Response to Commentaries on Sakaluk (2020)

John K. Sakaluk¹

Received: 18 March 2021 / Revised: 13 April 2021 / Accepted: 16 April 2021 / Published online: 9 June 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Archives of Sexual Behavior signaled its collective interest in joining the discourse regarding replicability of scientific findings—already ongoing both within sexual science (see Sakaluk, 2016; Sakaluk & Graham, 2018; Seto, 2019) and beyond (Nelson et al., 2018; Spellman, 2015; Vazire, 2018)—with its publication of the invited Guest Editorial by Lorenz (2020). Proximally, Wisman and Shrira (2020) published an article in *Archives* describing a trio of experiments ostensibly evidencing “sexual chemosignals” and their impact on directing men’s sexual arousal and motivation. I critiqued both of these papers (Sakaluk, 2020), out of concern that *Archives* was engaging in performative methodological reform: that *Archives* wanted to reap the optical benefits of talking the reform-talk, but wasn’t yet capable of (or even interested in) walking the reform-walk.

To summarize my critiques beginning with Wisman and Shrira (2020): I subjected the focal effects from their three experiments to 10 different tests of “evidential value,” a concept that for me is agnostic as to whether a given set of effects are “significant” or not (or how big or large they are), but rather attempts to describe how compelling or convincing a set of effects is, significant or not, big or small. Those looking for a one-paragraph rehashing of the outcomes of these tests can find them summarized in Sakaluk (2020, p. 2749), but the gist is they suggest that the results of Wisman and Shrira (2020) are irreproducible and un-credible.

Although some of the evidential value testing techniques I used to evaluate Wisman and Shrira (2020) were more novel and complex, many were not (e.g., eye-balling the implausibly repetitious “borderline/marginally significant” nature of the p values of their three focal tests, $p = .01$, $.044$, and $.054$, respectively). Indeed, distributions of p values, whether informally or formally analyzed, have been a go-to statistical canary in the replicability coal mine for years now (e.g., Simonsohn et al., 2015).

This is what stung about the Lorenz (2020) Editorial: that a paper failing such an obvious replicability-minded check could be published in *Archives* so soon after *Archives* was signaling that it now finally invested in appraising and increasing the replicability of effects. I took further umbrage with the exclusively individual-level focus of the replicability promoting strategies described by Lorenz, particularly in absentia of any discussion (or implementation) of systems-level strategies that would be much more effective. I could not put the sentiments I felt at the time better than they are described by Lorenz and Holland (2020): the Editorial felt entirely “milquetoast” (p. 2761) at best, and collectively deceptive, at worst, if *Archives* were not to adjust any of its standard operating procedures following its increasingly replication-friendly rhetoric.¹ And so, I engaged in a little scholarly catharsis, though not soon enough to pre-empt Wisman and Shrira (2020) from being featured in high-impact articles as a reliable source (see Hofer et al., 2020).

I have since had the pleasure of reading three response commentaries. The first, by Imhoff (2020), argues that salvation for “solid” sexual science is to be found through the pursuit of multi-study papers.² The second, by Lorenz and Holland (2020), lodges a counter-critique of my Commentary for having given short-shrift to qualitative methodologists and their place in the replicability/methods-reform conversation. The final, and most recent, by McCarthy et al. (2021), draws attention to the limits of heuristic appraisals of evidential value. I am grateful to both Imhoff, Lorenz and Holland, and McCarthy et al. for engaging constructively with my critical commentary. I wish to first respond to Imhoff and McCarthy et al., before engaging in more detail with the substance of Lorenz and Holland (as well as further related concerns of my own).

¹ The newly mandated “Declarations” section seems a (modest) step in the right direction.

² Imhoff described his proposal as promoting “solid” sexual science, and I am unsure as to what particular element(s) of research quality (if any) he intends to refer with this descriptor. My critique of Wisman and Shrira (2020) was preoccupied with replicability (but one way—and certainly neither the only, nor the most important—of thinking about quality research), and so I interpret the spirit of his commentary as also having replicability in mind.

✉ John K. Sakaluk
jsakaluk@uwo.ca

¹ Department of Psychology, Social Science Centre Rm 7418,
Western University, London, ON N6A 5C2, Canada

Are Multi-Study Papers the Way to “Solid” Sexual Science?

