Psychon Bull Rev (2017) 24:2037–2043
DOI 10.3758/s13423-017-1267-y

BRIEF REPORT

Meeting the challenge of the Psychonomic Society's 2012 *Guidelines on Statistical Issues*: Some success and some room for improvement

Peter E. Morris[1] · **Catherine O. Fritz**[2]

Published online: 17 March 2017

Abstract The Psychonomic Society (PS) adopted *New Statistical Guidelines for Journals of the Psychonomic Society* in November 2012. To evaluate changes in statistical reporting within and outside PS journals, we examined all empirical papers published in PS journals and in the Experimental Psychology Society journal, *The Quarterly Journal of Experimental Psychology* (QJEP), in 2013 and 2015, to describe these populations before and after effects of the Guidelines. Comparisons of the 2013 and 2015 PS papers reveal differences associated with the Guidelines, and QJEP provides a baseline of papers to reflect changes in reporting that are not directly influenced by the Guidelines. A priori power analyses increased from 5% to 11% in PS papers, but not in QJEP papers (2%). The reporting of effect sizes in PS papers increased from 61% to 70%, similar to the increase for QJEP from 58% to 71%. Only 18% of papers reported confidence intervals (CIs) for means; only two PS papers in 2015 reported CIs for effect sizes. Although variability statistics are important to understanding data, and to further analysis, they were only reported as numbers in just over half of the PS journal papers. Almost all PS and QJEP papers relied exclusively on null hypothesis significance testing to guide interpretation of the data. Changes associated with the Guidelines are in the desired direction with respect to reporting effect sizes and power analyses but are not yet reflected in researchers' practices in describing their data, addressing data assumptions, and thinking beyond the *p* value when interpreting their data.

Keywords Statistical inference · Statistics · Confidence intervals · Effect size

The Psychonomic Society's publications committee, ethics committee, and the editors-in-chief of the Society's six journals published new guidelines on statistical issues for papers to appear in the Society's journals (Psychonomic Society, 2012) in November 2012. In recent years there has been growing concern about research practices, at least in part related to statistical usage. These include problems with reproducibility (e.g., Open Science Collaboration, 2015), undisclosed flexibility in data collection, selection, and analysis (e.g., Simmons, Nelson, & Simonsohn, 2011), and reliance upon null hypothesis significance testing (NHST) combined with neglect of other properties of the data such as effect sizes (e.g., Cohen, 1994; Cumming, 2014; Cumming & Calin-Jageman, 2016; Fritz, Morris, & Richler, 2012; Kline, 2013). We believe that evaluating the success of the Guidelines provides a good starting point for reviewing the current state of reporting of statistics in several leading experimental psychology journals and as the foundation for future improvements in making the best use of research data.

We monitored reporting changes following the introduction of the Guidelines by examining papers from the empirical journals of the Society: *Attention, Perception, & Psychophysics* (AP&P); *Psychonomic Bulletin & Review* (PB&R); *Memory & Cognition* (M&C); *Cognitive, Affective, & Behavioral Neuroscience* (CA&BN); and *Learning & Behavior* (L&B). As a baseline, we examined how well the

Electronic supplementary material The online version of this article (doi:10.3758/s13423-017-1267-y) contains supplementary material, which is available to authorized users.

✉ Peter E. Morris
p.morris@lancaster.ac.uk

[1] Department of Psychology, Lancaster University, Lancaster LA1 4YF, UK

[2] Psychology Division, University of Northampton, Northampton, UK

Guidelines were already met by authors and editors in Psychonomic Society (PS) journals from 2013 (Morris & Fritz, 2014); these were papers accepted before the 2012 Guidelines were published. We compared these data with similar coding from papers published in 2015: papers that had been accepted for publication after the Guidelines had been adopted.

To take into account general changes in the reporting practices of authors extending beyond the PS Journals, we also coded empirical papers published in *The Quarterly Journal of Experimental Psychology* (QJEP) in 2013 and in 2015 (Morris & Fritz, 2017). QJEP is published by the Experimental Psychology Society (EPS), and its papers are similar in topics and status with those in the PS journals, but the EPS have not issued guidelines similar to the 2012 PS Guidelines.

