score

9.1. Other ways to select one school

The main text defined a paper's university as the one whose name appeared most often in the paper's text (i.e., the mode). Here, alternative approaches to assigning one university to papers were examined and submitted to the regression predicting fragile p-value percentages from university rankings (see Supplemental Materials 10). For the most part, multilevel linear regression retained a significant effect of higher ranking predicting more fragile p-values: This was for taking the maximum-ranked university ($\beta = .021$, $p < .0001$), the median-ranked university ($\beta = .017$, $p = .0001$), or the mean of the university ranks ($\beta = .024$, $p < .0001$). Taking the minimum-ranked school yielded an effect that fell short of the $\alpha = .001$ threshold ($\beta = .013$, $p = .002$). However, given that this is just one of the four alternative strategies tried, this is not taken to raise questions about the main-text conclusions. Note that the mean analysis required omitting the university-level and country-level random intercepts, given that the individual papers would be associated with multiple schools.

9.2. Yearly rankings

Times Higher Education has produced university rankings each year since 2011, although the number of universities ranked has increased over time. In 2011, they ranked 200 schools. From 2012-2015, they ranked 400-402 schools. In 2016, they ranked 800 schools. In 2017, they ranked 981 schools. From 2018 and after, they ranked over 1000 schools. Schools' ranks strongly correlate over time. Figure S6 below shows the Pearson correlation between schools' research scores in a given pair of years. Given this strong correlation across years, the main text analysis was done with respect to just a university's 2024 ranking; using a yearly score also limits the schools that could be analyzed each year.

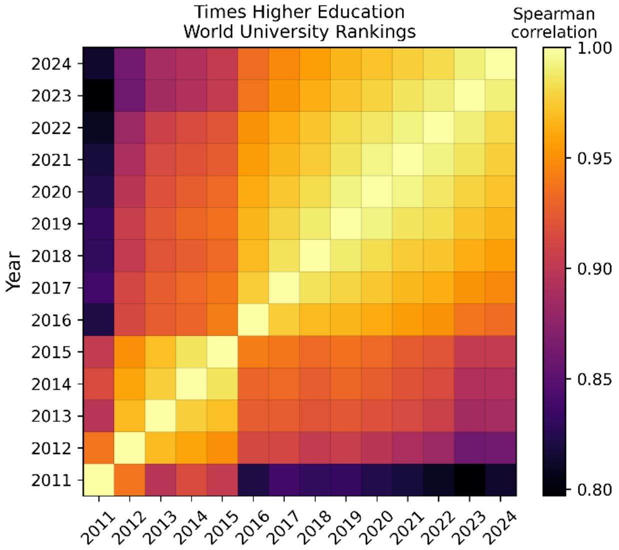


Figure S6. Strong temporal correlation in university ranks. The colors reflect the Pearson correlation between a given school's research score in a given pair of years. Schools ranked one year but not in the other are omitted from a given cell's correlation.

Nonetheless, the main text multilevel analysis of university ranking effects can also be performed while assigning papers the Times Higher Education research score of their universities at the time of the papers' publication years. This analysis is done using the same dataset used for the main text analyses, restricting the analysis to just the schools ranked in the top 1000 in 2024, regardless of their ranking in a paper's publication year. For papers published in a year before their university entered the rankings, these were assigned the most recent available year. For example, if a paper was published in 2012, but its university was only ranked in 2016, then the 2012 paper would be assigned the 2016 ranking.

The main text regression is premised on schools having a constant ranking, given that it defined a random intercept at the university level. Now, because a university's ranking can vary over time, the regression was adapted to include a random slope for the ranking effect. Hence, the effect of ranking will reflect both differences across universities and changes within a university over time. The R-style equation is as follows:

$$\textit{fragile\_p\_percentage} \sim 1 + \textit{ranking\_yearly} \times \textit{year} + \\ (1 \mid \textit{journal}) + (1 + \textit{ranking\_yearly} \mid \textit{school}) + (1 \mid \textit{country})$$

The link between p-values and ranking has weakened, falling below the α = .001 threshold described in the main text (*β* = .010, *p* = .008). Potentially, this weakening is because using research scores from different years for different papers creates a source of noise. Related to this, the p-value "*p* = .008" may seem concerning given the size of the dataset, although it's important to recognize that the degrees of freedom for this effect are just 200 (per Satterthwaite's method). In addition, it should be noted that the country-level intercept can limit the identification of between-university differences. If the country-level random intercept is omitted, then the effect of university ranking returns to strength (*β* = .017, *p* < .0001).

10. Multilevel regression structure

A series of multilevel regressions were performed linking papers' fragile p-value percentages with their SNIP, citation, and ranking scores. SNIP:

$$\begin{gathered} \textit{SNIP} \sim 1 + \textit{fragile p percentage} \times \textit{year} + \\ (1 + \textit{fragile p percentage} \mid \textit{journal}) + (1 + \textit{fragile p percentage} \mid \textit{school}) \end{gathered}$$

The use of random slopes ensures that associations with SNIP generalize across journals and schools. It was possible to include random slopes for the *journal* group because a journal's SNIP may vary over time. Citations:

$$\begin{gathered} \textit{citations} \sim 1 + \textit{fragile p percentage} \times \textit{year} + \textit{SNIP} \times \textit{year} + \\ (1 + \textit{fragile p percentage} \mid \textit{journal}) + (1 + \textit{fragile p percentage} \mid \textit{school}) \end{gathered}$$

Note that this regression assesses whether any effects of fragile p-values on citations go beyond any SNIP/journal effect – i.e., testing whether increases or decreases in citations cannot be explained by the journal in which a paper is published. A regression that does not control for *SNIP* and the *SNIP* × *year* interaction yields significant effects (*p* < .0001) in the same direction.

Finally, the university ranking analysis was as follows:

$$\begin{gathered} \textit{fragile p percentage} \sim 1 + \textit{ranking score} \times \textit{year} + \\ (1 \mid \textit{journal}) + (1 \mid \textit{school}) + (1 \mid \textit{country}) \end{gathered}$$

Note that the *ranking score* regression is somewhat different from the *SNIP* and *citations* regressions. Here, *ranking score* is the predictor rather than the dependent variable. The regression cannot be run with *ranking score* as the dependent variable because this measure is static for each school, meaning that the *school* random intercepts would entirely predict it. This static nature also means that the analysis is somewhat similar to a between-school correlation. The *country* random intercept was added here to ensure that rankings effects cannot be due to schools in one country. However, it was not feasible to add a *ranking score* random slope to *country* because this would cause a drastic reduction in the degrees of freedom when assessing statistical significance (just 6.8 degrees of freedom, per Satterthwaite's method). An alternative arrangement of the analysis, where a single large regression is performed, predicting fragile p-values based on all three other predictors, yields similar results to those reported, albeit with an insignificant *Citations* x *Year* interaction.

Because the regression covers interaction effects, the predictors were mean-centered prior to analysis; see Iacobucci et al. (2016) for a discussion on why this reduces collinearity and leads to coefficients that many see as more interpretable. For the year predictor, the mean was 2016.5 (not the exact middle of 2004-2024 because more papers were published in later years).

11. Considering possible sources of bias

11.1. Summary

First, instead of examining the percentage of a paper's p-values that are fragile, analyses were done considering the overall number of fragile p-values in a paper. These tests suggested that the percentage is a more valid measure, as the overall number is heavily impacted by the total quantity of p-values in a paper (i.e., extensive results inevitably lead to many fragile p-values).

