Perspectives on Behavior Science (2019) 42:1–11
https://doi.org/10.1007/s40614-019-00194-2

EDITORIAL

# Editorial: Replication and Reliability in Behavior Science and Behavior Analysis: A Call for a Conversation

Donald A. Hantula[1]

Published online: 11 March 2019


It has been over a decade since Ioannidis (2005) published a provocative indictment of medical research titled "Why Most Published Research Findings Are False." According to the *PloS Medicine* website, Iaonnidis' paper has been viewed nearly 3 million times and cited nearly 4,000 times. Clearly, it struck a nerve. What followed was a period of intense scrutiny, questioning and soul-searching not only in various fields of medicine but also in economics and especially in psychology as reports of large scale replication failures in these disciplines appeared (Bohannon, 2015; Camerer et al., 2016; Coyne, 2016; "Estimating the reproducibility of psychological science," 2015; Francis, 2012; Schmidt & Oh, 2016). This "replication crisis" in psychology cut deep and led to federal (USA) convened working groups to address the crisis such as NSF (National Science Foundation, 2015) and the Defense Advanced Research Projects Administration (DARPA) SCORE program (Rogers, 2019), replication initiatives, and changes in publication procedures, most notably by the America Psychological Society (APS) which began awarding "badges" to papers that conformed to certain replication enabling practices. Probable causes and consequences of the "replication crisis" in psychology are well documented and summarized in papers in this issue, especially in those by Branch, by Hales and colleagues, and by Laraway and colleagues.

## A Systems Approach

Behavioral systems analysis (Diener, McGee, & Miguel, 2009; McGee & Diener, 2010; Sigurdsson & McGee, 2015) teaches us that a system is perfectly designed to produce whatever it produces. In this context "design" is not necessarily intentional, but rather is the process of shaping and selection by consequences. A system that produces replicable and reproducible research, or failures thereof, has had ample opportunity to evolve in that direction. Many of these selection pressures contributing to the

✉ Donald A. Hantula
hantula@temple.edu

[1] Department of Psychology, Temple University, Philadelphia, PA 19122, USA

"replication crisis" are not unique to any discipline but have become part of the science enterprise writ large. Tincani and Travers (this issue) adopt this perspective in pointing out that workers in behavior science and behavior analysis are subject to many of the same contingencies that have produced the "replication crisis" in general, and in Psychology in particular.

Lilienfeld's (2017) analysis of the contingencies that contributed to the "replication crisis" takes a matching law approach to the problem. In Lillenfeld's view, behavioral scientists are faced with a continuous choice between two concurrent schedules: the reinforcers available from careful, scholarly science and those from participating in the "grant culture" of contemporary research, which he defines as a management system in which researchers are rewarded for grant dollars generated while their scholarly achievements are largely ignored by university administrators. Reinforcers available from careful, scholarly science are large, but they are delayed and often ephemeral. For individual scientists, these include the satisfaction of contributing to a scientific enterprise and the muted respect of colleagues. For the discipline, rigorous science may be relatively immune to replication failures. By contrast, individuals who participate in the "grant culture" of contemporary research gain relatively immediate and tangible reinforcers such as pay raises and continued employment. But for the discipline the consequences are unfortunate. Grant cycles are generally short (1-5 years) and granting agencies demand pilot data to secure a grant, and a bevy of publications based on the funded research. Per Lillienfeld, such academic and economic short-termism becomes a powerful motivator for questionable research practices (QRPs) (John, Loewenstein, & Prelec, 2012), fraud, confirmation bias, hyperspecialization and a dis-incentive for creativity, intellectual risk-taking, deep thinking, and replication.

