An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis

Article · March 2019

DOI: 10.1007/s40614-019-00193-3

CITATIONS
6

READS
291

4 authors:

Sean Laraway
San Jose State University
45 PUBLICATIONS **752** CITATIONS
SEE PROFILE

Susan Snycerski
San Jose State University
31 PUBLICATIONS **729** CITATIONS
SEE PROFILE

Sean Pradhan
Menlo College
7 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Bradley Huitema
Western Michigan University
226 PUBLICATIONS **1,725** CITATIONS
SEE PROFILE

Perspectives on Behavior Science
https://doi.org/10.1007/s40614-019-00193-3

An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis

Sean Laraway[1] · Susan Snycerski[1] · Sean Pradhan[2] · Bradley E. Huitema[3]

Published online: 22 March 2019

Abstract

For over a decade, the failure to reproduce findings in several disciplines, including the biomedical, behavioral, and social sciences, have led some authors to claim that there is a so-called "replication (or reproducibility) crisis" in those disciplines. The current article examines: (a) various aspects of the reproducibility of scientific studies, including definitions of reproducibility; (b) published concerns about reproducibility in the scientific literature and public press; (c) variables involved in assessing the success of attempts to reproduce a study; (d) suggested factors responsible for reproducibility failures; (e) types of validity of experimental studies and threats to validity as they relate to reproducibility; and (f) evidence for threats to reproducibility in the behavior science/analysis literature. Suggestions for improving the reproducibility of studies in behavior science and analysis are described throughout.

Keywords Reproducibility · Replication · Null hypothesis significance testing · Statistical power · Effect size measures · Statistical conclusion validity · Construct validity

> The crucial role of replication is established in science generally. The undetected equipment failure, the rare and possibly random human errors of procedure, observation, recording, computation, or report are known well enough to make scientists wary of the unreplicated experiment. When we add to the possibility of the random "fluke," common to all sciences, the fact of individual organismic differences and the possibility of systematic experimenter effects in at least the

✉ Sean Laraway
sean.laraway@sjsu.edu

[1] Department of Psychology, San José State University, San José, CA 95192-0120, USA

[2] Menlo College, Atherton, CA, USA

[3] Western Michigan University, Kalamazoo, MI, USA

> behavioral sciences, the importance of replication looms large still to the behavioral scientist (Rosenthal & Rosnow, 2009, p. 552)

Defining Reproducibility

In the last decade or so, researchers have expressed increasing concern about the *replicability* or *reproducibility* of studies in many fields, including the biomedical, behavioral, and social sciences (e.g., Begley & Ioannidis, 2015; Goodman, Fanelli, & Ioannidis, 2016; Ioannidis, 2005, 2015; McNeely & Warner, 2015; Open Science Collaboration [OSC], 2012, 2015, 2017; Pashler & Harris, 2012; Pashler & Wagenmakers, 2012). In the literature, there are varying definitions of *replicability* and *reproducibility*, and it is important to remember that these two terms (and alternatives) are constructs that can be operationalized in different ways. For example, Leek and Jager (2017) stated that a study is *replicable* "if an identical experiment can be performed like the first study and the statistical results are consistent" and *reproducible* "if all of the code and data used to generate the numbers and figures in the paper are available and exactly produce the published results" (p. 111). They further noted that "definitions of successful replications have included consistent effect sizes, consistent distributions, and consistent measures of statistical significance" (p. 114).

Goodman et al. (2016, pp. 2–4) identified three aspects of *reproducibility*: (a) *methods reproducibility*, or the "ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results" and "the provision of enough detail about study procedures and data so the same procedures could, in theory or in actuality, be exactly repeated"—this is consistent with Leek and Jager's (2017) definition of *reproducibility*; (b) *results reproducibility*, or the "production of corroborating results in a new study, having followed the same experimental methods" and "obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible"—this is consistent with many common definitions of *replication*, including that offered by Leek and Jager; and (c) *inferential reproducibility*, or "the making of knowledge claims of similar strength from a study replication or reanalysis" and "drawing qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study." However, this is not the same as the *replication of results* "because not all investigators will draw the same conclusions from the same results, or they might make different analytical choices that lead to different inferences from the same data." For convenience, throughout this article we will generally use the term *reproducibility* as defined by Goodman et al. to refer to the overall topic, unless we are discussing a specific aspect as defined by these or other authors. This choice should not limit the applicability of the topics we discuss to behavior science/analysis.[1]

[1] In keeping with *Perspectives on Behavior Science* guidelines (Hantula, 2018), we will use *behavior science/analysis* and *behavior scientists/analysts* to refer to the science/practice and scientists/practitioners, respectively, of the field formerly known as *behavior analysis*.

Concerns about Reproducibility in the Psychological Literature

The concern regarding the reproducibility of studies in psychology has heightened recently for a variety of reasons (see, for example, Earp & Trafimow, 2015; Maxwell, Lau, & Howard, 2015; Nelson, Simmons, & Simonsohn, 2018; Pashler & Wagenmakers, 2012; Schweinsberg et al., 2016; Yong, 2012), with some authors suggesting we have a "replication crisis in psychology" (RCP; e.g., Pashler & Harris, 2012; Pashler & Wagenmakers, 2012; Rotello, Heit, & Dubé, 2015). Apart from high-profile examples of scientific misconduct (e.g., see Hamblin, 2018), a recent and highly publicized reason for concern over the RCP is the publication of the OSC's (2015) large-scale effort to replicate studies in three psychology journals. The OSC reported that a large proportion of their studies "failed to replicate" the findings of the original studies. Gilbert, King, Pettigrew, and Wilson (2016) contested these findings, stating that the OSC article "contains three statistical errors and provides no support" for the conclusion that "the reproducibility of psychological science is surprisingly low" (p. 1037). With respect to the generalizability of the OSC's findings, Perone (2018) rightly noted that their findings may not be representative of psychological science in general and behavior science/analysis in particular. In addition to Gilbert et al., other authors have disagreed about the extent to which a "crisis" exists (e.g., Maxwell et al., 2015; Stroebe & Strack, 2014).

On a more positive note, Nelson et al. (2018) contended that an RCP *did* exist, but in the period prior to the current one, because psychologists are now paying increased attention to the issue of reproducibility and are exploring ways to improve the reproducibility of their studies. As a result, Nelson et al. termed the current period "psychology's renaissance." Perone (2018) expressed the optimistic opinion that when well-designed studies do not replicate prior research, this offers opportunities to improve our understanding of the phenomena of interest. Indeed, the increased attention to reproducibility could improve scientists' efforts at conducting and reporting reproducible studies, which should advance our knowledge of behavior and the variables that influence it (Hales, Wesselmann, & Hilgard, 2018).

The articles in this special section on replication/reproducibility in *Perspectives on Behavior Science* are examples of our field's reaction to the RCP, which will provide behavior scientists/analysts with much to consider as they conduct, consume, and review studies in behavior science. Regardless of what one feels about the extent and nature of the RCP, we feel this topic is worth our careful consideration. We suggest that behavior scientists/analysts not take the attitude that "it can't happen here" due to our methodological history of encouraging replication and experimental control and avoiding null hypothesis significance testing (NHST), which has been proposed as a contributing factor in replication failures (e.g., Branch, 2018; Killeen, 2018). Even if behavior science/analysis does not suffer from a reproducibility "crisis" (or some less dramatic form of reproducibility problems), a reminder of potential issues that could limit the reproducibility of our research studies and ways of avoiding them can only help our field. Later, we discuss manners in which behavioral studies conducted in ways consistent with our methodological tradition (e.g., as described by Branch, 2018; Perone, 2018) could face threats to reproducibility.

Concerns about Reproducibility in the Popular Press

The debate about the reproducibility "crisis" in science is not merely an academic concern, as it has garnered much attention from various media outlets. For example, websites such as *FiveThirtyEight* (https://fivethirtyeight.com/), a popular politics, sports, and culture hub specializing in quantitative analysis and forecasting, and *Vox* (https://www.vox.com/), a general news outlet examining trends in society, have criticized scientists for the lack of reproducibility of their investigations, as well as engaging in questionable research practices (QRPs, discussed in more detail later in this article; John, Loewenstein, & Prelec, 2012; Schmidt & Oh, 2016; Shrout & Rodgers, 2018). Aschwanden (2015) of *FiveThirtyEight* described the problem of *p-hacking*, the abusing of statistical analyses to achieve *p*-values below .05, the "holy grail" of statistical significance thresholds in NHST. *Vox*'s Resnick (2018) has argued that the replication crisis contributes to the distribution of misinformation to budding behavior science students due to textbooks reporting and emphasizing results from unsound yet "dramatic" studies. He points to cases of classic psychology studies being virtually nonreplicable such as the famed Stanford Prison Experiment (SPE; Haney, Banks, & Zimbardo, 1973) and Mischel's (1958) delayed gratification study using the "marshmallow test."

