Journal of the
Experimental Analysis of Behavior

JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 2019, **111,** 155–165 NUMBER 2 (MARCH)

Beyond inference by eye: Statistical and graphing practices in *JEAB*, 1992-2017

ELIZABETH G. E. KYONKA[1], SUZANNE H. MITCHELL[2], AND LEWIS A. BIZO[1]

[1]UNIVERSITY OF NEW ENGLAND
[2]OREGON HEALTH & SCIENCE UNIVERSITY

Debates about the utility of *p* values and correct ways to analyze data have inspired new guidelines on statistical inference by the *American Psychological Association* (*APA*) and changes in the way results are reported in other scientific journals, but their impact on the *Journal of the Experimental Analysis of Behavior* (*JEAB*) has not previously been evaluated. A content analysis of empirical articles published in *JEAB* between 1992 and 2017 investigated whether statistical and graphing practices changed during that time period. The likelihood that a *JEAB* article reported a null hypothesis significance test, included a confidence interval, or depicted at least one figure with error bars has increased over time. Features of graphs in *JEAB,* including the proportion depicting single-subject data, have not changed systematically during the same period. Statistics and graphing trends in *JEAB* largely paralleled those in mainstream psychology journals, but there was no evidence that changes to *APA* style had any direct impact on *JEAB.* In the future, the onus will continue to be on authors, reviewers and editors to ensure that statistical and graphing practices in *JEAB* continue to evolve without interfering with characteristics that set the journal apart from other scientific journals.

Key words: confidence intervals, error bars, graphs, null hypothesis significance testing, statistical reform

In 1999, the *American Psychological Association*'s (APA) *Task Force on Statistical Inference* published a set of guidelines advising researchers that $p < .05$ should not be used as an infallible indicator of the presence of a meaningful effect (Wilkinson, 1999). The guidelines further suggested that reporting confidence intervals around effect sizes for test statistics would improve the interpretability of many exploratory and hypothesis-driven results. They also recommended graphing data and "including graphical representations of interval estimates whenever possible" (p. 601). To assess whether those guidelines may have changed statistical practices, Cumming et al. (2007) examined empirical articles published in "leading international psychology journals that publish mainly empirical research," (p. 230) in 1998, 2003-2004, and 2005-2006. The *Journal of the Experimental Analysis of Behavior* (*JEAB*) was not included in the journals that Cumming and colleagues assessed, so the present study is a quantitative content analysis of articles published in *JEAB* before and after the publication of the Task Force's guidelines, to determine whether the way quantitative information is reported in *JEAB* has changed in the same manner as in other experimental psychology journals.

In publishing guidelines on statistical inference (Wilkinson, 1999) and subsequently incorporating many of those guidelines into the fifth edition of their Publication Manual (APA, 2001), the APA catalyzed a statistical revolution in psychology that continues today (Gigerenzer, 2018). In response, several psychology journals updated instructions to authors. Some specifically requested that authors report confidence intervals or other additional information beyond *p* values (e.g., Bakeman, 2005; Erdfelder, 2010; La Greca, 2005). The journal *European Psychologist* adopted a double-blind review policy in an attempt to combat publication bias (Greve, Bröder & Erdfelder, 2013). Also, as a matter of policy, the *Journal of Basic and Applied Social Psychology* banned null hypothesis significance testing (NHST) outright (Trafimow & Marks, 2015).

Cumming et al.' (2007) examination of statistical practices investigated whether changes to results sections following publication of the

Elizabeth G. E. Kyonka, Psychology, University of New England; Suzanne H. Mitchell, Behavioral Neuroscience, Psychiatry and the Oregon Institute of Health Sciences at Oregon Health & Science University; Lewis A Bizo, Psychology, University of New England.

Address correspondence to: the first author at School of Psychology, University of New England, Armidale, NSW 2351, Australia; ekyonka@une.edu.au

doi: 10.1002/jeab.509

APA Task Force's guidelines were in line with trends in changes to journal editors' instructions to authors. They coded articles from 10 journals (listed in Appendix A) that published empirical research mostly involving group designs. For each article, the authors recorded any use of NHST, confidence intervals, and figures with error bars, three practices that could be coded reliably. They acknowledged that effect size reporting is a potentially important statistical practice that might have been affected by the guidelines, but did not record whether articles included effect sizes. Almost all articles (97.5%) included NHST, and there was no change in the percentage of articles that reported NHST over time. Inclusion of confidence intervals and figures with error bars increased with time. Confidence intervals were most often reported in tables and text; error bars in figures typically showed standard error. Although statistical reporting practices had changed, Cumming et al. (2007, p. 232) concluded that confidence intervals and error bars were generally not interpreted correctly and that meaningful statistical reform would require "further detailed guidance, examples of good practice, and editorial or institutional leadership."

