2019

A snapshot look at replication and statistical reporting practices in psychology journals

Abdulrazaq Imam

Melissa Frate

A snapshot look at replication and statistical reporting practices in psychology journals

ABSTRACT

Current efforts started in 2012 by the Association for Psychological Science (APS) appear to be different from previous arguments against null hypothesis statistical testing (NHST), which remained largely rhetorical without specific actions for compliance by researchers in psychology. The APS advocacy involves specific promising implementation tactics. The present study examined the impact of those efforts on replication and statistical reporting practices in four psychology journals from 2011 and 2015. The results showed that amidst increased reporting of NHST statistics in 2015 compared to 2011 and an absence of power reporting in the behavioral journals, there was increased reporting of actual replications in *Psychological Science*, paradoxically surpassing *Journal of the Experimental Analysis of Behavior*, of CIs in all four journals, and of error bars on graphs in *Cognition* and *Behavioural Processes*. These trends suggest need for additional efforts at propagating the APS initiatives to ensure greater impact in the broader psychological community. Additionally, psychologists from all domains need to become advocates of best practices for sustainable impact.

Many introductory psychology textbooks open by describing the various approaches and perspective in psychology. They often conclude that the common thread that holds the various parts together is research method and the goals of the discipline as a science (e.g., Bernstein, Penner, Clarke-Stewart, & Roy, 2007; Griggs, 2017; see also Myer & Dewall, 2017). It is alarming and disheartening, therefore, to find that the field is embroiled in what amounts to a methodology crisis. The crisis has been brewing for decades now. There have been many warning calls from various people, both within and outside the field, about what is wrong with our research practices, most especially concerning excessive reliance on null hypothesis statistical testing (NHST) and the attendant publication bias (e.g., Bohannon, 2014; Fidler, Thomason, Cumming, Finch, & Leeman, 2004). Publication bias is just one part of Townsend's twin problems of persistent biases in psychology, the other being "the heavy bias against replication" (2008, p. 4) stemming from prejudicial treatment of direct

replications by publication editors. There has been, also, a variety of replication failures of important psychological research (e.g., Alcock, 2011; Bem, 2011; Cesario, 2014; Ritchie, Wiseman, & French, 2012). This is the backdrop against which one maps the research terrain in psychological science. A focal point of criticism besides the blind adoption of NHST, of course, is the pervasive misinterpretations of what the *p*-value represents, ranging from the misconception that it indicates an amount of certainty or confidence about the results of an experiment to the mistaken belief that it signifies replicability upon repetition of an experiment (Lambdin, 2012). Editors, reviewers, and authors commonly fall victim to such misconceptions, inadvertently contributing to publication bias, and further weakening the cumulative foundation of the science of psychology.

Broadly speaking, two disparate research approaches, namely, large-*N* group designs, with their various complexities, and single-subject (Gast & Ledford, 2014; Sidman, 1960; henceforth, small-*N*) designs, dominate psychology. This distinction probably is unique today to psychology and allied sciences, although it has a long history in medicine (Bernard, 1927/1957). The two approaches differ in very important ways and by tradition. One of the major differences is the heavy reliance on inferential statistics in the group designs than in small-*N* designs. Another important difference is the built-in emphasis on replication and reproducibility of effects in small-*N* designs than in the large-*N* group designs. These particular differences suggest a diametrically opposed impact of the methodology crisis in the different specializations within psychology that tend to adopt them, namely, group designs in cognitive psychology and small-N designs in behavioral psychology. One could say without prejudice that what the one is doing or experiencing because of the crisis is of no concern to the other, because of the apparent gulf between the two approaches in psychology. In fact, however, behavior analysts have been concerned and interested in the use of inferential statistics as attested to by the debates, albeit from a behavioral perspective, on the pages of one of their "trade" journal, namely, *The Behavior Analyst* (e.g., see Baron, 1999; Branch, 1999; Crosbie, 1999; Shull, 1999). The question is: if research method has been and continues to be the unifying force that binds psychology together as a science, can we afford to remain dispassionately indifferent to what goes on in the different domains of psychology?

There is a different kind of attention to the persistent crises of replication and dependency on NHST in psychological research that is underway. Previously, there has been an ongoing focus on the controversy of using and overusing, perhaps to the level of abuse (cf. Goldstein, 1984), of NHST in psychology (Harlow, Mulaik, & Steiger, 1997; Morrison & Henkel, 1970). Over the years, spanning many decades, many observers have cautioned, campaigned, and repudiated the over-reliance on NHST from various backgrounds in psychology (Falk & Greenbaum, 1995; Morrison & Henkel, 1970; Rozeboom, 1960) and beyond (e.g., Fidler et al., 2004; Kern, 2014). Arguments have ranged from firm support for (Chow, 1998; Frisk, 1996; Hagen, 1997) to suggestion of outright banning of NHST in psychological research (Schmidt & Hunter, 1997) and many more in between (e.g., Cohen, 1994; Mulaik, Raju, & Harshman, 1997; Serlin & Lapsley, 1985). It is safe to say that over half a century of clamors for change have yielded little or no change at all. Until recently, it has been just talking points, cautionary notes, rebukes, but little to no systematic action. As Thompson aptly noted, despite having devoted many pages over decades in its publications (e.g., Kline, 2013), the American Psychological Association (APA) official position as of the fourth edition of its *Publication Manual* amounted to

merely "encouraging" psychologists to report effect sizes (ESs) (Thompson, 1999, p. 68), for example (see also Fidler, 2010; Finch, Cumming, & Thomason, 2001). Subsequently, the fifth edition only provided tepid advocacy for the promotion of confidence interval (CI), graphically and otherwise (Fidler, 2010). The latest edition, improved on these only to the extent of specifying how to report CIs and ESs and emphasizing meta-analysis, and thus provides, at best, only words of "encouragements" to authors (see Fidler, 2010, p. 2; Thompson, 1999). In 2012, the Association for Psychological Science (APS) initiated a different kind of advocacy efforts to promote replications and better statistical practices. Two notable difference from previous attempts in the current efforts included 1) the specific devotion of special issues in *Perspectives in Psychological Science* (*PiPS*) dedicated to replicability and research practices in psychology (Pashler & Wagenmakers, 2012; Spellman, 2012) and 2) the adoption of the New Statistics (e.g., Cumming, 2014a) emphasizing the estimation approach over NHST in *Psychological Science* (*PS*). Editorial determination at refreshingly tangible actions appears to be the hallmark of the current efforts (Eich, 2014; Lindsay, 2015).

Prior to the inception of the efforts by the APS, others have reported on previous reporting practices in psychology (e.g., Cumming et al., 2007a; Fidler et al., 2004; Finch et al., 2001). Cumming et al., for example, found mixed results of reporting practices with persistent heavy use of NHST, low use of 95% confidence intervals (CI), increased use of standard error (*SE*) of the mean over CI, and continued use of unidentified error bars on graphs. Finch et al. (2004) reported similar trends in comparing authors who did or did not publish in *Memory & Cognition* during the editorship of Geoffrey Loftus who tried to enforce best practices in statistical reporting in that journal. Whereas 95% of these authors' works reported NHST, only 5% reported CIs. They noted that "only around 22% of articles [included CIs], about 40% . . . rely on NHST without interval estimation or visual displays of data, and another 36% included conventional figures without error bars" (Finch et al., 2004, p. 320). Finch et al. (2001) showed that historically entrenched practices of using relative as opposed to exact *p*-values continued to encourage verbal fudging of reporting nonsignificance (such as in "approaching significance," p. 201) and poor reporting and interpretation of CIs in journals like *Journal of Applied Psychology*.

Many observers of the various reporting practices of psychological research, including those of the aforementioned reports, have been critical about such practices while recommending alternative, specific best practices. Prominent among these alternative best practices, besides the strenuous objections to NHST, but some related to it, are: (1) report power (needed with or without NHST for group designs); (2), if you have to at all, report exact *p*, not relative *p*; (3) report ESs (as encouraged in the 6th edition of the APA manual); (4) use and identify error bars on graphs and do not confuse *SD*s, *SE*s, and CIs in doing so; (5) report CI instead of *SE* for inference purposes; and (6) report and interpret CIs instead of reporting *p*-values. Two of these best practices deserve further comments. First, ESs are not limited to the standard familiar ones like Cohen's *d* or Hedge's *g*. According to Cumming (2014b), ESs also include the mean and mean difference, percentage and percentage change, correlation (e.g., Pearson's *r*), proportion variance (e.g., R^2, η^2, ω^2, etc.), regression slope (*b* or β), etc. Second, often there is confusion about reporting and use of the *SD*, the *SE*, and CIs, perhaps because they are typically represented on graphs, albeit erroneously, as measures of variability on point estimates, such as the mean. Of the three, only the *SD* is a descriptive statistic of the sample, however; the CI is inferential, and SE is neither inferential. The *SE* is

the *SD* of the sampling distribution of the sample means, about twice the margin of error (or *MOE*). The CI, in contrast, represents the *MOE* below and above the point estimate and tells us something about where the population mean might reside, among other things (see Cumming, 2014b). Reporting choices among these statistics, therefore, are important in illuminating current reporting practices.