Note that the "grant culture" Lillienfeld (2017) calls out is but a subset of a larger movement toward short-term "accountability" (some would say "countability")[1] in all areas of academia and education that has produced a variety of corrosive practices such as the now discredited "value-added" models of K-12 teacher evaluation (Wasserstein, lazer & Goldhaber 2016; Goldhaber, 2015); unreliable, invalid and biased university instructor end of semester evaluations (Boysen, 2015; Boysen, Kelly, Raesly, & Casner, 2014; Carrell & West, 2010; Clayson, 2009; Ellis, Burke, Lomire, & McCormack, 2003; Greenwald, 1997; Kornell & Hausman, 2016; MacNell, Driscoll, & Hunt, 2015; Martin, 1984); untenable journal "impact factors" (Hantula, 2005; McGarty, 2000; Meyer & Evans, 2003; Moustafa, 2015; Seglen, 1997) rating and ranking schemes for institutions, fields of study, and individual research impact that rely on dubious methods and data (Bastedo & Bowman, 2009; Franceschini & Maisano, 2017; Shattock, 2015), and the "publish or perish" culture of academia at large (De Rond & Miller, 2005). The old management adage "What gets measured gets done" remains true, because in practice this means "what gets measured sets the occasion for

[1] This observation neither an argument against accountability in research, scholarship, and education, nor a suggestion that those who work in research, scholarship, and education are somehow beyond accountability. Rather it is an argument that accountability measures that do not meet with the highest standards of psychometrics, measurement and evaluation will do more harm than good. Any accountability measure must be reliable, valid and socially acceptable at the very least and adopted only after careful analysis of costs, benefits and externalities. Measures that are employed largely due to their ease of administration are not likely to be beneficial.

reinforcement and punishment." If short-term outcomes such as grant dollars, numbers of publications, journal rankings, instructor evaluations numbers, and college "selectivity" indices are what allow academicians to contact reinforcement, or more accurately avoid punishment, then these will be selected and increased, along with their correlates, including QRPs, fraud, "salami-slicing" publication, predatory publications[2], coercive editorial practices, and grade inflation. Perhaps the problem is a "countability culture."

The website Retraction Watch (https://retractionwatch.com/) that is produced by the Center for Scientific Integrity provides an illuminating insight into the one of the consequences of such short-termism. Its tagline is "Tracking retractions as a window into the scientific process" and as of January 2019 its database cataloged 19,660 scientific papers that were retracted since 2010. A retraction is a formal withdrawal of a paper by one or more authors. Papers may be retracted due to honest or inadvertent mistakes in the methods or data analysis, inability to replicate the results or they may be retracted due to misconduct. All three of these reasons for retraction can result from short-termism; especially QRPs[3] and willful transgression. Retractions are on the rise, with a faster growth rate than scientific publications, which provides perhaps the most general and persuasive that a systemic problem exists.

Oransaky and Macus (2016, p. 41) summarized the problem as it was seen in the early part of the 21st century:

> Between 2000 and 2010, the number of published papers in the sciences rose by 40 percent, from about 1 million per year to about 1.4 million. Over that same period, the number of retracted articles — the ultimate in academic take-backs — grew tenfold, from about 40 per year to about 400. The figure is now somewhere close to 700 papers retracted annually. Although retractions represent a small sliver of the total literature, accounting for roughly 0.05 percent of all articles, the

[2] Predatory publishers are an unfortunate externality of the publish or perish and countability culture coupled with the rise of open-access publishing and online journals (Beall, 2012, 2013; McLeod, Savage, & Simkin, 2018). A predatory publication has all of the external trappings of a legitimate scientific journal such as a lofty title, lists of editorial board members and tables of contents with serious sounding papers. However, in a typical predatory publishing arrangement, the author pays a hefty fee to the "journal" which then sends the manuscript out for cursory or often non-existent "peer review" after it is quickly accepted largely as-is and published soon after (for an example see Beall, J. (2014). Bogus journal accepts profanity-laced anti-spam paper [Electronic resource]. *Scholarly Open Access*: *Internet blog*. Retrieved http://scholarlyoa.com/2014/11/20/bogus-journal-accepts-profanity-lacedantispam-paper.). The safeguards of legitimate peer review are entirely absent in a predatory publication. Many of these alleged journals send out repeated spam calls for papers that promise a quick turnaround and guaranteed acceptance. However not all open access journals are predatory (for example the *PLoS* journals) and page charges in and of themselves are not indicative of a predatory journal.