As a result of the RCP, Reicher and Haslam (2006) sought to replicate the results of the SPE by partnering with the British Broadcasting Corporation (BBC) in a televised, documentary-style setting, which they termed *The Experiment*. In the original SPE, participants were randomly assigned to serve as prisoners or guards in a simulated prison. Haney et al. (1973) reported that subjects conformed to the social norms set by their roles, with "guards" exhibiting particularly harsh behaviors by punishing the "prisoners." In their 2006 replication, Reicher and Haslam did not reproduce the results of Haney et al. Rather, they found that those who served as prisoners banded together and "overthrew" the guards. Although there may be issues with the nature of the experimental conditions (e.g., the research being conducted in the context of a public broadcast), Reicher and Haslam asserted that the expectations (and experimenters' stated instructions) regarding roles may have confounded the results of the original experiment and overstated the value of the simulated prison situation. Other researchers have also questioned the validity and findings of the SPE (for a summary, see Bartels, 2015; Kulig, Pratt, & Cullen, 2017). Similar doubts have been discussed with respect to Milgram's classic studies on obedience (e.g., Haslam & Reicher, 2012; Perry, 2018; Romm, 2015).

Popular media have also noted that previous research may tend to exaggerate findings in the reported effect size measures (Resnick, 2017, 2018). For instance, Watts, Duncan, and Quan (2018) attempted to replicate the findings from Mischel's (1958) "marshmallow test" in a longitudinal study. Their results indicated that although similar trends between delay time and achievement as well as socioemotional behaviors were observed through the bivariate correlations, the obtained effect sizes were about half of those reported in earlier research. The apparent replication "failures" of such high-profile psychology studies in the popular press and academic literature serve to undermine confidence in the field as a whole, including behavior science/analysis (Branch, 2018). We do not think that laypeople will distinguish between behavior science/analysis and other areas within the broader field of psychology, so they could generalize the "RCP" to our field regardless of the extent to which it applies (Perone, 2018). Therefore, a purpose of *explicitly* conducting reproducible studies, apart from

finding "the truth" and adding to our knowledge base, is to enhance the believability of our science for the general public (Beck, 2017). The harmful consequences of the public being convinced that behavioral science is generally fraudulent, untrustworthy, or unreliable could have harmful effects for society, because it could hasten a turn away from evidence-based practices toward pseudoscience. As an example, consider the "anti-vaccination" movement that actively rejects the scientific evidence on the efficacy and relative safety of vaccines in favor of conspiratorial narratives regarding the alleged harmful effects of vaccines, which has led to outbreaks of vaccine-preventable diseases (Kata, 2010; Stein, 2017).

Purpose of this Article

It is not clear the extent to which the RCP actually represents a true "crisis," as the short- and long-term consequences of alleged failures to reproduce earlier studies have not been determined for psychology in general or for behavior science/analysis and its subfields in particular (e.g., applied behavior analysis, experimental analysis of behavior, behavioral pharmacology, organizational behavior management, clinical behavior analysis, consumer behavior analysis, behavioral economics). Although an emphasis on within- and between-subject "replication" has been part of our methodological tradition since the field's inception (e.g., Barlow & Hersen, 1984; Branch, 2018; Kazdin, 1982; Lanovaz, Turgeon, Cardinal, & Wheatley, 2018; Perone, 2018; Poling & Fuqua, 1986; Sidman, 1960), it behooves us to consider carefully the potential for our research studies to suffer from threats to reproducibility and to adopt practices that mitigate these threats. Therefore, in the remainder of this article, we discuss (a) factors involved in assessing the success of attempts to reproduce a study, (b) suggested reasons for failures of reproducibility, (c) types of validity of experimental studies and related threats thereto, and (d) evidence for threats to reproducibility in the behavior science/analysis literature. We will constrain our comments to experimental studies meant to identify causal (functional) relationships (i.e., treatment/intervention effects) as exemplified by most empirical studies seen in the major behavior science/analysis journals (e.g., *Journal of Applied Behavior Analysis* [*JABA*], *Journal of the Experimental Analysis of Behavior* [*JEAB*], *Journal of Organizational Behavior Management* [*JOBM*], *The Psychological Record* [*TPR*], *The Analysis of Verbal Behavior* [*TAVB*]). One of the primary purposes of this article is to describe scientific behaviors that could contribute to reproducibility problems. We will not attempt to provide a unified system for conducting reproducible studies, because the literature on reproducibility is too vast and varied in opinions to do so in a single article. Rather, we will discuss some suggestions for behavior scientists/analysts to improve the reproducibility of their research studies.

What Constitutes a Replication?

Unfortunately, there are no agreed-upon rules for determining whether or not a study has successfully reproduced another study or the most appropriate methods for assessing the evidence for such a "replication" (Anderson et al., 2016; Begley &

Ioannidis, 2015; Clemens, 2017; Hales et al., 2018; Maxwell et al., 2015; McElreath & Smaldino, 2015; Nosek & Errington, 2017; OSC, 2012). So, we will draw on Goodman et al.'s (2016) distinctions to discuss variables that relate to the *reproducibility* of studies in behavior science/analysis. We acknowledge that this is an arbitrary choice given the many possible definitions of *replication* and *reproducibility*. This underscores the role of "subjective" or investigator-based judgments involved in the topic of reproducibility. We might consider each of Goodman et al.'s three types of reproducibility as falling on a continuum rather than on a binary scale, and the goal would be to move our studies to the "more reproducible" end of each continuum. Regardless of the various forms of reproducibility and definitions thereof, Goodman et al. argue that the main point of scientific research is to provide "cumulative evidence and truth." Put differently, the goal of research should be to avoid false positives (or "false discoveries;" Leek & Jager, 2017) and false negatives. At the most general level, replication involves finding relationships in the same direction and at similar strength as reported in the original study, or what Goodman et al. termed *results reproducibility*. In describing what constitutes reproducibility, Begley and Ioannidis stated, "it is completely reasonable that the one or two big ideas or major conclusions that emerge from a scientific report should be validated and withstand close interrogation" (p. 117). Rosenthal and Rosnow (2009, ch. 18) proposed that:

> [a] general principle that a replication of [a single] experiment which obtains similar results is maximally convincing if it is maximally separated from the first experiment along such dimensions as time, physical distance, personal attributes of the experimenters [e.g., different theoretical approaches], experimenters' expectancy [with respect to the results], and experimenters' degree of personal contact with each other (p. 556)

As we alluded to earlier, the most common criterion for assessing results reproducibility is statistical significance, as determined by finding a *p*-value in NHST below a certain threshold (often called *alpha* and the *significance level*), usually set at .05 or .01, but occasionally adjusted downward to account for an increased false positive (Type I) error rate resulting from multiple comparisons (Maxwell et al., 2015; Rosenthal & Rosnow, 2009, ch. 18; Williams et al., 1999). When using the *p*-value as a criterion for determining replication or results reproducibility, it is possible that the reproducing study will find a *p*-value *above* the standard .05 threshold when the original study found a *p*-value *below* .05, even if that replication study was conducted faithfully to the original and there really is a relationship between/among the variables being studied. That is, the replicating study might find a *false negative*. This could occur because the *p*-value is a sample statistic that is subject to sampling error (random fluctuation). For example, if the original study reported a *p*-value of .01, the probability of reproducing a *p*-value below .05 in a second study is .73; the probability getting a *p*-value above .05 ("not significant") is .27 (Huitema, 2016). So, lack of statistical significance in a single replicating study does not necessarily mean that the first study reported false positives (Huitema, 2016; Leek & Jager, 2017).

We will not discuss their arguments here, but prominent behavior scientists/analysts have frequently critiqued inferential statistics (most notably NHST), questioned their relevance to our research, and advocated for visual analysis as a more effective

alternative (e.g., Branch, 1999, 2018; Branch & Pennypacker, 2013; Johnston & Pennypacker, 2009; Michael, 1974; Perone, 1999; Parsonson & Baer, 1986; Sidman, 1960). They are not alone in their criticisms, because problems with traditional NHST have been described for decades (Cohen, 1990, 1994; Greenwald, Gonzalez, Harris, & Guthrie, 1996; Kirk, 1996; Loftus, 1996; Nix & Barnette, 1998; Shadish, Cook, & Campbell, 2002; Wilkinson & the American Psychological Association Task Force on Statistical Inference Science Directorate, 1999; Valentine, Aloe, & Lau, 2015). Some psychologists have called for the elimination of NHST altogether (e.g., Cumming, 2014), and some journals have banned *p*-values from their pages (e.g., *Basic and Applied Social Psychology*; Lang, Rothman, & Cann, 1998; Trafimow & Marks, 2015). Although behavior scientists/analysts have traditionally eschewed inferential statistics, including NHST, this practice is not universal in the field. For example, studies published in *JEAB* have increasingly reported inferential statistics in their results (Foster, Jarema, & Poling, 1999; Zimmerman, Watkins, & Poling, 2015). So, behavior science/analysis as a whole is not immune from threats to reproducibility related to NHST, nor do we suggest that behavior scientists/analysts reject NHST as a potentially viable scientific practice. NHST can serve a useful purpose when conducted and reported properly, particularly when focusing on measures of effect size and confidence intervals, rather than just the *p*-value (Greenwald et al., 1996; Hales et al., 2018; Harris, 1997; Weaver & Lloyd, 2018). In addition, Jones and Tukey (2000) outlined what they termed "a sensible approach" to significance testing, which they claimed avoids the most common problems associated with traditional NHST (see also Harris, 1997, 2016; Hurlbert & Lombardi, 2009). As Huitema (1986b) noted, the broader scientific community generally expects statistical results, so reporting such results can improve the acceptability of our research to audiences outside of behavior science/analysis, including funding agencies and nonbehavioral journals. This does not mean that one has to conduct traditional NHST, as alternatives exist (e.g., Bayesian, "Neo-Fisherian" methods). We simply suggest that behavior scientists/analysts regularly use appropriate statistical tools at their disposal using "best practices" throughout the "data analysis pipeline" (Leek & Jager, 2017; Leek & Peng, 2015; Simmons, Nelson, & Simonsohn, 2011), while keeping in mind the expectations and/or requirements of our (sometimes nonbehavioral) audiences (Hales et al., 2018).