How widespread the problematic practices continue to be since 2012 when the APS reform efforts became widely publicized, therefore, is of interest. This is particularly so, in light of the seriousness of these efforts and the new rigor in their pursuit. Indeed, for example, a more recent study reported continued use of the notion of "marginally significant" to describe "*p* values between .05 and .10" in various psychological fields (Olsson-Colletine, van Assen, & Hartgerink, 2019, p. 1). What sorts of impact are these efforts having on reporting practices? Besides the editorial efforts noted above, there have been, in addition, recent educational efforts in support of changing these practices. At annual meetings of the APS, workshops on the new statistics, on open science fora, and on data handling, are frequently offered. How effective have they been? There are different approaches to answer such question. One way is to explore reporting practices of authors of articles appearing in various psychological journals for an extended period, including the period in question. This option would require in depth analyses of reporting practices say, over a decade, to examine yearly changes in those practices in multiple journals. Needless to say, that would require a great deal of resources in time and effort. Another approach is to select a few journals with varying diversity of focus and take a snapshot of the practices in select volumes. In the present study, we adopted the latter approach, examining research articles published in two behavioral (*Journal of the Experimental Analysis of Behavior* (*JEAB*) and *Behavioral Processes* (*BP*)) and two cognitive (*Cognition* and *PS*) journals right before (2011) and soon after (2015) the APS efforts became pronounced in 2012. We selected the first three journals based on the predominant subject-matter focus of the publications, two being behavioral and one being cognitive, whereas we chose the last one because, in addition to being mainly cognitive in focus, it provides a direct comparison to the other three non-APS journals, as a primary publication of the APS, the main organization advocating the current reform efforts in psychology.

The rationale for examining behavioral vs. cognitive journals is twofold. First, behavioral research tends to emphasize replication by virtue of the inherent comparisons required for experimental manipulations in small-*N* designs (Sidman, 1960; see Imam, 2018). For that reason, one would not expect the replication crisis in psychology at large to manifest in the area of behavioral research. It remains an empirical question, however, especially in light of the next consideration. Second, to the extent that authors publishing in the behavioral journals may deploy statistical analyses such as NHST that is commonly used in group designs (e.g., Zimmermann, Watkins, & Poling, 2015), in what way(s) have the concerns and tribulations about the over-reliance on NHST in psychology in general, with their pertinent recommended solutions, influenced reporting practices in behavioral research?

The present study examined how reporting of replication and statistics may have changed, in the year immediately preceding the wide dissemination of the APS efforts in 2012 and three years after, to assess the comparative impact of the efforts. We considered reports of replication, or its absence, of NHST in general, and of other specific statistical measures and considerations often recommended as essential to proper understanding of the role of statistics in psychological research; the latter included power, the use of error

bars on graphs, differentiating *SD*s and *SE*s and reporting CIs, and ESs, including those identified above for best practices. As indicated in the preceding paragraph, if replication is central to behavioral analytic research, one would expect superior replication reporting in the behavioral journals compared to the cognitive journals, on the one hand, and less reporting of statistical measures and practices that are more germane to group designs commonly used in the rest of psychology, on the other hand. If the APS efforts are having an impact, reporting and practices should reflect the recommended solutions typically offered to researchers in extant psychology; namely, more power-, CI-, error-bar reporting and less NHST-statistics in the mainstream cognitive journals.

Method

Data collection and analyses

We examined 1,157 research articles from four journals in psychology (see Table 1). Articles and entries in these journals that were excluded from analysis variously included memorials, corrigenda, errata and retractions, acknowledgements and letters, book reviews, technical notes, preface, simulations, some perspectives and commentaries, theoretical and review articles, editorials, and publisher note.

We tallied frequencies of "replication" and mentions of associated terms, including "replicable," "replicability," "reproducible," and "reproducibility," in the journals in the 2011 and 2015 volumes, excluding studies reporting "statistical replications." This approach is similar to that of Makel, Plucker, and Hegarty (2012) in which they used the "replicat*" search term. A hit on a search term in the present study triggered a check on whether an actual replication was conducted or not, and tallies of direct vs. systematic replications were made respectively (see supplements). In addition, tallies of various NHST and estimation statistics included, but were not limited to, the mean (*M*), median, and mode, the standard deviation (*SD*), range, minimum and maximum, proportions, and percentages (descriptive

Table 1. Summary of the journals analyzed with some NHST statistics including total t, F, others (χ^2, *z*, and Mann Whitney *U*).

				NHST Statistics		
Journal	Year	No. of Articles	No. of Vol/Issues	*t*	*F*	Others
Cognitive Journals						
COG	2011	149	2	9	11	0
	2015	196	2	18	19	12
Total		**345**	**4**	**24**	**30**	**12**
PS	2011	234	3	34	73	26
	2015	181	12	82	113	92
Total		**415**	**15**	**116**	**186**	**118**
Behavioral Journals						
JEAB	2011	43	4	82	104	54
	2015	50	12	130	117	85
Total		**93**	**16**	**212**	**221**	**139**
BP	2011	127	1 (12)	201	221	103
	2015	177	1 (12)	199	190	115
Total		**304**	**2 (24)**	**400**	**411**	**218**
Grand Total		**1157**		**752**	**848**	**487**

Notes. *COG: Cognition; PS: Psychological Science; JEAB: Journal of the Experimental Analysis of Behavior; BP: Behavioural Processes*

statistics); the χ^2, z, Mann Whitney U, t, and F statistics, p-value (exact vs. relative), and power (inferential statistics); and SE, CI, η^2, partial η^2, ω^2, Cohen's d, r, R^2, and variance accounted for (estimation approach). CI reporting tallies included interpretations of CI based on interval bounds, interval width, CIs overlap, or non-zero overlap. Statistics encountered but not counted included β, α, b, semi partial r^2, RMSEA, MSE, Spearman rho, Cohen's k, p_{rep}, and W.

To tally power reporting, search words included "power," "sample size," and/or "effect size." Explicit statements of power determinations were counted for each experiment in an article, usually before the results section in each journal. References to power or power-related issues, usually in results and/or discussion sections, counted as implicit reporting of power (see Fidler et al., 2004) and were specifically not counted as power (Simmons, Nelson, & Simonsohn, 2011). Many studies referred to sample-size determinations in terms of "same as in previous studies," referencing specific studies; these counted in a separate category reported here as SAP in the present study. A number of articles reported using "pilot" studies to determine sample size; one article employed "pretesting," and another used "post hoc." Encounters of "sample size determined in advance" without specifying how, or manipulations, statistical and otherwise, performed to "increase" power, did not count for power reporting. Tallies also included use of graphs and error bars, as well as reporting of missing data. The supplements accompanying the manuscript present the detail tallies for all journals.

To analyze the data, we computed the proportion of articles reporting each measure of interest and their respective 95% CIs using the ESCI program for the latest edition of Cumming's textbook (Cumming & Calin-Jageman, 2017), specifically the Two-Proportions tab of Chapters 10–16 of the software. The tab also allowed for computation of proportion differences for a given comparison of proportions of a measure. We report both of these in either graphical (for visual comparisons) and/or tabular (for completeness) forms. Two types of graphs are presented; first, those for presentation of proportions of reporting practices for each of the journals (Figures 1, 4, and 5) and those for pairwise comparisons of various journals on different measures (Figures 2, 3, and 6–7). Asterisks on the first set of graphs indicate proportion differences for the tagged comparisons did not contact or overlap with zero on the difference axis and therefore considered significantly different.

Results

In the graphs depicting differences in point and interval estimates throughout the paper, if the 95% CIs on the proportion differences do not overlap zero on the difference axis, the difference indicates a significant difference (Cumming & Calin-Jageman, 2017). The results are presented separately for replication reporting and statistical reporting.

Replication reporting

Reporting of replication was multidimensional across the four journals with three replication categories (none, actual, and a mention) in two years (2011 and 2015) as shown in Figure 1; the actual replications combined direct and systematic replications tallies. As such, multiple layers of comparisons are possible. First, Figure 1 shows that all journals reported higher no replications than actual replications and replication mentions, respectively, with the exception of *JEAB* in 2011 recording the highest actual replications compared to all other journals in

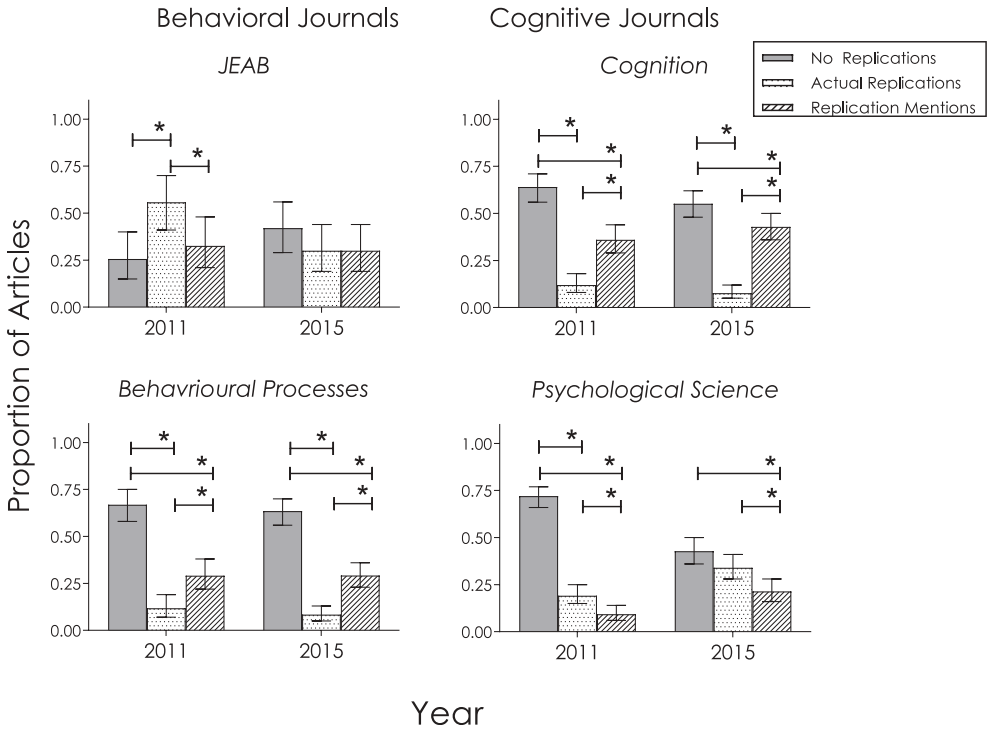

Figure 1. Proportion of articles (and their 95% CIs) reporting no replications, actual replications, and mentions of replication in the behavioral (left panel) and cognitive (right panel) journals published in 2011 and 2015. * indicates significant comparisons based on proportion differences during each publication year for each journal.