[3] QRPs identified by RetractionWatch [ https://replicationindex.wordpress.com/2015/01/24/questionable-research-practices-definition-detect-and-recommendations-for-better-practices/ ]: Selective reporting of (dependent) variables; Deciding whether to collect more data after looking to see whether the results will be significant; Deciding whether to collect more data after looking to see whether the results will be significant; excluding studies that did not work; rounding off a p-value just above .054 and claim that it is below .05; reporting an unexpected finding as having been predicted from the start**;** claiming that results are unaffected by demographic variables (e.g., gender) when one is actually unsure (or knows that they do)**;** Falsifying Data. Many of these QRPs seem to apply largely to NHST driven research. Identifying QRPs unique to SCRD is a worthwhile undertaking.

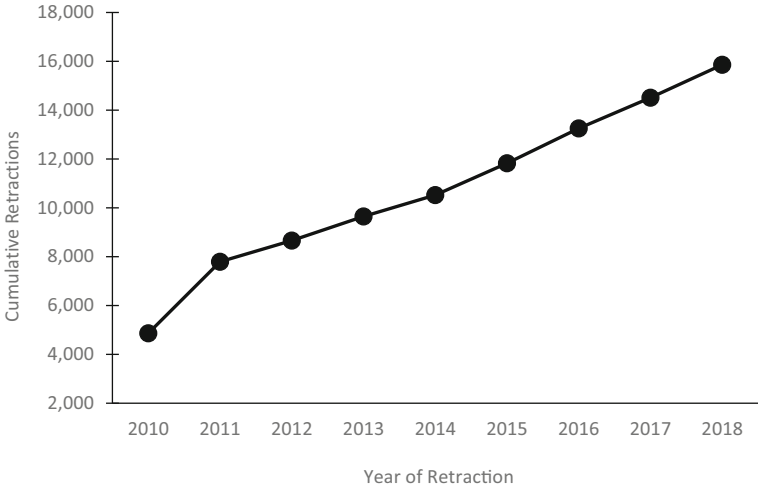


**Figure 1** Cumulative retractions, expressions of concern from 2010-2018 as indexed in http://retractiondatabase.org/. Data were generated by entering the first and last date of each year in the "Retractions or Other Notices" text box entry. Note that RetractionWatch started in 2010, hence there are no data prior to 2010.

> remarkable increase in the retraction rate has been seen by many as a symptom of sickness in the body scientific.

Figure 1 updates Oransaky and Macus. It was generated from the Retraction Watch database (http://retractiondatabase.org/), and shows the cumulative number of retractions by year from 2010-2018. Note the steady upward slope and overall mean of more than 1,700 retractions per year. Each retracted article represents a faulty or fallacious piece of evidence that may have been used as a basis for a later study, a treatment or therapy choice or a policy decision.

It may be heartening that only 66 articles are returned when Behavior* is entered as a search term in the Retraction Watch Database, but that does not absolve behavior science or behavior analysis. Retraction Watch's FAQ acknowledges that it is not a complete compendium but a work in progress and that its database has more coverage of life (medical) sciences than others. The problem lies in the general system of science, not a particular discipline[4].

## Replication and Reliability in Behavior Science and Behavior Analysis

The "replication crisis" in psychology has serious implications for replication and reliability in behavior science and behavior analysis. Despite certain metatheoretical disputes (Burgos & Killeen, 2018), behavior science, behavior analysis, and psychology have much more in common than differences. Hence the "replication crisis" in

[4] Perhaps the entire countability culture and its destructive effects may be best summed up by a nugget of career advice offered by a Nobel Laureate who told me" Remember, deans cannot read but most of them can count."

psychology could well be repeated in behavior science and behavior analysis. Even if it is not, it may hold some important lessons for both scientists and practitioners. However, from a scan of behavioral journals it appears as if the "replication crisis" has largely been ignored in behavior science and behavior analysis. Indeed, apart from the papers in this issue, there are few other articles discussing the "replication crisis" from a behavior science or behavior analysis perspective. Rare examples include Hanley (2017), who announced policy changes for the *Journal of Applied Behavior Analysis* that included a section dedicated to replications and Imam (2018), who recommended changes in editorial policies as well as collaboration with organizations outside of our field. But overall those in our field have been silent on this important topic, and remaining indifferent to the large scale social movement that has engulfed medical, social and behavioral sciences (broadly defined) is at best risky and at worst irresponsible. Where replication failures in behavior science and concerned, absence of evidence is not necessarily evidence of absence.

