Have reserachers increased reporting of outliers in response to the reproducability crisis?

K. D. Valentine[1] & Erin M. Buchanan[2]

[1] University of Missouri

[2] Missouri State University

Author Note

K. D. Valentine is a Ph.D. candidate at the University of Missouri. Erin M. Buchanan is an Associate Professor of Quantitative Psychology at Missouri State University.

Correspondence concerning this article should be addressed to K. D. Valentine, 210 McAlester Hall, Columbia, MO, 65211. E-mail: kdvdnf@mail.missouri.edu

Abstract

The social sciences have begun to take a careful look at the way we process and interpret data, as many famous experiments do not appear to replicate (Open Science Collaboration, 2015). The Open Science Foundation (OSF) was founded in 2013 to promote a transparent research process from formation of the hypotheses to completely reproducible papers (Nosek et al., 2015). This project examines the impact of the formation of OSF and changing research culture on the publication of information concerning data screening methods for outliers, as the impact of outliers can critically change the findings and interpretation of experiments.

Keywords: outlier, influential observation, replication

Have reserachers increased reporting of outliers in response to the reproducability crisis?

Psychology is undergoing a "renaissance" in which focus has shifted to the replication and reproducibility of current published reports (Nelson, Simmons, and Simonsohn, 2018; Etz & Vandekerckhove, 2016; Lindsay, 2015; Open Science Collaboration, 2015; van Elk et al., 2015). A main concern has been the difficulty in replicating phenomena, often attributed to publication bias (Brannick, 2012), the use and misuse of *p*-values (Gigerenzer, 2004; Ioannidis, 2005), and researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). In particular, this analysis focused on one facet of questionable research practices that affect potential replication (QPRs), specifically, the selective removal or inclusion of data points.

As outlined by Nelson et al. (2018), the social sciences turned inward to examine its practices due to the publication of unbelievable data (Wagenmakers et al., 2011), academic fraud (Simonsohn, 2013), failures to replicate important findings (Doyen et al., 2012), and the beginning of the Open Science Framework (Open Science Collaboration, 2012). These combined forces led to the current focus on QRPs and *p*-hacking and the investigation potential solutions to these problems. Recommendations included CAN YOU OUTLINE THESE VARIOUS PUBS HERE. (Cumming, 2008; Simmons, Nelson, & Simonsohn, 2011; Asendorpf et al., 2013; Lakens, 2013; Benjamin et al., 2017; Lakens et al., 2018; Valentine, Buchanan, Scofield & Beauchamp, 2018; Maxwell, Lau, & Howard, 2015; Nosek, Spies, & Motyl, 2012). Additionally, Klein et al. (2014) developed the Many Labs project to aid in data collection for increased power, while the Open Science Collaboration (2015) published their findings from a combined many labs approach about the replication of phenomena in psychology.

I THINK THIS SENTENCE WILL GET INTEGRATED ABOVE. These include an increased (encouragement/requirement) for researchers to be transparent about their research practices, including not only the design and execution of their experiments, but especially the data preparation and resulting analyses.

While we have seen vast discussion of the problems and proposed solutions, research has yet to determine how this new culture may have impacted reporting practices of researchers. Herein, we aim specifically to quantify the rates of reporting of outliers within psychology at two time points: 2012 when the replication crisis was born (Pashler & Wagenmakers, 2012), and 2017, after the publication of reports concerning QPRs, replication, and transparency (WHAT IS THIS CITATION GAH NOSEK? 2012).