both years. Figure 1 also shows that the journals reported at least 25% replication mentions in 2011, while reporting above 25% in 2015, except *PS* reporting only 9% replication mentions. Second, various other comparisons of proportions of the replication categories for 2011 and 2015 across the behavioral and cognitive journals revealed important differences, some of which are shown in separate data sets presented in the appendix. For example, data for the individual journals compared by publication year showed that only *JEAB* recorded a significant decline in actual replications in 2015 whereas *PS* recorded a significant decrease in no replications with a corresponding increase in actual replications and replication mentions in 2015; all other comparisons showed no significant differences (see Figure 1A in the appendix). Figure 2 compares the replication categories by year of publication for each journal. Figure 2 shows that whereas reporting differences in *JEAB* in 2011 were significant for no replication vs. actual replications on the one hand, and actual replications vs. replication mentions on the other, none of the differences in 2015 were significant. The significant differences in *JEAB* in 2011 both favored actual replications. In contrast to *JEAB*, all the differences in replication reporting in both years were significant in *BP*; none of them, however, favored actual replications or replication mentions. The same was true for both *Cognition* and *PS* in both years, except for the replication reporting difference in both years for *PS* in which actual replications surpassed replication mentions, but more so in 2015 (see Figure 2).

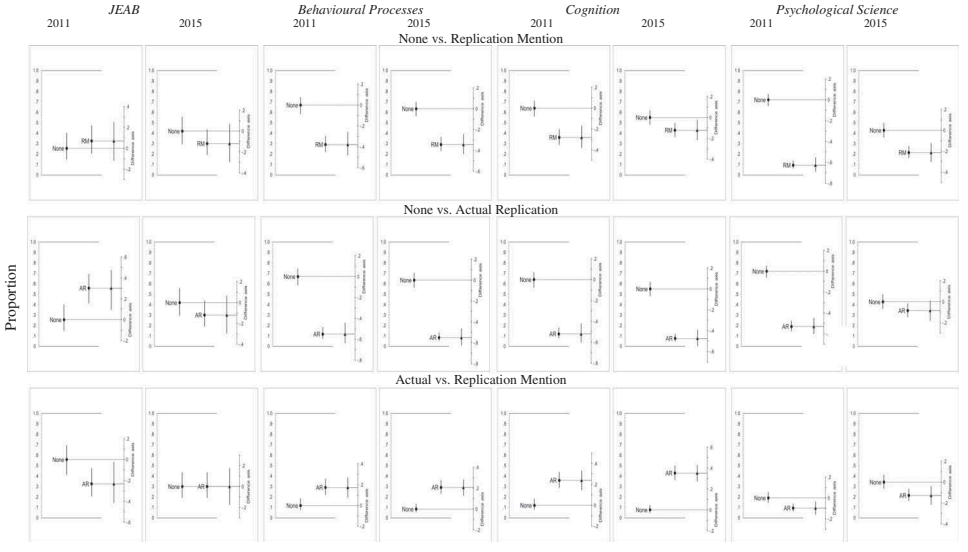

Figure 2. Cross comparisons of proportion of articles reporting no replication (None) vs. replication mention (RM) (top rows) no replication (None) vs. actual replications (AR) (middle row), and actual vs. replication mentions (bottom row) in all journals showing their proportion differences with the respective 95% CIs, which signify significant differences when they do not overlap zero on the difference axis.

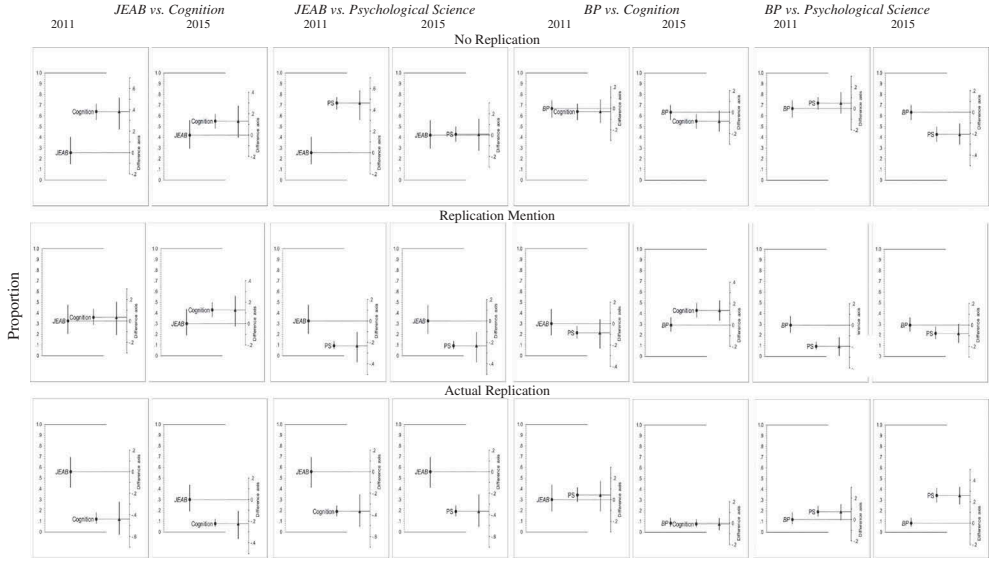

Figure 3. Cross comparisons of proportion of articles reporting no replication (top) replication mentions (middle), and actual replications (bottom) in *BP* vs. *cognition* (left of left-half panel) and *BP* vs. *PS* (right of left-half panel) and in *JEAB* vs. *cognition* (left of right-half panel) and *JEAB* vs. *PS* (right of right-half panel) showing their proportion differences with the respective 95% CIs, which signify significant differences when they do not overlap zero on the difference axis.

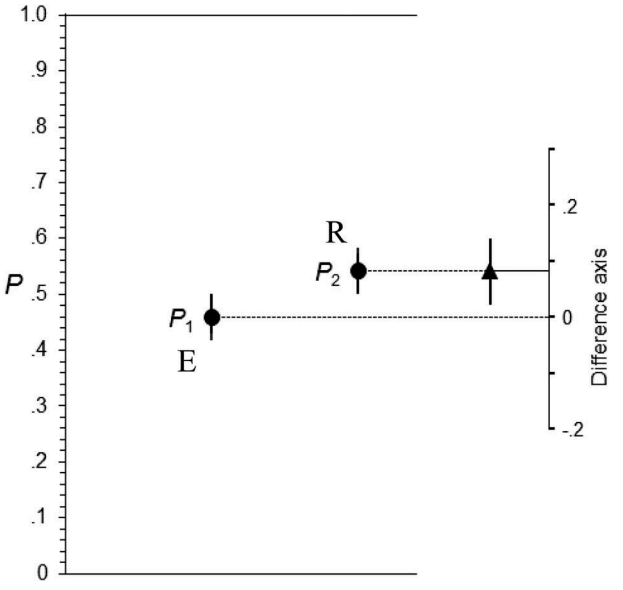

Figure 4. Proportions and proportion difference (with their respective 95% CI) of articles reporting exact (E) and relative (R) *p*-values across the behavioral journals and cognitive journals published in 2011 and 2015.

Comparisons of proportions of replication reporting by replication categories between the two behavioral and the two cognitive journals, respectively, revealed that the behavioral journals showed that reporting of no replication was significantly higher in *BP* than in *JEAB* but reporting of actual replications was higher in *JEAB* than in *BP* in both years (see Figure 2A, left panel in the appendix). Thus, although *JEAB* reported significantly more actual replications than *BP* in both years, the difference in replication reporting markedly declined in 2015. Whereas no replication reporting was significantly lower in *PS* than in *Cognition* in 2015 and in both years for replication mentions, actual replications were significantly higher in *PS* than in *Cognition* in 2015 (see Figure 2A, right panel, in the appendix).

Figure 3 presents cross comparisons between *JEAB* and the cognitive journals on the one hand, and between *BP* and the cognitive journals, on the other. Cross-comparisons between *JEAB* and *Cognition* showed that reporting no replication was significantly higher in *Cognition* than in *JEAB* in 2011, whereas actual replication reporting was higher in *JEAB* than in *Cognition* in both years (see Figure 3, left half). Cross-comparisons between *JEAB* and *PS* showed that all categories of replication-reporting differences were significant only in 2011, with higher no-replication difference in *PS* than in *JEAB*, but higher replication mentions and actual replications in *JEAB* than in *PS*. In 2015, there were virtually no differences on the three categories of replication reporting between the two journals (see Figure 3, right of left-half). Thus, whereas *JEAB* reported more actual replications than *Cognition* in both years, the difference not only declined in comparison to *Cognition* (Figure 3, left half,), it virtually disappeared in comparison to *PS* in 2015 (Figure 3, right of left-half). Finally, actual and no replication reporting were not different between *BP* and *Cognition* in both years (see Figure 3, left of right-half) and between *BP* and *PS* in 2011 (Figure 3, right half); replication mentions in 2011 were significantly higher in *PS* than in