Imhoff (2020) argued that promotion of multi-study papers is a “glaringly missing” (p. 2755) proposal from my suggestions for how to promote replicable sexual science at *Archives*, as “This overreliance on single studies is in my perspective one of the larger roadblocks on our way towards a more solid sexual science.” (p. 2755) I will not mince words; I strongly disagree with Imhoff. Before explaining why, however, there are a number of features of Imhoff’s commentary that I sincerely appreciate.

First, I am enthusiastic that Imhoff (2020) focusses on a structural proposal for improving replicability at *Archives*. Rather than leaving it the individual researcher(s) to overcome structural barriers, and conjure for themselves the requisite incentives, Imhoff seems to think as I do: journals must dangle the carrots (or wave the sticks) necessary to encourage the behavior change they wish to see in the field. Second, Imhoff comes equipped with meta-scientific data to inform his position, in the form of methodological review of the article types published in *Archives*. Although I find the evidence for his proposed solution lacking, I nevertheless applaud his collecting evidence to inform his characterization of what he sees as problematic. And finally, I am in total agreement with Imhoff’s argument for the need to change institutional incentive structures (e.g., decisions regarding hiring, raises, tenure, promotion, grants and awards).

Otherwise, I disagree almost entirely with Imhoff (2020): multi-study papers are not one of the—or *the*—larger obstacles between our field (or the papers at *Archives*, specifically) and a more replicable sexual science. The history of the replicability discourse teaches us that multi-study papers offered no safeguards against claims with poor evidential value; in fact, they encouraged many of the methodological problems our fields are now trying to stamp out! Indeed, two of the “poster children” papers that sparked—or featured centrally in—concerns about replicability within social psychology were extensive multi-study papers. Baumeister et al.’s (1998) original ego depletion paper, for example, reported on four experiments yet more than 6000 citations later, we now have not one, but *two* Registered Replication Reports in which entire communities of laboratories report being unable to detect the ego depletion effect (Hagger et al., 2016, consisting of 23 laboratories and 2,141 participants; Vohs et al., in press, consisting of 36 laboratories and 3531 participants). Bem’s (2011) paper espousing scientific support for extrasensory perception, meanwhile, included 9 experiments! Multi-study papers, therefore, have been at the heart of some of the most concerning examples of widely touted claims with poor evidential value.

Sterling (1959), long ago, articulated the underlying problem in the context of multi-study papers: reviewers and Editors prefer papers describing significant findings, thereby creating

selection pressure in favor of papers describing an “error of the first kind” (p. 30). Schimmack (2012) later described how this kind of selection pressure becomes particularly problematic in the context of multi-study papers in psychology: in these cases, reviewers and Editors prefer papers describing effects that are consistently significant across multiple studies. But given psychologists’ proclivity for running underpowered studies (Maxwell, 2004), this publishing standard became statistically unachievable without getting extraordinarily lucky or through the (ab)use of questionable research practices (or “researcher degrees of freedom”; John et al., 2012; Simmons et al., 2011). Thus, prioritizing multi-study papers in sexual science, as Imhoff (2020) would have us do, could create the very same kind of structural pressures faced by the social psychologists of yore that led to the production of less credible papers.

Imhoff (2020) would have us further believe that evidential value cannot be reliably gleaned from single-study papers. Here, too, I disagree. The power posing paper (Carney et al., 2010) is one of the most widely discussed single-study papers in the replicability discourse, and even without certain evidential tests that require multiple effects, there were plenty of clues that its evidential value may have questionable: (1) Three of its key statistical tests yielded borderline *p* values (.045, .01, .049, and .004); (2) many of its reported effect sizes grossly exceeded what is typical and plausible for psychological effects (e.g., $r = 0.43$, $d = 0.91$; see Funder & Ozer, 2019; Simonsohn, 2014), leaving us to (3) wonder (again) how humans would be able to navigate everyday life with such an easily provoked and powerful effect swimming (ostensibly) in the piranha tank of human psychology amidst other equally provokable and powerful determinants of our thoughts, feelings, and actions (Gelman, 2017).