In estimating the likelihood that our field is at risk of a "replication crisis," some readers may assume that the one in Psychology stems solely from misunderstandings and misuses of inferential statistics and misinterpretations of null hypothesis significance tests (NHST). This problem certainly exists (e.g., Branch, 1999), and most research in behavior science and behavior analysis does not commonly employ NHST, but rather relies on single case research designs (SCRD).[5] It would be problematic to view this as a panacea, for two reasons. First, SCRD does not define behavior science or behavior analysis research; it is a tool, not an epistemic foundation, and the tool exists as part of the broader systems of science in which researchers participate. Second, use of NHST in behavior science and behavior analysis research is increasing (Zimmermann, Watkins, & Poling, 2015) and many of the most well-known applications of behavior analysis such addictions, autism, and leadership (Komaki, Desselles, & Bowman, 1989; Komaki, Zlotnick, & Jensen, 1986; Lovaas, 1987; Silverman et al., 2002) are documented in studies using group designs and NHST. As Hyten (2017) observed, organizational behavior management has led that movement toward methodological pluralism and behavioral research. *JOBM* has published NHST -driven behavioral studies for decades (e.g., Goltz, 1999; Hantula & Crowell, 1994; Wikoff et al., 1982). The *Journal of the Experimental Analysis of Behavior* (JEAB) has published statistical papers in the recent past (e.g., Gilroy, Franck, & Hantula, 2017; Young, 2018), and as this journal appears in print, *JEAB* will feature a special issue on "modern statistical practices in behavior analysis." As Killeen points out in this issue, statistics can foster the fundamental goals of behavior science.

Importantly, just as is true of other design approaches, SCRD studies may also incorporate QRPs, "researcher degrees of freedom" (ambiguously or un-reported choices in the design, conduct and analysis of an experiment), misapplications and errors of interpretation that are not altogether different from those in group design studies using NHST. In any study, regardless of methodology, researchers may sometimes delete or exclude participants or data points on ad hoc grounds that only serve to support the researcher's expectations. Just like a typical NHST study, all SCRD studies rely on sampling and inference. Whether

[5] SCRD is not unique to behavior science and behavior analysis. It has been employed in efforts ranging studies ranging from Ebbinghaus' (1885) pioneering studies of memory to modern studies of cognition and visual psychophysics (Smith & Little, 2018).

one samples from a population (as in NHST) or from an individual's repertoire (as in SCRD), a researcher is drawing small samples of behavior that are then combined and analyzed in some manner to make an inference about an independent variable's effect on a dependent variable. Any kind of sampling can introduce error.

A common QRP identified in the "replication crisis" literature is p-hacking (Simonsohn, Nelson, & Simmons, 2014). In a study using NHST this may involve continuing to collect data until an effect is observed, or collecting multiple dependent variables but only reporting those whose values are statistically significant. Analogues in SCRD studies include running multiple, slightly different functional analysis trials but only reporting those that "worked," and not reporting on dependent variables that were collected but were not consistent with expected effects. Both sampling and p-hacking problems allow a researcher to capitalize on chance, rather than real effects. Sheer dumb luck may be one of the most potent forces in nature, but it is not a good basis for scientific veracity.

Because science rests on a foundation of replication, and evidence or empirically based practice demands replicative evidence of efficacy, any possible threats to the replicative foundation of science and practice must be addressed. The "replication crisis" is but a symptom of many underlying causes; some are recent, others are distal. There is not a single cure, nor are there any quick solutions. The "replication crisis" in psychology may best serve as a motivating operation or "teachable moment" for researchers and practitioners in behavior science and behavior analysis to reflect on their own potential failings and devise sustainable solutions.