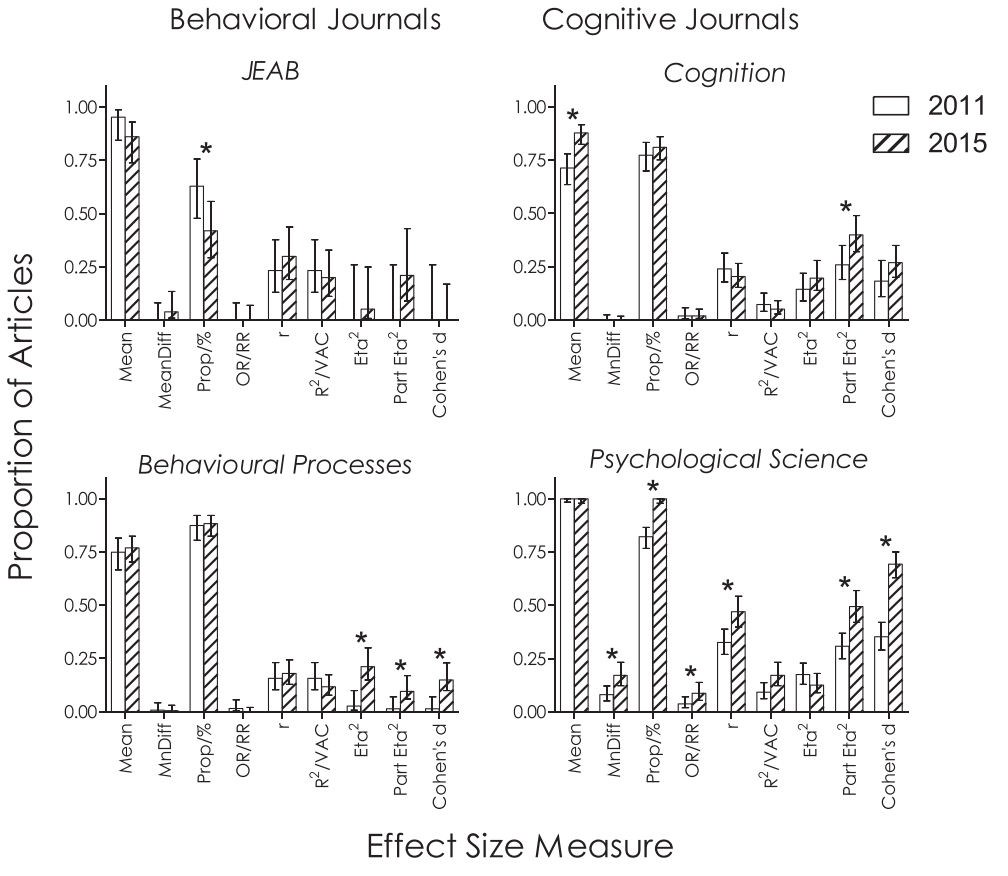

Figure 5. Proportion of articles reporting various effect size measures including mean, mean difference (MnDiff), proportion and percentage (Prop/%), odds ratio and relative risk (OR/RR), correlation (*r*), R^2 and variance accounted for (R^2/VAC), η^2 (eta^2), partial η^2 (part. eta^2), and Cohen's *d* in the behavioral journals (left panel) and the cognitive journals (right panel) published in 2011 and 2015. * indicates significant differences (see Table 1A in the appenndix for these and other differences discussed in text).

BP. In 2015, however, *PS* recorded significantly more actual replications and less no replication than *BP* did (Figure 3, right half).

All in all, then, replication reporting across the two years indicate two notable results. First, there was a decline in actual replications in 2015 in the lone star on replications in 2011, namely *JEAB*. Second, there was an increase in actual replications and replication mentions in *PS* in 2015.

Statistical reporting

Various statistical reporting patterns are discernable from the results, including on NHST, power, exact vs. relative *p*-value, ES measures, *SD* vs. *SE*, CIs and error bars on graphs. As shown in Table 1, reporting of key NHST statistics such as the *t*- and *F*-tests, as well the others including χ^2, *z*, and Mann Whitney *U* tests, increased in all journals except in *BP* in 2015 compared to 2011. For the cognitive journals, *Cognition* recorded greater increases in reporting of these statistics than *PS* in 2015, 50%, 58%, and 100% vs. 41%, 65%, and 28%,

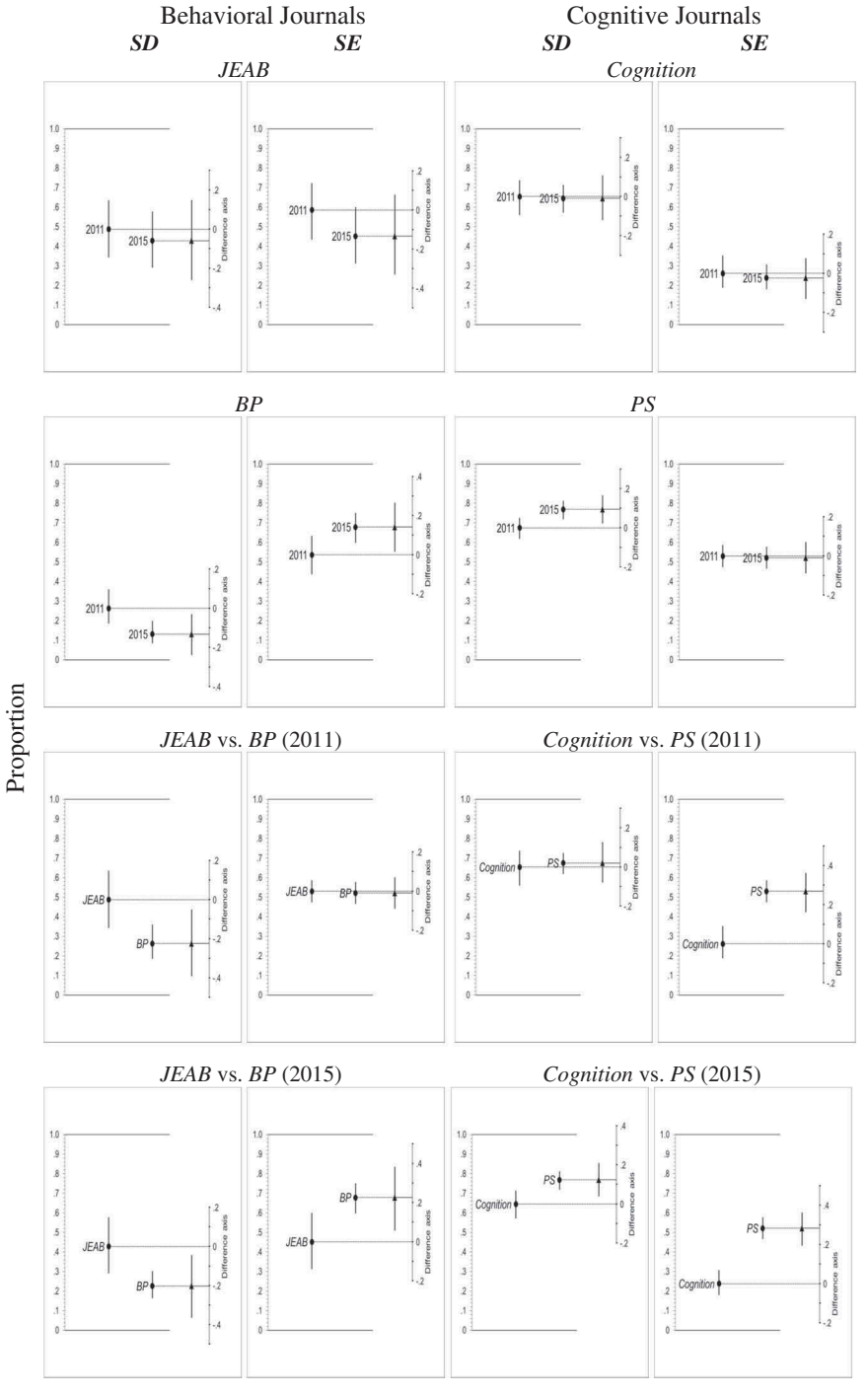

Figure 6. Proportions and proportion differences (with their respective 95% CI) of articles reporting *SD*s and *SE*s in the respective behavioral and cognitive journals (left of each panel) and their comparisons (right of each panel) during the 2011 and 2015 publication years.