As other authors in this section replication have discussed (e.g., Branch, 2018; Hales et al., 2018; Killeen, 2018; Kyonka, 2018; Perone, 2018), a primary problem with NHST is the focus on and misinterpretation of the *p*-value. Some common misinterpretations are that the *p*-value tells us: (a) the effect of the independent variable was large or important (or anything about the size of the treatment/intervention effect), (b) the probability that the null hypothesis is true, (c) the probability that the result was due to chance, and (d) that $1 - p$ is the probability that the results will be replicated (Cohen, 1990, 1994; Huitema, 1986b; Huitema & Urschel, 2014; Kirk, 1996). One proper interpretation of the *p*-value is that it is the conditional probability of obtaining a relationship between/among variables at least as large as the one found by the hypothesis test under the condition that there is no actual relationship (Huitema, 2011, p. 8). In other words, it is the probability of getting the data you got, given that the null hypothesis is true. Unfortunately, the *p*-value is often interpreted as the probability that the null is true/false given the data you got (Branch, 2018; Cohen, 1990, 1994; Killeen, 2018; Kyonka, 2018). However, misunderstandings of the meaning of the *p*-value does

not, in and of itself, make NHST invalid, nor are visual and inferential statistical techniques mutually exclusive.

Good data analysis will involve both visualization and quantitative evaluation, possibly including NHST or other inferential techniques. The extent to which significance testing (not necessarily of the traditional NHST variety) can be useful depends on its proper implementation and interpretation, consideration of the audience to whom the results are directed, the need for input for meta-analyses, and situations in which there is a discrepancy between visual analysis and statistical analysis (e.g., when the data are noisy—such as in unstable baselines—and clear patterns cannot be discerned by visual analysis alone; Barlow & Hersen, 1984; Hales et al., 2018; Huitema, 1986b; Kyonka, 2018). Of course, visual analysis should be informed by suggested best practices of data presentation so as to be effective judgment aids (e.g., Tufte, 1990, 1997, 2006, 2009; Tukey, 1977). Even when visual displays satisfy these principles, conclusions based on visual analysis may be contaminated by confirmation bias and subjectivity, in which the researchers "see" data that conform to their hypotheses even when a statistical analysis demonstrates poor evidence of a treatment effect (Bakker & Wicherts, 2011; Hales et al., 2018; Huitema, 2016). Of course, NHST may also be contaminated by subjective decisions made throughout the "data pipeline" (Leek & Peng, 2015).

As mentioned earlier, behavior scientists/analysts have advocated the superiority of visual (graphical) analysis over the use of statistical procedures like NHST. Scientists outside of behavior science/analysis have also noted the importance of visual analysis in analyzing quantitative data (e.g., Cleveland & McGill, 1985, 1986; Cumming & Finch, 2005; Tufte, 1990, 1997, 2006, 2009; Tukey, 1977; Valentine et al., 2015). However, it was first pointed out in the mid-1980s that reliance on visual analysis had been harmful to the field of behavior science/analysis because studies using only this form of analysis were not included in meta-analyses (Huitema, 1986b). Impressive (at least to those in our field) visual outcomes in many areas of behavior science/analysis were completely ignored by those in other areas of psychology who carried out meta-analyses. This is understandable because a meta-analysis cannot incorporate the results of a study that has no quantitative outcome metric. Three decades later, we find that behavior science/analysis researchers are increasingly aware of meta-analysis and the need to report some quantitative outcome metric (Beretvas & Chung, 2008; Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014; Horner, Swaminathan, Sugai, & Smolkowski, 2012). Unfortunately, careful inspections of meta-analyses performed on behavioral data tend to reveal the poor quality of these underlying analyses. A major problem permeating these analyses is the inconsistent nature of the outcome metrics subjected to the meta-analysis and the lack of quantitative metrics being reported in the behavior science/analysis literature.

A concept frequently invoked when describing the effect of a behavioral experiment is "level change." An ongoing study (Collini & Huitema, 2019) of single-case designs (SCD) methodology textbooks, chapters, and expository articles devoted to the visual analysis of SCDs leads to two conclusions regarding current presentations of level change: (a) most sources refer to *level change* as a key concept in the visual evaluation of outcomes; and (b) descriptions of exactly how level change should be determined and reported are both inconsistent and ambiguous. Some sources provide only a diffuse description of level or level change whereas others provide details, but these details

often differ in major ways from one source to another. It is interesting that the inconsistent descriptions of level change measures provided in the visual analysis literature are not the same as the various level-change measures often recommended and computed by behavioral statisticians. The inconsistencies among these statistical methods have been discussed extensively in the literature (Huitema, 1986b, 2004, 2018; Huitema & McKean, 2000; Huitema, McKean, & Laraway, 2008). It turns out that some of these discrepancies are massive. Examples can be found of one method estimating level change to be over 100 times larger using one method than using another. In some cases, the level-change estimates provided for a single study differ in sign as well as magnitude. The implications of carrying out a meta-analysis that rests on a foundation of these flawed metrics should be obvious. The finding that results of some behavioral experiments appear not to replicate is not an unexplainable crisis; in the case of inferences about level change it is a predictable outcome of an understandable but flawed process of measurement and analysis. Because "level change" is inconsistently defined within the visual analysis literature, as well as within the statistical analysis literature, it is not a surprise that results differ when carried out by various researchers using different methods. A similar case could be made with respect to choices of statistical procedures based on misunderstandings of the assumptions of those procedures (e.g., Huitema, 1986a, 1988; Huitema & McKean, 1998)

Reasons for Failures of Reproducibility

There are many reasons for a failed attempt to reproduce the results of an experiment that have nothing to do with the "truthfulness" of the original study, so apparent "negative" results do not automatically indicate that the first study reported a false positive (Huitema, 1986b; Leek & Jager, 2017; Lynch, Bradlow, Huber, & Lehmann, 2015; Maxwell et al., 2015; Schmidt & Oh, 2016; Simonsohn, 2015). It is possible, for example, that the replicating study was underpowered due to small sample size, was affected by error (e.g., sampling or measurement), or was not conducted closely enough to the original study such that it differed from it in important ways (see, e.g., Gilbert et al., 2016). Unfortunately, researchers who conduct replication studies sometimes commit the fallacy of interpreting nonsignificant results as evidence the null hypothesis is true (Maxwell et al., 2015; Schmidt & Oh, 2016). So, judging an original study based on the success or failure of a single replication study seems unwise. Rather, some authors have recommended the evaluation of the results of multiple studies using meta-analytic strategies (e.g., Braver, Thoemmes, & Rosenthal, 2014; Maxwell et al., 2015; Schmidt & Oh, 2016), although the results of meta-analysis in the same area of research may not always agree (de Vrieze, 2018). Other options include collaborations across multiple, independent laboratories to attempt to reproduce research findings before or after publication (e.g., Frank et al., 2017; OSC, 2015; Schweinsberg et al., 2016). It would be interesting to see behavior scientists/analysts conduct similar large-scale attempts at reproducing research findings.

Researchers have proposed a plethora of threats to reproducibility (e.g., Begley & Ioannidis, 2014). Reasons that we feel might be relevant to behavior science/analysis include (a) "variations in measurement technology, population characteristics, study protocols, and data analyses" (Leek & Jager, 2017, p. 114); (b) "seemingly irrelevant

factors" like the phrasing and font of experimental instructions to participants (Kahneman, 2014); (c) publication bias (e.g., authors' only publishing positive, "interesting," or statistically significant findings and/or selectively reporting statistically significant over nonsignificant results; reducing methodological detail due to journal page-length considerations; and insufficient specification of experimental conditions; Begley & Ioannidis, 2014; Huitema, 1986b, 2016; Lanovaz, Robertson, Soerono, & Watkins, 2013; OSC, 2015); (d) low statistical power in the replicating study (Maxwell et al., 2015; Rosnow & Rosenthal, 1989); (e) "improper execution of an experimental technique" (Nosek & Errington, 2017); (f) "insufficient, incomplete, or inaccurate reporting of methodologies" (Errington et al., 2014); (g) honest errors (i.e., nonmalicious or unintentional violations of scientific integrity, such as failing to select the most appropriate statistical test; Huitema, 2016; Resnick & Stewart, 2012, Shaw, 2018); (h) the notion that the same (or similar) *p*-value found in the original study will be found in the study attempting to reproduce that study (Huitema, 2016; Leek & Jager, 2017); (i) regression to the mean (Leek & Jager, 2017); and (j) contextual factors, including location of the study (e.g., rural vs. urban), the population from which the samples were drawn (e.g., racially diverse vs. mostly White; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016), and the selection of individuals from Western, educated, industrialized, rich, and democratic (WEIRD) societies (Henrich, Heine, & Norenzayan, 2010).

