1 Have psychologists increased reporting of outliers in response to the reproducibility crisis?

2 K. D. Valentine[1], Erin M. Buchanan[2], Arielle Cunningham[3], Tabetha Hopke[3], Addie

3 Wikowsky[3], & Haley Wilson[3]

4 [1] University of Missouri

5 [2] Harrisburg University of Science and Technology

6 [3] Missouri State University

7 Author Note

8 K. D. Valentine is a Ph.D. candidate at the University of Missouri. Erin M. Buchanan

9 is an Professor of Cognitive Analytics at Harrisburg University of Science and Technology.

10 Arielle Cunningham, Tabetha Hopke, Addie Wikowsky, and Haley Wilson are master's

11 candidates at Missouri State University.

12 Correspondence concerning this article should be addressed to K. D. Valentine, 210

13 McAlester Hall, Columbia, MO, 65211. E-mail: kdvdnf@mail.missouri.edu

14   Abstract

15   Psychology is currently experiencing a "renaissance" where the replication and

16   reproducibility of published reports are at the forefront of conversations in the field. While

17   researchers have worked to discuss possible problems and solutions, work has yet to uncover

18   how this new culture may have altered reporting practices in the social sciences. As outliers

19   and other errant data points can bias both descriptive and inferential statistics, the search

20   for these data points is essential to any analysis using these parameters. We quantified the

21   rates of reporting of outliers and other data within psychology at two time points: 2012

22   when the replication crisis was born, and 2017, after the publication of reports concerning

23   replication, questionable research practices, and transparency. A total of 2235 experiments

24   were identified and analyzed, finding an increase in reporting from only 15.7% of experiments

25   in 2012 to 25.0% in 2017. We investigated differences across years given the psychological

26   field or statistical analysis that experiment employed. Further, we inspected whether data

27   exclusion mentioned are whole participant observations or data points, and what reasons

28   authors gave for stating the observation was deviant. We conclude that while report rates

29   are improving overall, there is still room for improvement in the reporting practices of

30   psychological scientists which can only aid in strengthening our science.

31   *Keywords:* outlier, influential observation, replication

32   Have psychologists increased reporting of outliers in response to the reproducibility crisis?

33   Psychology is undergoing a "renaissance" in which focus has shifted to the replication

34   and reproducibility of current published reports (Etz & Vandekerckhove, 2016; Lindsay, 2015;

35   Nelson, Simmons, & Simonsohn, 2018; Open Science Collaboration, 2015; van Elk et al.,

36   2015). A main concern has been the difficulty in replicating phenomena, often attributed to

37   publication bias (Ferguson & Brannick, 2012), the use and misuse of p-values (Gigerenzer,

38   2004; Ioannidis, 2005), and researcher degrees of freedom (Simmons, Nelson, & Simonsohn,

39   2011). In particular, this analysis focused on one facet of questionable research practices

40   (QRPs) that affect potential replication; the selective removal or inclusion of data points.

41   As outlined by Nelson et al. (2018), the social sciences turned inward to examine their

42   practices due to the publication of unbelievable data (Wagenmakers, Wetzels, Borsboom, &

43   van der Maas, 2011), academic fraud (Simonsohn, 2013), failures to replicate important

44   findings (Doyen, Klein, Pichon, & Cleeremans, 2012), and the beginning of the Open Science

45   Framework (Nosek, 2015). These combined forces led to the current focus on QRPs and

46   p-hacking and the investigation into potential solutions to these problems. Recommendations

47   included integrating effect sizes into results (Cumming, 2008; Lakens, 2013), encouraging

48   researchers to be transparent about their research practices, including not only the design

49   and execution of their experiments, but especially the data preparation and resulting

50   analyses (Simmons et al., 2011), attempting and interpreting well thought out replication

51   studies (Asendorpf et al., 2013; Maxwell, Lau, & Howard, 2015), altering the way we think

52   about p-values (Benjamin et al., 2018; Lakens et al., 2018; Valentine, Buchanan, Scofield, &

53   Beauchamp, 2019), and restructuring incentives (Nosek, Spies, & Motyl, 2012). Additionally,

54   Klein et al. (2014) developed the Many Labs project to aid in data collection for increased

55   power, while the Open Science Collaboration (2015) utilized a many labs approach to

56   publish combined findings to speak to the replication of phenomena in psychology.

57 While we have seen vast discussion of the problems and proposed solutions, research

58 has yet to determine how this new culture may have impacted reporting practices of

59 researchers. Herein, we aim specifically to quantify the rates of reporting of outliers within

60 psychology at two time points: 2012, when the replication crisis was outlined (Pashler &

61 Wagenmakers, 2012), and 2017, after the publication of reports concerning QRPs, replication,

62 and transparency (Miguel et al., 2014). Because of the slow editorial, revision, and

63 publication process, publications with the year 2012 were likely completed in 2011 or earlier,

64 thus, a good starting time point for gathering data at or around the start of the "crisis".