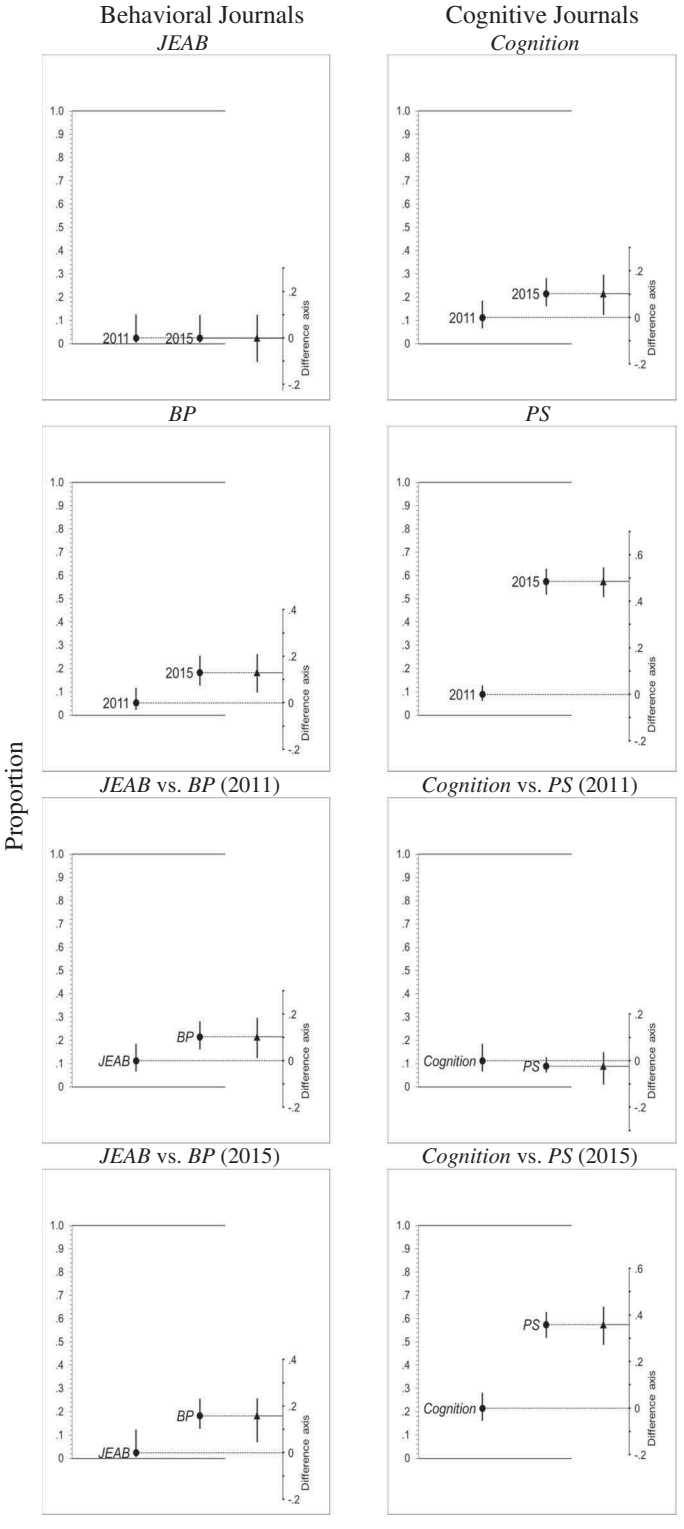

Figure 7. Proportions and proportion differences (with their respective 95% CI) of articles reporting CIs in the respective behavioral and cognitive journals (left panel) and their comparisons (right panel) published in 2011 and 2015.

respectively, for *t, F*, and others. For the behavioral journals, *JEAB* recorded greater increases than *BP* in 2015, 63%, 88%, and 64% vs. .99% decrease, .86% decrease, and .9% increase, respectively, for *t, F*, and others. Collectively, the four journals showed 76%, 93%, and 60% respectively for *t, F*, and the other statistical tests.

Power

Despite the substantial amount of reporting of the NHST statistics, reporting of statistical power usage, comparatively, were minimal (see Table 2). Table 2 shows that the generally low levels of power reporting nevertheless reflect increased reporting of the various power formats from 2011 to 2015, especially in the explicit, implicit, and same as in previous study (SAPS) categories; *JEAB* was an exception to these, showing no reporting at all of any of the categories except implicit in 2015. Table 2 also shows that whereas *Cognition* and *PS* reported more implicit power than explicit power in 2011, they reported the latter more than the former in 2015; both behavioral journals did not report explicit power at all in both years.

Exact vs. relative p-values

All journals reported both exact and relative *p*-values in both years about equally; only the reporting for 2011 in *PS* was significantly different with more relative than exact *p*-values reported. A comparison of relative vs. exact *p*-value across all journals showed a small but significant difference in reporting of relative over exact *p*-value as shown in Figure 4. None of the proportion differences on exact vs. relative *p*-value reporting across publication years in all journals was significant.

Effect size measures

Reporting of the various ESs mentioned in the introduction showed interesting commonalities and differences across the four journals. For one, all journals reported the

Table 2. Power reporting in each journal in 2011 and 2015 showing the numbers reported explicitly (explicit), implicitly (implicit), in terms of same as in previous study (SAPS), pilot, or others (pretesting and post hoc).

		Power Reporting				
Journals	Year	Explicit	Implicit	SAPS	Pilot	Others
Cognitive Journals						
COG	2011	1	8	0	0	0
	2015	13	14	10	1	0
Total		**14**	**22**	**10**	**1**	**0**
PS	2011	7	20	0	0	0
	2015	62	40	41	9	2
Total		**69**	**60**	**41**	**9**	**2**
Behavioral Journals						
JEAB	2011	0	0	0	0	0
	2015	0	4	0	0	0
Total		**0**	**4**	**0**	**0**	**0**
BP	2011	0	2	0	0	0
	2015	0	8	1	0	0
Total		**0**	**10**	**1**	**0**	**0**
Grand Total		**83**	**96**	**52**	**10**	**2**

mean and proportions/percentages a lot more than all other ESs in both years as shown in Figure 5, although only the proportion/percentage decline for *JEAB* and increases for *Cognition* and *PS* across 2011 and 2015 were significantly different. For another, the other ESs were reported relatively less frequently across all journals except for *PS* in which the reporting of *r*, partial η^2 and Cohen's *d* was comparatively higher than the other journals (see Figure 5). Of these other ESs, partial η^2 reporting was significantly better in 2015 than 2011 for *Cognition, PS*, and *BP*, but only so for Cohen's *d* in *PS* and *BP*, and only η^2 for *BP*. In addition, notably, there was virtual absence of reporting of η^2, partial η^2 and Cohen's *d* in the behavioral journals (see Figure 5, left panel) compared to the cognitive journals (see Figure 5, right panel) in 2011. The increased reporting in the behavioral journals in 2015 were only significant for *BP*. The two cognitive journals reported these effect sizes in both years, mostly showing increases in 2015 over 2011, but only those for partial η^2 in both journals and Cohen's *d* in *PS* were significant (see Figure 5, right panel). Table 1A in the appendix presents the proportion differences and their CIs that confirm these findings.

Standard deviation and standard error

Although the two cognitive journals reported more *SD*s in both years compared to the two behavioral journals, the proportion differences in the two years for all four journals were not significant except for *PS*, which recorded significantly more *SD*s in 2015 than in 2011 (see Figure 6, left panels). In both years, *JEAB* reported significantly more *SD*s than *PB*, but the *PS* reported significantly more *SD*s than *Cognition* only in 2015 (see Figure 6, right panels). Figure 6 also shows that whereas there was no difference in *SE*-reporting in both years in the cognitive journals, there was significantly more *SE* reporting in 2015 than in 2011 in *BP* (left panels). Reporting of *SE* was significantly higher in *PS* than in *Cognition* in both years, but only so for *BP* compared to *JEAB* in 2015 (see Figure 6, right panels). Although the two cognitive journals reported CIs more in 2015 than in 2011, *PS* reported CIs substantially more in 2015 than in 2011.

Confidence interval

Although CI reporting in *JEAB* was virtually nil in both years, of the two behavioral journals, only *BP* recorded more CI reporting in 2015 than in 2011 (see left panels, Figure 7). In both years, *BP* recorded significantly more CIs than *JEAB*, whereas *PS* did so over *Cognition* only in 2015 (see right panels, Figure 7).

The details of the CI reporting in these journals presented in Table 3 depict various outcomes. The table shows that more experiments in all four journals reported CIs in 2015 than in 2011. Collectively, the four journals reported more CIs in text (166/361 or 46%) than graphically (136/361 or 38%), with the least reporting in tabular form (59/361 or 16%). Individually, whereas *Cognition* reported more CIs graphically (8/12 or 67%) than both in text and in tables (2/12 or 17% each) in 2011, there were more textual (24/44 or 55%) than graphical (12/44 or 27%), then tabular (8/44 or 18%) in 2015. In *PS*, similarly, there was more CI reporting graphically (15/27 or 56%) than textual (10/27 or 37%) and tabular (2/27 or 7%) reporting in 2011, but more textual (120/241 or 50%) than graphical (81/241 or 33%), then tabular (40/241 or 17%) in CI reporting in 2015.

Table 3. Confidence interval (CI) reporting in each journal showing the number of experiments reporting, reporting format (graphical, tabular, or textual), and the number interpreting CIs (Total) in the form of interval bounds (IB), interval width (IW), overlap (OL), or non-zero (NZ) interpretation in 2011 and 2015.

			CI Reporting			CI Interpretation				
Journals	Year	No. of Experiments	Graph	Table	Text	IB	IW	OL	NZ	Total (%)
Cognitive Journals										
COG	2011	12	8	2	2	0	0	0	1	1(8)
	2015	36	12	8	24	1	2	1	3	7(16)
Total		**48**	**20**	**10**	**26**	**1**	**2**	**1**	**4**	**8(14)**
PS	2011	26	15	2	10	0	0	0	1	1(3)
	2015	181	81	40	120	2	1	1	11	15(6)
Total		**207**	**96**	**42**	**130**	**2**	**1**	**1**	**12**	**16(4)**
Behavioral Journals										
JEAB	2011	1	0	0	0	0	0	0	1	1(100)
	2015	2	2	0	1	0	0	0	0	0(0)
Total		**3**	**2**	**0**	**1**	**0**	**0**	**0**	**1**	**1(100)**
BP	2011	5	4	0	0	0	0	0	1	1(25)
	2015	25	14	7	9	0	0	3	0	3(17)
Total		**30**	**18**	**7**	**9**	**0**	**0**	**3**	**1**	**4(12)**
Grand Total		**288**	**136**	**59**	**166**	**3**	**3**	**5**	**18**	**29(8)**

Together, the cognitive journals tended to report CIs textually (156/324 or 48%) than graphically (116/324 or 36%) or in tables (52/324 or 16%). In contrast, the behavioral journals tended to report CIs graphically (20/37 or 54%) than in tabular (7/37 or 19%) or textual (10/37 or 27%) forms. Whereas CI reporting in *JEAB* was virtually nil with only one reporting in 2011 and very limited with only three in 2015, it was a little better in *BP* in both years. All CI reporting in 2011 was graphical (4/4 or 100%) in *BP*, in contrast to 2015 with 47% (14/30) graphical, 30% (9/30) tabular, and 23% (7/30) textual CI reporting.

Of the total CI reporting in each journal, in 2011, there was only one (8% and 3% respectively) CI interpretation in *Cognition* and *PS*, increasing to seven and 15 (16% and 6%, respectively) in 2015; *BP* reported only one and three representing 25% and 17%, respectively, in 2011 and 2015. Collectively, CI interpretation was more non-zero (62%) than overlap (24%) interpretations, and least based on interval bounds or width interpretations, at 10% each, in all four journals. Whereas cognitive journals recorded more non-zero (67%) than any other interpretations, the behavioral journals reported using more overlap interpretations (60%) than any other form of interpretation.