These threats to reproducibility can be categorized as those involving nature, the researcher, or the publication process. With respect to nature, uncontrollable and unexpected variables as well as sampling error may influence our results. There is only so much researchers can do about this. A commitment to transparency in reporting the conditions under which an experiment was conducted should help readers to determine the extent to which variables and conditions idiosyncratic to a particular study might have influenced the results (Hales et al., 2018; Leek & Jager, 2017). Although the *ideal* of behavior science/analysis research involving extensive manipulation and control of nuisance variables promulgated by Sidman (1960) is appealing, the truth is that time and other resources are limited. Even if it were possible to do so (and it is not clear that it is possible), we may not be able to spend the necessary time and energy to control many possible nuisance variables that might influence our results. In applied settings, the need to get things done may compete with the desire to exert the best possible level of experimental control (Ferron & Jones, 2006). Simply put, most of the world does not resemble an operant conditioning chamber—it is messy. Although experimental control is an ideal to which we should strive, it may not always be practically obtainable (Barlow & Hersen, 1984, ch. 2). Experimental control over nuisance variables helps us reduce the noise (Huitema, 2011, pp. 20–21), but we cannot eliminate the noise altogether. So, statistical techniques, including NHST supplemented with confidence intervals and effect size measures, can help us cut through that noise to find the signal.

In discussing researcher-related threats to reproducibility, we will not focus on honest errors or deliberate scientific misconduct. The former cannot be avoided entirely (but can be mitigated, for example, by better statistics and methodology education for researchers) and the latter should require no comment from us. Between acceptable and blatantly unethical (fraudulent) practices by researchers, there is a "gray area" of QRPs identified by John et al. (2012); some of these overlap with the reasons for replication failure mentioned above. The most common self-reported QRPs among nearly 6,000 academic psychologists in the United States were (in order of self-

admission rate): (a) failing to report all of a study's dependent measures; (b) deciding to collect more data after looking to see whether the results were significant; (c) selectively reporting studies that "worked"; (d) deciding whether to exclude data after looking at the impact of doing so on the results; (e) failing to report all of a study's conditions; (f) reporting unexpected findings as having been predicted from the start; (f) "rounding off" a *p*-value (e.g., reporting that a *p*-value of .054 as .05 or less); and (g) stopping data collection earlier than planned because one found the results that one had been looking for (p. 525). Some of these QRPs fall under the term "*p*-hacking," which can include "choosing to report a subset of multiple independent variables or adding observations [data points] until the effect of interest is significant" (Bruns & Ioannidis, 2016). Simonsohn, Nelson, and Simmons (2014) gave other examples of *p*-hacking: "adding a covariate [post hoc], data peeking (adding new observations), data exclusions (e.g., dropping "outliers"), choosing among several correlated dependent variables, and choosing among several experimental conditions" (p. 546). Another QRP involves reporting unexpected findings as if they were predicted from the start; this has been termed HARKing (Hypothesizing After the Results are Known; Kerr, 1998). Kerr defined this QRP as "presenting a *post hoc* hypothesis in the introduction of a research report as if it were an *a priori* hypothesis" (p. 197). In addition, Bakker and Wicherts (2011) found evidence that research studies in psychology journals sometimes misreport their statistical results "often in line with the researchers' expectations" (p. 666); these incidents could be honest (possibly "unconscious") errors or scientific misconduct. Behavior scientists/analysts should consider these QRPs and avoid engaging in them.

Many of these QRPs occur because of "researcher degrees of freedom," which refers to the influence that a number of different decisions that researchers can make in the data collection and analysis phases have on the likelihood of reporting a false positive (Hales et al., 2018). The more researcher degrees of freedom, the higher the probability of a false positive. Simmons et al. (2011) noted that "despite empirical psychologists' nominal endorsement of a low rate of false-positive findings ($\leq .05$), flexibility in data collection, and reporting dramatically increases actual false-positive rates" (p. 1359). These authors proposed requirements for researchers (in Table 2, p. 1362). Although not all of these apply to behavior science/analysis research as traditionally conducted, they should be considered whenever behavior scientists/analysts design and conduct studies that use NHST. We will briefly discuss a few of these requirements that we feel apply to behavior science/analysis research. Their first requirement that "authors must decide the rule for terminating data collection before data collection begins and report this rule in the article" is consistent with the recommendations of behavior scientists/analysts regarding choosing and reporting stability criteria in SCDs (e.g., Kazdin, 1982; Kratochwill et al., 2010; Perone, 1991; Poling, Methot, & LeSage, 1995). So, behavior scientists/analysts should have no problem meeting this requirement or reporting why they could not do so. We should note that there are no firm decision rules for determining stability in SCDs, so this does add a researcher degree of freedom. Simmons et al. (2011, Table 2, pp. 1362–1363) also recommend that "authors should list all variables collected in a study" and "authors must report all experimental conditions, including failed manipulations"; this should also not be a problem for behavior scientists/analysts. To the extent that this is relevant, "if observations are eliminated, authors must also report what the statistical results are if those observations are included." This includes outlier elimination procedures and other data-cleaning

methods. Researchers should also compute and report effect size measures and confidence intervals for comparisons of interests, whether or not they also report NHST results. Nonsignificant results should be published as part of an entire study with all tested variables included in the analysis. Authors should describe plausible explanations for nonsignificant findings and suggest potential improvements for future studies. Authors should be explicit about their choices, and this might involve more first- or third-person narratives about these choices, with defending data and/or logic. Again, this should not prove difficult for behavior scientists/analysts, but it is worth remembering. Hales et al. (2018) describe additional strategies that individual researchers can take to improve the reproducibility of their research; we recommend readers consider these strategies when designing, conducting, and reporting their research, as well as when consuming or reviewing published studies.

With respect to the publication process, Simmons et al.'s (2011, p. 1363) suggestions for reviewers and editors include: (a) ensuring that authors follow Simmons et al.'s requirements and "prioritize transparency over tidiness"; (b) being more tolerant of imperfections in results; (c) requiring authors to describe and defend any arbitrary decision in the data analysis process and show that the conclusions are not dependent on such decisions; and (d) requesting exact replications of experiments when the authors' justifications of data collection and analysis decisions are not compelling. Begley and Ioannidis (2014) stated that the failures of reproducibility are "a consequence of a system that is willing to overlook and ignore lack of scientific rigor and instead reward flashy results that generate scientific buzz or excitement" (p. 118). In their cultural evolutionary analysis of "bad science," Smaldino and McElreath (2018) described institutional reasons for the continued use of poor methodology and misuse of statistical techniques in published research. These reasons involve incentives that reward the number of publications and publications that are deemed "innovative," both of which can contribute to publication bias. From the literature on reproducibility, we can deduce additional suggestions. Journals should not only publish "exciting" and "novel" findings and ignore seemingly mundane or statistically nonsignificant results (i.e., the classic "file drawer problem"; Rosenthal, 1979). Direct replications should be encouraged with authors providing explicit descriptions of the similarities and differences between the original and replicating study. Reviewers should consider these issues when evaluating articles and make appropriate suggestions to authors whose work they review. At the level of institutional incentive structures, universities should consider ways to evaluate candidates for hiring, tenure, and promotion that avoid mere counting of publications and publications of "exciting" or "newsworthy" findings (Smaldino & McElreath, 2018).

Obviously, our traditional methodology based on the works of Sidman and Skinner does not protect us from publication bias, so it goes without saying that behavior science/analysis journals should seek to publish high-quality replications (and meta-analyses) and should encourage the submission of these types of articles. Whether this occurs in the normal course of a journal's publication process or in a special section on replications will depend on the journal's editors to decide. Journals should provide incentives for honest reporting and judge a study's results not just based on *p*-values or single researcher's judgments of visual analysis. Nonsignificant results of NHST should be publishable if the study has methodological rigor, describes a priori power determinations, explores potential reasons for those nonsignificant results, avoids claiming such results prove the null hypothesis is true, and emphasizes confidence intervals and

effect size measures. Journals should require authors to discuss practical significance in addition to statistical significance (Kirk, 1996), which seems like it would be less of a problem in applied studies than in fundamental science studies. We should collect data on the results of efforts to improve reproducibility and encourage researchers to evaluate the effects of such efforts through systematic literature reviews and meta-analyses. Ioannidis (2014) described several ways that researchers, journals, universities, and other research institutions can work to "make more published research true." Likewise, the OSC (2017) discussed ways to improve the likelihood that our research is reproducible, as did Shrout and Rodgers (2018) and Hales et al. (2018).