## In This Issue

This issue is dedicated to a careful and critical analysis of the ways that we conduct research in behavior science and behavior analysis. Such a reappraisal and reassessment has been long overdue. Science is all about change and advancement. Although the "replication crisis "was the initial motivating operation, the papers in this special issue approach not only the "replication crisis," but also take a broader look at the craft of behavioral research. The article by Hale et al is a solid explanation of the replication crisis, while Laraway et al explore its relevance for behavior science and behavior analysis research. Tincani and Travers point to some potential cracks in the foundation of applied behavior analysis and allegedly empirically supported treatments as revealed by publication bias. Branch looks at the "replication crisis "from the perspective of behavior science and suggests some solutions from this research tradition. Perone turns the pejorative implication of "replication failure' on its head with an account of how replication failures in his own work prompted more interesting experiments to solve the riddle. Indeed, Perone's paper is an eloquent testimony to Sidman's (1960) observation that a replication failure is not an end, but rather it is a beginning; it is not enough to report a replication failure – the researcher must also explain it. Killeen shows how statistics play an important role in fostering reproducible research. Kyonka offers a gentle but rigorous tutorial on power analysis for SCRD, Lanovaz, et al ask whether within subject replications are necessary in SCRD in purely applied settings and Kaplan et al describe an R package for performing behavioral economic analyses. Their software is open source and freely available to all.

Mechner closes the issue with a memorial tribute to graduate school colleague, friend and former ABAI president Kurt Salzinger. Dr. Salzinger's intellectual contributions ranged widely, from his well-regarded and very behavioral work in schizophrenia (Lam, Marra, & Salzinger, 2005; Leibman & Salzinger, 1998; Salzinger, 1980, 1983, 1996, 1998; Salzinger & Serper, 2004; Serper, Goldberg, & Salzinger, 2004) to an operant method for using goldfish behavior as a warning system for water pollution (Salzinger, Fairhurst, Freimark, & Wolkoff, 1973).

## A Call for a Conversation

The papers in this special issue may seem like a random, or perhaps contradictory collection. A behavior science journal addressing replication and reproducibility problems, but with papers on statistics and power analyses, and statistical software? A paper arguing that replication is a luxury in applied settings? I suggest viewing these papers as voices in broad conversation that, as Coyne (2016) points out, speaks to the very credibility of science. Currently, there is no final word on the current status of replication in behavior science and behavior analysis, not in the present issue or elsewhere. There is only the conversation, one that we hope will continue amongst readers and in future issues of *PoBS*, and in other journals in the field. This must be a multi-faceted conversation, because different areas of behavioral research have their own unique challenges and solutions that should be shared widely.

A good way to begin this conversation is to focus on our own publication practices as Imam (2018) suggested. While other fields struggle to make replication mainstream (Zwaan, Etz, Lucas, & Donnellan, 2018), Hanley (2017) has provided an explicit invitation and space for replications in *JABA*. All journals in our field should follow suit. As Ioannidis (2018) argues and Perone (this issue) shows, replication efforts can be a source of new knowledge. For example, the "Many Labs 2" project (Klein et al. 2018) found that diversity in samples and settings was not a factor in replication failures across 190 investigators; that is population characteristics had little to no bearing on the failure of a finding to replicate. The oft-invoked "well that only applies to…." criticism of research was not supported. We should begin talking about replication (or lack thereof) in the spirit of Sidman (1960) as a motivating operation for discovery rather than an S-delta for punishment. Finally, we must bring more transparency to our publication process.

Public repositories of data and methods should become the norm, not the exception. For example, Kaplan et al (this issue) made their beezdemand software publicly available on GitHub. Other *PoBS* papers have made software available as an online supplement (Bullock, Fisher, & Hagopian, 2017; Kaplan et al., 2016) and in the future, all studies should make custom software, R scripts and other tools available online. For empirical articles, all raw data should be available as a supplement. While SCRD studies appear to show 'all the data ' in graphs, what we do not know are factors such as date, time, length of sessions, or excluded data points. Imagine a study in which the x-axis is labeled "sessions" from 1-10 There is an important difference between a study in which sessions occurred on 10 consecutive days or if they occurred multiple times on the same day, maybe 5 times on one day and 5 on the next day. Copies of all IOA and procedural fidelity data sheets, copies of scripts/protocols should also be appended as

online supplements. Adopting a ‘cinema verite’ approach to documentation (Suls, 2013) in which videos of representative sessions, or reenactments of sessions are included as supplemental material. More detailed information will enable better meta-analyses and research summaries.