65 **Outliers**

66 Bernoulli first mentioned outliers in 1777 starting the long history of examining for

67 discrepant observations (Bernoulli & Allen, 1961), which can bias both descriptive and

68 inferential statistics (Cook & Weisberg, 1980; Stevens, 1984; Yuan & Bentler, 2001;

69 Zimmerman, 1994). Therefore, the examination for these data points is essential to any

70 analysis using these parameters, as outliers can impact study results. Outliers have been

71 defined as influential observations or fringliers but herein, we specifically use the definition of

72 "an observation which being atypical and/or erroneous deviates decidedly from the general

73 behavior of experimental data with respect to the criteria which is to be analyzed on it"

74 (Muñoz-Garcia, Moreno-Rebollo, & Pascual-Acosta, 1990, pg. 217). This definition was used

75 to capture a wide range of what one might consider "deviant": participant errors in an

76 experiment, unusable data, and data may be found at the tail ends of a distribution. The

77 removal of any data point for discrepant reasons should be transparently conveyed in a study,

78 and therefore, we used a more broad definition to include these different scenarios.

79 Additionally, the definition of outliers can vary from researcher to researcher, and a wide

80 range of graphical and statistical options are available for deviant data detection (Beckman

81 & Cook, 1983; Hodge & Austin, 2004; Orr, Sackett, & Dubois, 1991; Osborne & Overbay,

82 2004). For example, Tabachnick and Fidell (2012) outline several of the most popular

83 detection methods including visual data inspection, residual statistics, a set number of

84 standard deviations, Mahalanobis distance, Leverage, and Cook's distances. Participants

85 who do not complete the study correctly and/or unusable data is often found with these

86 types of detection techniques, and therefore, a broad definition of outliers is necessary to

87 capture researcher behavior.

88   Researchers have separated outliers into categories in many ways over the years

89 (Beckman & Cook, 1983; Hodge & Austin, 2004; Muñoz-Garcia et al., 1990; Orr et al., 1991;

90 Osborne & Overbay, 2004). Some of the most pervasive categories include experimenter error

91 (e.g., an error in the way the data was collected, coded, or prepared), participant behaviors

92 (e.g., intentional or motivated misreporting), and natural variability (including legitimate

93 data that are interesting because they do not fit the expected scheme). Just as there are

94 different categories of outliers, there are different ways to handle outliers. For instance, an

95 outlier that is a legitimate data point that does not fit into the expected scheme should not

96 necessarily be removed. However, an outlying data point that arose due to a coding error

97 should be corrected, not necessarily removed from an analysis.

98   Therefore, it is important to understand how outliers were detected, what type of

99 outlier they may be, and a justification for how the outliers were handled. Before the serious

100 focus on QRPs, the information regarding outlier detection as part of data screening was

101 often excluded from publication, particularly if a journal page limit requirement needed to

102 be followed. Consider, for example, Orr et al. (1991), who inspected 100

103 Industrial/Organizational Psychology personnel studies and found no mention of outliers

104 whatsoever.

105   However, while outliers may not be publicized, outlier detection and removal is likely

106 part of a researchers data screening procedure. LeBel et al. (2013) found that 11% of

107 psychology researchers stated that they had not reported excluding participants for being

outliers in their papers. Fiedler and Schwarz (2016) suggested that more than a quarter of researchers decide whether to exclude data only after looking at the impact of doing so. Bakker and Wicherts (2014) investigated the effects of outliers on published analyses, and while they did not find that they affected the surveyed results, they did report that these findings are likely biased by the non-reporting of data screening procedures in some articles, as sample sizes and degrees of freedom often did not match. These studies indicate that a lack of transparency in data manipulation and reporting is problematic.

By keeping outliers in a dataset, analyses are more likely to have increased error variance (depending on sample size, Orr et al., 1991), biased estimates (Osborne & Overbay, 2004), and reduced effect size and power (Orr et al., 1991; Osborne & Overbay, 2004), which can alter the results of the analysis and lead to falsely supporting (Type I error), or denying a claim (Type II error). Inconsistencies in the treatment and publication of outliers could also lead to the failures to replicate previous work, as it would be difficult to replicate analyses that have been p-hacked into "just-significant" results (Leggett, Thomas, Loetscher, & Nicholls, 2013; Nelson et al., 2018). The influence of this practice can be wide spread, as non-reporting of data manipulation can negatively affect meta-analyses, effect sizes, and sample size estimates for study planning. On the other hand, outliers do not always need to be seen as nuisance, as they will often be informative to researchers as they can encourage the diagnosis, change, and evolution of a research model (Beckman & Cook, 1983). Taken together, a lack of reporting of outlier practices can lead to furthering unwarranted avenues of research, ignoring important information, creating erroneous theories, and failure to replicate, all of which serve to weaken the sciences. Clarifying the presence or absence of outliers, how they were assessed, and how they were handled, can improve our transparency and replicability, and ultimately help to strengthen our science.

The current zeitgeist of increased transparency and reproducibility applies not only to the manner in which data is collected, but also the various ways the data is transformed,

cleaned, pared down, and analyzed. Therefore, it can be argued that it is just as important for a researcher to state how they identified outliers within their data, how the outliers were handled, and how this choice of handling impacted the estimates and conclusions of their analyses, as it is for them to report their sample size. Given the timing of the renaissance, we expected to find a positive change in reporting rates for outliers in 2017, as compared to 2012. This report spans a wide range of psychological sub-domains; however, we also expected the impact of the Open Science Collaboration (2015) publication to affect social and cognitive psychology more than other fields.