Finally, Imhoff (2020) seems to think that—irrespective of whether they promote evidential value or not—multi-study papers will make it easier for researchers to assess their evidential value. For multi-study papers, to facilitate inspection of evidential value, however, depends on researchers conforming to reporting standards (e.g., sufficient descriptive statistics, full test statistics, effect sizes, e.g., Appelbaum et al., 2018) or (failing that) transparently sharing their data and materials. Meanwhile, it is often surprisingly difficult to identify which effects in their articles authors believe are “focal” (see Sakaluk et al., 2019). That Wisman and Shrira (2020) conducted—and sufficiently reported—three direct replications of their (easily determined) focal effect is not a typical case of evidential value sleuthing, and I would not anticipate other cases to be as straightforward.

Although Imhoff (2020) comes equipped with compelling meta-scientific data to suggest that single-study papers are common at *Archives*, he comes with no meta-scientific data to suggest that they are problematic, and the historical record within neighboring fields provides compelling evidence that expecting multi-study can harm the evidential value of the field. Indeed, when statistical illiteracy drives Editors and reviewers to demand a fantastical level of consistency in statistical testing (a

la Schimmack, 2012), the research community becomes all too quick to oblige. Further, there remains a very real possibility that any expectations for multi-study papers could quickly become a structural barrier to publishing research at *Archives* on populations for which recruitment is not so easily and quickly done as is recruiting large convenience samples online (e.g., via MTurk, Prolific, etc.); both *Archives* and the field would be worse for wear were this risk to be realized. Multi-study papers should therefore not be the thought of as a promising first line of intervention in replicability promotion. What benefits multi-study papers therefore might bring—in terms of directly assessing replicability and the testing of conceptual boundary conditions—could be more reliably gained if the system-level interventions I described (Sakaluk, 2020) were first implemented: enforcing statistical reporting standards and a bare minimum standard of transparency, offering replicability-friendly reports and policies, and promoting stronger statistical reasoning among quantitative researchers in our field.

Heuristics are Heuristics

McCarthy et al. (2021), meanwhile, expressed concern “that readers [of Sakaluk, 2020] may infer that these heuristics [effect size plausibility, sample size, and the “piranha problem”] are *always* [emphasis in the original] appropriate for critiquing the evidence presented within a study” (p. 773). I lament that my enthusiasm for the substance of this critique is even less than my enthusiasm for the idea that structurally promoting multi-study articles will save us, but I wish to first acknowledge what I do appreciate about McCarthy et al.’s critique. Specifically, their (1) identification of Wisman and Shrira’s (2020) error in effect size calculation, (2) verification of this error through computational reproducibility, and (3) correction of their estimates through using the appropriate formulae is, in its entirety, a very clever piece of data sleuthing.

But beyond the value of their effect size corrections, I must admit that I find myself confused by the goals of McCarthy et al. (2021). In their own words, they argued that “Heuristics are useful, but they can be blunt” (p. 773). No disagreement there; that is why I labeled the first three tests as “Heuristic Tests of Evidential Value” and supplemented my critique with seven tests lest subject to the intraocular whims of the critic. If readers wish to understand tests labeled as heuristic as “Infallible Tests of Evidential Value,” then I am unable to save them from that leap, and I doubt that McCarthy et al.’s elaborated cautions will make the difference between those who understand that heuristics are heuristics and not something else.

But so that the record is clear: I am aware of the precariousness of any one way of defining research quality, and further aware that heuristics of the sort that I employed have limitations and boundary conditions on their usefulness. Sample size is not synonymous with power, and even if it were, there are good

reasons to urge caution in prioritizing larger sample sizes above all else (see Kenny & Judd, 2019).

And true enough, large effect sizes need not always be cause for suspicion. There are large effect sizes in the social world. COVID-19-induced job losses for women and Black and Hispanic people (McKinsey & Company, 2021)—those are large effects. But if we are to trust the effect sizes yielded by social psychological laboratory experiments (particularly when they bear out some of the broader patterns identified within the crisis), randomness must be given its long-run opportunity to do its good work (either in larger n studies, or many [read: more than 3] smaller n replications) if it is to wipe out the noisy impact of the outside world on participants, who do not come to our laboratories quietly *tabula rasa*.