Types of Validity and Threats Thereto

In addition to the considerations above, a brief discussion of research validity and its threats can provide us with suggestions for reducing the chances of finding and reporting false positives and/or false negatives. Methodologists (e.g., Shadish et al., 2002) have described different types of validity that a given research study can possess and have enumerated ways that these types of validity can be threatened. These threats to validity can negatively affect the reproducibility of a given research study. In the next two sections, we will focus on two of Shadish et al.'s (2002) four types of validity (statistical conclusion and construct) and describe how certain threats to these types of validity could compromise the replicability of studies in behavior science/analysis. This is not meant to be an exhaustive discussion but rather a starting point for behavior scientists/analysts to consider how their research might be affected by these threats to validity. Our limited discussion is not meant to convey the notion that the rest of Shadish et al.'s coverage is irrelevant to behavior science/analysis research; quite the contrary, because we feel that all behavior scientists/analysts could benefit from considering their presentation. For a more thorough examination of the relevance of Shadish et al.'s types of threats to validity in single-case studies in behavior science/analysis using visual analysis, see Petursdottir and Carr (2018). Regardless of the type of design and analysis method used, we encourage behavior scientists/analysts to consider all four types of validity and relevant threats when designing, conducting, reporting, and evaluating research in the field.

Statistical Conclusion Validity

Shadish et al. (2002) stated that: "Statistical conclusion validity [SCV] concerns two related statistical inferences that affect the covariation component of causal inferences: (1) whether the presumed cause and effect covary and (2) how strongly they covary" (p. 42). In other words, researchers decide the extent to which their data provide evidence of a treatment effect (in a particular direction) and the strength or size of that effect. Likewise, Kratochwill et al. (2010) stated that researchers using SCDs "traditionally have relied on visual analysis of the data to determine: (a) whether evidence of a relation between an independent and an outcome variable exists; and (b) the strength or magnitude of that relation" (p. 17); this is nearly identical to the definition of SCV. So, it should be clear that scholars conducting research in the tradition of behavior science/analysis should be concerned with SCV even if they use SCDs and visual analysis.

In assessing SCV, the first decision involves statistical and/or visual analysis to determine the extent to which the dependent variable changes as a function of manipulations of the independent variable (i.e., *covariation*). The most common method of evaluating covariation in the behavioral and social sciences is NHST. In making such decisions about covariation, it is always possible that the researchers' judgments about their data are in error. That is, we may claim (a) the treatment was effective when it really was not (a false positive, false discovery, or *Type I error*), sometimes labeled an error of "gullibility or overeagerness" (Rosnow & Rosenthal, 1989); or (b) the treatment was not effective when it really was (a false negative, or *Type II error*).

Although language regarding Type I and II errors is used mostly in the context of NHST, the notion that researchers may draw incorrect conclusions about their results does not require operating within this framework (e.g., Hanley, 2012; Rooker, Iwata, Harper, Fahmie, & Camp, 2011; Shirley, Iwata, & Kahng, 1999). This is because all experimental designs have a basic logic that typically involves a comparison of a baseline/control condition/phase (no treatment, current treatment, etc.) with at least one treatment condition/phase (Barlow & Hersen, 1984; Kazdin, 1982; Kratochwill et al., 2010). If behavior changes in *detectable*, orderly, and meaningful ways from the baseline phase to the treatment phase(s), then we (tentatively) conclude that the intervention had an effect; otherwise, we conclude that we do not have compelling evidence of a treatment effect (not that there is *no* effect).

This logic is not too different in principle from the logic of comparing a null hypothesis (a nonzero but scientifically uninteresting treatment effect) with an "alternative" hypothesis (there was a scientifically interesting treatment effect). Of course, the specifics differ, but the logic is similar (White, Rusch, Kazdin, & Hartmann, 1989). So, false positives (Type I errors) and false negatives (Type II errors) can occur any time researchers make a conclusion about their data, regardless of their data-analytic methods (Ferron & Jones, 2006; Matyas & Greenwood, 1990). Simply put, researchers are humans who make fallible judgments, and our assessment of our data will not lead us to the correct conclusion with 100% accuracy. This can occur with both statistical and visual analyses (Brossart, Parker, Olson, & Mahadevan, 2006; Fisch, 1998).

Previous studies have questioned the reliability of visual analysis to detect treatment effects (e.g., Brossart et al., 2006; Fisch, 1998; Matyas & Greenwood, 1990), although others have found more promising results (Bobrovitz & Ottenbacher, 1998). Nonetheless, it is possible that behavior scientists/analysts using visual analysis could make decision errors with respect to covariation in the independent and dependent variables. Researchers have provided several suggestions for evaluating the evidence of such covariation in SCDs using visual and/or statistical methods (e.g., Barlow & Hersen, 1984; Fisher, Kelley, & Lomas, 2003; Huitema, 2011; Kazdin, 1982; Kratochwill et al., 2010; Lane & Gast, 2014; Lanovaz et al., 2017). Fortunately, it is possible to train researchers to reliably visually evaluate graphed data. Further, it has been shown (e.g., Huitema, 1979) that it is possible to train researchers to provide effect estimates from visual analyses that are similar to those provided by complex statistical analyses (see also, Fisher, Anderson, Peng, & Leek, 2014; Fisher et al., 2003). Better training in both visual analysis and statistical methods appropriate for behavioral research is a worthwhile goal for the field and should be included in behavior science/analysis curricula at the undergraduate and graduate levels. This should improve the SCV and *inferential reproducibility* of behavioral studies (Goodman et al., 2016).

As noted previously, the appropriate use of statistics is cited as a major factor in replication problems. Although *p-hacking* and the overestimation of effect sizes describe a mere subset of such issues, recent literature has extended this notion through an inquiry into conventional error rates. Although Type I and II error rates are obtained through the near arbitrary selection of alpha (α) and power ($1 - \beta$) values, existing research has posited the use of novel error rates. As an example, work by Gelman and Carlin (2014) offered the calculation of error rates termed: *Type S (sign)* and *Type M (magnitude)*, also termed the *exaggeration ratio* and *error rates*, respectively. Type S error deals with the "probability that the replicated estimate has the incorrect sign, if it is statistically significantly different from zero" (p. 644), whereas Type M error refers to the "expectation of the absolute value of the estimate divided by the effect size, if statistically significantly different from zero" (p. 644). Gelman and Carlin have streamlined calculations of Type S and M errors with an R function, `retrodesign()`. Put simply, these errors are determined retrospectively using the effect size, standard errors, set alpha level (e.g., $\alpha = .05$), and degrees of freedom. We encourage researchers using NHST to follow developments in this area, explore reporting these new statistics, and perform rigorous design analyses both a priori and a posteriori.

The second decision described by Shadish et al. (2002) involves quantitative description of *how much* the dependent variable changed under various treatment conditions (i.e., the effect size). As we noted earlier, one of the problems with NHST is that that researchers may interpret the *p*-value as a measure of the size of a treatment effect. We can avoid this problem by proper reporting and interpretation of effect size measures. Regardless of one's choice to use inferential statistics like NHST and confidence intervals in behavior science/analysis studies, reporting effect size measures is essential and does not require NHST (Valentine et al., 2015). Effect size measures for various research designs and statistical analyses are widely available (e.g., Cohen, 1992; Ellis, 2010; Kirk, 1996; Rosenthal, Rosnow, & Rubin, 2000), and researchers have described several such measures for SCDs (e.g., Beretvas & Chung, 2008; Brossart et al., 2006; Harvey, Boer, Meyer, Evans, 2009; Heyvaert et al., 2014; Huitema, 2011; Kratochwill et al., 2010; Olive & Smith, 2005; Parker & Vannest, 2009; Shadish, Hedges, & Pustejovksy, 2014a; Shadish et al., 2014b). At this point, however, there is no clear consensus on what effect size measures for SCDs have the most desirable statistical properties (Horner et al., 2012), so this remains an exciting area for further research and discussion. Such research will surely benefit our field and perhaps others.

Irrespective of the metric used, computing effect size measures require proper data collection and interpreting them requires strong methodologies to ensure high-quality data. However, Harvey et al. (2009) observed that in their meta-analysis of behavioral treatments for challenging behavior: "Research reports are still being published that may appear visually to demonstrate a treatment effect but which lack the requisite data needed for the calculation of effect-size algorithms with sufficient statistical power to allow valid conclusions" (p. 78). Several guides to experimental design in behavior science/analysis have been available for years (e.g., Barlow & Hersen, 1984; Huitema, 2011, Ch. 18-21; Johnston & Pennypacker, 2009; Kazdin, 1982; Poling & Fuqua, 1986; Poling et al., 1995; Sidman, 1960). Nevertheless, studies in behavior science/analysis do not always have high methodological rigor. For example, in their review of habit reversal techniques for treating tics, Carr and Chong (2005) noted: "However, during our initial review of the literature, it became apparent that the quality of the

studies in the area was quite variable" (p. 860) and "Results indicate that although research has been conducted in this area for almost three decades, the majority of studies contain considerable methodological shortcomings" (p. 858). Likewise, in their meta-analysis of the influence of preintervention functional behavioral assessment on the effectiveness of interventions for problem behavior, Hurl, Wightman, Haynes, and Virues-Ortega (2016) noted that the "methodological quality of the studies varied greatly," with the main concerns being too few data points in study phases and "high within-phase variability and high data point overlap across study phases" (p. 76).