Second, tests were performed considering the possibility that results may be confounded by researchers' ages. The tests showed that older researchers tend to be employed at higher-end institutions but produced no evidence that age would confound the analysis of fragile p-values.

Third, the present analyses do not accommodate papers that report significance entirely through figures or tables. Among papers containing a Results section, at least one p-value was extracted for 72% of them. In principle, the remaining 28% could induce a selection bias. Although studying these papers is limited, the existence of bias was assessed. Analyses were done that extrapolated this 28% of papers' fragile p-value percentages by assuming that papers from a given journal in a given publication year will tend to be held to similar standards of p-value strength. Although the extrapolated data cannot be used to assess the effects of SNIP or citations, it is informative for the university rankings effect. Ultimately, the dataset with extrapolation yielded similar results on university rank.

Fourth, by focusing on p-values, the present research targets frequentist papers. Yet, the use of Bayesian and/or machine learning methods may also vary with respect to the variables studied. The results notably showed that top universities were more likely to publish empirical papers containing keywords related to Bayesian statistics or machine learning. However, these papers were also found to just be a small minority of psychological literature (altogether roughly just 5% of empirical papers) and are unlikely to bias the multilevel regression conclusions.

11.2. Examining the overall number of fragile p-values

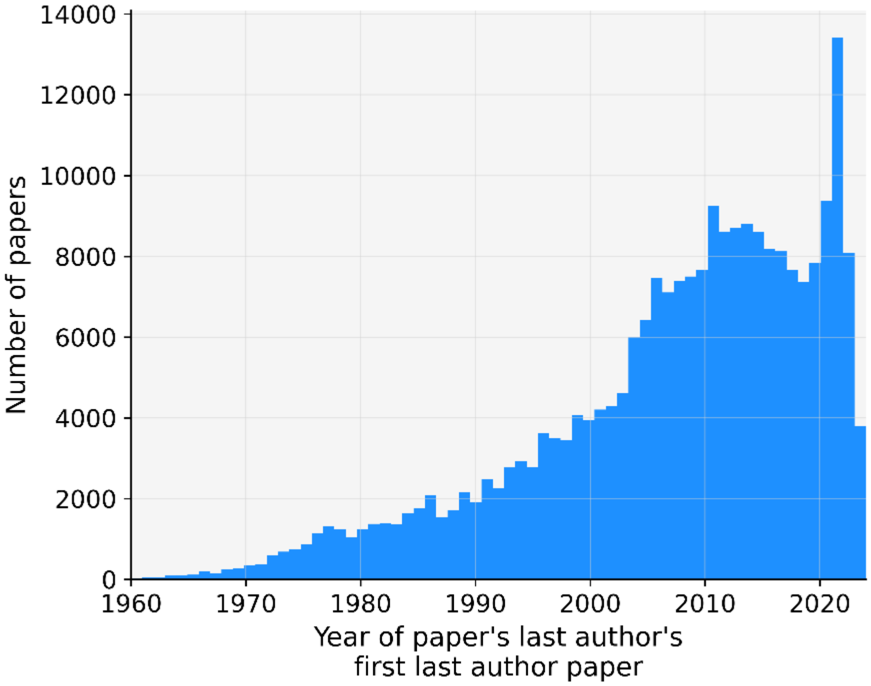
The main text analyses focused on the percentage of a paper's p-values that are fragile rather than the *number* of a paper's p-values that are fragile. This choice was made because the latter depends primarily on the overall number of p-values in a report. Specifically, the number of fragile p-values correlates more strongly with the overall number of p-values (r = .66) than it does with the percentage of a paper's p-values that are fragile (r = .47). That is, most papers that simply report extensive results will, by chance, have many fragile p-values. Additionally, because the number of p-values in a paper is strongly right-skewed (see main text Figure 1C), this produces outliers that may have outsized effects on the analysis. Hence, the main text analysis focused on the percentage of p-values that are fragile rather than the overall quantity.

Nonetheless, reproducing the main text results while instead predicting the overall number of fragile p-values would prove to show the robustness of the conclusions. Hence, the main multilevel regression was performed again as so, albeit now while excluding 1.8% of papers that reported an exceedingly high number of fragile p-values to account for the rightward skew (dropping z-standardized counts over 3.0). The same regressions were used as described in Supplemental Materials 10, but with fragile p-value percentages being changed to fragile p-value counts.

The fragile p-values x Year interaction effect on SNIP reported in the main text, wherein papers with fewer fragile p-values are published in higher journals, was reproduced here ($\beta = -.022$, $p < .0001$). Likewise, the links to university rankings were reproduced, wherein highly ranked institutions tend to have more fragile p-values ($\beta = .015$, $p = .0001$). Interestingly, the link between fragile p-values and citations shifted to falling just under the $\alpha = .001$ threshold mentioned in the main text ($\beta = -.002$, $p = .002$). However, there remains a significant fragile p-values x Year interaction effect on citations ($\beta = -.029$, $p < .0001$). Regressing all three incentive variables to predict the overall number of significant p-values in a paper (fragile or robust) reveals that papers with more citations have more p-values overall ($\beta = .010$, $p < .0001$). This latter effect, notably, is stronger than the effect on the fragile p-value count, suggesting that this preference for many results is causing it. Hence, this analysis overall does not cast doubt on the validity of the main text findings.

11.3. Ruling out the possibility of age-related confounds

Analyses considered the possibility that some effects on fragile p-values may be confounded by researchers' ages. Hence, the "academic age" of a paper's senior author was calculated and defined as the year in which the author first published a paper in the senior position (Figure S7). Identifying this first year used the full pool of 1.7M records downloaded from Lens.org, including the data on papers published by the APA and Taylor & Francis, even though these publishers could not be used in the main analyses requiring full-text papers.[4]



[4] Note that multiple authors may share the same name, and this would influence the academic age calculation. However, basic checks suggest that this is a small effect. Among the 23,283 last-position authors on at least 10 papers in the dataset, there are 4,886 unique first names/initials and 14,836 unique last names. Beyond first and last names, 55% of names contain at least three elements (e.g., contain a middle initial). It is difficult to estimate the exact number of overlapping names, given that first and last names are not independent, although the number of possible combinations is extremely large (72.5 million first-last name combinations).

Figure S7. Histogram of identified author ages. There are some interesting patterns. There is an uptick in 2004. Per manual inspections, this may be because 2004 was when many journals began publishing web versions of articles, which may have slightly influenced the Lens.org database construction. There is a local maximum near 2010. Per manual inspections, this may be because 2010 is when publishing web versions became standard across every publisher, which likewise may have impacted the Lens.org database constructions. Finally, there is a spike in 2021. This may be related to the COVID-19 pandemic slowing publication rates in 2020. These patterns point to limitations in this way of modeling age, although overall, the patterns are unsurprising (i.e., a clear upward trend over time, consistent with the increase in the number of papers published over time).

Regressing age on year, SNIP, citations, and ranking scores using the random intercepts above revealed that older researchers do not publish in higher/lower journals (β = .005, p = .51) nor receive more/fewer citations (β = -.004, p = .051). However, older researchers are employed significantly more top-ranked universities (β = .05, p < .001); the negative Ranking Score x Year interaction also shows that prestigious university's preferences for older researchers have decreased slightly over time (β = -.007, p = .001). Yet, regressing fragile p-values on age and year did not yield any significant effect of age (β = .002, p = .53), meaning that any effect of age is null or small. Hence, age is not a problematic confound.