Likewise, it is true: generalizability is not always the goal. We could go further: it is not often the goal, even among method reformers. For example, there is nary a generalizability-focused guideline added to TOP Guidelines since Nosek et al., 2015, despite the availability of ostensibly beloved generalizability-focused transparency interventions (see Simons et al., 2017). But even for proof-of-laboratory-purified-concept demonstrations, experiments are only useful insofar as they are credible.

I could engage further on these three points, but I think it wasteful. Heuristics remain heuristics. Readers must read and think carefully, for themselves. And we should not spill more ink carving out nuance in critiques that ultimately provide accommodation for an article with many impactful errors and concerning features. The original authors have not emerged to set the record straight, and until such time that they do, I feel there is more important discourse in which to engage.

Replicability and Qualitative Research: Warning Shots Across the Epistemological Bow

Quite separately from questions of the merit in promoting multi-study articles (Imhoff, 2020) or the value of heuristics (McCarthy et al., 2021), Lorenz and Holland (2020) have critiqued my commentary for “focusing so closely on proper statistical testing and reporting...”, which “...left out our qualitative colleagues of the conversation about open science practices.”

My responsibility to provide the first salvo in the qualitative/sexual science/replicability discourse is, I think, a complicated matter to adjudicate. Doubly so is the complexity of resolving to what extent the qualitative contingency of sexual scientists ought to engage with—and what they can draw from and contribute to—the reform discourse. But resolved it must be (and I wholeheartedly support Lorenz and Holland for arguing it must be a priority), and with unfortunate expediency, as the reform movement moves itself into more epistemologically threatening territory for many qualitative researchers. I therefore offer my

thoughts as a relative outsider to the qualitative paradigm, but as an insider who is in solidarity with qualitative researchers in sexual science and the valuable work that they do.

It is true that my Commentary omitted any engagement with implications of methods reform for qualitative method users—a possibility that I declined to engage with during peer review. From my perspective, the galvanizing exemplar (Wisman & Shrira, 2020) was quantitative, and therefore so too was my focus. Moreover, I have previously attempted to call qualitative (and quantitative) sexual scientists into the replicability and methods reform discourse (see Sakaluk, 2016)—to no avail—for similar reasons (e.g., problems with one-size-fits-all reforms, sensitivity of data) that Lorenz and Holland (2020) articulate.

It is also true, as Lorenz and Holland (2020) note, that I was gun-shy to speak to the intersection qualitative methods and replicability in part because of the dearth of training I have in qualitative methods.³ However unsatisfying it may have been to Lorenz and Holland, I continue to consider my abstention an attempt to “stay in my lane”—a practice I think that I (and other members of my demographic) could stand to put to use more frequently than we do. Indeed, in social-psychology-land, “qualitative research” is often practiced (and thereby its breadth corrupted) to be synonymous with “count the words!” I therefore know that while there is emerging scholarship on matters of replicability in the qualitative sphere (see Langford, 2020, for collection), I also know enough to know that I don’t know enough to say much of anything with (deserving) confidence about what qualitative sexuality researchers ought to do (if anything)—especially for matters that might shape editorial policy. To quote a family idiom, “A closed mouth gathers no feet.”

Still, I think Lorenz, Holland, and myself are in agreement for the need for conversations about methods reform—and any subsequent editorial policy changes—to be sensitive of issues unique to qualitative researchers, and to be mindful of how policies are implemented even when the targeted issues (e.g., preregistration) could conceivably span both qualitative and quantitative modalities. To this end, although I cynically believe sexual scientists have largely surrendered what opportunity they had to help shape broader standards of practice as a result of the replicability discourse (i.e., at the time of Sakaluk, 2016), they maintain control of—and should exercise their power over—the rules of the game called science (Bakker et al., 2012) within their own editorial “house”.⁴ These conversations, however, should feature and be facilitated by those with bona fide qualitative and mixed-method expertise, lest they lead to unhelpful (or genuinely harmful) methodological policy.