In considering SCV, we should not just be concerned about the final step of computing descriptive statistics, inferential statistics (e.g., NHST, confidence intervals), and effect size measures. The data move through many phases before ending up in an article or presentation. As we alluded to earlier, Leek and Peng (2015, p. 612) described a "data pipeline" that includes steps that eventually lead to summary statistics (e.g., means, standard deviations, effect size measures), and, if inferential statistics are used, the *p*-value and/or confidence interval. Errors or misconduct at each step could occur in one or more of the phases of the data pipeline: experimental design, data collection, data cleaning yielding "tidy data"; exploratory data analysis yielding potential statistical models; and statistical modeling that lead to summary statistics (e.g., measures of central tendency, variability, slope). Apart from the importance of behavior scientists/analysts to engage in appropriate behaviors at each step in the data pipeline, we have a lot to offer in examining the behaviors involved. Leek and Peng suggested that "the ultimate goal is evidence-based data analysis" and noted "Statistical research largely focuses on mathematical statistics, to the exclusion of the behaviour and processes involved in data analysis. To solve this deeper problem, we must study how people perform data analysis in the real world" (p. 612). Given our expertise in studying human behavior "in the real world," behavior scientists/analysts could contribute to the effort of developing an evidence-based data analysis. We encourage researchers to do so at both the level of behavioral and cultural analysis (see Smaldino & McElreath, 2018); here, we have another interesting area for additional research to which our field could contribute.

As we mentioned earlier, a common threat to SCV is low statistical power, which reduces a test's ability to detect statistically significant effects (i.e., find evidence of covariation). This may lead the researcher to incorrectly declare that the treatment had no effect (false negative or Type II error). Low power may result from a combination of several factors, including small effect sizes, small sample sizes, measurement error in the dependent variable, and error variation that obscures effect sizes (Huitema, 2011, p. 8; Shadish et al., 2002, pp. 45–48). When behavior scientists/analysts (and other researchers) plan on using NHST or confidence intervals, they should consider ways to increase the power of their techniques and assess power a priori for the techniques they plan on using. Even if researchers use visual analysis and not NHST, the number of data points is still important for making correct decisions about the effects of an intervention (Kratochwill et al., 2010; Lanovaz et al., 2018). As with NHST, in general more is better when using visual analysis (up to a point). Unfortunately, in their review of the SCD literature (including those behavior science/analysis studies), Shadish and Sullivan (2011) found that nearly half of the studies examined had fewer than 5 data points in the first baseline phase. Kratochwill et al. (2010) recommended that there should be *at least* 5 data points in this phase; Lanovaz, Huxley, and Dufour (2017) suggested that a minimum of 3 and 5 data points should be included in the first baseline

(A_1) and treatment (B_1) phases in ABAB designs, respectively. Huitema (2011, p. 389) listed 10 details to consider in determining the number of data points in each phase of an SCD but stated that the initial baseline phase should have *at least* 6 data points for proper statistical analysis using his regression-based approach.

Apart from the importance of collecting and reporting more data points, another variable that influences the power to detect treatment effects in both NHST and visual analysis is error variation. In inferential statistics, this is generally quantified by measures of variability, including standard errors and similar statistics. In visual analysis of SCDs, the lack of stability and/or high variability ("noisy data") in one or more phases also represents a source of error variation, which could produce a similar result as does variability in inferential statistics (i.e., possibly obscuring treatment effects due to noisy data). As we mentioned earlier, despite the admirable Skinnerian and Sidmanian ideal of comprehensively controlling most relevant nuisance variables to reveal the effects of the independent variable, this may not always be possible, so we may need some statistical way of dealing with error variation (Weaver & Lloyd, 2018).

Construct Validity

Shadish et al. (2002) stated that "construct validity is fostered by . . . starting with a clear explication of the person, setting, treatment, and outcome constructs of interest. . ." (p. 66) and that "Construct validity involves making inferences from assessments of *any* of the sampling particulars in a study to the higher-order constructs they represent" (p. 70; emphasis in original). Using Goodman et al.'s (2016) terms, proper explication of constructs is necessary for *methods reproducibility*. Although one might find examples of many of these threats to construct validity in the behavior science/analysis literature, we will discuss only a few. One threat that has been examined repeatedly in reviews of behavior science/analysis studies is *inadequate explication of constructs*, which involves properly defining the participants, research setting, independent variables, and dependent variables. Reviews have shown that research studies in behavior science/analysis suffer from this threat to their construct validity. For example, Peterson, Homer, and Wonderlich (1982) identified problems with the integrity of independent variables in *JABA* studies, noting "the majority of [*JABA*] articles published do not use any assessment of the actual occurrence of the independent variable and a sizeable minority do not provide operational definitions of the independent variable" (p. 477). Likewise, Gresham, Gansle, and Noell (1993) reported that two-thirds of studies in *JABA* with children participants failed to provide clear operational definitions of the independent variable. Although more recently McIntyre, Gresham, DiGennaro, and Reed (2007) found improvements in reporting of operational definitions of the independent variable, they also noted that "Reporting of treatment integrity data has been relatively stable and low over the years" (p. 666), and "Just under half of the included studies (n = 69; 45%) were considered to be at high risk for treatment inaccuracies" (p. 664). Armstrong, Ehrhardt, Cool, and Poling (1997) found that over three-quarters of the 39 studies they reviewed in the *Journal of Developmental and Physical Disabilities* did not provide treatment integrity data. Apart from posing a threat to construct validity (by not accurately describing the independent variable *as it was actually applied*), failure to ensure treatment integrity could also affect the SCV and internal validity of a study (i.e., the demonstration of a causal relationship between the independent and dependent variables; Shadish et al., 2002).

Another threat to construct validity found in the behavioral literature involves the inadequate description of subjects/participants and their characteristics. Poling and colleagues found a lack of reported information about participants' medication regimen in applied studies (Poling, Grossett, Karas, & Breuning, 1985; Weeden et al., 2011) and the techniques to identify reinforcers in people with autism (Weeden & Poling, 2011). So, for example, the construct "person with autism" may not accurately describe a person diagnosed with autism who is receiving pharmacotherapy while participating in an ABA study that uses stimuli identified through a certain type of reinforcement assessment. This could also have implications for internal validity and SCV because drugs may interact with behavioral treatments, rendering them more/less effective (e.g., methylphenidate can serve as an abolishing operation for edibles; Northup, Fusilier, Swanson, Roane, & Borrero, 1997). Not knowing, for instance, how a "reinforcer" was identified in the presence of a drug with potential motivational effects (i.e., serving as a motivating operation) could complicate efforts to establish covariation between the independent and dependent variables and interpret the results. This would also affect a study's external validity (i.e., the extent to which research results generalize to other persons, settings, treatments, and outcomes; Shadish et al., 2002) because the reader would be missing information about key functional variables, making generalizing difficult and/or inaccurate.

Apart from concerns regarding the validity of research findings, adequate descriptions of participant characteristics are also important for reasons related to social justice and equity (e.g., health disparities). It is vital that behavior scientists use diverse samples that better represent humanity and meet the needs of the stakeholders of the interventions. We cannot assess how well and to what extent we are meeting the needs of various underserved groups if we do not adequately collect and report the characteristics of our participants in applied studies. Given the importance of culturally sensitive interventions, the inclusion of people with different characteristics should be a priority. We should also routinely collect and report social validity data to measure the extent to which stakeholders and the community find our interventions, goals, and results important, acceptable, and viable (Wolf, 1978; Schwartz & Baer, 1991). Unfortunately, in some cases, social validity data for behavioral modification studies have not been regularly reported (Armstrong et al., 1997). Obviously, for conducting socially valid research with diverse populations, a clear specification of what participant characteristics are relevant to our interventions is important (Fong, Catagnus, Brodhead, Quigley, & Field, 2016). Identifying these characteristics will require empirical research. As behavior scientists/analysts, we are in a good position to do this type of work.

Conclusion

In this article, we discussed the so-called RCP and concepts related to the reproducibility of research studies, including scientific practices and validity threats. We also identified ways that we as a field could improve the reproducibility of our studies. We provided examples from the behavior science/analysis literature that suggest our studies are not immune from threats to reproducibility or validity. The point of these examples is not to cast doubt on the validity of most behavior science/analysis studies. Instead, we hoped to show that concerns about reproducibility, replication, and validity are reasonable when considering behavior science/analysis research as practiced rather than

merely an ideal based on our historical mythology. Whether we use SCDs or other designs and visual and/or statistical analyses, we must think about ways to increase the validity of our scientific findings/conclusions. We can never reduce the probability of our decision errors to zero, but we can take actions to improve our chances by contemplating the threats to reproducibility and validity briefly summarized above. We described some of these actions, but we have only scratched the surface. The interested reader will have no trouble finding relevant resources using the articles we have cited, including those in this special section. Although our conclusions may not agree completely with our esteemed colleagues who have contributed to this special section, a diversity of opinion can only strengthen the field by offering alternative scientific narratives for further testing, application, and refinement.