11.4. Papers with Results sections but no p-values

Analyses were done investigating the nature of papers for which a Results section was extracted but not even a single p-value was identified. This analysis uses 221,471 articles containing a Results section along with an identified affiliation and SNIP score (i.e., of the 372,633 empirical papers, dropping ones without a SNIP score or identified school). Among these articles, 166,191 contained at least one p-value (significant or insignificant). This suggests, for the most part, that the present p-value-focused approach taps into the robustness of most empirical psychology papers. Nonetheless, the p-valueless papers and how they may differ from papers with p-values deserve consideration. A multilevel regression was tested which mirrors the one used for the main text analysis but is now a logistic regression predicting whether a paper contains at least one p-value or not. Its R-style equation is as follows:

$$\begin{gathered} has\_p \sim 1 + SNIP \times year + citations \times year + ranking\ score \times year + \\ (1 + SNIP \mid journal) + (1 \mid school) + (1 \mid country) \end{gathered}$$

Top journals have become more likely to publish papers containing at least one p-value in recent years (SNIP x Year interaction: β = .03, p < .0001; SNIP main effect: β = .04, p = .05). Papers with at least one p-value also tend to receive fewer citations (β = -.05, p < .0001), although the effect over time is unclear (Citations x Year: β = -.015, p = .01, note the weak p-value, which is below the α = .001 threshold in the main text). Whether papers had at least one p-value was unrelated to the ranking of its authors' universities and that effect did not interact with time (ps > .33).

Although the link to university ranking was insignificant, further tests were done to consider the possibility that the papers without p-values are relevant to the main text finding that higher ranking scores are linked to less robust p-values. Specifically, the multilevel regression predicting the fragile p-value percentage from university ranking (Supplemental Materials 10) was performed again after extrapolating the fragile p-value percentage for papers without p-values. The extrapolation essentially assumes that all papers from the same journal and same year have similar robustness (e.g., strict journals will largely have strict standards across the board). Hence, the fragile p-value percentage for papers with missing p-values was defined as the mean fragile p-value percentage of a given journal's papers in the paper's publication year. Using this dataset where some p-fragile percentages are extrapolated, the regression still yields a effect, whereby higher-ranking universities tend to report a higher percentage of fragile p-values in their papers (β = .015, p < .0001).

11.5. Shifts away from null-hypothesis significance testing

This analysis considers papers that employed Bayesian and/or machine learning methods. The analysis was done using 221,471 articles containing a Results section along with a SNIP score and an identified university (i.e., including many papers without a single p-value). Each paper's Results section was searched for the presence of at one least Bayesian or machine learning keyword. The keywords are listed below and were generated manually and via prompting a large language model (Claude.ai) with "Please produce words you would expect to see in a Psychology paper using [Bayesian/machine learning] methods." An initial list of words was pruned to remove non-specific terms – e.g., the word "prior" may also be used to simply refer to a sequence of events, or "model fit" is also regularly used in frequentist papers. This pruning was done by manually inspecting papers to see where a word was being regularly triggered outside of a Bayesian or machine learning context.

This strategy generated the following Bayesian words (search was case-insensitive):

> Bayesian, information criterion, bayes, log-likelihood, credible interval, Markov chain Monte Carlo, Gibbs sampling, Metropolis-Hastings algorithm, hierarchical modeling, conjugate prior, information criterion, beta distribution, Dirichlet distribution, Gaussian distribution, Bayes factor, marginal likelihood, posterior distribution, No-U-Turn Sampler, Hamiltonian Monte Carlo, gaussian process, Kullback-Leibler, KL divergence, Jeffrey's prior, maximum a posteriori, variational inference, informative prior, posterior likelihood, convergence diagnostics, posterior odds, prior odds, posterior probability, prior probability

The strategy generated the following Machine Learning words (again, case-insensitive):

> machine learning, deep learning, neural network, artificial intelligence, support vector machine, random forest, gradient boosting, xgboost, k-means clustering, k-nearest neighbors, principal component analysis, natural language processing, convolutional neural network, recurrent neural network, long short-term memory, gated recurrent unit, transformer network, autoencoder, generative adversarial network, unsupervised learning, supervised learning, semi-supervised learning, transfer learning, ensemble learning, hyperparameter tuning, confusion matrix, distributed stochastic neighbor embedding, t-SNE, word embedding, ensemble methods, training set, test set, validation set, feature selection, f1 score, testing set, train set

For each paper, dummy variables were defined, *baye_paper* and *ML_paper*, representing whether the paper contained at least one Bayesian or machine learning word, respectively. Overall, 3.8% of papers with a Results section contained at least one Bayesian word, and 2.0% contained at least one machine-learning word.[5] To assess how such papers changed in frequency

[5] For reference, producing frequentist keywords and evaluating their frequency yielded at least one hit for 52.1% of papers with a Results section (see released repository code for those keywords).

over time and how they are linked to the incentive predictors, multilevel logistic regressions were tested, mirroring the multilevel linear regressions from the main text analyses. The R-style equations are as follows:

$$\textit{baye\_paper} \sim 1 + \textit{SNIP} \times \textit{year} + \textit{citations} \times \textit{year} + \textit{ranking score} \times \textit{year} + \\ (1 + \textit{SNIP} \mid \textit{journal}) + (1 \mid \textit{school}) + (1 \mid \textit{country})$$

$$\textit{ML\_paper} \sim 1 + \textit{SNIP} \times \textit{year} + \textit{citations} \times \textit{year} + \textit{ranking score} \times \textit{year} + \\ (1 + \textit{SNIP} \mid \textit{journal}) + (1 \mid \textit{school}) + (1 \mid \textit{country})$$

Bayesian papers became more frequent over time (standardized $\beta = .43$, $p < .0001$). Also, over time, Bayesian papers started to be published in higher reputation journals (SNIP x Year interaction: $\beta = .06$, $p = .0001$; insignificant SNIP main effect: $\beta = .04$, $p = .33$). Bayesian papers may receive fewer citations ($\beta = -.03$, $p = .01$); note that this weak effect does not cross the $\alpha = .001$ threshold mentioned in the main text, although for these analyses, statistical power is much weaker given that the number of Bayesian papers is fairly small. Bayesian papers may also be more likely to be published by authors from high-ranked universities ($\beta = .06$, $p = .004$), although this result again does not cross the $\alpha = .001$ threshold. The interaction effects between a paper's publication year and citations or university ranking were insignificant (*p*s > .04).

Regarding machine learning papers, these have also become more frequent over time ($\beta = .17$, $p < .0001$) and have begun to be published in higher journals (SNIP x Year interaction: $\beta = .12$, $p < .0001$; insignificant SNIP main effect: $\beta = .05$, $p = .06$). Machine learning papers receive more citations ($\beta = .12$, $p < .0001$). There was no significant effect of ranking score ($\beta = .03$, $p = .13$), although note that this is after top-university authors being more likely to publish in high journals and receive many citations is accounted for. The interaction effects between a paper's publication year and citations or ranking were insignificant (*p*s > .08).