I close on an optimistic note. Writers like Branch (this issue) have suggested that behavior science and behavior analysis have much to offer those who are concerned with the "replication crisis." And this may well be true, but few from this community have bothered to engage with Psychology and other fields to share their expertise (for an exception, see Normand, 2016). For example, buried in the paper that brought the “replication crisis “in Psychology to the fore is an observation that within-subject designs yielded results that were much more readily replicable than between-subject designs. What more can be offered? *Behavioral and Brain Sciences* recently dedicated an entire issue (with much commentary) to making replications mainstream (Zwaan et al., 2018) and Smith and Little (2018) published an excellent case for SCRD in *Psychonomic Bulletin & Review*. Doing so requires not just making the effort; behavior scientists and behavior analysts must first work out the empirical and conceptual bases of their views on replication than repeating decades-old tropes about the superiority of behavioral research. The same conversations that will allow readers of *PoBS* and everyone in the ABAI to better understand the role of replication in their own discipline also will prepare them to contribute to interdisciplinary dialogues in a thoughtful, respectful, and helpful manner. Let’s talk. Let’s listen. Let’s learn.

**Acknowledgements** I thank Tom Critchfield, Dave Jarmolowicz, and Erin Rasmussen for their comments on this editorial.



## References

Bastedo, M. N., & Bowman, N. A. (2009). US News & World Report college rankings: Modeling institutional effects on organizational reputation. *American Journal of Education, 116*(2), 163–183.

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature News, 489*(7415), 179.

Beall, J. (2013). Predatory publishing is just one of the consequences of gold open access. *Learned Publishing, 26*(2), 79–84.

Bohannon, J. (2015). Many psychology papers fail replication test. *Science, 349*(6251), 910–911. https://doi.org/10.1126/science.349.6251.910.

Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology, 42*(2), 109–118. https://doi.org/10.1177/0098628315569922.

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641–656. https://doi.org/10.1080/02602938.2013.860950.

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.

Bullock, C. E., Fisher, W. W., & Hagopian, L. P. (2017). Description and validation of a computerized behavioral data program: 'BDataPro. *The Behavior Analyst, 40*(1), 275–285. https://doi.org/10.1007/s40614-016-0079-0.

Burgos, J. E., & Killeen, P. R. (2018). Suing for peace in the war against mentalism. *Perspectives on Behavior Science*. https://doi.org/10.1007/s40614-018-0169-2.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409–432.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–30. https://doi.org/10.1177/0273475308324086.

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology, 4*(1), 28.

De Rond, M., & Miller, A. N. (2005). Publish or perish: bane or boon of academic life? *Journal of Management Inquiry, 14*(4), 321–329.

Diener, L. H., McGee, H. M., & Miguel, C. F. (2009). An integrated approach for conducting a behavioral systems analysis. *Journal of Organizational Behavior Management, 29*(2), 108–135. https://doi.org/10.1080/01608060902874534.

Ebbinghaus, H. (1885). *Über das Gedächtnis*. Leipzig: Dunker.

Ellis, L., Burke, D. M., Lomire, P., & McCormack, D. R. (2003). Student grades and average ratings of instructional quality: The need for adjustment. *The Journal of Educational Research, 97*(1), 35–40. https://doi.org/10.1080/00220670309596626.

Estimating the reproducibility of psychological science. (2015). *Science, 349*(6251), aac4716. doi:https://doi.org/10.1126/science.aac4716

Franceschini, F., & Maisano, D. (2017). Critical remarks on the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics, 11*(2), 337–357.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science, 7*(6), 585–594. https://doi.org/10.1177/1745691612459520.

Gilroy, S. P., Franck, C. T., & Hantula, D. A. (2017). The discounting model selector: Statistical software for delay discounting applications. *Journal of the Experimental Analysis of Behavior, 107*(3), 388–401. https://doi.org/10.1002/jeab.257.

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher, 44*(2), 87–95.

Goltz, S. M. (1999). Can't stop on a dime: The roles of matching and momentum in persistence of commitment. *Journal of Organizational Behavior Management, 19*(1), 37–63. https://doi.org/10.1300/J075v19n01_05.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182–1186. https://doi.org/10.1037/0003-066x.52.11.1182.

Hanley, G. P. (2017). Editor's note. *Journal of Applied Behavior Analysis, 50*, 3–7. https://doi.org/10.1002/jaba.366.

Hantula, D. A. (2005). The impact of JOBM: ISI impact factor places the Journal of Organizational Behavior Management third in applied psychology. *Journal of Organizational Behavior Management, 25*(3), 1–15. https://doi.org/10.1300/J075v25n03_01.