Apart from working to improve the reproducibility of our research, we feel that behavior scientists/analysts could contribute to a *science of science*, which could incorporate, for example, the metascience approach advocated by Schooler (2014), metaresearch and journalology (Couzin-Frankel, 2018; Stokstad, 2018), and/or the cultural evolutionary approach described by Smaldino and McElreath (2018). We have an intellectual tradition of (a) studying scientific practices as *behavior* (e.g., Michael, 1974; Skinner, 1957/2014), and (b) focusing on the actual practice of science rather than on an idealized version of it (e.g., Skinner, 1956). It is interesting that it seems that we sometimes hold an idealized Skinnerian/Sidmanian version of behavior science/analysis that does not always resemble the science as it is practiced, particularly in applied settings. This does not invalidate the tradition or methodological ideas of Sidman or Skinner, but it should lead us to consider how scientists *actually behave* scientifically given the contexts in which they conduct their research (e.g., university incentive structures, journal requirements, resource constraints, stakeholder needs). This seems a perfect opportunity to apply our focus to the behaviors and cultural practices that lead to reproducible and valid research. A behavioral and cultural analysis of evolving best practices in behavioral science could contribute much to our science and to society as a whole. It would also benefit behavior scientists/analysts to contact the methodological literature outside of the Skinnerian/Sidmanian tradition (e.g., Shadish et al., 2002; Tukey, 1977), because this literature has helpful ideas that could strengthen our research practices, especially with respect to the topic of reproducibility and quantification of research results.

References

Anderson, C. J., Bahnik, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., et al. (2016). Response to Comment on Estimating the reproducibility of psychological science. *Science, 351*(6277), 1037c. https://doi.org/10.1126/science.aad9163.

Armstrong, K. J., Ehrhardt, K. E., Cool, R. T., & Poling, A. (1997). Social validity and treatment integrity data: Reporting in articles published in the *Journal of Developmental and Physical Disabilities*, 1991–1995. *Journal of Developmental & Physical Disabilities, 9*(4), 359–367.

Aschwanden, C. (2015). Science isn't broken. *FiveThirtyEight.* Retrieved from https://fivethirtyeight.com/features/science-isnt-broken/

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods, 43*(3), 666–678.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change*. New York: Pergamon.

Bartels, J. M. (2015). The Stanford prison experiment in introductory psychology textbooks: A content analysis. *Psychology Learning & Teaching, 14*(1), 36–50.

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation research, 116*(1), 116–126.

Beck, J. (2017). The challenge of fighting mistrust in science. *The Atlantic Monthly.* Retrieved from https://www.theatlantic.com/science/archive/2017/06/the-challenge-of-fighting-mistrust-in-science/531531/

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment & Intervention, 2*(3), 129–141.

Bobrovitz, C. D., & Ottenbacher, K. J. (1998). Comparison of visual inspection and statistical analysis of single-subject data in rehabilitation research. *American Journal of Physical Medicine & Rehabilitation, 77*(2), 94–102.

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst, 22*(2), 87–92.

Branch, M. N., & Pennypacker, H. S. (2013). Generality and generalization of research findings. In G. J. Madden (Ed.), *APA Handbook of Behavior Analysis* (Vol. *1*, pp. 151–175). Washington, DC: American Psychological Association.

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.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9*(3), 333–342.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*(5), 531–563.

Bruns, S. B., & Ioannidis, J. P. (2016). *P*-curve and *p*-hacking in observational research. *PLoS One, 11*(2), e0149144.

Carr, J. E., & Chong, I. M. (2005). Habit reversal treatment of tic disorders: A methodological critique of the literature. *Behavior Modification, 29*(6), 858–875.

Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys, 31*(1), 326–342.

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science, 229*(4716), 828–833.

Cleveland, W. S., & McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies, 25*(5), 491–500.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–1312.

Cohen, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003.

Collini, S. A., & Huitema, B. E. (2019). Effect metrics for behavioral data. Paper to be presented at the Association for Behavior Analysis International Conference, Chicago.

Couzin-Frankel, J. (2018). Journals under the microscope. *Science, 361*(*6408*), 1180–1183. https://doi.org/10.1126/science.361.6408.1180.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*(*2*), 170–180.

de Vrieze, J. (2018). The metawars. *Science, 361*(6408), 1184–1188.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*, 621. https://doi.org/10.3389/fpsyg.2015.00621.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). Science forum: An open investigation of the reproducibility of cancer biology research. *Elife, 3*, e04333.

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education, 75*(1), 66–81.

Fisher, A., Anderson, G. B., Peng, R., & Leek, J. (2014). A randomized trial in a massive online open course shows people don't know what a statistically significant relationship looks like, but they can learn. *PeerJ, 2*, e589. https://doi.org/10.7717/peerj.589.

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*(3), 387–406.

Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21*(1), 111–123.

Fong, E. H., Catagnus, R. M., Brodhead, M. T., Quigley, S., & Field, S. (2016). Developing the cultural awareness skills of behavior analysts. *Behavior Analysis in Practice, 9*(1), 84–94.

Foster, T. M., Jarema, K., & Poling, A. (1999). Inferential statistics: Criticised by Sidman (1960), but popular in the *Journal of the Experimental Analysis of Behavior*. *Behaviour Change, 16*(3), 203–204.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421–435.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science, 351*(6277), 1037–1037.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine, 8*(341), 1–6.

Greenwald, A., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology, 33*(2), 175–183.

Gresham, F. M., Gansle, K. A., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis, 26*(2), 257–263.

Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2018). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, 1–19. https://doi.org/10.1007/s40614-018-00186-8.

Hamblin, J. (2018). A credibility crisis in food science. *The Atlantic Monthly.* Retrieved from https://www.theatlantic.com/health/archive/2018/09/what-is-food-science/571105/

Haney, C., Banks, W. C., & Zimbardo, P. G. (1973). A study of prisoners and guards in a simulated prison. *Naval Research Review, 30*, 4–17.

Hanley, G. P. (2012). Functional assessment of problem behavior: Dispelling myths, overcoming implementation obstacles, and developing new lore. *Behavior Analysis in Practice, 5*(1), 54–72.

Hantula, D. A. (2018). Behavior science emerges. *Perspectives on Behavior Science, 41*(1), 1–6.

Harris, R. J. (1997). Significance tests have their place. *Psychological Science, 8*(1), 8–11.

Harris, R. J. (2016). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger, (Eds.) *What if there were no significance tests?* (pp. 179–206). New York: Routledge.

Harvey, S. T., Boer, D., Meyer, L. H., & Evans, I. M. (2009). Updating a meta-analysis of intervention research with challenging behaviour: Treatment validity and standards of practice. *Journal of Intellectual & Developmental Disability, 34*(1), 67–80.

Haslam, S. A., & Reicher, S. D. (2012). Contesting the "nature" of conformity: What Milgram and Zimbardo's studies really show. *PLoS Biology, 10*(11), e1001426.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral & Brain Sciences, 33*(2–3), 111–135.

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 35*(10), 2463–2476.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education & Treatment of Children, 35*(2), 269–290.

Huitema, B. E. (1979). *Graphic vs. statistical methods of evaluating data: Another look and another analysis*. Dearborn: Paper presented at the meeting of the Association for Behavior Analysis.

Huitema, B. E. (1986a). Autocorrelation in behavioral research. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 187–208). New York: Plenum.

Huitema, B. E. (1986b). Statistical analysis and single-subject designs: Some misunderstandings. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and Advances* (pp. 209–232). Boston: Springer.

Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10*(3), 253–294.

Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics: Statistical Issues in Psychology, Education, & the Social Sciences, 3*(1), 27–46.

Huitema, B. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Hoboken: Wiley.

Huitema, B. E. (2016, May). Final fusilillade. Paper presented at the meeting of the Association for Behavior Analysis International, Chicago.

Huitema, B. E. (2018). *The effect. Unpublished Department of Psychology Technical Report*. Kalamazoo: Western Michigan University.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods, 3*(1), 104–116.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational & Psychological Measurement, 60*, 38–58.

Huitema, B. E., McKean, J. W., & Laraway, S. (2008). Time-series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods, 6*, 367–379.

Huitema, B.E., & Urschel, J. (2014). Elementary statistics courses fail miserably in teaching the *p*-value. Paper presented at the meeting of the Association for Behavior Analysis International, Chicago.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46*(5), 311–350.

Hurl, K., Wightman, J., Haynes, S. N., & Virues-Ortega, J. (2016). Does a pre-intervention functional assessment increase intervention effectiveness? A meta-analysis of within-subject interrupted time-series studies. *Clinical Psychology Review, 47*, 71–84.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS Medicine, 11*(10), e1001747. https://doi.org/10.1371/journal.pmed.1001747.

Ioannidis J. P. (2015). Failure to Replicate: Sound the Alarm. *Cerebrum: The Dana forum on brain science*, *2015*, cer-12a-15. City of publication is NY, NY. The editor is Glovin, B.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532.

Johnston, J. M., & Pennypacker, H. S., Jr. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York: Routledge/Taylor & Francis Group.

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*(4), 411–414.

Kahneman, D. (2014). A new etiquette for replication. *Social Psychology, 45*(4), 310–311.

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine, 28*(7), 1709–1716.

Kazdin, A. (1982). *Single-case research designs: Methods for Clinical and Applied Settings*. New York: Oxford University Press.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality & Social Psychology Review, 2*(3), 196–217.

Killeen, P. R. (2018). Predict, control, and replicate to understand: How statistics can foster the fundamental goals of science. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-018-0171-8.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational & Psychological Measurement, 56*(5), 746–759.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from what works clearing house website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kulig, T. C., Pratt, T. C., & Cullen, F. T. (2017). Revisiting the Stanford Prison Experiment: A case study in organized skepticism. *Journal of Criminal Justice Education, 28*(1), 74–111.