Method

Fields

A list of psychological sub-domains was created to begin the search for appropriate journals to include. The authors brainstormed the list of topics (shown in Table 1) by first listing major research areas in psychology (i.e., cognitive, clinical, social, etc.). Second, a list of common courses offered at large universities was consulted to add to the list of fields. Last, the American Psychological Association's list of divisions was examined for any potential missed fields. The topic list was created to capture large fields of psychology with small overlap (i.e., cognition and neuropsychology) while avoiding specific sub-fields of topics (i.e., cognition overall versus perception and memory only journals). Sixteen fields were initially identified; however, only thirteen were included in final analysis due to limitations noted below.

Journals

Once these fields were agreed upon, researchers used various search sources (Google, EBSCO host databases) to find journals that were dedicated to each broad topic. Journals

were included if they appeared to publish a wide range of articles within the selected fields. A list of journals, publishers, and impact factors (as noted by each journal's website in Spring of 2013 and 2018) were identified for each field. Two journals from each field were selected based on the following criteria: 1) impact factors over one at minimum, 2) a mix of publishers, if possible, and 3) availability due to university resources. These journals, impact factors, and publishers are shown in the online supplemental materials at https://osf.io/52mqw/.

Articles

Fifty articles from each journal were manually examined for data analysis: In the Spring of 2013, 25 articles were collected from each journal from 2012 backward, then, in Fall 2017, 25 articles were collected from 2017 backward. Data collection of articles started at the last volume publication from the given year (2012 or 2017) and progressed backwards until 25 articles had been found. Thus, while some journals may only include articles from 2012, other journals will include articles from previous years in order to fulfill the 25 article goal. Articles were included if they met the following criteria: 1) included data analyses, 2) included multiple participants or data-points, and 3) analyses were based on human subjects or stimuli. Therefore, we excluded theory articles, animal populations, and single subject designs. Based on review of the 2012 articles, three fields were excluded. Applied Behavior Analysis articles predominantly included single-subject designs, evolutionary psychology articles were primarily theory articles, and statistics related journal articles were based on user simulated data with a specific set of characteristics. Since none of these themes fit into our analysis of understanding data screening with human subject samples, we excluded those three fields from analyses.

Data Processing

Each article was manually reviewed for key components of data analysis. Each experiment in an article was coded separately. For each experiment, the type of analysis conducted, number of participants/stimuli analyzed, and whether or not they made any mention of outliers were coded by research assistants.

Analysis types. Types of analyses were broadly defined as basic statistics (descriptive statistics, z-scores, t-tests, and correlations), ANOVAs, regressions, chi-squares, non-parametric statistics, modeling, and Bayesian/other analyses.

Outlier coding. For outliers, the project team used a dichotomous yes/no coding regarding whether or not they were mentioned in an article. Outliers were not limited to simple statistical analysis of discrepant responses, but we also coded for specific exclusion criteria that were not related to missing data or study characteristics (i.e., we did not consider it an outlier if they were only looking for older adults). If outliers were mentioned, we coded information about outliers into four types: 1) people, 2) data points, 3) both, or 4) none found. Data that were coded as data points refer to the identification of individual trials being outlying while those coded as people referred to identification of the participant's entire row of data being outlying. We found that a unique code for data points was important for analyses with response time studies where individual participants were not omitted but rather specific data trials were eliminated.

Then, for those articles that mentioned outliers, the author's decision for how to handle the outliers was hand coded into whether they removed participants/data points, left these outliers in the analysis, or winsorized the data points. Experiments were coded for whether they tested the analyses with, without, or both for determination of their effect on the study. If they removed outliers, a new sample size was recorded. However, this data was not analyzed, as we determined it was conflated with removal of other types of data

unrelated to the interest of this paper (e.g., missing data). Lastly, we coded the reasoning for outlier detection as one or more of the following: 1) Statistical reason (e.g., used numbers to define odd or deviant behavior in responding, such as z-score or Mahalanobis distance scores), 2) Participant error (e.g., failed attention checks, did not follow instructions, or low quality data because of participant problems), and 3) Unusable data (e.g., inside knowledge of the study or experimenter/technological problems).

Results

Data Analytic Plan

Because each article constituted multiple data points within the dataset which were each nested within a particular journal and article, a multilevel model (MLM) was used to control for correlated error (Gelman, 2006). The Pinheiro, Bates, Debroy, Sarkar, and Team (2017) *nlme* package in *R* was used to calculate these analyses. A maximum likelihood logistic multilevel model was used to examine how the year in which the experiment was published predicted the likelihood of mentioning outliers (yes/no) while including a random intercept for journal and article. This model was analyzed over all of the data, as well as broken down by sub-fields or analyses in order to glean a more detailed account of the effect of year on outlier reporting. Additionally, three MLMs were analyzed attempting to individually predict each outlier reason (i.e., statistical reason yes/no; unusable data yes/no; participant reason yes/no) given the year while including a random intercept for journal and article. We did not use publication year as a dichotomous variable, as not all articles were from 2012 or 2017 because of publication rates (i.e., number of articles and issues per year) and article exclusions. Publication year ranged from 2001 to 2013 for articles collected in 2012, and 2015 to 2018 for articles collected in 2017 (several articles were considered online first with publication dates officially in 2018, and the official data was used for each article). Therefore, we treated this variable as continuous to capture the differences in years present

230 across each subfield and time point collected. Data is presented in tables dichotomously to

231 preserve space. We further explored whether these outliers were people or data points, how

232 outliers were handled, and the reasons data were named outliers with descriptive statistics.