Behavior analysts generally have different attitudes to statistics and data analysis than other psychological scientists, so *JEAB* authors may have responded to the Task Force guidelines differently, or not at all. The concern with NHST that prompted the statistical revolution is its susceptibility to misinterpretation and misuse. In six principles outlined in a recent statement on *p*-values, the *American Statistical Association* (Wasserstein & Lazar, 2016, p. 132) described what *p*-values are and how they are often misinterpreted, concluding "by itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis." The earliest insights from experimental analyses of behavior (e.g., Skinner, 1938) relied on interpretations of cumulative records that showed each response from every subject. This approach formed the foundation of behavior analysis as a natural science that is distinct from psychology and related disciplines, with its own methodological strategies (Sidman, 1960). Behavior analysts have traditionally been more skeptical of and less likely to use NHST (e.g., Branch, 1999; 2014; Michael, 1974), preferring to establish the reliability and validity of results through experimental control (Perone, 1999) and replication (Branch, 2018).

Documented skepticism about NHST does not mean that behavior analysts never use it. Psychological scientists have always been aware of the problems with NHST, but until the onset of the statistical revolution in psychology it was the expected data-analytic approach (Cohen, 1994; Nickerson, 2000). Zimmermann, Watkins and Poling (2015) reported that more than half of all articles published in *JEAB* in 2005-2010 included "an inferential statistic." Zimmermann et al. did not specify NSHT, so presumably some of those inferential statistics were reported in the course of curve fitting and other forms of quantitative model development and comparison. Therefore, recording the rate of NHST in *JEAB* articles is important for assessing trends in *JEAB* statistical practices.

Even if an experiment is not hypothesis-driven and does not involve NHST, reporting precision and variability is a necessary component of any comprehensive summary or synthesis of quantitative empirical results. Confidence intervals estimate the precision of parameter estimates, effect sizes and other measurements (Cumming & Finch, 2005). Error bars, which typically represent the standard error of the mean, standard deviation, interquartile range or 95% confidence interval, all represent variability graphically (Lane & Sándor, 2009). All are potentially useful for determining whether responding is stable, gauging the amount of variability in behavior within or between subjects, and for evaluating the efficacy of interventions or experimental manipulations. Behavior analysts often rely on visual inspection of single-subject graphs, but agreement across individuals interpreting those graphs is often low (DeProspero & Cohen, 1979; Diller, Barry & Gelino, 2016), though it is generally higher among behavior analysts with more training and recognized expertise (Kahng et al., 2010; Vanselow, Thompson & Karsina, 2011). Like other researchers, behavior analysts would likely benefit from guidance, examples and leadership when it comes to improving statistical practice. Surely, any such guidance will be most effective if it is evidence-based.

The objective of the present study was to assess trends in reporting statistics and

variability in *JEAB* before and during the ongoing statistical revolution in psychology. Like Cumming et al. (2007), we recorded the inclusion of NHST, confidence intervals and figures with error bars. Although we considered recording reported effect sizes, ultimately, we decided against it. The *APA* has provided specific guidelines about how to report NHST and confidence intervals and their appearance in *APA*-style manuscripts tends to be fairly uniform. Compared to *p* values and confidence intervals, there is wide variety in the possible effect size statistics that can be reported for any particular quantitative analysis. There is little standardization in the manner in which they appear. At a glance, it can be difficult to discriminate an unstandardized effect size from a sample statistic. This degree of variety and lack of standardization makes effect sizes less suited to the type of analysis used here.

Visual inspection of graphs is a prominent means of presenting and interpreting results in the experimental analysis of behavior, so we also recorded the number of figures per article and number of panels per figure. We classified each by type using previously established categories (Best, Smith & Stubbs, 2001; Peden & Hausmann, 2000). According to the journal's masthead, *JEAB* "is primarily for the original publication of experiments relevant to the behavior of individual organisms." To assess trends in the illustration of single-subject data, we recorded whether each figure included data from individual organisms and when present, whether error bars illustrated variability in individual or group data.

Method

Samples

To examine statistical and graphing practices in *JEAB*, we conducted a quantitative content analysis of empirical articles published in *JEAB* over six time periods before and during the ongoing statistical revolution in psychology. Following Cumming et al. (2007), we coded the first 40 empirical articles published in *JEAB* in 1992, 1997, 2002, 2007, 2012 and 2017. An "empirical article" was defined as any article that reported quantitative data in any format, regardless of the article's classification (e.g., Commentary, Original Research, Technical Article, Theoretical & Conceptual Review). The data reported in the article did not need to be original or have been generated from experiments or observations involving animals or humans—some articles reported simulated data and secondary data analyses.

Electronic copies of each article were either obtained from the PubMed Central archive for the journal (https://www.ncbi.nlm.nih.gov/pmc/journals/299/) or directly from the Wiley Online Library (https://onlinelibrary.wiley.com/journal/19383711). Interobserver agreement on the 40 articles to be included for each of four time periods was 97.5%. One of the discrepancies was due to a mislabeled article in the PubMed table of contents. The other three were articles that included very few quantitative data points; in each case one coder did not initially identify them as empirical articles. In 1997 and 2012, *JEAB* published fewer than 40 articles that matched our operational definition of empirical. To keep sample sizes the same for all time periods, we included the first two empirical articles *JEAB* published in 1998 with the 1997 sample, and the first three empirical articles from 2013 with the 2012 sample.

Procedure

Content analysis and operational definitions. Coding an article involved recording whether it included any NHST, confidence intervals, figures with error bars, the number of figures, and several details about each figure. An article was coded as including NHST if it included any test statistic with a *p* value or any written statement about the statistical significance of an effect. An article was coded as including confidence intervals if confidence intervals were reported anywhere in the article, including text, figure captions and figures. It was coded as including a figure with error bars if any part of any figure had error bars.