It was not possible to examine every topic addressed in the Guidelines, but we were able to explore many of the issues raised. These included the reporting of a priori power analyses (e.g., Faul, Erdfelder, Lang, & Buchner, 2007) to estimate the number of participants required to have a given probability (e.g., 80%) of obtaining a significant result for a particular size of effect, if the effect does exist. We also recorded whether there was any discussion of power in the papers. The Guidelines emphasize the benefits of going beyond NHST by routinely reporting effect sizes (e.g., Fritz et al., 2012; Morris & Fritz, 2013a, b) and their confidence intervals (CIs; e.g., Cumming, 2012, 2014; Masson & Loftus, 2003; Smith & Morris, 2015). We therefore coded the papers for these practices. The Guidelines state: "It is important to report appropriate measures of variability around means and around effects (e.g., confidence intervals around means and/or around standardized effect sizes)." We surveyed the reporting of measures of variability both of the sample data, such as standard deviations (*SD*s), and of the sample means, through standard errors (*SE*s) and confidence intervals (CIs). There are two principle ways in which variability is reported: in error bars in figures or in numerals within the text or tables. Error bars can visually convey the likely values of means in other data samples, but the figures are often too small to allow the error bars to be translated into numbers for further analysis (e.g., Morris & Fritz, 2013a). Therefore, we coded both the use of error bars and numbers in reporting variability within each article. We also took the opportunity to survey the types of statistical tests being reported in the papers surveyed, and we catalogued the types of effect size measures reported.

Finally, we noted the types of figures used in presenting means and variability. Newman and Scholl (2012) demonstrated that the use of bar charts to present means leads to a within-the-bar bias such that values within the bar are perceived as more likely than values outside (e.g., above) the bar (see also Fritz, Morris, Cherchar, Smith, & Roe, 2015; Okan, Garcia-Retamero, Cokely, & Maldonado, 2017).

Our purpose in this research was to document recent practices in the conduct and reporting of experimental research in both Psychonomic Society journals and another experimental psychology journal. Where practice falls short of the Guidelines, we hope to encourage improvement.

Method

Articles

Every empirical paper published in 2013 and in 2015 in *Attention, Perception, & Psychophysics* (AP&P); *Psychonomic Bulletin & Review* (PB&R); *Memory & Cognition* (M&C); *Cognitive, Affective, & Behavioral Neuroscience* (CA&BN); *Learning & Behavior* (L&B); and *The Quarterly Journal of Experimental Psychology* (QJEP) was coded. This comprised a total of 1,272 articles. Table 1 reports the numbers of papers from the 2013 and 2015 issues of the PS journals and the QJEP that were coded.

Statistics coded

The following statistics were coded if they were reported at least once in each article: A priori power analyses and references to power; the inclusion of standardized effect sizes, the type of effect size; and any discussion of these effect sizes. Where papers reported eta squared and were using factorial designs, we checked by calculation whether the statistic reported was eta squared (η^2) or was actually partial eta squared (η_p^2), which could be seriously misinterpreted (see Fritz et al., 2012; Morris & Fritz, 2013a). Papers giving partial eta squared values were counted as reporting partial eta squared, despite having been mislabelled; this error occurred in 7% of both the QJEP and PS papers for 2015.

The reporting in the Results sections, graphically or numerically (including tables), of means, standard errors, standard deviations, mean square errors, and confidence intervals was coded. We also identified papers where figures representing means appeared and the types of error bars provided, if any. The statistical tests reported were also coded.

Table 1 The numbers of surveyed empirical articles from 2013 and 2015 in the Psychonomic Society Journals and the *Quarterly Journal of Experimental Psychology*

		Psychonomic Society journals					
Year	QJEP	AP&P	PB&R	M&C	CA&BN	L&B	PS Overall
2013	141	148	120	99	60	36	463
2015	127	208	152	91	60	30	541

Note. QJEP = *Quarterly Journal of Experimental Psychology*; AP&P = *Attention, Perception, & Psychophysics*; PB&R = *Psychonomic Bulletin & Review*; M&C = *Memory & Cognition*; CA&BN = *Cognitive, Affective, & Behavioural Neuroscience*; L&B = *Learning & Behavior*

Coding was carried out by searching each paper for a set of keywords or terms and supplementing these searches by reading through the method and results sections to check that no terms had been missed because of unusual names or spelling.