Error bars

Error bars appeared on graphs in both the cognitive and the behavioral journals, but on some more than on others; generally more in the cognitive than in the behavioral journals. For the behavioral journals, in both years, error bars appeared significantly more on *BP* than on *JEAB* graphs. In contrast, for the cognitive journals, graphs in *Cognition* contained more error bars than those in *PS*, but significantly so only in 2011 (see Figure 8, right panels). Figure 8 also shows that the proportion difference in error-bar usage in both years for each journal except *PS*, were not significant, although there was significant increase in error-bar use in *PS* in 2015 compared to 2011, as shown in the right panels of Figure 8.

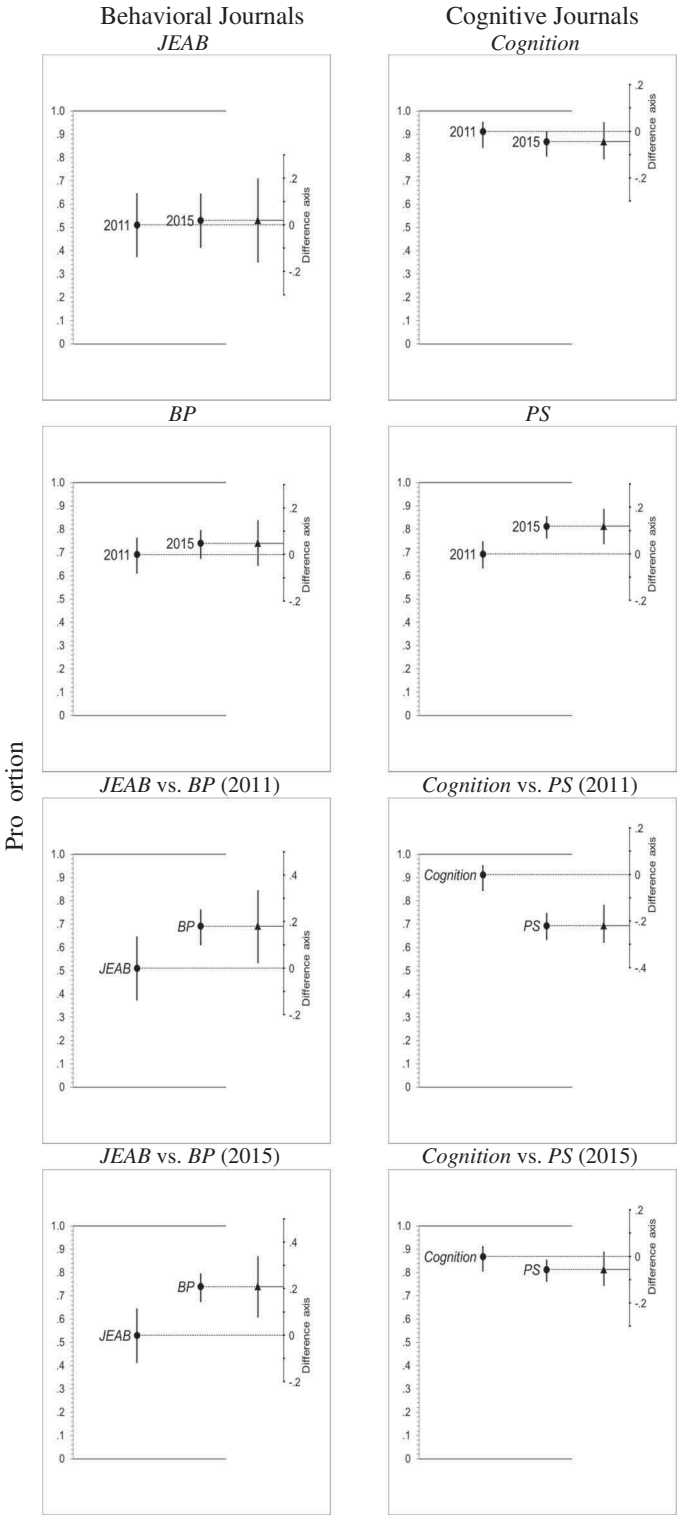

Figure 8. Proportions and proportion differences (with their respective 95% CI) of articles reporting error bars on graphs in the respective behavioral and cognitive journals (left panel) and their comparisons (right panel) published in 2011 and 2015.

Discussion

Has the recent APS efforts on promoting better replication and statistical reporting practices began to have a demonstrable impact on journal publications in psychology? How widespread is such an impact? The present study adopted a snapshot approach to answer these questions, by selecting a couple of behavioral and of cognitive journals (one of which is an APS publication) published before and after the APS efforts became widely disseminated in 2012. Of the four journals included in the present study, pre-initiative (2011) reporting of actual replications was highest only for *JEAB*; post-initiative (2015), however, this reversed in favor of *PS* only, as all the other journals declined in their reporting of actual replications in 2015 compared to 2011. Across the years, actual replications remained high for *JEAB* than for *Cognition*, although less so in 2015 than in 2011. In all cross-comparisons of actual replications, replications tended to be lower in *Cognition* than in all other journals, except in *BP*; the huge difference in actual replications between *JEAB* and *PS* in 2011 had collapsed by 2015 and the gap in actual replications between *PS* and *BP* in 2011 grew in 2015.

One cannot help but notice the paradox of *JEAB* recording less replication than *PS* in 2015 given *JEAB*'s reputation on use of small-*N* designs that requires replicating experimental conditions as a matter of course. Whereas *JEAB*'s replication decline that year remains inexplicable, *PS*'s replication increase is attributable to the current reform efforts of the APS. In both years, the decline in actual replications recorded in 2015 by the other three journals dampens the good news in the burst in actual replications observed in *PS*, being the only APS journal included in the present study. The relatively high levels of mentions of replication in these three journals in both years did not make up for this laxity in actual replications; if only actual replications could rise to the levels of replication mentions!

As indicated by Makel et al. (2012), it is difficult to characterize accurately the level of replication in the literature if authors do not refer to the fact that their study was actually a replication study. Relying heavily on search terms could be a shortcoming of the present study, although actual replications were assessed during a hit on the search terms. Actual replications, therefore, provided an important index on the impact of the APS efforts on replication reports. In this regard, the decline in actual replications in *JEAB* in 2015 is alarming in light of the emphases usually placed on replications in behavior analytic research. A small-*N* approach to research would, more likely than not, warrant a declaration of replication of experimental conditions in the course of describing the procedures. The decline in actual replications that year, therefore, is not likely due to nondeclaration by authors, as implied by Makel et al.'s noted observation. Perhaps the culprit is more likely the growing use of human participants in behavior analytic research that may have been accompanied by the use of group designs being published in *JEAB* that peaked during the 2010s (Zimmermann et al., 2015). Indeed, Zimmermann et al. reported a concomitant increase in the use of inferential (NHST) statistics in the journal as well. Whether 2015 was just an aberration in *JEAB* regarding the recorded decline in actual replications reporting remains an empirical question.

Contrary to what one might expect given the common differences outlined in the introduction between cognitive and behavioral approaches in psychology, the results of the present study showed substantial reporting of NHST statistics in the behavioral journals in both years in the present study (see Table 1). In addition, the generally high

reporting of NHST across all journals of 76%, 93%, and 60% for *t, F*, and the other statistics surveyed in the present study, and the percent of *F*-reporting in particular is consistent with those reported by Finch et al. (2004) for *Memory & Cognition*. Furthermore, all journals, behavioral and cognitive alike, showed increased reporting of HNST statistics in 2015 compared to 2011 and yet with little corresponding increase in reporting of power, especially in the behavioral journals; at least, the cognitive journals relatively increased reporting of various categories of power. The generally low levels of power reporting are similar to those reported by Fidler et al. (2004) for two medical journals with similarly high NHST reporting records. Given the aforementioned substantially high reporting of NHST statistics in the behavioral journals in both years, consistent with previous reports of growing use of group designs in *JEAB* in recent years (Zimmermann et al., 2015), for example, it is concerning that power reporting in the behavioral journals were virtually non-existent except for the implicit category for *BP*. The virtual absence of power reporting in the behavioral journals perhaps may be due to the generally ad hoc nature of the use of NHST and/or inadequate training in the use of group-design statistics in behavioral psychology, which usually is more disposed to small-*N* design approach to the evaluations of data (Gast & Ledford, 2014; Sidman, 1960).

All journals reported *p*-values as exact or relative values indiscriminately in both years, except *PS* reporting more relative than exact *p*-values in 2011. The indiscriminate reporting of exact and relative *p*-values remains problematic across the board and is consistent with Finch et al.'s (2001) report of practices in the *Journal of Applied Psychology*. Authors seem not to have gotten the message about the oft-cited preference for exact p-values as best practice in reporting NHST statistics. This is one area of best practices where editorial imperative (Imam, 2018) can have an immediate impact by simply requesting reporting of exact *p*s.

Of the nonstandardized ESs recorded in the present study, both behavioral and cognitive journals showed high frequency reporting of the mean and proportion/percentage measures in both years compared to all other ES measures, although *JEAB, Cognition*, and *PS* involved significant differences in these measures. These results are very consistent with those reported particularly for *Epidemiology* and in part for the *American Journal of Public Health* by Fidler et al. (2004); in the latter journal, there was less reporting of means and mean differences than in *Epidemiology*. The three standardized ESs (η^2, partial η^2, and Cohen's *d*) recorded in the present study appear to have revealed better prospects in showing increases in 2015 compared to 2011, especially in *BP* and *Cognition*, but particularly in *PS*, in support of the current efforts initiated by the APS.