This time, however, our qualitative colleagues may need to commune with greater urgency. Whereas for most of the methods reform movement’s past ~8 years, the discourse largely spoke to issues peripheral to qualitative method users (and, therefore, it could be more safely ignored), the discourse now finds itself taking positions that are, frankly, alarming with how they position qualitative work. Take the recent review by Nosek et al. (2021), and their description and characterization of reproducibility and its place in determining credibility:

Reproducibility refers to testing the reliability of a prior finding using the same data and analysis strategy... In principle, *all reported evidence* [my emphasis] should be reproducible. If someone applies the same analysis to the same data, then the same result should occur. Reproducibility tests can fail for two reasons. A *process reproducibility error* [emphasis in original] occurs when the original data cannot be repeated to verify the original evidence because of unavailability of the data or code, information about the analysis to recreate the code, or unavailability of the software or tools needed to conduct the reproducibility test. An *outcome reproducibility failure* [emphasis in original] occurs when the reanalysis obtains a different result than reported originally. This can occur because of an error in the original analysis or reporting, the reproduction analysis or reporting, or fraudulent behavior. (p. 4)

There is much here to digest, but I offer the following summary: here is a characterization of credible science that would position a hefty amount of qualitative work—whereby the researchers and participants are thought to interactively co-construct what knowledge is generated from a given study (e.g., Gergen et al., 2015)—as a form of one or more reproducibility error(s). From my perspective, this is the beginning of a rhetorical path in which some of worst fears about the reform movement’s view of qualitative work (anticipated by Lorenz & Holland, 2020) might be realized.⁵

Many qualitative researchers, for example, would situate themselves in terms of instrumentalist views of concepts and theories (Godfrey-Smith, 2009) and as adhering to some level of commitment to constructionist views (Hacking, 2000). The resulting epistemological stance and philosophy of science from this position is difficult, if not impossible, to reconcile with the belief that one ought to be able to reproduce—and later replicate—a scientific finding (or set of findings) to a consistently appreciable degree. How can strict reproducibility and eventual replicability of sexual science be expected when some sexuality-related concepts cannot be trusted to exist outside of our perception, some sexual science theories are thought to be useful (but

³ One published paper (including a singleton qualitative study) does not a qualitative methodologist make.

⁴ Journals and funders, for example, have made non-trivial changes to their values and workflows after years of the discourse, during which our voices have been largely absent.

⁵ And I, for one, would not be surprised to see this citation a regular feature used by those attempting to take aim at the qualitative paradigm.

never true), and the sexuality scientist (along with their participants) is sometimes thought to actively create the “facts” their scholarly endeavors generate? Yet in spite of its contrast with the characterization of credibility offered in Nosek et al. (2021), this is an approach to sexual science that has paid dividends in the context some of the world’s largest problems (e.g., the HIV epidemic).

I feel it important to stress that there is no later qualification of Nosek et al.’s (2021) statement regarding the necessity of reproducibility for credible evidence; “all reported evidence should be reproducible” (p. 4) is not later characterized to mean, more precisely, that “all quantitative [and not necessarily all qualitative] evidence should be reproducible” Further, one should not so quickly doubt the influence of this movement’s ability to persuade the gatekeepers of scientific business to adopt this view of what is and isn’t “credible”; in relatively short order, Nosek’s Center for Open Science has convinced over 5,000 journal editorial boards, professional societies, and publishers to adopt their TOP standards (see <https://www.cos.io/initiatives/top-guidelines>),⁶ while having successfully created and promoted a new bibliographic metric on which journals now compete (see TOP Factor here, <https://www.topfactor.org>).⁷

Regardless, therefore, of whether the movement speaks of reproducibility’s meaning and virtues in a way that is careless or intentional in its positioning of qualitative work when it writes “all reported evidence should be reproducible,” qualitative sexuality researchers will need to move quickly to determine if and how they are to engage with the reform discourse, before returning to (if they think it valuable) discerning whether the mainstream methods reform movement is epistemic *ramen* or *varelse* (Card, 1986). Questions like, “How is rigor constructed and defined and by whom?”, “How are the constructions of rigor similar and/or different between qualitative and quantitative work?”, and “How can the pursuit of credibility enrich—and be enriched by—pluralistic notions of rigor across qualitative and quantitative approaches?” could all be generative.⁸ And as Lorenz and Holland (2020) noted, qualitative researchers have developed their own rich traditions of fostering and evaluating rigor (some of which would epistemically challenge a realist-leaning reformer, e.g., criteria of trustworthiness), and unique insights into how features of the researcher can become embedded—for

better and/or for worse—in the research. As Lorenz and Holland highlight, there is wisdom here, to spare and potentially share with the reform discourse. Fostering a truly collaborative discourse of this sort, however, will be challenged if members of the reform discourse insist that those outside of it must first capitulate exclusively to their epistemic view, or risk having their approach to science framed (intentionally or inadvertently) as un-credible (see Steltenpohl et al., 2021, for a similar sentiment).