The coding was carried out by the first author. However, as a check on reliability, a randomly selected 10% sample from each of the 2013 PS journals was independently coded by the second author, recording the results using the same spreadsheet columns recording the data (data are available at osf.io/589by). There was a 99% correspondence, indicating a high degree of reliability. The small number of disagreements almost always involved unusual or ambiguous phrasing, and all were satisfactorily resolved.

Results and discussion

The summary results of our coding of the journal articles for 2013 and 2015 are reported as percentages. We restrict our comments here to the overall results for the PS journals and for QJEP, except where particular PS journals differ markedly in the pattern of the percentages. Details for the individual PS journals can be found in the online supplementary materials.

The types of statistical analyses carried out in the papers have implications for what other statistics can be expected; for example, statistics such as standard error of the mean and standard deviation are appropriate for a normal distribution and suggest a parametric analysis. We therefore begin with a brief summary of the statistical analyses reported in the papers (see Fig. 1), which we found to be almost always parametric. ANOVA and related tests (MANOVA and ANCOVA) were the most prevalent for all journals, followed by *t* tests. The *t* tests were often follow-up tests after an ANOVA, or were used in the analysis of multiple regressions, but they also appeared quite frequently as the main statistical test.

Pearson product moment correlations were reported in 20% to 33% of papers, increasing across the 2 years, and slightly more frequently in QJEP. Linear or multiple regressions were less frequently employed (11%–23%) and appeared more often in QJEP than PS journals. Other tests were even less frequently reported, as summarized in Fig. 1.

Power

The Guidelines begin with a strong encouragement for researchers to conduct an a priori power analysis before carrying out research. Such an analysis ensures that adequate numbers of participants are tested and requires a clear definition of the research design prior to data collection, which, in turn, easily enables preregistration of the research (Cumming & Calin-Jageman, 2016). However, power was infrequently addressed, as shown in the top sections of Fig. 2. For 2015, power was mentioned (a priori or otherwise) in only 16% of PS and QJEP papers, usually merely as comments that a failure to find a significant effect might be attributed to low power. A priori power analyses were very rare in the 2013 journals but for PS journals increased from 5% to 11% in 2015. This improvement in reporting of the a priori power analyses was most noticeable in M&C and CA&BN.

Effect size

The Guidelines emphasize the benefits of going beyond simple NHST by including effect size estimates and their CIs. Effect size