The common refrain about the use of *SD*s and *SE*s, in text or on graphs by authors, is that they tend to be used interchangeably as measures of variability and as if they are both descriptive statistics (see Cumming, Fidler, & Vaux, 2007b). As Cumming et al. elucidated, however, one (*SD*) is a descriptive statistic and the other (*SE*) is neither descriptive nor inferential, only CI is inferential. Accordingly, use of error bars on graphs can be very confusing and deceptive, therefore, especially when they are not identified appropriately, if at all. In the present study, only *BP* and *PS* showed significant increases respectively in *SE*- and *SD*- reporting in 2015 compared to 2011. Although error-bar reporting was recorded by most of the journals in the present study, they were not as high as those reported previously for various journals by Cumming et al. (2007a). With respect to CI reporting, whereas *JEAB* reporting of CI was abysmally low at near-zero in both years, the

other three journals increased CI reporting in 2015 over 2011, the greatest increase being in *PS*, a most encouraging development in light of the journal's home and providing further evidence for effectiveness of the current reform efforts. Although the present analysis did not differentiate reporting of *SD*s and *SE*s by graphical, textual, or tabular formats as was done with CIs, the use of error bars other than CI remains a concerning practice in light of persistent recommendations to the contrary (Cumming, Fidler, Kalinowski, & Lai, 2012; Cumming et al., 2007a; Cumming & Finch, 2005). It is one thing to use error bars at all and another to identify what they represent on the graphs they accompany. Yet another though, and perhaps more importantly, is the preference for error bars to be CI rather than *SE* (see Cumming, 2009), or anything else for that matter, due to what Cumming et al. refer to as its "inferential information" value (2012, p. 144). That CIs are not "descriptive" statistics as *SD*s are, needs better appreciation in psychological research reporting and cannot be overemphasized (see Cumming et al., 2007b). Finally, CI interpretations were largely of the nonzero variety in the two cognitive journals that showed an increase in CI-interpretation reporting in 2015 in the present study.

The foregoing suggests that whereas there have been certain areas of improvement in reporting of both replication and statistical results in both the cognitive and the behavioral journals examined in the present study, a number of areas remain problematic toward achieving best practices in psychological science reporting. Together, they implicate collective lines of action by relevant institutions and/or organizations in psychology (see Imam, 2018). For example, the Loftus editorial experience in *Memory & Cognition* of having to correct authors' errors on submissions (Fidler et al., 2004) reveals the limits of rules and rule-governed behaviors compared to contingency-managed behavior; what is needed is beyond editorial policies (e.g., Fan & Thompson, 2001) and practices. The APA Manual and an editorial policy represent rules that alone or together are inadequate in maintaining sustained impact on reporting practices (see Fidler et al., 2004). Editorial accept-reject decisions on manuscripts represent the point of exerting contingency on what is appropriate or acceptable to the research community. There have to be consequences as a bridge between the rules and the outcomes or practices. Debate and focus should be on identifying effective strategies for contingency management at all levels, from the classroom to the grant funding bodies. Lilienfeld (2017) is exemplary on delineating what goes on in awarding grants, just as Koole and Lakens (2012) have outlined workable incentive regimes that would promote replications in extant psychological research. Ator (1999) provides cogent examples why there might be a drift, even in behavior analysis at that, to the adoption of NHST. Pinpointing what works and how well, to what is the best way to advance the science of psychology regardless of its various domains and subdivisions is what is required at this juncture. Deliberate outreach also is called for in the current efforts in light of increasing use of NHST by authors publishing in behavioral journals such as *JEAB* (see Zimmermann et al., 2015), lest there remain pockets of resistance in adopting acceptable NHST practices that may linger. Furthermore, efforts have to extend to training that promotes curricula emphases (e.g., Fidler, 2010) for both current and future researchers in psychology and associated disciplines, including undergraduate and graduate training, in order to ensure adequate acquisition of desirable requisite research behaviors and reporting practices.

The results of the present study also suggest that the new efforts of the APS are beginning to have some impact as illustrated by some increased reporting of replications especially in *PS*, of CIs in all four journals, and of error bars on graphs in *Cognition* and *BP*. Thus, change appears to be happening already, but it should not be limited to the pages of APS journals. A concerted effort is required to expand the reach of the impact of the new approaches spearheaded by the APS to tackling the menace that NHST poses to psychological science. All psychologists in all domains of psychology should own these efforts and become advocates for these new practices in a sustainable version. Only then can psychology as a discipline hope to survive the scourge of mindless applications of NHST to the design, analysis, and interpretation of psychological research. The fact is that, contrary to popular believe, as noted earlier, behavior analysts have been engaged in the debates on the use of inferential statistics and how it impacts psychological research (e.g., Baron, 1999; Branch, 1999; Crosbie, 1999; Shull, 1999). As modern psychology struggles to find and consolidate its new identity on its quest for self-examination and self-discovery (see Barrett, 2016, for example), at some point, psychologists have to come to grips with whether and when experimental control (Perone, 1999) weighs more than statistical control in psychological research. Future research on topics like this should consider a look at other journals in the field that might shed further light on the growth and advancement of best practices in conducting and reporting sound researchin psychological science.

References

Alcock, J. (2011, Jan 6). Back from the future: Parapsychology and the bem affair. Center for Skeptical Inquiry. Retrieved from http://www.csicop.org/specialarticles/show/back_from_the_future

Ator, N. A. (1999). Statistical inference in behavior analysis: Environmental determinants? *The Behavior Analyst*, *22*, 93–97.

Baron, A. (1999). Statistical inference in behavior analysis: Friend or foe? *The Behavior Analyst*, *22*, 83–85.

Barrett, L. (2016). Why brains are not computers, why behaviorism is not satanism, and why dolphins are not aquatic apes. *The Behavior Analyst*, *39*, 9–23.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Bernard, C. (1927/1957). *An introduction to the study of experimental medicine*. New York, NY: Dover.

Bernstein, D. A., Penner, L. A., Clarke-Stewart, A., & Roy, E. J. (2007). *Psychology* (7th ed.). Boston, MA: Houghton Mifflin.

Bohannon, J. (2014). Replication effort provokes praise-and 'bullying' charges. *Science*, *344*, 788–789.

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

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*, 40–48.

Chow, S. L. (1998). Precis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences*, *21*, 169–239.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychology*, *49*, 997–1003.

Crosbie, J. (1999). Statistical inference in behavior analysis: Useful friend. *The Behavior Analyst*, *22*, 105–108.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. *Statistics in Medicine*, *28*, 205–220.

Cumming, G. (2014a). The new statistics: Why and how? *Psychological Science*, *25*, 7–29.

Cumming, G. (2014b, May 22). The new statistics: Effect sizes and confidence intervals (Part 3: Research integrity and the new statistics) [video file]. Retrieved from https://www.psychologicalscience.org/members/new-statistics

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. New York, NY: Routledge.

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology*, *64*, 138–146.

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

Cumming, G., Fidler, F., & Vaux, D. L. (2007b). Error bars in experimental biology. *The Journal of Cell Biology*, *177*, 7–11. doi:10.1083/jcb.200611141.

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

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Philosophy*, *5*, 75–98.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Educational and Psychological Measurement*, *61*, 517–531.

Fidler, F. (2010). The American psychological association publication manual sixth edition: Implications for statistics education. In C. Reading (Ed.), *Data and context in statistics*

education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics. Voorburg, The Netherlands: International Statistical Institute.
Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, *15*, 119–126.
Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *journal of applied psychology*: Little evidence of reform. *Educational and Psychological Measurement*, *61*, 181–210.
Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., … Goodman, O. (2004). Reform of statistical inference in psychology: The case of *memory & cognition*. *Behavior Research Methods, Instruments, & Computers*, *36*, 312–324.
Frisk, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.
Gast, D. L., & Ledford, J. R. (Eds). (2014). *Single case research methodology: Applications in special education and behavioral sciences*. New York, NY: Routledge.
Goldstein, H. (1984). Present position and potential developments: Some personal views statistics in the social sciences. *Journal of Royal Statistical Society*, *147*, 260–267.
Griggs, R. A. (2017). *Psychology: A concise introduction*. New York, NY: Worth Publishers.
Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24.
Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum.
Imam, A. A. (2018). Place of behavior analysis in the changing culture of replication and statistical reporting in psychological science. *European Journal of Behavior Analysis*, *19*, 2–10.
Kern, S. E. (2014). Inferential statistics, power estimation, and study design formulation continues to suppress biomedical innovation. *arXiv Preprint*, arXiv:1411.0919.
Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, D. C.: American Psychological Association.
Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614.
Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—Significant tests are not. *Theory & Psychology*, *22*, 67–90.
Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives in Psychological Science*, *12*, 660–664.
Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832.
Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science*, *7*, 537–542.
Morrison, D. E., & Henkel, R. E. (Eds.). (1970). *The significance test controversy: A reader*. Chicago: Aldine.
Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–116). Mahwah, NJ: Lawrence Erlbaum.
Myer, D. G., & Dewall, C. N. (2017). *Psychology in everyday life*. New York, NY: Worth.
Olsson-Colletine, A., van Assen, M. A. L. M., & Hartgerink, C. H. J. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, 1–11. doi:10.1177/0956797619830326.
Pashler, H., & Wagenmakers, E., (Eds.). (2012). Special section on replicability in psychological science: A crisis of confidence? [Special issue]. *Perspectives on Psychological Science*, *7*, 528–530.
Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, *22*, 109–116.
Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's retroactive facilitation of recall effect. *PloS One*, *7*(3), e33423.
Rozeboom, W. W. (1960). The fallacy of null hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Lawrence Erlbaum.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 73–83.

Shull, R. L. (1999). Statistical inference in behavior analysis: Discussant's remarks. *The Behavior Analyst*, *22*, 117–121.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Boston, MA: Authors Cooperative.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Spellman, B. A., (Ed.). (2012). Special section on research practices. [Special issue]. *Perspectives on Psychological Science*, *7*, 655–689.