233 All code and data can be viewed inline with the manuscript, which was written with the

234 *papaja* package (Aust & Barth, 2017).

Overall Outliers

236   Data processing resulted in a total of 2235 experiments being coded, 1085 of which

237 were from 2012 or prior, with the additional 1150 being from 2017 or prior. Investigating

238 reporting of outliers, we found that in 2012, 15.7% of experiments mentioned outliers, while

239 in 2017, 25.0% of experiments mentioned outliers. Actual publication year was used to

240 predict outlier mention (yes/no) with a random intercept for journal and article, as described

241 above. We found that publication year predicted outlier mentions, $Z = 2.74$, $p = .006$. Each

242 year, experiments were 12.2% more likely to report outliers as the previous year.

Fields

244   Further exploration reveals that differences in reporting between years arise between

245 fields which can be seen in Table 1. Figure 1 displays the percentage of outlier mentions of

246 each field colored by year examined. A MLM was analyzed for each field using journal and

247 article as a random intercept to determine the influence of year of publication on outlier

248 reporting rates. Specifically, if we look at the change in reporting for each field analyzed at

249 the level of the experiment, we find the largest changes in forensic (43.6% more likely to

250 report), social (41.4%), and I/O (33.9%), followed by developmental (22.5%) and cognitive

251 (16.9%). In support of our hypothesis, we found that social and cognitive fields showed

252 general increases in their outlier reporting; however, it was encouraging to see positive trends

in other fields as well. These analyses show that in some fields, including overview and neurological fields, we found a decrease in reporting across years, although these changes were not significant.

The analyses shown below were exploratory based on the findings when coding each experiment for outlier data. We explored the relationship of outlier reporting to the type of analysis used in each experiment, reasons for why outliers were excluded, as well as the type of outlier excluded from the study.

Analyses Type

Table 2 indicates the types of analyses across years that mention outliers, and Figure 2 visually depicts these findings. An increase in reporting was found for non-parametric statistics (33.3%), basic statistics (23.6%), modeling (17.1%), ANOVA (15.1%), and regression (13.3%). Bayesian and other statistics additionally showed a comparable increase, 25.1%, which was not deemed a significant change over years.

Type of Outlier

In our review, the majority of outliers mentioned referred to people (65.9%) as opposed to data points (25.4%), or both people and data points (5.7%), and a final (3.1%) of experiments mentioned outliers but did not specify a type, just that they searched for outliers and found none. The trends across years were examined for mentioning outliers (yes/no) for both people and data points, dropping the both and none found categories due to small size. Therefore, the dependent variable was outlier mention where the "yes" category indicated either the people or data point categories separately. The mentions of excluding entire participants increased across years, 15.2%, $Z = 3.00$, $p = .003$, while the mention of data trial exclusion was consistent across years, 6.6%, $Z = 0.68$, $p = .495$. Overall, when

276 handling these data, some experiments chose to winsorize the data (0.7%), most analyzed

277 the data without the observations (88.6%), some analyzed the data with the observations

278 (7.4%), and some conducted analyses both with and without the observations (3.4%).

Reason for Exclusion

280   We found that researchers often used multiple criterion checks for outlier coding, as

281 one study might exclude participants for exceeding a standard deviation cut-off, while also

282 excluding participants for low effort data. Therefore, reason coding was not unique for each

283 experiment, and each experiment could have one to three reasons for data exclusion.

284 Statistical reasoning was the largest reported exclusion criteria of papers that mentioned

285 outliers at 58.0%. Next, participant reasons followed with 50.3% of outlier mentions, and

286 unusable data was coded in 6.3% of experiments that mentioned outliers. To examine the

287 trend over time, we used a similar MLM analysis as described in the our data analytic plan,

288 with journal and article as a random intercept, year as the independent variable, and the

289 mention of type of outlier (yes/no for participant, statistical, and unusable data) as the

290 dependent variables separately. Statistical reasons tended to decrease about 8.5% each year,

291 $Z = -0.65$, $p = .518$. Participant reasons increased by 17.2% each year, $Z = 1.45$, $p = .147$.

292 Unusable data increased by about 6.7% each year, $Z = 0.69$, $p = .491$. None of these trends

293 would be considered "significant"; however, their pattern is an interesting finding to see that

294 traditional deviant data points for statistical reasons was decreasing, while there was

295 increased reporting for other types of deviant data.