Figure S8 plots the relationship between a university's ranking and the percentage of its papers that include Bayesian or machine learning keywords. For this, the journal and citation effects are no longer accounted for, explaining why positive trends appear in the figure even though the prior multilevel regression yielded an insignificant link between machine-learning papers and ranking.

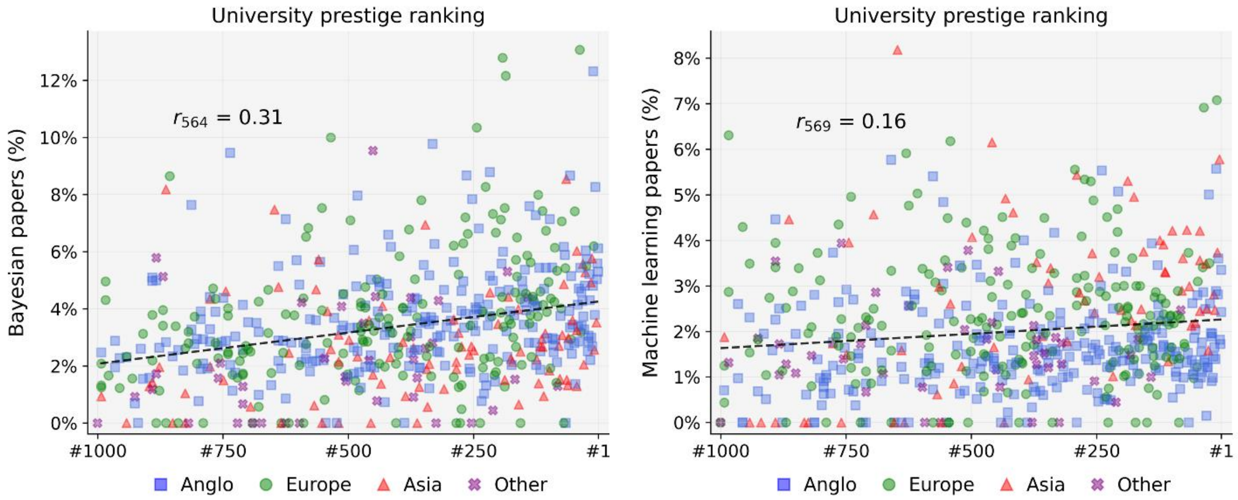


Figure S8. Linking university ranking and a tendency to publish papers with Bayesian statistics or machine learning. Reported coefficients reflect Spearman correlations.

Given that top universities are particularly likely to publish Bayesian and machine-learning papers, it is worth considering whether this is relevant to the association between fragile p-values and ranking described in the main text. To investigate this possibility, the ranking two new variables were defined at the university level: *school_M_baye* and *school_M_ML*. These correspond to the percentage of papers containing a Results section from a given university that included a Bayesian or machine-learning term, respectively; all papers from a given university are assigned the same *prop_Bayesian* value and the same *prop_machine_learning* value. Then, the multilevel regression predicting p-values from university rankings (Supplemental Materials 10) was updated to incorporate these new variables:

$$\begin{gathered} \textit{p\_fragile\_percentage} \sim 1 + \textit{ranking score} \times \textit{year} + \\ \textit{has\_baye} + \textit{ML\_paper} + \textit{school\_M\_baye} + \textit{school\_M\_ML} + \\ (1 \mid \textit{journal}) + (1 \mid \textit{school}) + (1 \mid \textit{country}) \end{gathered}$$

In this context, the *baye_paper* and *ML_paper* variables may be 1 if a paper reporting p-values also includes a Bayesian or machine learning term in their Results. The regression found that these types of papers that mix p-values with Bayesian/machine-learning methods tend to have fewer fragile p-values (*baye_paper*: $\beta = -.13$, $p < .0001$, *ML_paper*: $\beta = -.20$, $p < .0001$). However, *school_M_baye* and *school_M_ML* were not significant predictors (*p*s > .44), meaning that university-level shifts toward these methods have little impact on papers' p-values. Accordingly, the effect of rankings on fragile p-values remained significant ($\beta = .017$, $p = .0003$). Indeed, the coefficient hardly changed (original: $\beta = .016$), suggesting that even if the employed keywords do not entirely capture papers using Bayesian or machine-learning techniques, then a more refined approach would nonetheless yield similar results. These small effects may be because, overall, Bayesian and machine-learning papers remain a fairly small minority of the psychological literature, and university-level differences remain nominally slight (see Figure S8 and note that the median ranking across all papers is #192; it is not #500.5 because higher ranked universities produce more papers).

12. Further details on the text extraction

12.1. Sentence extraction

Text analysis investigated the sentences preceding or surrounding reported significant p-values. For example, a paper's Results section may contain the following three sentences:

"The conditions yielded different ratings (F[3, 50] = 5, p < .01). The first condition's ratings were above-baseline (t = 2.1, p = .04), which was interesting. This study is not real."

The employed scripts attempted to extract one sentence for each of the two p-values reported. For the first p-value, the extracted sentence would be: "*The conditions yielded different ratings*". For the second p-value, the extracted sentence would be: "*The first condition's ratings were abovebaseline which was interesting*" (non-letter characters, such as hyphens, were removed; "*abovebaseline*" is not a typo). Sentence extraction was implemented by scanning 512 characters before and after p-values for the presence of periods while ignoring numbers/decimals, parentheticals, and the word "Fig.". Not every sentence could be parsed (see Supplemental Materials 12.2), and in many cases, parsing may not have exactly identified the true sentence surrounding a p-value (e.g., due to abbreviations). Nonetheless, every extracted sentence necessarily represents the words near reported p-values, and a failure to parse some sentences was expected to only be a source of noise and not a bias because the analysis never compared frequencies between words.

Each extracted sentence was split into words, and for each of the 2500 most common words, normalized usage scores were calculated for each paper. This score was calculated by counting how many of the paper's sentences a word appeared in and dividing by the total number of top-2500 words across all the paper's sentences. For example, if a paper included one 8-word sentence containing the word "*the*" and one 12-word sentence containing the word "*the*", then the paper's "*the*" score would be 2 / 20 = 0.1. After calculating these scores, a separate linear regression was fit for each word predicting its usage on the paper's fragile p-values, as described in the main text.

12.2. Sentence quantities

By looking at different types of statistics, this permits insight into a wider range of results. For the analysis of a specific statistic type, the fragile p-value percentage used in the regression was computed based solely on the p-values associated with said statistic (e.g., when evaluating t-value word usage, only the p-values nearby t-values are used for calculating the fragile percentage). In total, the dataset contains significant p-values reported alongside 434k t-values (with or without degrees of freedom). After dropping sentences that could not be parsed and discarding repeat sentences, there remained 201k usable sentences. Similarly, there were 932k F-value entries yielding 437k usable sentences, 142k chi-square entries yielding 68k usable sentences, and 298k correlation/regression (r/b/B/β) entries yielding 132k sentences. Note that statistics selected in these cases were not required to be reported with degrees of freedom. For the analysis of all p-values, irrespective of one specific test statistic, there were 1.67M usable sentences.