Outliers

Bernoulli first mentioned outliers in 1777 starting the long history of examining for discrepant observations (Bernoulli, 1961), which can bias both descriptive and inferential statistics (Cook & Weisberg, 1980; Stevens, 1984). Therefore, the examination for these data points is essential to any analysis using these parameters, as outliers can impact study results. Outliers have been defined as influential observations or fringliers but specifically we use Munoz-Garcia, Moreno-Rebollo, and Pascual-Acosta (1990, pg SOMETHING)'s definition of "an observation which being atypical and/or erroneous deviates decidedly from the general behavior of experimental data with respect to the criteria which is to be analyzed on it". However, the definition of outliers can vary from researcher to researcher, as a wide range of graphical and statistical options are available for outlier detection (Beckman & Cook, 1983; Hodge & Austin, 2004; Orr, Sackett, & Dubois, 1991; Osborne & Overbay, 2004). For example, Tabachnick and Fidell (2012) outline several of the most popular detection methods including visual data inspection, residual statistics, a set number of standard deviations, Mahalanobis distance, Leverage, and Cook's distances. Before the serious focus on QRPs, the information regarding outlier detection as part of data screening was often excluded from publication, particularly if a journal page limit requirement needed to be considered. Consider, for example, Orr, Sackett, and Dubois (1991), who inspected 100 Industrial/Organizational Psychology personnel studies and found no mention of outliers.

However, outlier detection and removal is likely part of a researchers data screening procedure, even if it does not make the research publication. Lebel et al. (2013) that 11% of 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 do report that these findings are likely biased by the non-reporting of data screening procedures, as sample sizes and degrees of freedom often did not match. The 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, Sackett, & Dubois, 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 (Nelson, Simmons, & Simonsohn, 2018; Legget, 2013). The influence of this practice can be wide spread, as non-reporting of data manipulation can negatively affect meta-analyses, effect size, 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 ratings 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. Lastly, 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, perception, and memory). 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 journals 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) high impact factors over one at minimum, 2) a mix of publishers, if possible, and 3) availability due to university resources. These journals are shown in the online supplemental materials at https://osf.io/52mqw/.

Articles

Fifty articles from each journal were examined for data analysis: 25 articles were collected beginning in Spring 2013 for 2012 and in Fall 2017. 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. We excluded online first publications and started in 2012 to ensure time for errata and retraction of articles. 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 for 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 created data with specific set 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 then 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.

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, we used a dichotomous yes/no for if they were mentioned in an article. Outliers were not limited to simple statistical analysis of discrepant responses, but we also checked 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 so, we coded information about outliers into four types: 1) people, 2) data points, 3) both, or 4) none found. The distinction between people and data points was if individual trials were eliminated or if entire participant data was eliminated. 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 coded into whether they removed participants/stimuli, 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; although, 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 (i.e., missing data). Lastly, we coded the reasoning for outlier detection as one or more of the following: 1) Statistical reason (i.e., used numbers to define odd or deviant behavior in responding, such as z-score or Mahalanobis distance scores), 2) Participant error (i.e., failed attention checks, did not follow instructions, or low quality data because of participant problems), and 3) Unusable data (i.e., inside knowledge of the study or experimenter/technological problems).

Results

Data Analytic Plan

KD IS GOING TO DO A THING HERE. Be sure to explain the MLMS.

Chi-Square tests were used to analyze the associations between variables. When a chi-square analysis suggested that the approximation may be incorrect a Fisher's Exact test will be analyzed as well and both will be reported. Frequencies were used to describe the outliers mentioned.

Overall Outliers

Data processing resulted in a total of 2234 experiments being coded, 1085 of which were from 2012 or prior, and the additional 1149 being from 2017 or prior. Investigating reporting of outliers, we found that in 2012, 15.7% of experiments mentioned outliers, while in 2017 25.1% of experiments mentioned outliers. Actual publication year was used to predict outlier mention (yes/no) with a random intercept for journal, as described above. We found that publication year predicted outlier mentions, $Z = 5.82$, $p < .001$. Each year, experiments were 13.6% more likely to report outliers as the previous year.

Fields

Further exploration reveals that differences in reporting between years arise between fields which can be seen in Table 1. Figure 1 displays the percentage of outlier mentions of each field colored by year examined. A MLM was analyzed for each field using journal as a random intercept to determine the influence of year of publication on outlier reporting rates. Specifically, if we look at the change in reporting for each field analyzed at the level of the experiment, we find the largest changes in forensic (44.9% more likely to report), social (33.7%), and I/O (34.9%), followed by developmental (19.6%), counseling (19.8%), and cognitive (15.2%). In support of our hypothesis, we found that social and cognitive fields showed 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 our overall 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 to support research hypotheses, reasons for why outliers were excluded, as well as the type of outlier excluded from the study. MAY HAVE TO REORDER HERE.