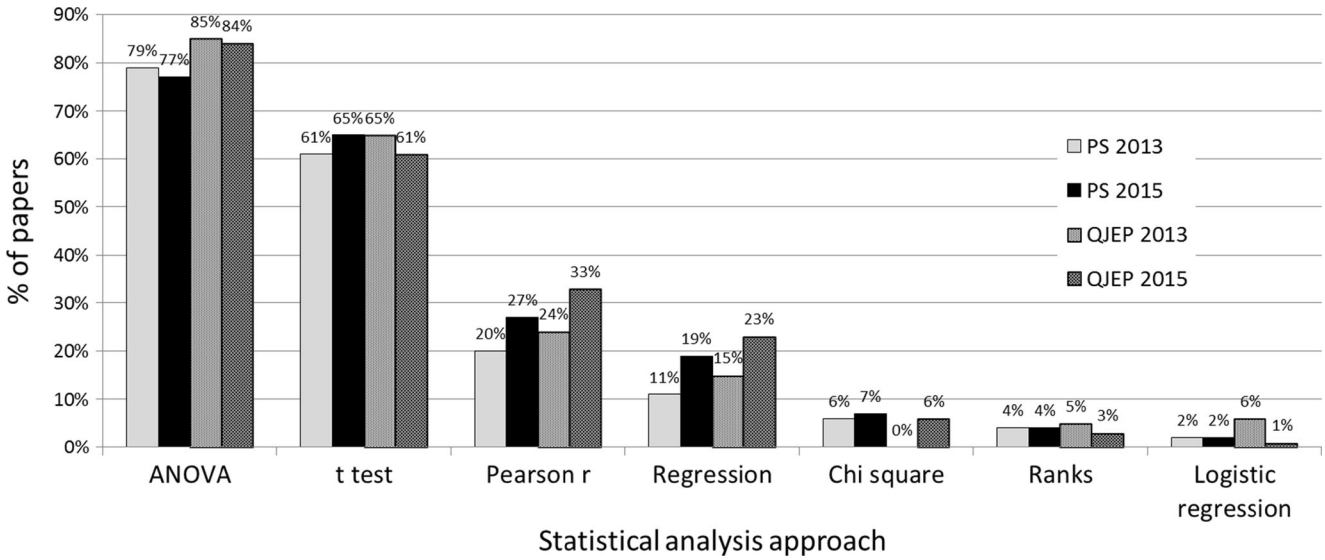

Fig. 1 Statistical tests reported in psychonomics journals and QJEP in 2013 and 2015. Other tests appeared far less frequently. *Note.* PS = Psychonomic Society journals; QJEP = *Quarterly Journal of Experimental Psychology*

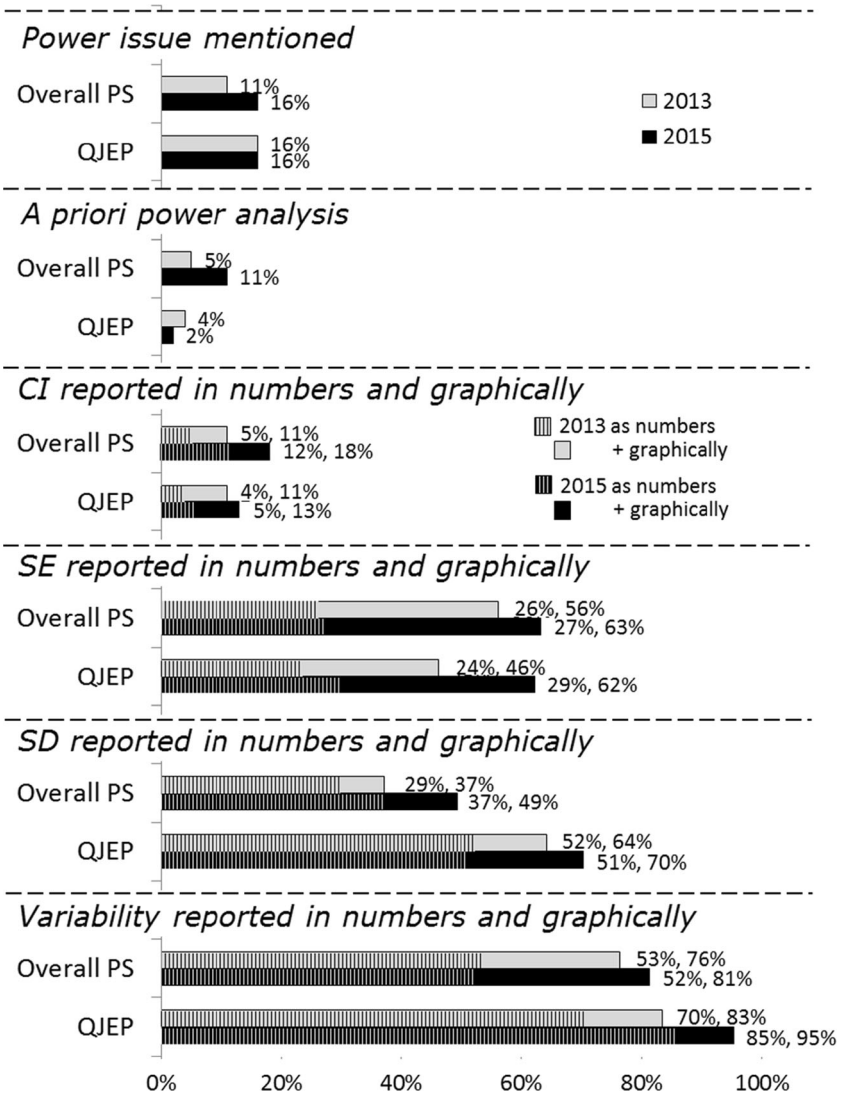

Fig. 2 Percentage of papers reporting power considerations and variability information. *Note.* PS = Psychonomic Society journals; QJEP = *Quarterly Journal of Experimental Psychology*; CI = confidence intervals; *SE* = standard error; *SD* = standard deviation

estimates add a great deal to any report (e.g., Cumming, 2012; Cumming & Calin-Jageman, 2016; Fritz et al., 2012; Smith & Morris, 2015). The reporting of standardized effect sizes was quite widely adopted in the journals that we surveyed; some effect size estimates were included in three fifths of the Psychonomic Society papers in 2013, and this rose to nearly three quarters by 2015. For the QJEP, half of the papers were reporting standardized effect sizes in 2012, rising to three fifths in 2015. Details are in the online supplementary materials.