13. Analysis with implied p-values

13.1. Implied p-values also strongly predict replicability

Among the 113 papers in the dataset with replicability data, 99 provided test statistics for calculating implied p-values (47 successful replications and 52 conclusive failures). There was a significant difference in the fragile implied p-value percentage between the 47 papers that replicated (M = .31, SE = .04) and those that did not (M = .48, SE = .04) (*t*[98] = 3.02, *p* = .003, *d* = .61; Figure S9); note that this effect does not cross the α = .001 threshold mentioned in the main text, but the sample size here is just 99 papers. In addition, a classifier fit using the same strategy as in Supplemental Materials 7.3 yielded a cross-validation accuracy of 63.4%.

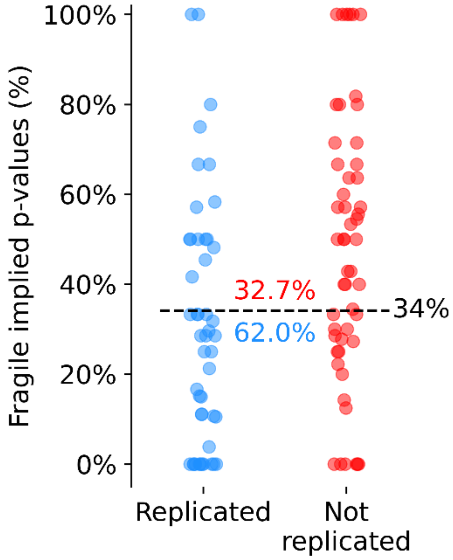


Figure S9. Reproduction of Figure S5 using implied p-values. This figure reproduces Figure S5 while now calculating the fragile percentage using p-values implied from test statistics. Each dot represents the fragile implied p-value percentage for one paper and whether the paper replicated or not. The dashed line at 34% represents the cutoff for whether a paper has over or under a 50% chance to replicate given its fragile p-value percentage, as identified using a logistic regression. The red "32.7%" indicates that 32.7% of papers above the dashed line did not replicate. Accordingly, the green "62.0%" indicates that 62.0% of papers below the dashed line replicated.

13.2. Implied p-value gap

In the main text, Figure 2 shows a clear difference between the mean percentage of fragile p-values and the mean percentage of fragile implied p-values (i.e., p-values inferred from nearby test statistics). Figure S10A plots the exact difference between those main text lines. Additionally, Figure 10B represents this comparison while only considering papers for which implied p-values can be computed – i.e., accounting for a selection bias related to many papers reporting p-values but not test statistics. For the most part, Figure 10B shows a smaller gap than Figure 10A, suggesting that some of the differences seen can be explained by selection effects. That is, papers that tend to report more robust p-values are more likely to report test statistics. When these selection effects are accounted for, Figure 10B suggests that there is roughly a 2% gap between fragile p-values and fragile implied p-value percentages. Before 2010, the differences were larger for some subjects, but this time period was associated with fewer papers and a smaller pool of usable journals, meaning that the differences may be noise and/or non-generalizable.

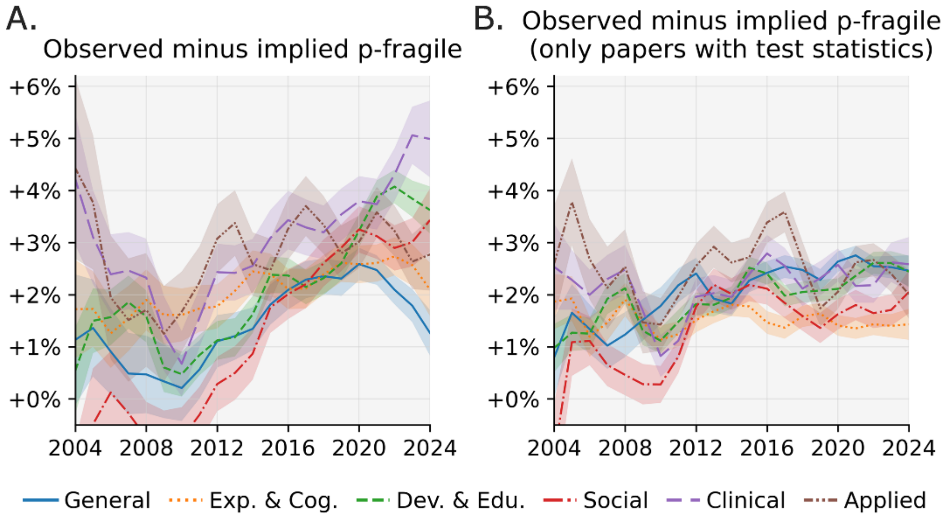


Figure S10. Differences between mean fragile p-value and fragile implied p-value percentages. A. These lines reflect the difference between the lines of main-text Figures 2A and 2B. Note that many of the papers reporting fragile p-values do not report any test statistics, and thus different papers contribute to each side of this difference. The shaded interval represents one standard error for a comparison between two samples. B. These lines reflect comparisons between fragile p-value percentages and implied fragile p-value percentages among only papers that report at least one test statistic and thus for which implied p-values can be computed. The shaded intervals now represent one standard error for a paired comparison; this notably causes the errors to become tighter than in the A plot.

13.3. P-values and incentives

The multilevel regressions used for the main text analysis were also done while calculating papers' fragile percentages using implied p-values. The effect of fragile p-values on citations remains significant (β = -.03, p < .0001), although the fragile p-values x Year interaction effect on citations falls below α = .001 significant threshold (β = -.03, p = .005). The link between fragile p-values x Year and SNIP also shows a trend but falls below the significance threshold (β = -.046, p = .005). The link between university ranking scores and fragile p-values remains robust (β = .016, p = .0007). Note that although some links fell to insignificance (per α = .001), the analysis of implied p-values uses a dataset that contains just about half as many papers (86k) as what was used for the analysis of p-values. In addition, the individual paper estimates become less reliable as there are fewer p-values used. Hence, the weakening of the associations below an α = .001 threshold is not taken to challenge the main text conclusions.