We classified each figure in every article by type. Following previous content analyses of psychology textbooks (Peden & Hausmann, 2000) and journal articles (Best et al., 2001), we classified figures as line graphs, bar graphs, scatterplots, or frequency distributions (including histograms and cumulative frequency distributions). Visual displays of quantitative information that did not fit any of those four categories were classified as 'other data visualization,' and the coder recorded a brief

description of the figure. When a figure included a combination of different types of graphs it was also classified as 'other data visualization,' and the coder noted the types of graph. Figures showing procedural diagrams, sample stimuli, equipment schematics, or any illustration that did not include graphical representation of quantitative data were classified as 'other, not data visualization.' Because the objective of this study was to examine text and graphics related to statistics and variability, we did not record any additional details about those figures.

For each figure, we also recorded the number of panels, whether the figure illustrated single-subject data, and whether there were any error bars shown in the figure. In multipanel figures, the number of panels was determined based on the number of unique sets of axes. A graph showing multiple phases separated by phase change lines was one panel; a figure with similar information organized into multiple graphs, each with their own axes, was multiple panels. Any figure that included any data drawn from a single subject, including but not limited to figures that showed both aggregated and individual-subject data, was coded as illustrating single-subject data. Because some figures showed data from single subjects without error bars and group mean data with error bars (e.g., Fig. 1 of Beeby & Alsop, 2017), if a figure contained both single-subject data and error bars, we also recorded whether the error bars were attached to a single-subject data point.

Interobserver agreement. Eighty-nine articles, 37% of the whole sample, were independently coded by two observers. Agreement on whether articles included NHST, confidence intervals, figures with error bars and number of figures was 93.5%. The 89 articles coded by two observers contained 541 figures (36% of all figures). Agreement on the type of figure, number of panels, whether the figure showed single-subject data, and whether the figure included error bars was 88.7%. Instances in which the two observers coded different values were resolved through discussion prior to data analysis.

Data analysis. We used regressions to evaluate trends in statistical practices over time rather than *t* tests to compare statistical practices before and after the publication of Task

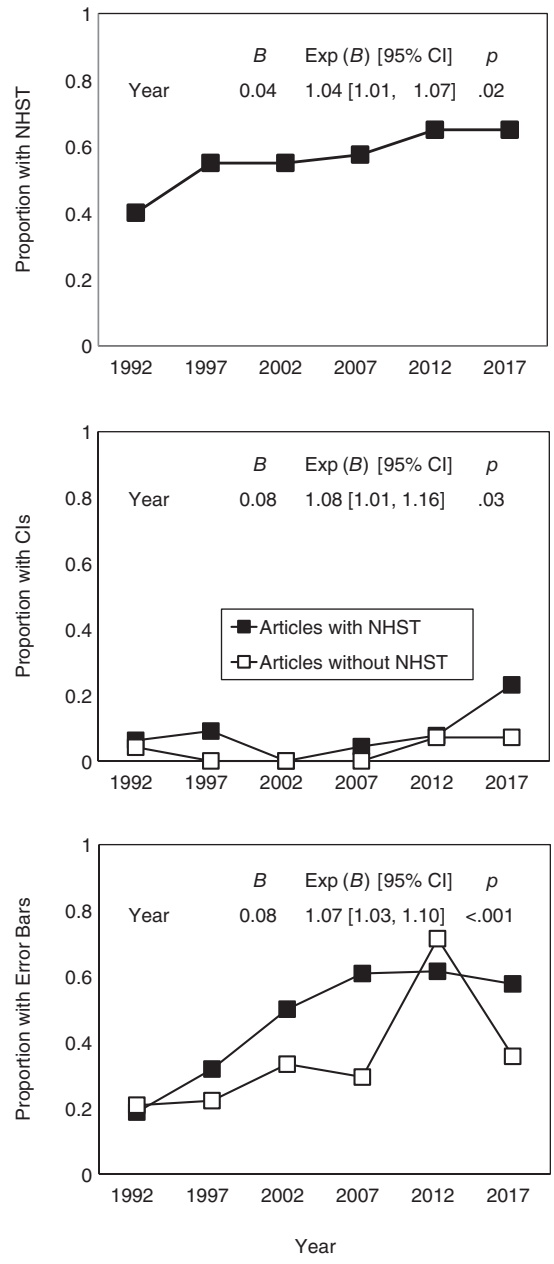

Fig. 1. *Top:* Proportion of empirical articles published in the *Journal of the Experimental Analysis of Behavior* (*JEAB*) that included null hypothesis significance testing (NHST) by year, $R^2 = .02$ (Cox & Snell), .03 (Nagelkerke), model $\chi^2(1) = 5.88$, $p = .02$. *Middle:* Proportion of empirical articles in *JEAB* with confidence intervals (CI) by year, $R^2 = .02$ (Cox & Snell), .06 (Nagelkerke), model $\chi^2(1) = 5.43$, $p = .02$. *Bottom:* Proportion of empirical articles in *JEAB* that included any figures with error bars by year, $R^2 = .07$ (Cox & Snell), .09 (Nagelkerke), model $\chi^2(1) = 16.98$, $p < .001$.