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 (38.2%), basic statistics (22.6%), regression (15.1%), ANOVA (14.5%), and modeling (11.7%). Bayesian and other statistics additionally showed a comparable increase, 23.5%, which was not a significant change over years.

Reasons

We found that researchers often used multiple criterion checks for outlier coding, as one study might exclude participants for exceeding a standard deviation cut-off, while also excluding participants for low effort data. Therefore, reason coding was not unique for each experiment, and each experiment could have one to three reasons for data exclusion. Statistical reasoning was the largest reported exclusion criteria of papers that mentioned outliers at 57.9%. Next, participant reasons followed with 50.4% of outlier mentions, and unusable data was coded in 6.3% of experiments that mentioned outliers. To examine the trend over time, we used a similar MLM analysis as described in the our data analytic plan, with Journal as a random intercept, year as the independent variable, and the mention of type of outlier (yes/no for participant, statistical, and unusable data) as the dependent variables separately. Statistical reasons decreased by 8.4%, Z = -1.91, p = .056. Participant reasons increased over time by 13.7%, Z = 2.92, p = .003. Unusable data increased by about 5.3%, Z = 0.59, p = .557.

Type of Outlier

In our review, the majority of outliers mentioned referred to people (65.9%) as opposed to data points (25.3%), 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 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 participants increased across years, 17.2%, $Z = 6.03$, $p < .001$, while the mention of data point exclusion was consistent across years, 4.5%, $Z = 1.11$, $p = .268$. When handling these data, some experiments chose to winzorize the data (0.7%), most analyzed the data without the observations (86.2%), some analyzed the data with the observations (7.2%), and some conducted analyses both with and without the observations (3.3%).

Discussion

While modest improvements in reporting can be seen in some fields, overall the outlook is still bleak. (OH GOOD GRIEF YOU). Some may argue that use of the precious word limit dictated by journals to describe such acts as outlier identification and handling may be irrational. However, we contest that given the current availability and use of online supplements and appendixes, as well as the invent of the Open Science Framework (OSF; cite) which allows researchers the ability to upload any number of additional resources and supplements that can be easily referenced in their manuscripts in as little as 4 words: See supplement at osf.io/52mqw.

References

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	.665
Cognitive	31.1	164	49.6	135	1.15	3.05	.002
Counseling	14.3	56	28.1	57	1.20	1.91	.056
Developmental	20.0	70	34.4	61	1.20	2.20	.028
Educational	8.9	56	12.1	58	1.07	0.54	.586
Environmental	12.1	58	12.1	58	1.01	0.05	.957
Forensics	3.2	62	18.6	70	1.45	2.50	.012
IO	5.8	104	19.4	124	1.35	3.04	.002
Methods	13.6	66	11.7	60	1.01	0.08	.933
Neuro	30.5	59	17.9	56	0.87	-1.55	.121
Overview	21.9	114	18.9	132	0.96	-0.64	.523
Social	9.8	164	33.8	231	1.34	5.08	$< .001$
Sports	6.9	58	12.3	57	1.08	0.64	.522

Table 2

Outlier Reporting by Analysis Type Across Years

Analysis	% 12	N 12	% 17	N 17	OR	Z	p
Basic Statistics	15.0	407	31.0	507	1.23	5.86	$< .001$
ANOVA	19.6	469	31.8	466	1.14	4.40	$< .001$
Regression	12.0	209	22.1	244	1.15	2.62	.009
Chi-Square	19.6	112	23.8	172	1.04	0.59	.557
Non-Parametric	6.2	64	25.5	47	1.38	2.67	.008
Modeling	12.0	217	21.9	407	1.12	2.20	.028
Bayesian or Other	13.2	53	25.9	143	1.23	1.61	.107