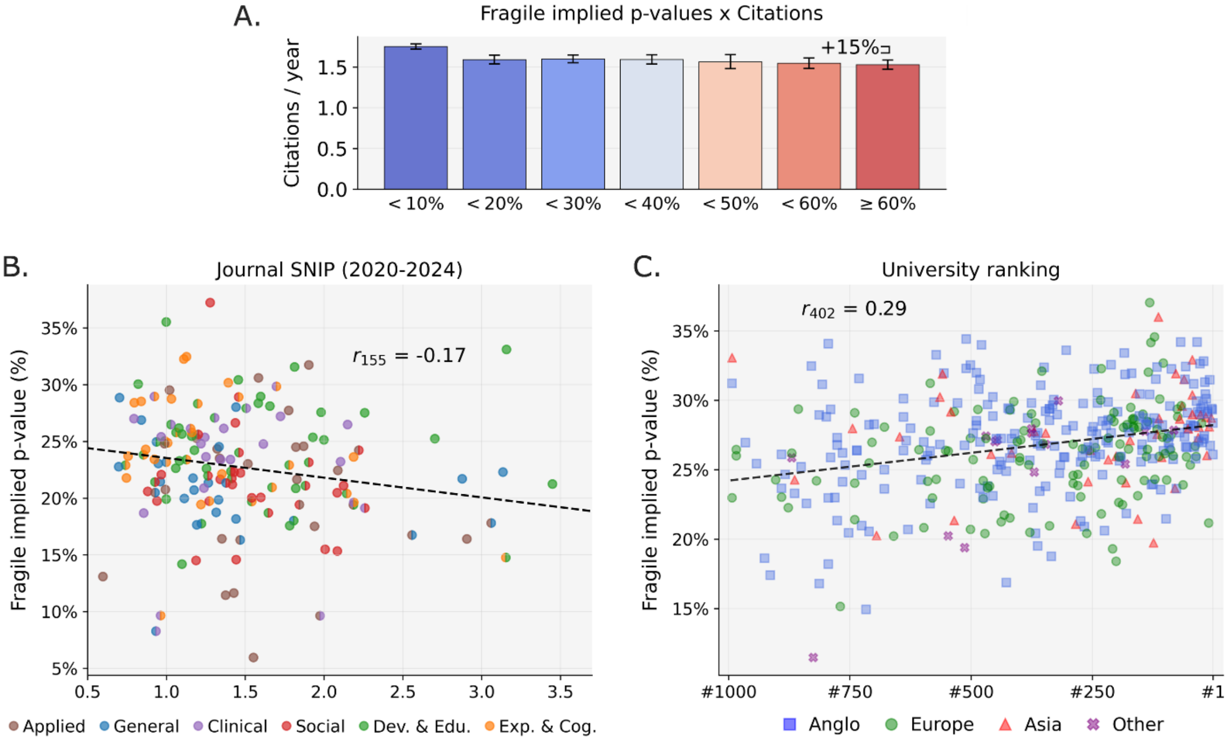


Figure S11. Reproduction of main text Figure 5 using implied p-values. This figure reproduces Figure 5 from the main text but now while calculating the fragile percentage using p-values implied from test statistics. See the Figure 3 caption in the main text. Note that the number of journals and universities plotted here is less than in Figure 3 because fewer papers report test statistics than report p-values. Hence, the exclusion criteria (requiring at least 10 papers and a standard error of the mean p-fragile percentage that is under .04) led to more drops.

14. Language-use-overlap tables for university ranking and p-value effects

Words linked *fewer* fragile p-values and to *lower-ranked* universities						
machiavell.	entreprene.	cyberbully.	procrastin.	dark	narcissism	pillais
trace	innovation	bartletts	swb	passion	alexithymia	hope
kmo	burnout	sphericity	lambda	has	adequacy	discrimina.
kaisermeye.	adaptabili.	organizati.	wilk	usefulness	factorial	exhaustion
theoretical	wilkss	national	resilience	autonomy	confirmato.	satisfacto.
homogeneity	capital	bullying	leadership	obtained	sem	dimension
dissatisfa.	normality	explaining	workplace	according	career	beside
onefactor	mediating	adequate	construct	victimizat.	manova	intrinsic
invariance	brand	problematic	internet	acceptable	path	threefactor
multivaria.	profession.	pearsons	bootstrap	respective	sampling	intention
goodnessof.	influence	hypothesis	supervisor	direct	measurement	structural
supported	loading	step	online	psychologi.	teacher	convergent
employee	directly	accounted	value	job	proposed	chisquare
can	follow	partially	solution	satisfacti.	excellent	standardiz.
impact	good	statistic	indirect	academic	mediation	ranged
covid	engagement	student	mediator	index	therefore	table
motivation	profile	mindfulness	finally	variable	through	indicate
equation	latent	lastly	show	assumption	relationsh.	explained
fit	hierarchic.	indicator	coefficient	work	variance	medium
trust	dependent	moreover	factor	are	personality	positive
positively	furthermore	descriptive	criterion	perception	correlation	subscale
support	attitude	considered	indicated	grade	predictive	moderate
related	datum	its	test	hypothesi	very	five
suggested	based	well	total	life	model	scale
figure	second	regression	predictor	social	factors	item
shown	sample	and	the	that		

Table S2. Words associated with fewer fragile p-values and with lower-ranked universities. Much like for Table 1 in the main text , the words listed here represent words showing significant associations for both fragile p-values an university rankings in a regression (*word usage ~ 1 + fragile p-value percentage + ranking*). Also like Table 1, the present table seeks to shed light on the association reported in the main text wherein higher-ranked universities tend to produce papers with more fragile p-values. However, unlike main-text Table 1, here the focus is on words used less by authors from top universities and which are linked to fewer fragile p-values. Long words that have been cut off to fit into the table end with ".". All of the tables below follow this same structure mirroring Table 1. For all of the language overlap tables, a small number of words associated with specific nationalities have been omitted.

<table>
<tr><th colspan="7"><b>Words linked <i>fewer</i> fragile p-values and to <i>higher-ranked</i> universities</b></th></tr>
<tr><td>perceiver</td><td>judged</td><td>judgments</td><td>lag</td><td>policy</td><td>human</td><td>character</td></tr>
<tr><td>united</td><td>midpoint</td><td>varied</td><td>greatest</td><td>true</td><td>estimate</td><td>above</td></tr>
<tr><td>moral</td><td>feature</td><td>prevalence</td><td>much</td><td>adult</td><td>random</td><td>majority</td></tr>
<tr><td>across</td><td>similarity</td><td>average</td><td>worse</td><td>person</td><td>rating</td><td>relatively</td></tr>
<tr><td>could</td><td>slightly</td><td>even</td><td>better</td><td>my</td><td>expected</td><td>improved</td></tr>
<tr><td>confirmed</td><td>again</td><td>still</td><td>belief</td><td>including</td><td>strongly</td><td>similar</td></tr>
<tr><td>highly</td><td>most</td><td>study</td><td>each</td><td>these</td><td>this</td><td>from</td></tr>
<tr><td>were</td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><th colspan="7"><b>Words linked <i>more</i> fragile p-values and to <i>lower-ranked</i> universities</b></th></tr>
<tr><td>dog</td><td>serum</td><td>muscle</td><td>con</td><td>regarding</td><td>balance</td><td>moderation</td></tr>
<tr><td>regard</td><td>statistical</td><td>tukey</td><td>marital</td><td>women</td><td>nonsignifi.</td><td>exercise</td></tr>
<tr><td>stage</td><td>significan.</td><td>product</td><td>experiment.</td><td>difference</td><td>moderated</td><td>betweengro.</td></tr>
<tr><td>posthoc</td><td>except</td><td>female</td><td>significant</td><td>girl</td><td>emerged</td><td>between</td></tr>
<tr><td>family</td><td>level</td><td>sex</td><td>found</td><td>analysis</td><td>group</td><td></td></tr>
</table>

Table S3. Word usage linked to rankings but inconsistent with identified rankings-p-values link. The words here were identified by searching for overlap in significance, much like for main-text Table 1 and Table S2, but now the words represent trends that run counter to the identified link between rankings and fragile p-values shown in the main-text regression analysis. That is, the words in the top half of the table are more so used by authors at top-ranking universities and linked to less fragile p-values. The words in the bottom half are less used by authors at top-ranking universities and are linked to more fragile p-values.