Force guidelines for two reasons. First, the publication of Task Force guidelines in 1999 was not an isolated event. The formation of the Task Force was a reaction to the perception that psychologists' attitudes towards statistics were evolving. Some editorial policies and statistics textbooks adapted to the recommendations quickly, others took longer. Comparing practices before and after 1999 could mask a delayed reaction or misattribute changes that were part of a longer-term trend. Second, using regressions facilitated a more direct comparison to the results of Cumming et al. (2007).

Results

Trends in Statistical Practices in *JEAB* Over Time

In total, 135 (56.3%) of the 240 *JEAB* articles included in the sample reported NHST, 15 (6.3%) included confidence intervals, and 101 (42.1%) included at least one figure with error bars. Figure 1 shows the proportion of articles that included NHST, confidence intervals, and figures with error bars for each year. It also includes the results of binary logistic regressions of (1) NHST on year, (2) confidence intervals on year and NHST, and (3) inclusion of figures with error bars on year and NHST. Separate binary logistic regressions indicated significant increases in each statistical practice over time. The proportion of articles reporting NHST increased by 4% every 5 years. The magnitude of the trend is consistent with the increase in articles that included inferential statistics reported by Zimmermann et al. (2015). Although few articles included confidence intervals, the expected increase in the proportion of articles reporting confidence intervals every 5 years was 2.5%. Inclusion of NHST did not predict whether the article included confidence intervals or error bars.

The proportion of articles that included figures with error bars increased monotonically (though nonlinearly) over time by 6% every 5 years on average. A total of 327 figures with error bars appeared in 101 different empirical articles. Error bars were described as confidence intervals in 19 figures, standard deviation in 77 figures and standard error in 152 figures. Error bars were related to range in 46 figures (i.e., they represented total range, interquartile range or semi-interquartile range) and represented some other measure of variability in six figures. Error bars were not labeled in the figure, figure caption or text for 33 figures (10.9% of all figures with error bars). The frequency of such omissions varied over time (13.6%, 11.4%, 3.2%, 18.4%, 9.2%, and 9.8%, in 1992, 1997, 2002, 2007, 2012 and 2017, respectively).

Figure 2 shows the number of figures per article and number of panels per figure for articles from each year. We conducted separate Poisson regressions using year as predictor with (1) the number of figures per article and (2) number of panels per figure as outcome variables. The number of figures per article ($M = 6.30$, $SD = 3.93$) increased from $M = 5.00$ (95% $CI = [3.98, 6.25]$) in 1992 to $M = 7.05$ (95% $CI = [5.73, 8.37]$) in 2012. The number of panels per figure ($M = 5.42$, $SD = 5.61$) decreased from $M = 6.72$ (95% $CI = [5.90, 7.54]$) in 1997 to $M = 4.60$ (95% $CI = [4.08, 5.11]$) in 2017. Although the regressions indicated that changes from 1992 to 2017 were statistically significant, they were very small.

Figure 3 shows the frequency of each type of figure by year. The marginal percentage of figure types changed over time, $\chi^2 (25) = 94.05$, $p < .001$. As in other journals (Best et al., 2001) and psychology textbooks (Peden & Hausmann, 2000), line graphs were the most frequently used, comprising 37-51% of the figures in the time period. The number of bar charts published increased every year except 2017. The number of scatterplots was fairly stable; each set of 40 articles included 35-44 figures showing scatterplots. The number of frequency distributions, other types of data visualizations and figures that did not show data varied from year to year without systematic trends over time.

Figures from Articles with and without NHST

Associations between different statistical and graphing practices may exist regardless of whether there were changes in those practices over time. To determine if the statistical practice of reporting NHST was associated with any differences in graphing practices in *JEAB*, we compared figures from articles that included NHST with those from articles that