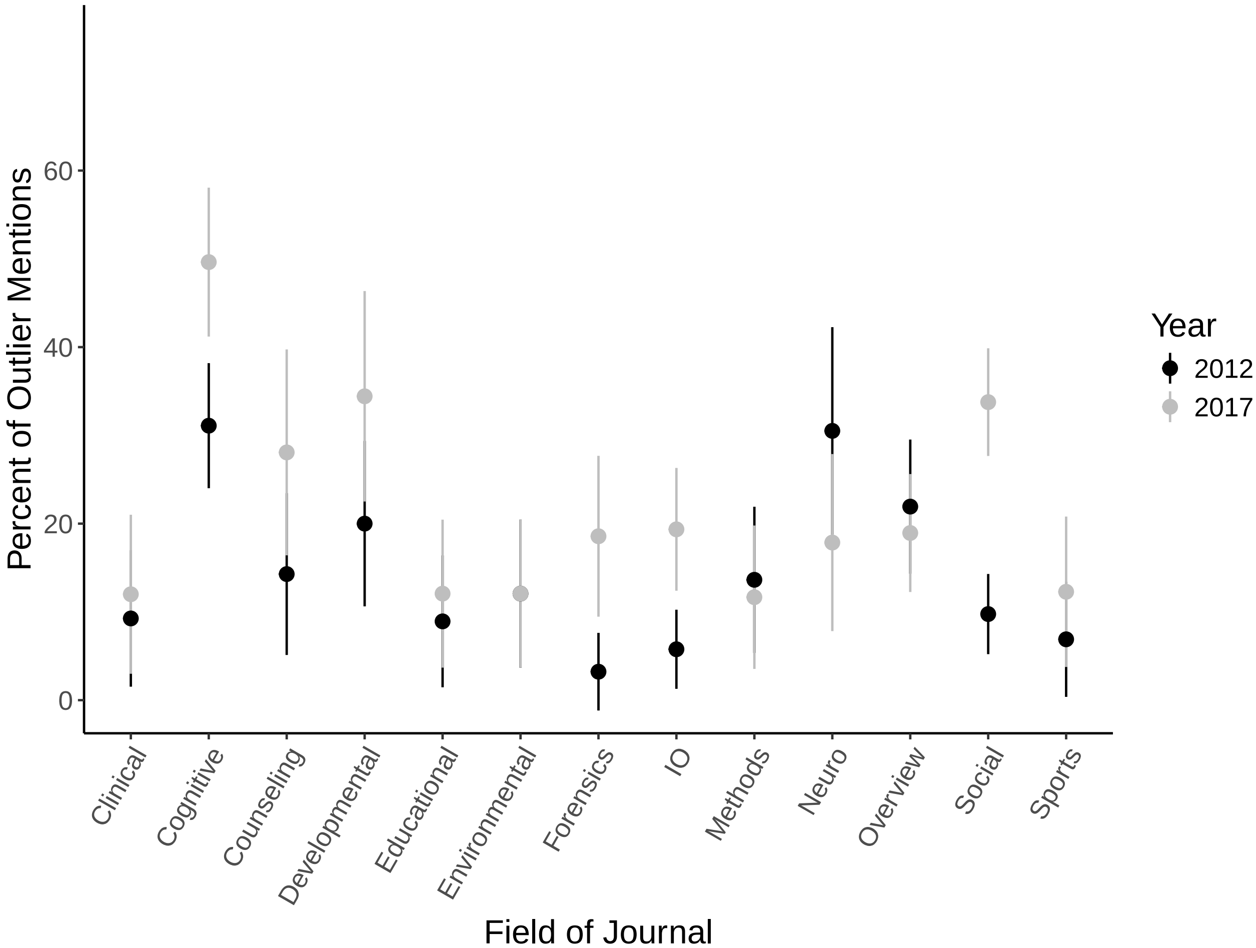

Figure 1. Percent of outlier mentions by sub-domain field and year examined. Error bars represent 95% confidence interval.