Discussion

297   We hypothesized that report rates for outliers would increase overall in experiments

298 from 2012 to 2017, and we largely found support for this hypothesis. We additionally

²⁹⁹ hypothesized larger increases in report rates of outliers for the domains of social and

³⁰⁰ cognitive psychology because of the overwhelming response to the Open Science

³⁰¹ Collaboration (2015) publication. This hypothesis was supported, with increases for both

³⁰² areas, along with most other sub-domains in our study. Social and cognitive psychology

³⁰³ publications included the most experiments in their papers, and reporting outliers for each

³⁰⁴ experiment and analysis will be crucial for future studies or meta-analyses. While

³⁰⁵ improvements in reporting can be seen in almost all fields, it is worthwhile to note that in

³⁰⁶ 2017 the average proportion of experiments reporting outliers was still only 25.0%, with

³⁰⁷ some fields as low as approximately 12%. While the effort of many fields should not be

³⁰⁸ overlooked, we suggest that there is still room for improvement overall.

³⁰⁹   All analytic techniques presented in these experiments showed increased reporting over

³¹⁰ time, ranging from 17.1% for modeling to 33.3% for nonparametric statistics. Of all outliers

³¹¹ reported, we found that the majority discussed were people (65.9%), and that while

³¹² reporting of exclusion of people as outliers increased from 2012 to 2017, reporting of

³¹³ exclusion of outlying data points remained consistent across time. Most experiments cited

³¹⁴ outliers as those found through statistical means (e.g., Mahalanobis distance, leverage, or a

³¹⁵ standard deviation rule) and/or participant reasons (e.g., failed attention checks or failure to

³¹⁶ follow instructions), but a small subset were cited as unusable data (e.g., individuals who

³¹⁷ believed the procedure was staged or participants whose position in a room was never

³¹⁸ recorded by the experimenter). The trends across years indicate that potentially, the

³¹⁹ detection of outliers as only a statistical technique is decreasing, while transparently

³²⁰ presenting information about the exclusion of other errant data is increasing. These findings

³²¹ suggest that not only is discussion of outliers important for the study at hand, but also for

³²² future studies. Given insight into ways data can become unusable, a researcher is better

³²³ equipped to prepare for and deter unusable data from arising in future studies through

³²⁴ knowledge of past failures that can improve their research design.

₃₂₅   Given the frequentist nature of most psychological work, and the impact those outliers

₃₂₆ can have on these statistics (Cook & Weisberg, 1980; Stevens, 1984), research would be well

₃₂₇ served if authors described outlier data analysis in their reports. One confounding issue may

₃₂₈ be journal word limits. The Open Science Framework provides the option to publish online

₃₂₉ supplemental materials that can be referenced in manuscripts with permanent identifiers

₃₃₀ (i.e., weblinks and DOIs). Potentially, if journal or reviewer comments indicate shortening

₃₃₁ data analyses sections, the detailed specifics of these plans can be shifted to these online

₃₃₂ resources. While the best practice may be to include this information in the published article

₃₃₃ (as Bakker and Wicherts (2014) note that sample size and degrees of freedom are often

₃₃₄ inconsistent and difficult to follow in publications), online materials can be useful when that

₃₃₅ option is restricted.

₃₃₆   We recommend that those researchers who create reporting guidelines and checklists

₃₃₇ ensure that they address the need to discuss if outliers were identified, and if so, how outlier

₃₃₈ were identified, how these outliers were handled, and any differences that arose in

₃₃₉ conclusions when outliers were excluded/included. We believe this information should be

₃₄₀ included in every single publication that includes human subjects data (even simply to

₃₄₁ report that no outliers were found). We implore researchers not to overlook the importance

₃₄₂ of visualizing their data and identifying data that may not fall within the expected range or

₃₄₃ pattern of the sample. We suggest that all researchers implement relevant checklists or

₃₄₄ guidelines (a list of which can be found at https://osf.io/kgnva/wiki/home/) when handling,

₃₄₅ analyzing, and reporting their data as following these types of checklists has been shown to

₃₄₆ make a modest improvement in some reporting practices (Caron, March, Cohen, & Schmidt,

₃₄₇ 2017; Macleod & The NPQIP Collaborative Group, 2017). Additionally, there are online

₃₄₈ tools that can assist even the most junior researcher in the cleaning of data, including outlier

₃₄₉ detection and handling, for almost any type of analysis. From online courses (e.g.,

₃₅₀ Coursera.org, DataCamp.com), free software with plugins (e.g., JASP and jamovi; JASP

₃₅₁ Team, 2018; Jamovi project, 2018), and YouTube tutorials that detail the step by step

procedures (available from the second author at StatsTools.com), inexperienced researchers can learn more and better their reporting and statistical practices. Further, we recommend that those who are reviewers and editors consider data screening procedures when assessing research articles and request this information be added when it is absent from a report. This article spotlights the positive changes that are occurring as researchers are actively reshaping reporting practices in response to the conversation around transparency. We believe that these results present a positive outlook for the future of the psychological sciences, especially when coupled with the training, reviewer feedback, and incentive structure change, that can only improve our science.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K.,
 . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in
 psychology. *European Journal of Personality*, *27*(2), 108–119. doi:10.1002/per.1919

Aust, F., & Barth, M. (2017). papaja: Create APA manuscripts with R Markdown.
 Retrieved from https://github.com/crsh/papaja

Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors
 and quality of psychological research. *PLoS ONE*, *9*(7), 1–9.
 doi:10.1371/journal.pone.0103360

Beckman, R. J., & Cook, R. D. (1983). [Outlier..........s]: Response. *Technometrics*, *25*(2),
 161. doi:10.2307/1268548

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk,
 R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human
 Behaviour*, *2*(1), 6–10. doi:10.1038/s41562-017-0189-z

Bernoulli, D., & Allen, C. G. (1961). The most probable choice between several discrepant
 observations and the formation therefrom of the most likely induction. *Biometrika*,
 48(1-2), 3–18. doi:10.1093/biomet/48.1-2.3

Caron, J. E., March, J. K., Cohen, M. B., & Schmidt, R. L. (2017). A survey of the
 prevalence and impact of reporting guideline endorsement in pathology journals.
 American Journal of Clinical Pathology, *148*(4), 314–322. doi:10.1093/ajcp/aqx080

Cook, R. D., & Weisberg, S. (1980). Characterizations of an empirical influence function for
 detecting influential cases in regression. *Technometrics*, *22*(1), 495–508.
 doi:10.2307/1268187

Cumming, G. (2008). Replication and p intervals. *Perspectives on Psychological Science*, *3*(4), 286–300. doi:10.1111/j.1745-6924.2008.00079.x

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. doi:10.1371/journal.pone.0029081

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2), 1–12. doi:10.1371/journal.pone.0149794

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128. doi:10.1037/a0024445

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

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, *48*(3), 432–435. doi:10.1198/004017005000000661

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. doi:10.1016/j.socec.2004.09.033

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85–126. doi:10.1007/s10462-004-4304-y

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Jamovi project. (2018). jamovi (Version 0.8)[Computer software]. Retrieved from https://www.jamovi.org

JASP Team. (2018). JASP (Version 0.8.6)[Computer software]. Retrieved from
https://jasp-stats.org/

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., . . .
Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication
project. *Social Psychology, 45*(3), 142–152. doi:10.1027/1864-9335/a000178

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a
practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4.*
doi:10.3389/fpsyg.2013.00863

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . .
Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour, 2*(3), 168–171.
doi:10.1038/s41562-018-0311-x

LeBel, E. P., Borsboom, D., Giner-Sorolla, R., Hasselman, F., Peters, K. R., Ratliff, K. A., &
Smith, C. T. (2013). PsychDisclosure.org. *Perspectives on Psychological Science,
8*(4), 424–432. doi:10.1177/1745691613491437

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p:
"Just significant" results are on the rise. *Quarterly Journal of Experimental
Psychology, 66*(12), 2303–2309. doi:10.1080/17470218.2013.863371

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science, 26*(12),
1827–1832. doi:10.1177/0956797615616374

Macleod, M. R., & The NPQIP Collaborative Group. (2017). *Findings of a retrospective,
controlled cohort study of the impact of a change in Nature journals' editorial policy
for life sciences research on the completeness of reporting study design and execution.*
doi:10.1101/187245

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a
    replication crisis? What does "failure to replicate" really mean? *American
    Psychologist*, *70*(6), 487–498. doi:10.1037/a0039400

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., . . . Laan, M.
    van der. (2014). Promoting transparency in social science research. *Science*,
    343(6166), 30–31. doi:10.1126/science.1245317

Muñoz-Garcia, J., Moreno-Rebollo, J. L., & Pascual-Acosta, A. (1990). Outliers: A formal
    approach. *International Statistical Review / Revue Internationale de Statistique*,
    58(3), 215–226. doi:10.2307/1403805

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual
    Review of Psychology*, *69*(1), 511–534. doi:10.1146/annurev-psych-122216-011836

Nosek, B. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.

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. doi:10.1177/1745691612459058

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

Orr, J. M., Sackett, P. R., & Dubois, C. L. Z. (1991). Outlier detection and treatment in I /
    O psychology: A survey of researcher beliefs and empirical illustration. *Personnel
    Psychology*, *44*(3), 473–486. doi:10.1111/j.1744-6570.1991.tb02401.x

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should
    always check for them). *Practical Assessment, Research & Evaluation*, *9*(6), 1–12.

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on

replicability in psychological science. *Perspectives on Psychological Science*, *7*(6), 528–530. doi:10.1177/1745691612465253

Pinheiro, J., Bates, D., Debroy, S., Sarkar, D., & Team, R. C. (2017). nlme: Linear and nonlinear mixed effects models. Retrieved from https://cran.r-project.org/package=nlme

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. doi:10.1177/0956797611417632

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*(10), 1875–1888. doi:10.1177/0956797613480366

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, *95*(2), 334–344. doi:10.1037/0033-2909.95.2.334

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (Sixth.). Boston, MA: Pearson.