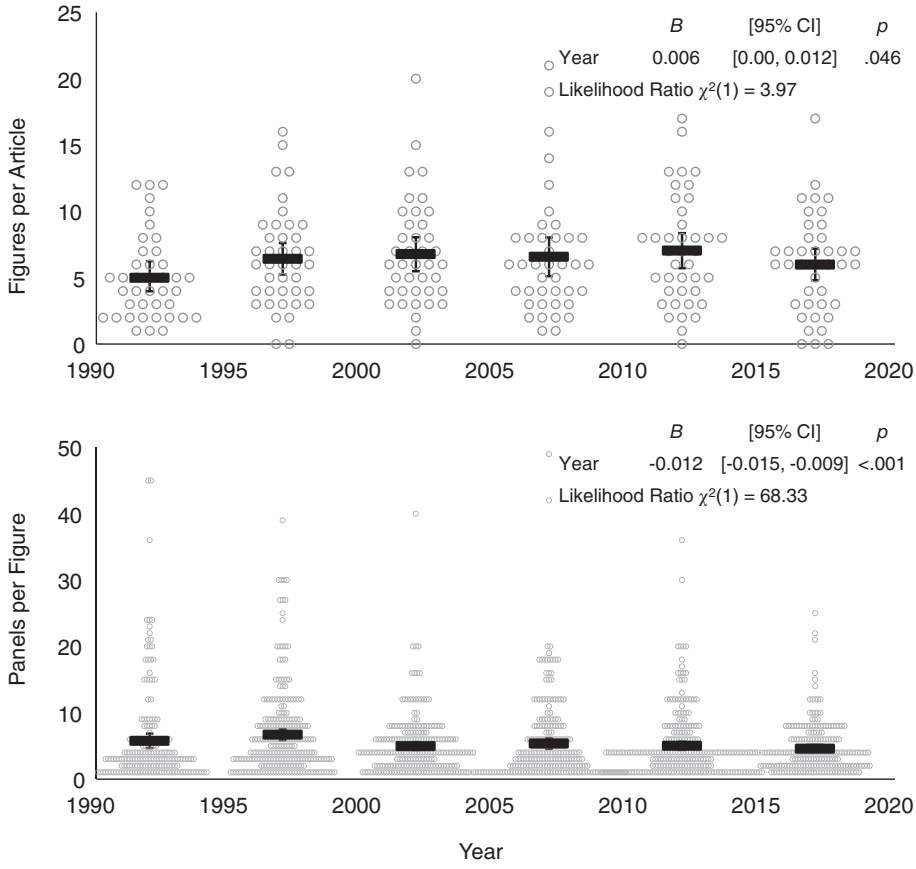

Fig. 2. *Top:* Number of figures included in empirical articles published in the *Journal of the Experimental Analysis of Behavior* (*JEAB*) by year. Grey open circles represent individual articles. Parameter estimate (*B*) and 95% confidence intervals (CI) are from Poisson regression of figures per article on year. Pearson's $\chi^2(238) = 580.71$, with a value/df ratio of 2.44, which indicates that the model fitted the data well. *Bottom:* Number of panels included in each figure of *JEAB* empirical articles by year. Grey open circles represent individual figures. Parameter estimate (*B*) and 95% confidence intervals (CI) are from Poisson regression of panels per figure on year. Pearson's $\chi^2(1318) = 7415.45$ with a value/df ratio of 5.63, which indicates that the model fitted the data well. In both panels, black lines with error bars show mean and 95% CI.

did not. Articles that included NHST had different types of figures, but the same number of figures per article and panels per figure as articles without NHST. There was no difference between the number of figures in articles with NHST ($M = 6.44$, $SD = 4.00$) and the number in articles without NHST ($M = 6.13$, $SD = 3.85$), a difference between means of 0.30 figures, 95% CI [-0.71, 1.31]. The number of panels in figures from articles with NHST ($M = 5.36$, $SD = 5.24$) was similar to the number of panels in figures from articles without NHST ($M = 5.50$, $SD = 6.12$), a difference between means of 0.14 panels, 95% CI -0.48, 0.76]. Scatterplots and line graphs were more likely to come from articles with NHST, with odds ratios of 1.23 and 1.38, respectively. Frequency distributions, the category that included histograms and cumulative records, were about equally likely to come from articles with and without NHST (*Odds Ratio (OR)* = 0.95). 'Other' data visualizations, including figures that combined multiple graph types, were as well ($OR = 0.84$). Non-data figures ($OR = 0.56$), and bar charts ($OR = 0.78$) were less likely to come from articles with NHST. Table 1 shows the number of figures of each type from articles with and without NHST.

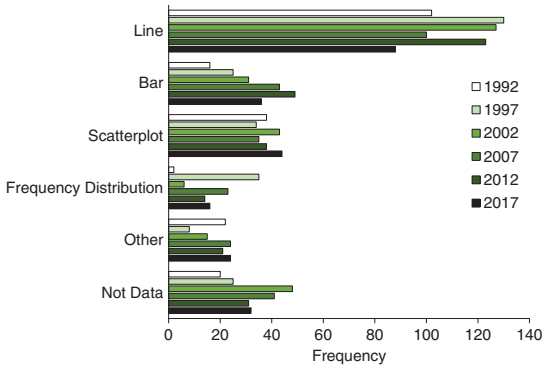

Fig. 3. Frequency of different types of figures included in empirical articles published in the *Journal of the Experimental Analysis of Behavior* (*JEAB*) by year.

Single-Subject Data in JEAB Figures

An emphasis on single-subject data is a distinguishing feature of *JEAB*, so we compared the relative frequency of graphical features in figures that did and did not show any single subject data. Table 2 shows the frequency of various features in these two types of figures. Figures that did not show any data at all, such as illustrations of sample stimuli, procedural diagrams and equipment schematics, were excluded from these analyses. Data visualizations that did not show any single-subject data were 1.54 times more likely to come from articles that included NHST, χ^2 (1) = 12.53, $p < .001$, but were neither more nor less likely to include error bars, χ^2 (1) = 0.29, $p = .86$, $OR = 1.02$ than figures that included single-subject data. Of the 210 figures that included both single-subject data and error bars, 35 presented error bars only on aggregated data, so error bars were somewhat less likely to be used to illustrate variability in single-subject samples or precision of parameters fitted to individual data. Nevertheless, the independence of error bars and single-subject data in figures is evidence that the increase in error bars has not occurred at the expense of illustrations of single-subject results. Figure type was also independent of whether the figure showed single-subject data, χ^2 (4) = 8.41, $p = .08$.