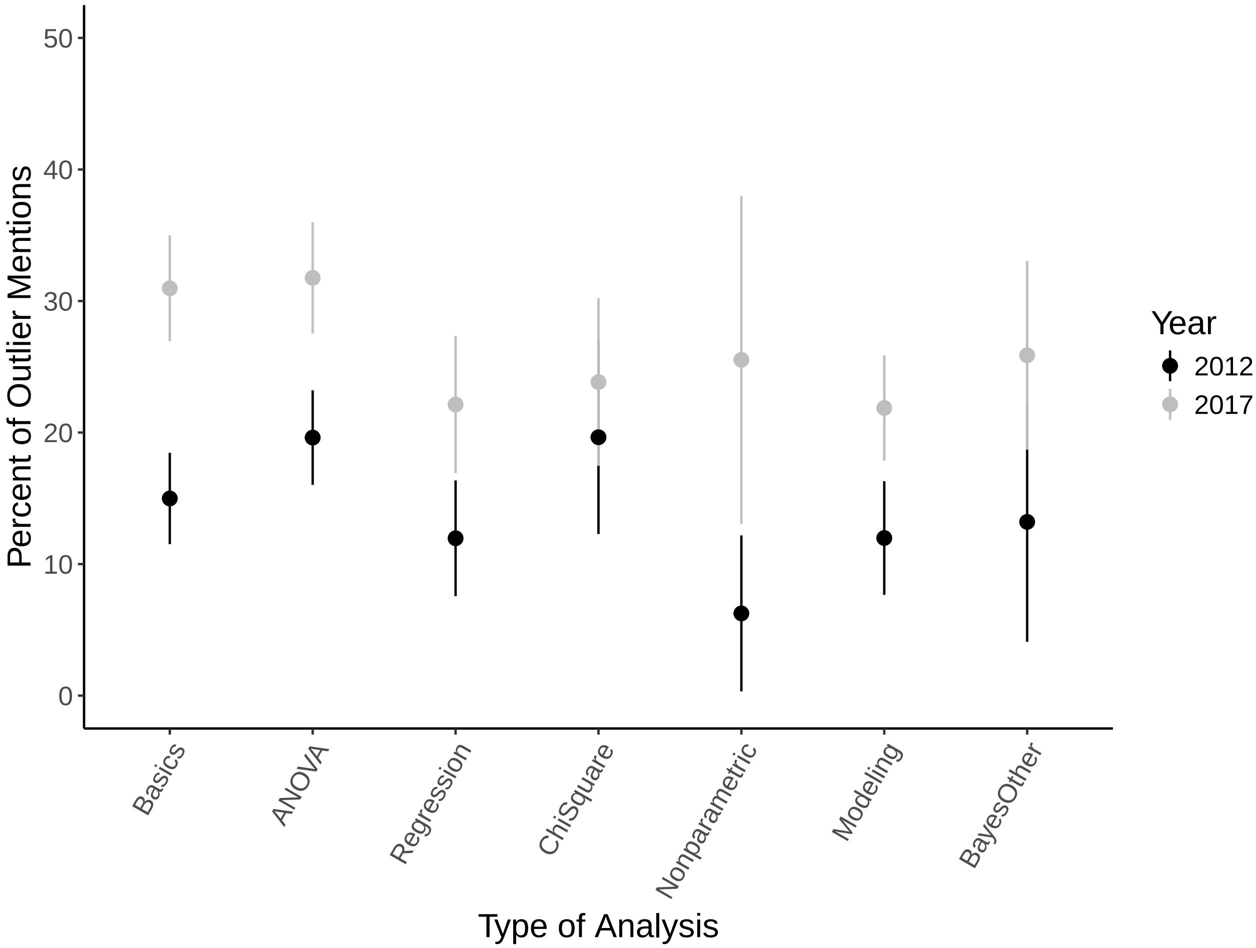

Figure 2. Percent of outlier mentions by analysis type and year examined. Error bars represent 95% confidence interval.