Kyonka, E. G. (2018). Tutorial: small-N power analysis. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-018-0167-4.

Lang, J. M., Rothman, K. J., & Cann, C. I. (1998). That confounded P-value. *Epidemiology, 9(1)*, 7–8.

Lanovaz, M. J., Huxley, S. C., & Dufour, M. M. (2017). Using the dual-criteria methods to supplement visual inspection: An analysis of nonsimulated data. *Journal of Applied Behavior Analysis, 50*(3), 662–667.

Lanovaz, M. J., Robertson, K. M., Soerono, K., & Watkins, N. (2013). Effects of reducing stereotypy on other behaviors: A systematic review. *Research in Autism Spectrum Disorders, 7*(10), 1234–1243.

Lanovaz, M. J., Turgeon, S., Cardinal, P., & Wheatley, T. L. (2018). Using single-case designs in practical settings: Is within-subject replication always necessary? *Perspectives on Behavior Science*, 1–10. https://doi.org/10.1007/s40614-018-0138-9.

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation, 24*(3–4), 445–463.

Leek, J. T., & Jager, L. R. (2017). Is most published research really false? *Annual Review of Statistics & Its Application, 4*, 109–122. https://doi.org/10.1146/annurev-statistics-060116-054104.

Leek, J. T., & Peng, R. D. (2015). Statistics: P values are just the tip of the iceberg. *Nature News, 520*(7549), 612.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*(6), 161–171.

Lynch, J. G., Jr., Bradlow, E. T., Huber, J. C., & Lehmann, D. R. (2015). Reflections on the replication corner: In praise of conceptual replications. *International Journal of Research in Marketing, 32*(4), 333–342.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*(3), 341–351.

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

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLoS One, 10*(8), e0136088.

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the *Journal of Applied Behavior Analysis* 1991–2005. *Journal of Applied Behavior Analysis, 40*(4), 659–672.

McNeeley, S., & Warner, J. J. (2015). Replication in criminology: A necessary practice. *European Journal of Criminology, 12*(5), 581–597.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*(4), 647–653.

Mischel, W. (1958). Preference for delayed reinforcement: An experimental study of a cultural observation. *Journal of Abnormal & Social Psychology, 56*(1), 57.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual review of Psychology, 69*, 511–523. https://doi.org/10.1146/annurev-psych-122216-011836.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5*(2), 3–14.

Northup, J., Fusilier, I., Swanson, V., Roane, H., & Borrero, J. (1997). An evaluation of methylphenidate as a potential establishing operation for some common classroom reinforcers. *Journal of Applied Behavior Analysis, 30*(4), 615–625.

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *Elife, 6*, e23383.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*(2–3), 313–324.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660.

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

Open Science Collaboration. (2017). Maximizing the reproducibility of your research. In S. O. Lilienfeld & I. D. Waldmen (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 1–21). New York: Wiley.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367.

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and Advances* (pp. 157–186). New York: Plenum.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531–536.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530.

Perone, M. (1991). Experimental design in the analysis of free-operant behavior. In I. H. Iversen & K. A. Lattal (Eds.), Techniques in the behavioral and neural sciences: Vol. 6. *Experimental Analysis of Behavior: Part I* (pp. 135–171) Amsterdam: Elsevier.

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22*(2), 109–116.

Perone, M. (2018). How I learned to stop worrying and love replication failures. *Perspectives on Behavior Science.* https://doi.org/10.1007/s40614-018-0153-x.

Perry, G. (2018). The shocking truth of Stanley Milgram's obedience experiments. *New Scientist.* Retrieved from https://www.newscientist.com/article/mg23731691-000-the-shocking-truth-of-stanley-milgrams-obedience-experiments/

Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis, 15*(4), 477–492.

Petursdottir, A. I., & Carr, J. E. (2018). Applying the taxonomy of validity threats from mainstream research design to single-case experiments in applied behavior analysis. *Behavior Analysis in Practice, 11*(3), 228–240.

Poling, A., & Fuqua, R. W. (1986). *Research methods in applied behavior analysis: Issues and Advances*. New York: Plenum.

Poling, A., Grossett, D., Karas, C. A., & Breuning, S. E. (1985). Medication regimen: A subject characteristic rarely reported in behavior modification studies. *Applied Research in Mental Retardation, 6*(1), 71–77.

Poling, A., Methot, L. L., & LeSage, M. G. (1995). *Fundamentals of behavior analytic research*. New York: Plenum Press.

Reicher, S., & Haslam, S. A. (2006). Rethinking the psychology of tyranny: The BBC prison study. *British Journal of Social Psychology, 45*(1), 1–40.

Resnick, B. (2017, July). What a nerdy debate about *p*-values shows about science—and how to fix it. *Vox*. Retrieved from https://www.vox.com/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-0005

Resnick, B. (2018). The Stanford Prison Experiment was massively influential. We just learned it was a fraud. *Vox*. Retrieved from https://www.vox.com/2018/6/13/17449118/stanford-prison-experiment-fraud-psychology-replication

Resnik, D. B., & Stewart, C. N. (2012). Misconduct versus honest error and scientific disagreement. *Accountability in Research, 19*(1), 56–63.

Romm, C. (2015). Rethinking one of psychology's most infamous experiments. *The Atlantic Monthly*. Retrieved from https://www.theatlantic.com/health/archive/2015/01/rethinking-one-of-psychologys-most-infamous-experiments/384913/

Rooker, G. W., Iwata, B. A., Harper, J. M., Fahmie, T. A., & Camp, E. M. (2011). False-positive tangible outcomes of functional analyses. *Journal of Applied Behavior Analysis, 44*(4), 737–745.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(*3*), 638–641.

Rosenthal, R., & Rosnow, R. L. (2009). *Artifacts in behavioral research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*. New York: Oxford University Press.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American psychologist, 44*(10), 1276–1284.

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review, 22*(4), 944–954.

Schmidt, F. L., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology, 4*(1), 32–37.

Schooler, J. W. (2014). Turning the lens of science on itself: Verbal overshadowing, replication, and metascience. *Perspectives on Psychological Science, 9*(5), 579–584.

Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis, 24*(2), 189–204.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., & Srinivasan, M. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55–67.

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014a). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147.

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2014b). A *d*-statistic for single-case designs that is equivalent to the usual between-groups *d*-statistic. *Neuropsychological Rehabilitation, 24*(3–4), 528–553.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971–980.

Shaw, D. (2018). The quest for clarity in research integrity: A conceptual schema. *Science & Engineering Ethics*, 1–9. https://doi.org/10.1007/s11948-018-0052-2

Shirley, M. J., Iwata, B. A., & Kahng, S. (1999). False-positive maintenance of self-injurious behavior by access to tangible reinforcers. *Journal of Applied Behavior Analysis, 32*(2), 201–204.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*, 487–510. https://doi.org/10.1146/annurev-psych-122216-011845.

Sidman, M. (1960). *Tactics of scientific research*. Oxford: Basic Books.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534–547.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11*(5), 221–233.

Skinner, B. F. (2014). *Verbal behavior*. Cambridge: B. F. Skinner Foundation (Original work published 1957).

Smaldino, P. E., & McElreath, R. (2018). The natural selection of bad science. *Royal Society Open Science, 3*(9), 160384.

Stein, R. A. (2017). The golden age of anti-vaccine conspiracies. *Germs, 7*(4), 168–170.

Stokstad, E. (2018). The truth squad. *Science, 361*(*6408*), 1189–1191. https://doi.org/10.1126/science.361.6408.1189.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59–71.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic & Applied Social Psychology, 37*, 1–2.

Tufte, E. R. (1990). *Envisioning information*. Cheshire: Graphics Press.

Tufte, E. R. (1997). *Visual explanations*. CT: Cheshire.

Tufte, E. R. (2006). *Beautiful evidence*. CT: Cheshire.

Tufte, E. R. (2009). *The visual display of quantitative information* (2nd ed.). CT: Cheshire.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after NHST: How to describe your data without "*p*-ing" everywhere. *Basic & Applied Social Psychology, 37*(5), 260–273.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113*(23), 6454–6459.

Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science, 29*(7), 1159–1177.

Weaver, E. S., & Lloyd, B. P. (2018). Randomization tests for single case designs with rapidly alternating conditions: An analysis of *p*-values from published experiments. *Perspectives on Behavior Science*, https://doi.org/10.1007/s40614-018-0165-6.

Weeden, M., & Poling, A. (2011). Identifying reinforcers in skill acquisition studies involving participants with autism: Procedures reported from 2005 to 2009. *Research in Autism Spectrum Disorders, 5*(1), 388–391.

Weeden, M., Porter, L. K., Durgin, A., Redner, R. N., Kestner, K. M., Costello, M., et al. (2011). Reporting of medication information in applied studies of people with autism. *Research in Autism Spectrum Disorders, 5*(1), 108–111.

Wilkinson, L. & American Psychological Association Task Force on Statistical Inference Science Directorate. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Williams, V. S., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24*(1), 42–69.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta analysis in individual-subject research. *Behavioral Assessment, 11*(3), 281–296.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*(2), 203–214.

Yong, E. (2012). In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature, 485*(7398), 298–300.