Figures that showed only group data were more likely to be from articles that included NHST. Reporting hypothesis tests on group data instead of analyzing the behavior of individual organisms would be a move away from the unique research strategy developed in the early days of the experimental analysis of behavior, so we counted the number of articles from each year that included NHST but no figures with single-subject data. There were 2, 4, 4, 7, 5, and 3 such articles in successive time periods, constituting 10.4% of the total sample. Changes over time were unrelated to the steady small increase in the proportion of articles with NHST, which suggests that that trend has not occurred at the expense of including single-subject graphs.

Table 1

Figures from articles with and without NHST by type and year

	Type						
Year	Line	Bar	Scatterplot	Frequency Distribution	Other	Not Data	Total
Figures from articles without NHST ($N = 105$)							
1992	50	6	24	2	16	17	115
1997	61	12	13	16	4	15	121
2002	49	20	17	2	12	24	124
2007	24	17	22	10	5	21	99
2012	42	19	10	8	14	16	109
2017	30	22	2	4	2	15	75
Total (% of all figures)	*256 (17.0)*	*96 (6.4)*	*88 (5.8)*	*42 (2.7)*	*53 (3.5)*	*108 (7.2)*	
Figures from articles with NHST ($N = 135$)							
1992	52	10	14	0	6	3	85
1997	69	13	21	19	4	10	136
2002	78	11	26	4	3	24	146
2007	76	26	13	13	19	20	167
2012	81	30	28	6	7	15	167
2017	58	14	42	12	22	17	165
Total (% of all figures)	*414 (27.4)*	*104 (6.9)*	*144 (9.5)*	*54 (3.6)*	*61 (4.0)*	*89 (5.9)*	

Table 2

Features of figures that did and did not show single-subject data

		Type							
Year	Total	Line	Bar	Scatter	Freq	Other	No Error Bars	Error Bars	Mean panels [95% CI]
Figures showing single-subject data (N = 891)									
1992	122	75	14	19	2	12	105	17	6.75 [5.29, 8.21]
1997	184	92	19	32	34	7	160	24	8.11 [7.14, 9.09]
2002	145	74	27	27	5	12	106	39	5.81 [5.01, 6.61]
2007	117	57	23	19	13	5	99	18	6.37 [5.25, 7.48]
2012	172	87	32	27	14	12	100	72	6.10 [5.30, 6.91]
2017	151	61	22	38	8	22	107	44	5.28 [4.66, 5.91]
Total	891	446	137	162	76	70	677	214	
Data visualizations that did not include any single-subject data (N = 420)									
1992	58	27	2	19	0	10	52	5	3.71 [2.51, 4.90]
1997	47	38	6	2	1	0	36	11	2.32 [1.74, 2.90]
2002	77	53	4	16	1	3	53	24	3.43 [2.79, 4.06]
2007	108	43	20	16	10	19	77	31	4.31 [3.19, 5.42]
2012	73	36	17	11	0	9	47	26	2.49 [1.95, 3.04]
2017	57	27	14	6	8	2	40	17	2.77 [2.05, 3.50]
Total	420	224	63	70	20	43	305	114	

Note. Scatter = Scatterplot; Freq = Frequency Distribution; CI = confidence interval.

Discussion

The statistical practices included in *JEAB* have changed since 1992. There are more articles including null hypothesis tests, confidence intervals, and figures with error bars in *JEAB* than there used to be. At the same time, the number and types of figures that appear in *JEAB* articles has remained relatively stable. Consistent with tradition and reputation, the *JEAB* articles in our sample typically included several figures; only seven (3%) did not include any figures. Most figures, 87%, depicted data, and of those, 68% graphed data from individual subjects.

Comparison with Leading Mainstream Psychology Journals

In spite of the philosophical and methodological differences between the experimental analysis of behavior and mainstream psychology, the changes in *JEAB* statistical practices are generally similar to those reported by Cumming et al. (2007) in other psychology journals.

It would be reasonable to predict that the Task Force guidelines and other critiques of NHST would function to decrease the rate of NHST appearing in peer-reviewed journals, but that was not the case in mainstream psychology (Cumming et al., 2007) nor in *JEAB*. We found that the proportion of *JEAB* articles containing NHST has risen continuously since 1992. Other research (Zimmermann et al., 2015) suggests that the proportion of articles that include p values has been increasing since the publication of the first issue. Nonetheless, NHST appeared at a much lower rate in *JEAB* between 1992 and 2017 than it did in leading psychology journals 1998-2006, where it was nearly ubiquitous (Cumming et al., 2007). Figures from *JEAB* articles with NHST were somewhat less likely to include single-subject data. Articles that report inferential statistics from group data instead of analyses of the behavior of individual organisms constitute a move away from the unique research strategy developed in the early days of the experimental analysis of behavior, one that many behavior analysts might well regret (Skinner, 1976). However, the proportion of *JEAB* articles that included figures with single-subject data has not decreased as NHST has increased. In addition, *JEAB* articles that included NHST were similar to articles that did not in most other respects. They were no more or less likely to include confidence intervals or figures with error bars, and they included the same number of figures with the same number of panels per figure. These similarities suggest that NHST augments other analytic approaches in *JEAB* rather than replacing them. To the

extent that scientific communities benefit from greater variety in quantitative analyses, the increase in NHST in *JEAB* could be considered to enrich the journal if it is done in a logically consistent manner (Haig, 2017). At any rate, we contend the lack of a decrease in NHST use is not the cause for concern Cumming et al. (2007) viewed it to be.