Figure 3 plots the percentages of papers reporting particular effect size statistics for PS journals and QJEP in both years. By far the most frequently reported statistic was partial eta squared, which appeared in 41% of the PS papers in 2013, rising to 53% in 2015. Very similar frequencies for partial eta squared occurred in QJEP. The dominance of partial eta squared has the appearance of some authors addressing the need for reporting effect sizes in the easiest way—by copying partial eta squared from statistical software ANOVA output. Papers in which partial eta squared was reported for ANOVA frequently provided no effect size estimates for the comparisons between pairs of conditions when these were subsequently analysed by post hoc or *t* tests. So, for example, three fifths of PS papers in 2015 used *t* tests, which could have been accompanied by *d* or *r* or other effect-size statistics, but only a fifth of the papers reported *d* and a tiny handful partial eta squared for those comparisons. Although effect sizes were quite frequently reported, they were very rarely used when interpreting the findings.

Figure 3 shows that effect size statistics other than partial eta squared were far less frequently reported. Cohen's *d* was the most common. Many regression analyses reported neither R^2 nor any other standardized effect-size measure. For example, in 2015, 19% of PS papers used regression analyses (see Fig. 1), but only 9% reported R^2 or R_{adj}^2 (see Fig. 3). Standardized regression weights (β) are another relevant effect size statistic for regression

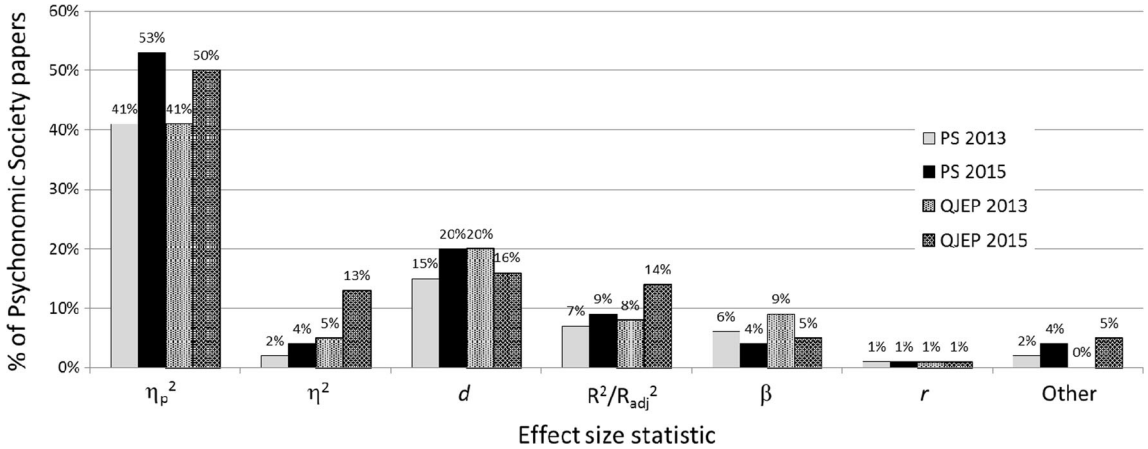

Fig. 3 The reporting of standardized effect sizes in the 2013 and 2015 journals. *Note.* PS = Psychonomic Society journals; QJEP = *Quarterly Journal of Experimental Psychology*

analyses, but, as Fig. 3 shows, they were reported in less than half of the regression analyses. Only 4% of PS papers in 2015 reported standardized regression weights, despite 19% of them employing regression. Other effect-size statistics were very rarely reported. Omega squared and partial omega squared are highly recommended measures of population effect size (e.g., Fritz et al., 2012; Grissom & Kim, 2011; Keppel & Wickens, 2004) but were almost never used.