15. Language-use-overlap tables for p-value and year effects

Words liked to *fewer* fragile p-values and are *more frequent* in recent years						
covid	entreprene.	conspiracy	smartphone	meanwhile	resilience	cyberbully.
passion	innovation	burnout	loneliness	autoregres.	kmo	mediating
addiction	heterogene.	government	machiavell.	authentici.	adaptabili.	humor
online	capital	engagement	metric	mindfulness	profile	customer
problematic	mobile	normality	creativity	medium	has	bartletts
configural	employee	impact	bootstrap	fixed	psychologi.	directly
pooled	safety	climate	direct	exhaustion	positively	dark
science	academic	gratitude	phone	ethical	sense	hope
procrastin.	invariance	moral	brand	trust	mathematics	intrinsic
swb	through	political	love	narcissism	sphericity	sem
excellent	violated	economic	indirect	motivation	leadership	vehicle
autonomy	pathway	generalized	accordingly	moderate	maltreatme.	according
acceptable	positive	moreover	bullying	mediation	adequacy	figure
random	class	dataset	latent	intention	career	victimizat.
loaded	descriptive	work	suggest	unconditio.	suggested	socioecono.
perceived	intercept	weakly	turn	student	resource	environmen.
hypothesis	ideation	shown	life	table	sampling	adequate
path	predictive	occupation.	show	adding	structural	confirmato.
subjective	organizati.	weaker	satisfacti.	model	assumption	environment
study	national	estimated	null	coefficient	relationsh.	social
good	support	value	identity	prevalence	measurement	pearsons
distress	suicidal	based	achievement	country	standardiz.	size
similarly	could	indicator	its	stronger	full	school
including	robust	influence	large	highest	community	likewise
next	person	dimension	confidence	supported	loading	fit
average	index	can	belief	are	hypothesi	slightly
that	test	related	finally	and		

Table S4. Words associated with fewer fragile p-values that have become more frequent over time. See the main-text Table 1 and supplemental materials Table S2 captions for details.

Words liked to *more* fragile p-values and are *less frequent* in recent years						
bipolar	restraint	subject	montholds	caucasian	depressed	obese
primed	smoking	drink	prime	abstinence	priming	quit
experiment.	twotailed	episode	looked	drug	planned	genotype
novel	her	administra.	verb	latency	consumed	made
ethnicity	subtest	mood	stories	threeway	overweight	hsd
cause	attempt	either	recall	infant	onset	meal
presentati.	unrelated	pretreatme.	weight	object	muscle	whereas
movement	familiar	cortisol	emerge	eat	angry	medication
differ	intake	lifetime	slower	substance	did	lsd
span	history	salience	girl	placebo	fruit	longer
fearful	previously	fishers	mannwhitney	took	completer	possible
patient	look	main	none	preference	interacted	event
peak	separately	exception	trial	woman	twoway	contrast
pair	energy	looking	performed	number	response	neutral
chose	ancova	than	they	face	tukey	emerged
continued	site	diagnosis	reaction	inversely	delay	questionna.
only	month	read	bmi	happy	sex	cue
followup	times	reached	out	fewer	report	more
percentage	choice	revealed	versus	prior	days	analysis
problem	period	received	condition	one	interaction	task
female	younger	group	treatment	mothers	status	less
not	however	significan.	rate	their	over	likely
posthoc	performance	children	control	there	age	greater
comparison	difference	significan.	score	significant		

Table S5. Words associated with more fragile p-values that have become less frequent over time. See the main-text Table 1 and supplemental materials Table S2 captions for details.

Words liked to *more* fragile p-values and are *more frequent* in recent years						
connectivi.	moderation	selfregula.	regarding	multivaria.	moderated	gyrus
postinterv.	additional.	con	learner	caregiver	spring	sleep
unadjusted	interventi.	inhibitory	association	obesity	rsa	roi
skills	parenting	household	associated	team	executive	pupil
brain	exercise	band	adjusting	working	probing	specifical.
tau	expressive	balance	pairwise	helping	post	marginal
interactive	count	vocabulary	driven	listening	observed	favor
diversity	caregivers	asd	odds	product	adherence	education
message	actor	term	symptom	subgroup	negative	adhd
such	fluency	threshold	identified	compared	risk	increases
language	decreased	sensitivity	lower	increase	externaliz.	exhibited
showing	right	exposure	amplitude	higher	left	covariate
power	experienced	parents	outcome	maternal	state	level
bias	found	family	after	decrease	baseline	had
with	activity	simple	increased	between	time	participant
Words liked to *fewer* fragile p-values and are *less frequent* in recent years						
quite	testretest	entered	selfesteem	wilkss	clearly	highly
increment	manova	wilk	important	anovas	lambda	yielded
discrimina.	pure	rejected	equation	chisquare	youngest	criterion
spelling	computed	simultaneo.	considerab.	zeroorder	set	rater
equally	produced	yielding	scale	intercorre.	somewhat	judgments
beyond	accounted	once	multivaria.	obtained	attraction	substantia.
judged	clear	revised	weights	these	zero	again
function	delinquency	dependent	resulting	because	modified	much
equivalent	step	actual	measure	together	rating	confirmed
form	criterium	justice	respective.	variance	twofactor	reason
account	this	six	similarity	almost	commitment	expected
subscale	answer	culture	contributed	calculated	most	validity
four	even	latter	five	mean	first	item
resulted	each	initial	above	hierarchic.	second	third
three	respondent	personality	additional	for	strongly	were
different	hypothesiz.	both	overall	better	other	variable
factor	similar	also	was	the		

Table S6. Word patterns inconsistent with the decrease in fragile p-values over time. See the supplemental materials S3 caption for details.

16. Language-use-overlap tables for p-value and journal SNIP effects

Words liked to *fewer* fragile p-values and *higher* SNIP journals						
heterogene.	withinpers.	usefulness	innovation	pooled	cyberbully.	online
customer	internet	leadership	weighted	secondorder	autoregres.	political
supervisor	mobile	employee	smartphone	perceiver	environmen.	improve
policy	mathematics	supported	study	climate	investment	trust
achievement	hypothesis	medium	authentici.	fourfactor	capital	vehicle
size	estimate	weaker	science	random	intercept	substantia.
hypothesi	substantial	organizati.	creativity	united	ethical	maltreatme.
my	party	job	instrument	phone	judgments	tend
consumer	safety	has	somewhat	proposed	nonetheless	threefactor
stronger	career	onto	brand	alternative	simultaneo.	engagement
fixed	cohort	view	robust	dataset	moral	hypothesiz.
influence	resource	environment	constrained	intention	suggest	work
provide	autonomy	respective	judged	norms	example	similarity
perceived	next	impact	confidence	turn	worse	teacher
varied	majority	prevalence	source	across	estimated	shown
profile	path	mediated	class	latent	support	grade
again	indicate	provided	show	thus	related	even
indirect	initial	through	much	large	average	better
produced	its	above	are	positive	relationsh.	this
similar	which	expected	moderate	student	including	slightly
perception	coefficient	fit	strong	social	datum	positively
model	that	overall				

Table S7. Words associated with fewer fragile p-values that are also associated with higher SNIP journals. See the main-text Table 1 and supplemental materials Table S2 captions for details.