Thompson, B. (1999). Five methodology errors in educational research: A pantheon of statistical significance and other faux pas. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 23–86). Stamford, CT: JAI Press.

Townsend, J. T. (2008). Mathematical psychology: Prospects for the 21st century. *Journal of Mathematical Psychology*, *52*, 269–280.

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.

APPENDIX

Table 1A. Proportion differences (Prp dif) and their 95% confidence intervals (CI) for proportions of articles' reporting of mean, mean difference, proportion (prop) and percentage (%), odds ratio (OR), relative risk (RR), correlation (*r*), and R^2 and variance accounted for (VAC), η^2, partial η^2, and Cohen's *d* in the cognitive (top panel; *cognition (COG)* and *psychological science (PS)*) and behavioral (bottom panel; *journal of the experimental analysis of behavior (JEAB)* and *behavioural processes (BP)*) journals in 2011 compared to 2015 (see Figure 5).

	Mean		Mean Difference		Prop/%		OR/RR		r		R^2/VAC		η^2		partial η^2		Cohen's *d*	
	Prp dif	CI	Prp dif	CI	Prp dif	CI	Prp dif	CI	Prp dif	CI	Prp dif	CI	Prop diff	CI	Prop diff	CI	Prop diff	CI
Cognitive Journals																		
COG	.164	[.079, .25][a]	.00	[−.025, .019]	.038	[−.047, .126]	.00	[−.039, .034]	−.036	[−.126, .051]	−.022	[−.08, .029]	.052	[−.049, .15]	.142	[.017, .259][a]	.086	[−.033, .193]
PS	.00	[−.021, .016]	.091	[.027, .159][a]	.178	[.13, .232][a]	.05	[.004, .104][a]	.143	[.049, .235][a]	.078	[.013, .147]	−.05	[−.119, .02]	.187	[.092, .278][a]	.34	[.244, .427][a]
Behavioral Journals																		
JEAB	−.093	[−.22, .036]	.04	[−.047, .137]	−.208	[−.388, −.005][a]	.00	[−.082, .071]	.067	[−.114, .238]	−.033	[−.202, .132]	.053	[−.21, .246]	.211	[−.077, .433]	.00	[−.259, .168]
BP	.022	[−.074, .121]	−.002	[−.038, .024]	.008	[−.065, .087]	−.016	[−.056, .008]	.022	[−.066, .105]	−.04	[−.123, .037]	.185	[.091, .272][a]	.084	[.01, .153][a]	.137	[.056, .215][a]

[a] Non-zero CI overlap

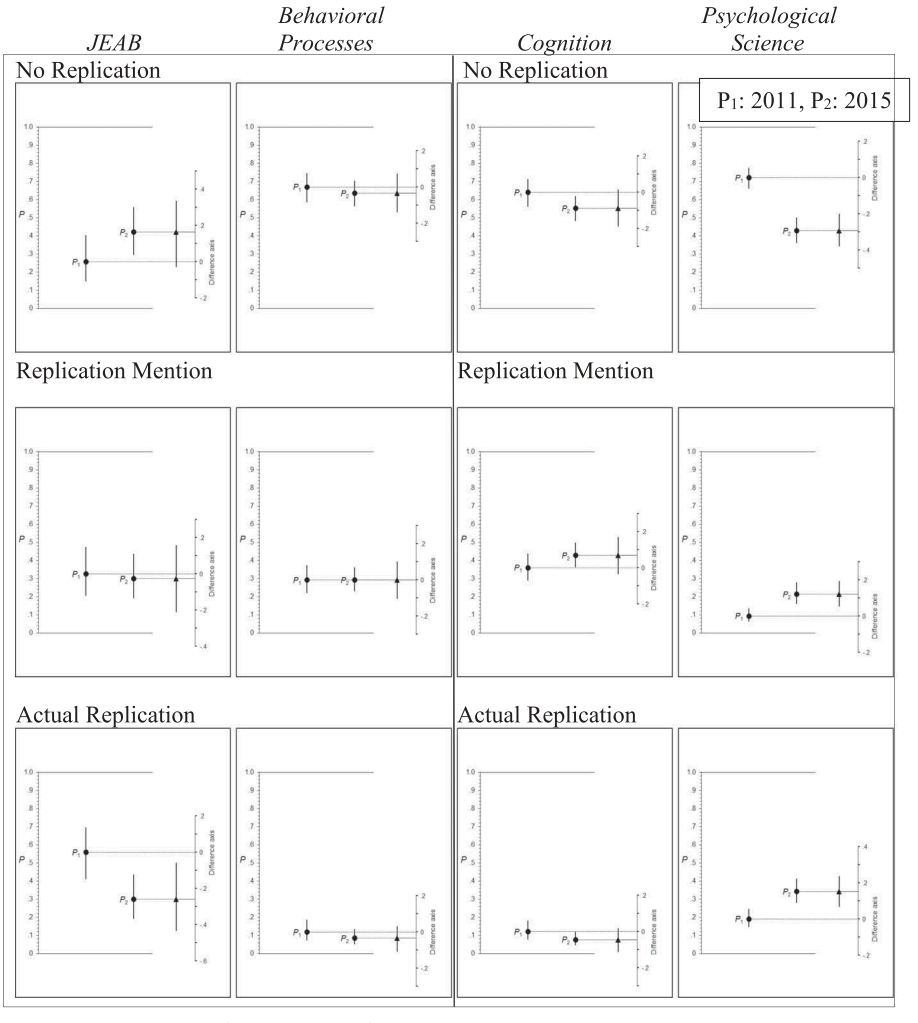

Figure 1A. Comparisons of proportion of articles reporting no replication, replication mention, and actual replications in the two behavioral journals (left panel) and the two cognitive journals (right panel) across the two publication years (2011 vs. 2015) showing their proportion differences with the respective 95% CIs, which signify significant differences when they do not overlap zero on the difference axis.

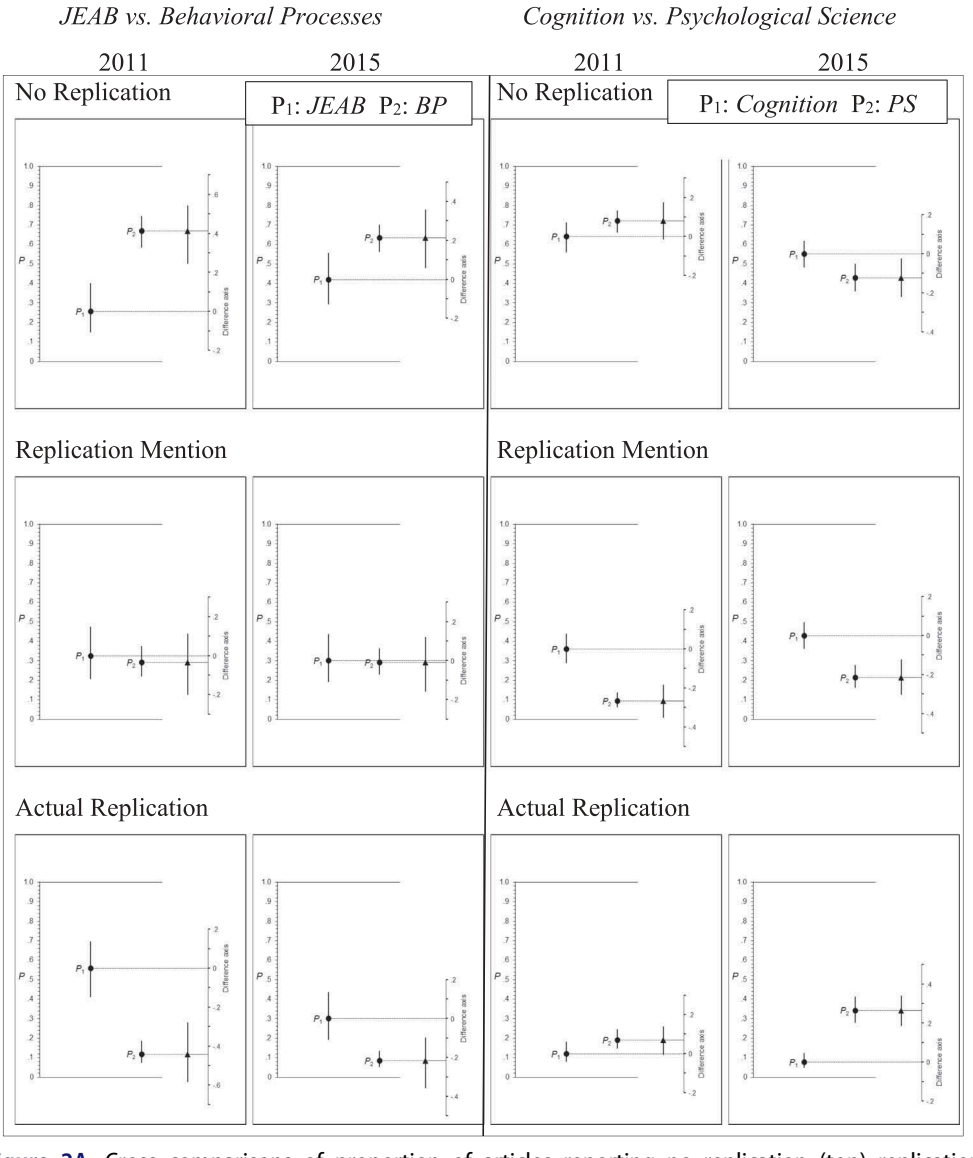

Figure 2A. Cross comparisons of proportion of articles reporting no replication (top) replication mentions (middle), and actual replications (bottom) in *JEAB* vs. *BP* (left panel) and *cognition* vs. *PS* (right panel) showing their proportion differences with the respective 95% CIs, which signify significant differences when they do not overlap zero on the difference axis.