Valentine, K. D., Buchanan, E. M., Scofield, J. E., & Beauchamp, M. T. (2019). Beyond p values: utilizing multiple methods to evaluate evidence. *Behaviormetrika*, 1–29. doi:10.1007/s41237-019-00078-4

van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E.-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology*, *6*, 1365. doi:10.3389/fpsyg.2015.01365

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why

475   psychologists must change the way they analyze their data: The case of psi:

476   Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3),

477   426–432. doi:10.1037/a0022790

478   Yuan, K. H., & Bentler, P. M. (2001). Effect of outliers on estimators and tests in covariance

479   structure analysis. *British Journal of Mathematical and Statistical Psychology, 54*(1),

480   161–175. doi:10.1348/000711001159366

481   Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and

482   nonparametric tests. *The Journal of General Psychology, 121*(4), 391–401.

Table 1

Outlier Reporting by Field Across Years

Field	% 12	*N* 12	% 17	*N* 17	OR	*Z*	*p*
Clinical	9.3	54	12.0	50	1.06	0.43	.666
Cognitive	31.1	164	49.6	135	1.17	0.92	.357
Counseling	14.3	56	28.1	57	1.20	1.80	.072
Developmental	20.0	70	34.4	61	1.22	1.41	.158
Educational	8.9	56	12.1	58	1.07	0.55	.585
Environmental	12.1	58	12.1	58	1.01	0.05	.957
Forensics	3.2	62	18.6	70	1.44	2.32	.020
IO	5.8	104	18.5	124	1.34	2.02	.043
Methods	13.6	66	11.5	61	1.00	0.01	.992
Neuro	30.5	59	17.9	56	0.88	-1.22	.224
Overview	21.9	114	18.9	132	0.94	-0.39	.695
Social	9.8	164	33.8	231	1.41	2.54	.011
Sports	6.9	58	12.3	57	1.06	0.49	.625

Note. *N* represents number of experiments for each category.

Table 2

Outlier Reporting by Analysis Type Across Years

Analysis	% 12	N 12	% 17	N 17	OR	Z	p
Basic Statistics	15.0	406	31.0	507	1.24	3.04	.002
ANOVA	19.6	469	31.5	466	1.15	2.51	.012
Regression	12.0	208	22.1	244	1.13	2.13	.033
Chi-Square	19.6	112	23.8	172	1.10	0.80	.424
Non-Parametric	6.2	64	25.5	47	1.33	2.03	.043
Modeling	12.0	217	21.8	408	1.17	2.12	.034
Bayesian or Other	13.2	53	25.9	143	1.25	0.86	.389