Hantula, D. A., & Crowell, C. R. (1994). Intermittent reinforcement and escalation processes in sequential decision making: A replication and theoretical analysis. *Journal of Organizational Behavior Management, 14*(2), 7–36. https://doi.org/10.1300/J075v14n02_03.

Hyten, C. (2017). OBM is already using the 'fuzzy concept' criteria for applied behavioral research: Commentary on Critchfield and Reed. *The Behavior Analyst, 40*(1), 179–182. https://doi.org/10.1007/s40614-017-0096-7.

Imam, A. A. (2018). Place of behavior analysis in the changing culture of replication and statistical reporting in psychological science. *European Journal of Behavior Analysis, 19*(1), 2–10. https://doi.org/10.1080/15021149.2018.1463123.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124. https://doi.org/10.1371/journal.pmed.

Ioannidis, J. P. A. (2018). Why replication has more scientific value than original discovery. *Behavioral and Brain Sciences, 41*. doi:https://doi.org/10.1017/S0140525X18000729.

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. https://doi.org/10.1177/0956797611430953.

Kaplan, B. A., Amlung, M., Reed, D. D., Jarmolowicz, D. P., McKerchar, T. L., & Lemley, S. M. (2016). Automating scoring of delay discounting for the 21- and 27-item Monetary Choice Questionnaires. *The Behavior Analyst, 39*(2), 293–304. https://doi.org/10.1007/s40614-016-0070-9.

Klein, et al. (2018). Many Labs 2, Investigating variation in replicability across sample and setting. Advances in Methods and Practices in Psychological Science. Preprint, doi:https://doi.org/10.1177/0956797618805750.

Komaki, J. L., Zlotnick, S., & Jensen, M. (1986). Development of an operant-based taxonomy and observational index of supervisory behavior. *Journal of Applied Psychology, 71*(2), 260–269. https://doi.org/10.1037/0021-9010.71.2.260.

Komaki, J. L., Desselles, M. L., & Bowman, E. D. (1989). Definitely not a breeze: Extending an operant model of effective supervision to teams. *Journal of Applied Psychology, 74*(3), 522–529. https://doi.org/10.1037/0021-9010.74.3.522.

Kornell, N., & Hausman, H. (2016). Do the Best Teachers Get the Best Ratings? *Frontiers in Psychology, 7*(570). doi:https://doi.org/10.3389/fpsyg.2016.00570

Lam, K., Marra, C., & Salzinger, K. (2005). Social reinforcement of somatic versus psychological description of depressive events. *Behaviour Research and Therapy, 43*(9), 1203–1218. https://doi.org/10.1016/j.brat.2004.09.003.

Leibman, M., & Salzinger, K. (1998). A theory-based treatment of psychotic symptoms in schizophrenia: Treatment successes and obstacles to implementation. *The Journal of Genetic Psychology: Research and Theory on Human Development, 159*(4), 404–420. https://doi.org/10.1080/00221329809596161.

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: Righting the ship. *Perspectives on Psychological Science, 12*(4), 660–664. https://doi.org/10.1177/1745691616687745.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology, 55*(1), 3–9. https://doi.org/10.1037/0022-006X.55.1.3.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303.

Martin, E. (1984). Power and authority in the classroom: Sexist stereotypes in teaching evaluations. *Signs, 9*(3), 482–492. https://doi.org/10.1086/494073.

McGarty, C. (2000). The citation impact factor in social psychology: a bad statistic that encourages bad science? *Current Research in Social Psychology*, *5*(1), 1–16.

McGee, H. M., & Diener, L. H. (2010). Behavioral systems analysis in health and human services. *Behavior Modification, 34*(5), 415–442. https://doi.org/10.1177/0145445510383527.

McLeod, A., Savage, A., & Simkin, M. G. (2018). The ethics of predatory journals. *Journal of Business Ethics, 153*(1), 121–131.

Meyer, L. H., & Evans, I. M. (2003). Motivating the professoriate: Why sticks and carrots are only for donkeys. *Higher Education Management & Policy, 15*(3), 151–168.

Moustafa, K. (2015). The disaster of the impact factor. *Science and Engineering Ethics, 21*(1), 139–142.

National Science Foundation (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf.