Words liked to *more* fragile p-values and *lower* SNIP journals						
con	restraint	dog	cigarette	abstinence	exercise	dietary
heavy	subtest	smoking	eat	muscle	obese	bmi
rat	dose	motor	father	caucasian	consumed	intake
hiv	overweight	alcohol	feeding	quit	hunger	meal
arm	fat	pre	weight	drink	snack	substance
coordinati.	fruit	food	obesity	drug	consumption	taste
pain	paternal	posthoc	attentional	attendance	mother	hsd
parenting	fathers	childs	inversely	selfreport.	span	mannwhitney
adherence	peak	impulsivity	score	caregivers	amplitude	marital
ethnicity	twoway	parents	movement	reaction	mothers	session
sex	questionna.	days	happy	bdi	incongruent	parent
care	lifetime	healthy	maternal	female	median	woman
family	medication	phase	delay	group	betweengro.	tukey
duration	asd	difference	shorter	girl	main	having
area	reached	neutral	sleep	statistical	cue	face
followup	physical	patient	symptom	characteri.	frequency	education
regard	faster	pairwise	problem	higher	during	baseline
significant	observed	there	compared	only	age	comparison
lower	revealed	after	reported	significan.	with	

Table S8. Words associated with more fragile p-values that are also associated with lower SNIP journals. See the main-text Table 1 and supplemental materials Table S2 captions for details.

Words liked to *more* fragile p-values and *higher* SNIP journals						
bar	computer	diversity	team	elaboration	montholds	looked
verb	moderated	novel	learner	saw	bilingual	spring
composition	primed	asymmetry	game	gesture	efficiency	read
remission	moderation	reliable	subgroup	comprehens.	connectivi.	interactive
infant	message	consistent	look	expressed	vocabulary	prime
interact	effective	rather	later	dyad	chose	reactivity
money	matched	looking	production	bias	performance	mdd
gaze	feedback	prior	outcome	immediate	took	word
exposure	choose	object	became	product	larger	earlier
region	assigned	whether	threeway	increases	driven	condition
planned	receiving	interacted	preference	did	qualified	relative
versus	presence	either	received	remained	spent	they
power	experienced	exhibited	when	longer	reduced	such
contrast	over	risk	children	participant	less	but
trial	more	interaction	greater	control	than	
Words liked to *fewer* fragile p-values and *lower* SNIP journals						
variablesf.	perfection.	bartletts	kmo	kaisermeye.	subscale	swb
wilkss	sphericity	normality	hope	lambda	adequacy	testretest
alexithymia	mindfulness	gambling	pearson	wilk	burnout	dark
resilience	configural	onesample	factorial	step	scale	life
responsibi.	convergent	pillais	validity	trace	discrimina.	accounted
total	profession.	bootstrap	manova	bullying	version	disability
multivaria.	dissatisfa.	pearsons	reliability	assumption	college	health
obtained	correlated	school	five	correlation	internal	contributed
hierarchic.	factors	explained	component	lowest	together	predictive
dimension	figure	equation	standardiz.	index	highest	variance
variable	regression	psychologi.	added	factor	presented	according
satisfacti.	final	demonstrat.	were	mean	indicated	predictor
three	addition	other	all			

Table S9. Word patterns inconsistent with the inverse relationship between fragile p-values and journal SNIP. See the supplemental materials S3 caption for details.

17. Language-use-overlap tables for p-value and citation effects

Words linked *fewer* fragile p-values and to *more* citations						
conspiracy	usefulness	heterogene.	smartphone	addiction	pooled	internet
mobile	cyberbully.	innovation	adaptabili.	leadership	environmen.	online
weighted	customer	covid	phone	ethical	procrastin.	exhaustion
burnout	employee	climate	bullying	resilience	gratitude	brand
pandemic	capital	trust	creativity	narcissism	workplace	organizati.
engagement	job	vehicle	mediated	secondorder	problematic	intention
study	big	prevalence	homogeneity	loneliness	proposed	adoption
entreprene.	mindfulness	structural	authentici.	fourfactor	has	mathematics
machiavell.	career	size	intrinsic	internaliz.	dark	supervisor
hypothesis	substantial	mediating	alexithymia	achievement	medium	path
science	ideation	work	influence	supported	suicidal	construct
hypothesiz.	government	mediator	sem	indirect	criterium	simultaneo.
mediation	fits	teacher	constrained	safety	country	constraini.
satisfacti.	stronger	direct	impact	dissatisfa.	measurement	support
revised	yielding	robust	autonomy	victimizat.	resource	need
academic	satisfacto.	substantia.	environment	behavioural	social	partially
weaker	onefactor	directly	explain	equation	alternative	psychologi.
related	fit	loading	residual	excellent	maltreatme.	indicator
political	strongest	through	motivation	well	fitted	hypothesi
strongly	pathway	moral	life	latent	respective	statistic
good	relationsh.	tested	willingness	positively	standardiz.	perceived
my	turn	explained	distress	adequate	adding	measure
student	worse	accounted	datum	acceptable	together	added
model	provided	show	slightly	moderate	estimate	improved
finally	attitude	positive	belief	suggest	coefficient	very
indicate	predictor	sample	additional	still	grade	variance
shown	large	moreover	school	health	five	including
table	general	similarly	yielded	overall	regression	variable
are	which	this	correlated	all	from	that

Table S10. Words associated with fewer fragile p-values that are also associated with more citations. See the main-text Table 1 and supplemental materials Table S2 captions for details.

Words linked *more* fragile p-values and to *fewer* citations						
con	abstinence	cue	priming	electrode	amplitude	movement
hiv	rat	presentati.	alliance	quit	smoking	pair
abstract	twoway	incongruent	genotype	shorter	prime	faster
spouse	red	offer	recall	interferen.	subtest	pain
object	hsd	reaction	central	occurrence	main	father
delay	drug	latency	threeway	inattention	phase	slower
response	trial	posthoc	peak	partner	unrelated	caregiver
pairwise	caregivers	fast	duration	tukey	mannwhitney	reached
attended	substance	neutral	side	childs	median	visual
viewing	drink	mother	window	instruction	session	alcohol
betweengro.	member	stage	attentional	temporal	face	memory
race	simple	expression	reach	task	condition	mothers
rate	performed	times	choice	statistical	percentage	made
interaction	care	preference	members	right	area	observed
qualified	revealed	longer	there	experiment.	comparison	difference
number	parents	group	following	frequency	baseline	than
participant	score	likely	significant	compared	lower	however
their	analysis	higher	only			

Table S10. Words associated with more fragile p-values that are also associated with fewer citations. See the main-text Table 1 and supplemental materials Table S2 captions for details.

Words linked *more* fragile p-values and to *more* citations						
asymmetry	bilingual	monolingual	remission	hamd	computer	mdd
connectivi.	diversity	product	game	moderation	tau	subgroup
completer	moderated	dropout	balance	responder	interventi.	secondary
eat	ancovas	clinically	posttreatm.	spent	depressed	asd
infant	sleep	reduction	post	team	looked	bar
composite	reduced	food	outcome	increases	exposure	controlling
engaged	parenting	later	home	experienced	girl	remained
working	disorder	looking	out	use	presence	improvement
reported	physical	language	activity	control	symptom	consistent
change	performance	did	associated	but	with	
Words linked *fewer* fragile p-values and to *fewer* citations						
perceiver	lag	auc	tablethe	normality	violated	kmo
onesample	sphericity	validity	kaisermeye.	character	adequacy	answer
consistency	convergent	agreement	disability	bartletts	responsibi.	judged
pearsons	midpoint	rejected	version	item	assumption	reliability
judgments	pearson	scenario	lowest	rating	presented	figure
different	factors	mean	subscale	test	scale	were
was	the					

Table S12. Word patterns inconsistent with the inverse relationship between fragile p-values and citations. See the supplemental materials S3 caption for details.