Note. N represents number of experiments for each category.
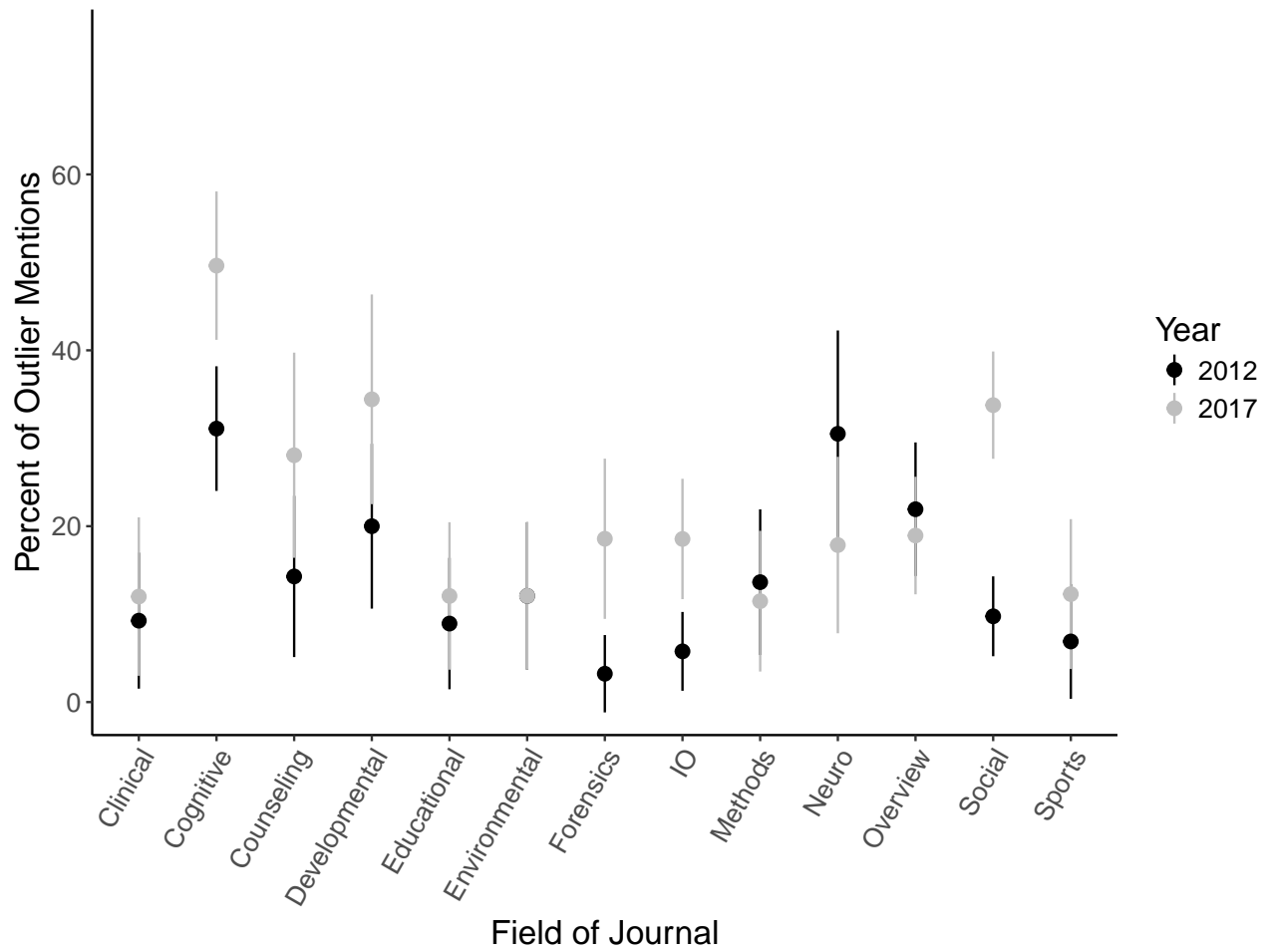
Figure 1. Percent of outlier mentions by sub-domain field and year examined. Error bars represent 95% confidence interval.
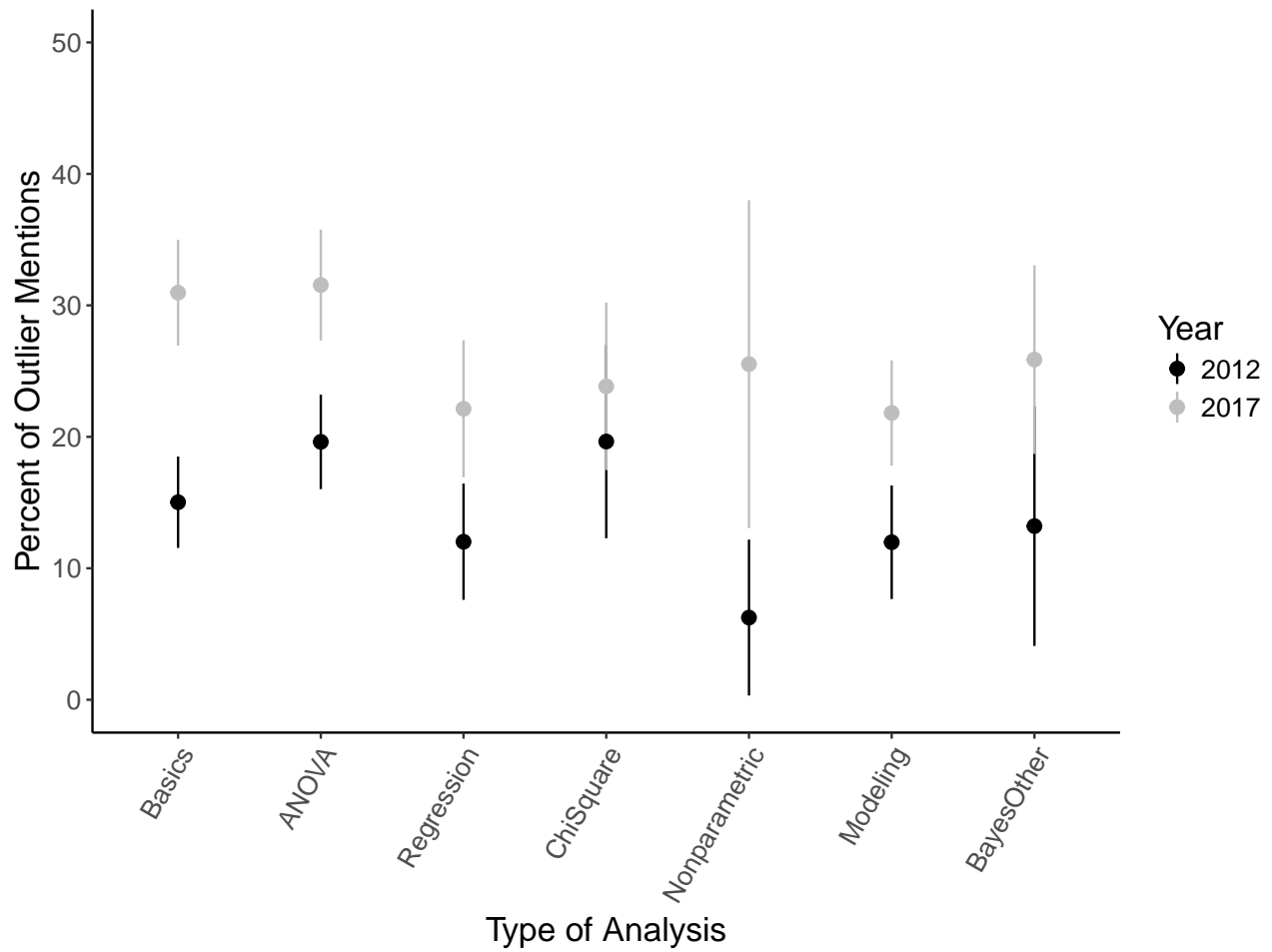
Figure 2. Percent of outlier mentions by analysis type and year examined. Error bars represent 95% confidence interval.