Normand, M. P. (2016). Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology, 7*, 934. https://doi.org/10.3389/fpsyg.2016.00934.

Oransky, I, & Marcus, A. (2016). Two Cheers for the Retraction Boom. *The New Atlantis*, Number 49, Spring/Summer, pp. 41–45.

Rogers, A. (2019). DARPA wants to solve science's reproducibility crisis with AI. *WIRED* https://www.wired.com/story/darpa-wants-to-solve-sciences-replication-crisis-with-robots/?mod=djemAIPro.

Salzinger, K. (1980). The behavioral mechanism to explain abnormal behavior. *Annals of the New York Academy of Sciences, 340*, 66–87. https://doi.org/10.1111/j.1749-6632.1980.tb35161.x.

Salzinger, K. (1983). The immediacy hypothesis in a theory of schizophrenia. *Nebraska Symposium on Motivation, 31*, 231–282.

Salzinger, K. (1996). Reinforcement history: A concept underutilized in behavior analysts. *Journal of Behavior Therapy and Experimental Psychiatry, 27*(3), 199–207. https://doi.org/10.1016/S0005-7916(96)00037-7.

Salzinger, K. (1998). Schizophrenia: From behavior theory to behavior therapy. In J. J. Plaud & G. H. Eifert (Eds.), *From behavior theory to behavior therapy* (pp. 98–115). Needham Heights, MA: Allyn & Bacon.

Salzinger, K., & Serper, M. (2004). Schizophrenia: The immediacy mechanism. *International Journal of Psychology & Psychological Therapy, 4*(2), 397–409.

Salzinger, K., Fairhurst, S. P., Freimark, S. J., & Wolkoff, F. D. (1973). Behavior of the goldfish as an early warning system for the presence of pollutants in water. *Journal of Environmental Systems, 3*(1), 27–40. https://doi.org/10.2190/XHV6-X934-KK1U-VEXE.

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. https://doi.org/10.1037/arc0000029.supp(Supplemental).

Seglen, P. O. (1997). Why the impact factor should not be used for evaluating research. *BMJ, 314*, 498–502.

Serper, M. R., Goldberg, B. R., & Salzinger, K. (2004). Behavioral assessment of psychiatric patients in restrictive settings. In S. N. Haynes & E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment, Vol. 3: Behavioral assessment* (pp. 320–345). Hoboken, NJ: John Wiley & Sons Inc..

Shattock, M. (2015). The impact of the UK research assessment exercise. *International Higher Education*, (56).

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

Sigurdsson, S. O., & McGee, H. M. (2015). Organizational behavior management: Systems analysis. In H. S. Roane, J. E. Ringdahl, & T. S. Falcomata (Eds.), *Clinical and organizational applications of applied behavior analysis* (pp. 627–647). San Diego, CA: Elsevier Academic Press.

Silverman, K., Svikis, D., Wong, C. J., Hampton, J., Stitzer, M. L., & Bigelow, G. E. (2002). A reinforcement-based Therapeutic Workplace for the treatment of drug abuse: Three-year abstinence outcomes. *Experimental and Clinical Psychopharmacology, 10*(3), 228–240. https://doi.org/10.1037/1064-1297.10.3.228.

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

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review, 25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8.

Suls, J. (2013). Using 'Cinéma Vérité' (truthful cinema) to facilitate replication and accountability in psychological research. *Frontiers in Psychology, 4*. doi:https://doi.org/10.3389/fpsyg.2013.00872.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133.

Wikoff, M., Anderson, D. C., & Crowell, C. R. (1982). Behavior management in a factory setting: Increasing work efficiency. *Journal of Organizational Behavior Management, 4*(1–2), 97–127.

Young, M. E. (2018). Discounting: A practical guide to multilevel analysis of choice data. *Journal of the Experimental Analysis of Behavior, 109*(2), 293–312. https://doi.org/10.1002/jeab.316.

Zimmermann, Z. J., Watkins, E. E., & Poling, A. (2015). JEAB research over time: Species used, experimental designs, statistical analyses, and sex of subjects. *The Behavior Analyst, 38*(2), 203–218. https://doi.org/10.1007/s40614-015-0034-5.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*. doi:https://doi.org/10.1017/S0140525X17001972.