Confidence intervals were rarely reported in leading mainstream psychology journals (Cumming et al., 2007) or in *JEAB,* but the practice increased over time at a similar rate in both. Cumming and colleagues expressed concern about the way confidence intervals were reported in the articles they evaluated. Confidence intervals were not presented as estimates of the width or precision of effects, as encouraged by proponents of the "new statistics" (Cumming & Calin-Jageman, 2016; Cumming & Finch, 2005). When confidence intervals were reported in psychology journals, they were most frequently not interpreted at all, or they were used to support the conclusions of NHST (e.g., a confidence interval that did not include zero might be used to support the conclusion that two means are statistically significantly different). In *JEAB,* authors' use of NHST and confidence intervals were independent, which suggests that confidence intervals were not reported exclusively in relation to NHST. Even if they had been, supporting hypothesis test results with confidence intervals and other quantitative details, rather than relying on *p* values alone, is very much in line with current recommendations of the APA (APA, 2010; Wilkinson, 1999) and the American Statistical Association (Wasserstein & Lazar, 2016).

The proportion of articles that included at least one figure with error bars increased moderately in leading mainstream psychology journals (Cumming et al., 2007) and in *JEAB.* Error bars are useful visual summaries that can aid interpretation of data when presented appropriately (Cumming & Finch, 2005). Of course, including error bars does not guarantee they will be interpreted correctly. There is some evidence (Belia, Fidler, Williams & Cumming, 2005) that behavioral scientists struggle to interpret error bars correctly even when all the necessary information is made available, and that within-subject designs pose additional difficulties. A clear description of what the error bars are (e.g., confidence interval, standard error, interquartile range) is essential for correct interpretation. Cumming et al. (2007) reported that one third of figures with error bars were not labeled in their sample of empirical psychology research, whereas we found that only 10.9% of error bars in *JEAB* figures were not labeled. Labels do not guarantee that error bars were used appropriately, but the lower proportion of unlabeled error bars in *JEAB* suggests that, compared to authors who publish in other journals, *JEAB* authors were less likely to include error bars that readers could not interpret at all.

The Role of the APA Task Force

If increases in reporting confidence intervals, showing error bars, or (perhaps paradoxically) using NHST were a function of the publication of the APA Task Force on Statistical Inference's recommendations (Wilkinson et al., 1999) and their subsequent incorporation into the fifth edition of the APA's publication manual (APA, 2001), one would expect that in our sample, the greatest changes would occur between 1997 and 2002, or perhaps in the following period. There was no indication of particularly dramatic changes to any of the statistical practices we observed during those 10 years. Instead, the changes in the proportion of *JEAB* articles with NHST, confidence intervals, and error bars have been relatively gradual and steady, both before and after the Task Force was convened. In the absence of evidence to the contrary, we suspect that the changes in *JEAB* are likely a reflection of a zeitgeist and adoption of these methods in scientific research generally, rather than resulting from a specific reaction to any particular APA publication.

Although the Task Force recommended including error bars "whenever possible" (Wilkinson, 1999, p. 601), error bars are not universally useful or necessarily the best way to illustrate intervals. Many experimental analyses of behavior involve absolute or relative frequencies for which error bars would not provide additional useful information (e.g., Fig. 1). Moreover, there are other, more information-rich ways of representing variability and precision that are sometimes preferable to error bars (Lane & Sándor, 2009). Although we did not formally record them, we noted several recent examples of beeswarm

plots (e.g., Fig. 2) and other types of graphs that plotted every observation. We contend that when it is possible to do so without obscuring meaning, graphing every observation makes better use of available space (Tufte, 2001) than showing error bars alone can do.

The APA's instructions to include estimates of variability because they are necessary for the reader to "corroborate the analyses conducted" (APA, 1994, p. 16) predate the establishment of the *Task Force on Statistical Inference*. In each of the six years we evaluated, there were single-subject figures that showed means without any associated measure of variability. *JEAB*'s longstanding tradition of showing quantitative summaries of each subject's behavior in separate panels of the same figure dates back to the journal's first issue (Conrad, Sidman, & Herrnstein, 1958; Hearst, 1958). When this kind of figure illustrates data that were analyzed using a group design, graphing the behavior of each subject in a separate panel is a valid (if somewhat unusual) way of illustrating variability. When the research design involves an experimental analysis of the behavior of individual organisms, estimates of variability should be provided for each mean or other quantitative summary reported for every subject—something *JEAB* authors do not currently do consistently.

Summary and Conclusion

So far, statistical and graphing practices in *JEAB* seem to be evolving along with those in mainstream psychology without sacrificing the journal's emphases on data visualization or single-subject analysis. The analyses presented here suggest a few potential areas of improvement for behavior analysts (and mainstream psychologists). Currently, behavior analysts do not report confidence intervals as a means of demonstrating the precision of results (Cumming & Finch, 2005), integrate statistical copy with figures (Lane & Sándor, 2009), or explicitly evaluate the power and severity of inferential hypothesis tests (Haig, 2017) as often as might be useful. Any of these practices has potential to improve the transparency and interpretability of some experimental analyses published in *JEAB*. Authors, reviewers, and editors might particularly consider whether estimates of variability are included for every parameter estimate or other quantitative summary, including response rates. Doing so is as important for single-subject analyses as it is for group designs.