Confidence intervals were provided for effect size estimates in just two of the PS 2015 journal papers, despite the recommendations of the Guidelines that they should be reported. The confidence intervals for the effect-size measures remind readers that the reported data are just one possible sample from those that might be obtained if the research was repeated; they suggest a range within which the vast majority of further sample means would probably lie. Confidence intervals in experimental psychological research are often wide because of small samples and a great deal of uncontrolled variability; this lack of precision has important implications for interpretation of the results in terms of theory, future research, and applications.

Variability

Turning to measures of variability, two types were coded: those indicating the precision of the estimate for the population mean (CIs and *SE*s) and those reporting the variability of the sample (*SD*s). Reports of variability are summarized in the middle parts of Fig. 2. A distinction is made between reporting variability as numbers and reporting it graphically because, in practice, only papers reporting values numerically can be used accurately for further analysis by anyone interested in drawing more information from the reported data. Numerical values for the individual variability statistics (shaded areas in the bars) were reported for only a minority of papers, except for standard deviation reporting in QJEP.

Confidence intervals for means were provided in some form in very few papers (see Fig. 2), with PS journals showing a larger increase from 2013 than QJEP. The improvement in the PS papers mainly resulted from the increased reporting of confidence intervals in M&C from 15% to 22% in 2015, and an even larger increase of 11% to 20% in PB&R. Confidence intervals were reported as numbers in roughly half as many papers as those providing confidence intervals in any form. All but one of the confidence intervals reported were for 95% confidence intervals, with the one exception of a 90% confidence interval.

The majority of PS papers reported standard errors (see Fig. 2). Reporting of standard errors increased for both journals, more so for QJEP. Almost two thirds of PS and QJEP papers reported standard errors in some form in 2015; in just under half of those papers standard error was specified in numbers.

Although standard deviations in some form (see Fig. 2) were reported more frequently for QJEP than for PS journals, a substantial increase occurred for both. More papers reported standard deviation in QJEP than in PS journals. Standard deviation in numbers was reported in roughly one third of PS papers and just over half of QJEP papers.

The bottom part of Fig. 2 gives the percentages of papers reporting *any* variability statistics; papers are included if at any point the results section provided confidence interval, standard error, and/or standard deviation. A small but nevertheless surprising number of papers failed to report any measure of variability. In 2013, this was true for 24% of PS journal articles; this percentage dropped to 19% in 2015. Some PS journals were more likely than others to omit all measures of variability: In 2015, 30% of M&C papers and 25% of PB&R papers neglected to report the variability of the data. QJEP achieved a much better standard, with 95% of papers providing some measure of variability in 2015. The shaded part of each bar also shows how many papers had reports of variability in a numerical form that could be used in any further analysis. This was the case for just over half of the PS papers in 2013 and 2015. QJEP papers were more suitable to further analysis, with 70% having numerical reports of variability in 2013, increasing to 85% in 2015.

In ANOVA reports, mean square error combined with *F* ratio and degrees of freedom provide another way to express variability in the data and to reconstruct the analysis. In PS journals, mean square errors were reported in only 20% of papers using ANOVA in both 2013 and 2015. Mean square errors were more frequently reported in QJEP, with 34% in 2013 and 33% in 2015.

Graphical display of means and variability

Most papers in both PS journals and QJEP included a figure when reporting means and the use of graphs increased slightly between 2013 and 2015. Bar charts were most often used; for PS journals, they appeared in 44% of the papers in 2013, increasing to 53% in 2015. QJEP percentages were similar at 47% and 55%, respectively. Bar charts may lead to a within-the-bar bias so that values within the bar are perceived as more likely than values outside the bar, even though points equidistant above and below the top of the bar are equally likely to represent the population mean (Newman & Scholl, 2012). Alternatives to bar charts, typically representing the mean as a point, often with error bars and frequently connected by lines to other levels in the variable, were reported in around a quarter of papers. Line graphs appeared in 23% of PS journal papers in both years; for QJEP they appeared in 21% and 25% of papers, respectively. Graphs representing means as points (with or without error bars) not connected by lines were very rarely used.

The vast majority of figures included error bars (see Table 2), with little change in frequency between 2013 and 2015. PS papers were more likely to include error bars than were QJEP papers. Error bars were sometimes (12%–21% of papers) present but not defined. There was considerable variation among PS journals. Standard error bars were the most common across all journals, appearing in more than half of the papers with means graphs, followed by confidence interval error bars and a few papers with standard deviation error bars.