Best practices for reporting statistics and graphing results are not static. They can be expected to evolve as research priorities shift, new software is developed, and philosophical debates among scientists and statisticians continue. We found no evidence that the changes that have occurred since 1992 were specifically related to the recommendations of the APA Task Force on Statistical Inference. If statistical or graphing practices in *JEAB* are to change course in the future (e.g., to reverse trends in NHST or to ensure that estimates of variability are reported), behavior analysts may need to intervene directly and not rely on the evolution of statistical thinking in the broader scientific community. It is often said that what gets measured gets managed, and what gets managed gets done. Will the publication of a special issue on modern statistical practices in behavior analysis have more of an impact on the behavior of *JEAB* authors than the APA Task Force did? It's an empirical question.

References

American Psychological Association (1994). *Publication manual of the American Psychological Association.* (4th ed.). Washington, DC: Author.

American Psychological Association (2001). *Publication manual of the American Psychological Association.* (5th ed.). Washington, DC: Author.

American Psychological Association (2010). *Publication manual of the American Psychological Association.* (6th ed.). Washington, DC: Author.

Bakeman, R. (2005). Editorial note: Infancy asks that authors report and discuss effect sizes. *Infancy, 7,* 5-6, https://doi.org/10.1207/s15327078in0701_2

Beeby, E., & Alsop, B. (2017). Choosing among multiple alternatives: Relative and overall reinforcer rates. *Journal of the Experimental Analysis of Behavior, 108,* 204-222.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10,* 389-396.

Best, L. A., Smith, L. D., & Stubbs, D. A. (2001). Graph use in psychology and other sciences. *Behavioural Processes, 54,* 155-165.

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst, 22,* 87-92.

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology, 24,* 256-277. https://doi.org/10.1177/0959354314525282

Branch, M. N. (2018). The "Reproducibility Crisis:" Might the methods used frequently in behavior-analysis

research help? *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-018-0158-5

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Conrad, D. G., Sidman, M., & Herrnstein, R. J. (1958). The effects of deprivation upon temporally spaced responding. *Journal of the Experimental Analysis of Behavior, 1*, 59-65.

Cumming, G., & Calin-Jageman, R. (2016). Introduction to the new statistics: Estimation, open science, and beyond. Routledge.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., ... Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*, 230-232.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.

Diller, J. W., Barry, R. J., & Gelino, B. W. (2016). Visual analysis of data in a multielement design. *Journal of Applied Behavior Analysis, 49*, 980-985.

Erdfelder, E. (2010). A note on statistical analysis. *Experimental Psychology, 57*, 1-4. https://doi.org/10.1027/1618-3169/a000001

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science, 1*, 198-218.

Greve, W., Bröder, A., & Erdfelder, E. (2013). Result-blind peer reviews and editorial decisions. *European Psychologist, 18*, 286-294. https://doi.org/10.1027/1016-9040/a000144

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement, 77*, 489-506.

Hearst, E. (1958). The behavioral effects of some temporally defined schedules of reinforcement. *Journal of the Experimental Analysis of Behavior, 1,* 45-55.

Kahng, S. W., Chung, K. M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 43*, 35-45.

La Greca, A. M. (2005). Editorial. *Journal of Consulting and Clinical Psychology, 73,* 3–5.

Lane, D. M., & Sándor, A. (2009). Designing better graphs by including distributional information and integrating words, numbers, and images. *Psychological Methods, 14*, 239-257.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7,* 647-653. https://doi.org/10.1901/jaba.1974.7-647

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241.

Peden, B. F., & Hausmann, S. E. (2000). Data graphs in introductory and upper level psychology textbooks: A content analysis. *Teaching of Psychology, 27*, 93-97.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22,* 109-116.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York, NY: Basic Books.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

Skinner, B. F. (1976). Farewell, my LOVELY! *Journal of the Experimental Analysis of Behavior, 25*, 218.

Trafimow, D., & Marks, M. (2015). Publishing models and article dates explained. *Basic Applied Social Psychology, 37*, 1.

Tufte, E. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Vanselow, N. R., Thompson, R., & Karsina, A. (2011). Data-based decision making: The impact of data variability, training, and context. *Journal of Applied Behavior Analysis, 44*, 767-780.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *The American Statistician, 70*, 129-133.

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Zimmermann, Z. J., Watkins, E. E., & Poling, A. (2015). *JEAB* research over time: Species used, experimental designs, statistical analyses, and sex of subjects. *The Behavior Analyst, 38*, 203-218.

Received: December 2, 2018
Final Acceptance: January 31, 2019
Editor in Chief: Michael Young
Associate Editor: Sarah Cowie

Appendix A

Journals assessed by Cumming et al. (2007):
Acta Psychologia
Child Development
Cognition
Journal of Abnormal Child Psychology
Journal of Abnormal Psychology
Journal of Consulting and Clinical Psychology
Journal of Experimental Psychology, General
Journal of Personality and Social Psychology
Psychological Science
Quarterly Journal of Experimental Psychology
(Section A only, before 2006)