Table 2 Error bars appearing in 2013 and 2015 Psychonomic Society Journals and the *Quarterly Journal of Experimental Psychology*; data are percentages of papers with graphs and the difference in percentages between years

Error bar type	2013		2015		Change	
	QJEP	PS	QJEP	PS	QJEP	PS
CI error bars	10	11	12	16	+2	+5
SE error bars	51	67	61	62	+10	-5
SD error bars	2	2	1	3	-1	+1
Undefined bars	21	13	12	12	-9	-1
Any error bars	83	96	86	93	+3	-3

Note. CI = confidence intervals; *SE* = standard error; *SD* = standard deviation; PS = Psychonomic Society journals; QJEP = *Quarterly Journal of Experimental Psychology*

An occasional paper provided box plots marking the median, quartiles, and range. Comparing 2013 to 2015, for PS journal papers there was a small shift from standard error to confidence interval error bars. Some papers depicted the error bars in only the upward direction, even though the margin of error extends in both directions from the mean. Both PS journals and QJEP had fewer of these cases in 2015 (13%) than in 2013 (19% and 17%, respectively).

Conclusions

Overall, the Guidelines appear to fill a need, as evidenced by the papers accepted for publication before the Guidelines' adoption. After the publication of the Guidelines, there were changes to some practices in the direction recommended by the Guidelines, but those changes were limited to a small proportion of the papers and left big gaps between the practices suggested in the Guidelines and the actual practice in most papers. If the overall level of reporting of data in PS journals (and the QJEP and other journals) is to be improved to reflect good practice and to make the most of the statistical techniques that are available for the analysis and interpretation of data, then more needs to change. We hope that our summary and analysis will encourage authors and editors to reflect on their practice and to embrace the Guidelines more fully, leading to further improvements in research planning, reporting, and data interpretation.

References

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge/Taylor & Francis.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7–29. doi:10.1177/0956797613504966

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond*. New York, NY: Routledge/Taylor & Francis.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. doi:10.3758/BF03193146

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations and interpretation. *Journal of*

Experimental Psychology: General, 141, 2–18. doi:10.1037/a0024338

Fritz, C. O., Morris, P. E., Cherchar, A., Smith, G., & Roe, C. (2015). *The bar graph bias*. Paper presented at the summer meeting of the Experimental Psychology Society, Lincoln, UK. Slides available at http://bit.do/Fritz-et-al-EPS-2015

Grissom, R. J., & Kim, J. J. (2011). *Effect sizes for research: A broad practical approach* (2nd ed.). New York, NY: Psychology Press.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Masson, M. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220. doi:10.1037/h0087426

Morris, P. E., & Fritz, C. O. (2013a). Effect sizes in memory research. *Memory, 21,* 832–842. doi:10.1080/09658211.2013.763984

Morris, P. E., & Fritz, C. O. (2013b). Methods: Why are effect sizes still neglected? *The Psychologist, 26,* 580–583.

Morris, P. E., & Fritz, C. O. (2014). The challenge of the Psychonomic Society Guidelines on Statistical Issues (2012). Poster presented at the 55th annual meeting of the Psychonomic Society, Long Beach, CA. Available at http://nectar.northampton.ac.uk/713*9/7/Morris20147139b.pdf

Morris, P. E., & Fritz, C. O. (2017, January). Statistics reported in *The Quarterly Journal of Experimental Psychology*, the journals of the Psychonomic Society, and the *Journal of Experimental Psychology: General*. Paper presented at the London Meeting of the Experimental Psychology Society. Slides available at http://bit.do/MorrisFritz-QJEP-Psychonomic-EPS2017

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review, 19,* 601–607. doi:10.3758/s13423-012-0247-5

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2017). *Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy.* Manuscript under review.

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

Psychonomic Society. (2012). *New statistical guidelines for journals of the Psychonomic Society.* Retrieved from http://www.springer.com/?SGWID=0-102-2-1390050-preview

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, 20,* 1–8. doi:10.1177/0956797611417632

Smith, G., & Morris, P. E. (2015). Building confidence in confidence intervals. *The Psychologist, 28,* 476–479.