My hope is that the reform discourse not only gives room (should it need to be “taken” by qualitative researchers) to these ways of thinking about knowledge, but instead actively creates this space, which may first require walking back some of the stronger positions taken on what is required for research to be considered credible. Failing that, I would not begrudge qualitative researchers in sexual science to keep their own council on what is rigorous, resist one-size-fits-all methodologizing, and remain in search of greener pastures of discourse.

Declarations

Conflict of interest Dr. Sakaluk is a statistical consultant for other sexuality journals and the author of the original target Commentary.

References

- Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist*, 73, 3–25. <https://doi.org/10.1037/amp0000191>
- Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554. <https://doi.org/10.1177/1745691612459060>
- Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, 74, 1252–1265
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407–425. <https://doi.org/10.1037/a0021524>
- Card, O. S. (1986). *Speaker for the dead*. Tom Doherty Associates LLC.
- Carney, D. R., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21, 1363–1368.
- Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2, 156–168. <https://doi.org/10.1177/2515245919847202>
- Gelman, A. (2017). *The piranha problem in social psychology/behavioral economics: The “take a pill” model of science eats itself*. Retrieved February 1, 2021 from <https://statmodeling.stat.columbia.edu/2017/12/15/piranha-problem-social-psychology-behavioral-economics-button-pushing-model-science-eats/>.
- Gergen, K. J., Josselson, R., & Freeman, M. (2015). The promises of qualitative inquiry. *American Psychologist*, 70, 1–9. <https://doi.org/10.1037/a0038597>

⁶ I myself have advocated, with some ambivalence, for the adoption of TOP at *Journal of Sex Research*.

⁷ This is all to say, that being familiar with the typical ways quantitative researchers sometimes wish to “hand wave” away qualitative sexual science, the rhetoric and potential impact of the reform movement on qualitative methodology, as well as their level of belief in their mission, makes for an entirely different—and more science-politics capable—animal.

⁸ To anyone inclined to clutch pearls at the idea that quantitative findings are actively constructed, I would point you toward the *Many Analysts* paper (Silberzahn et al., 2018).

- Godfrey-Smith, P. (2009). *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press.
- Hacking, I. (2000). *The social construction of what?* Harvard University Press.
- Hagger, M. S., Chatzisarantis, N. L. D., Alberts, C. O., Anggono, C., Batallier, A. R., Zerhouni, O., ... Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11, 546–573. <https://doi.org/10.1177/1745691616652873>
- Hofer, M. K., Chen, F. S., & Schaller, M. (2020). What your nose knows: Affective, cognitive, and behavioral responses to the scent of another person. *Current Directions in Psychological Science*. <https://doi.org/10.1177/0963721420964175>
- Imhoff, R. (2020). Assessment of evidential value requires more than a single data point [Commentary]. *Archives of Sexual Behavior*, 49, 2755–2759. <https://doi.org/10.1007/s10508-020-01836-2>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532. <https://doi.org/10.1177/0956797611430953>
- Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, 24, 578–589. <https://doi.org/10.1037/met0000209>
- Langford, P. (2020). *Open qualitative research resource and ideas collection*. Retrieved from <http://psychbrief.com/open-qualitative-research/>
- Lorenz, T. K. (2020). Reproducibility and registration in sexuality research [Guest Editorial]. *Archives of Sexual Behavior*, 49, 367–372. <https://doi.org/10.1007/s10508-020-01650-w>
- Lorenz, T. K., & Holland, K. J. (2020). Response to Sakaluk (2020): Let's get serious about including qualitative researchers in the open science conversation [Commentary]. *Archives of Sexual Behavior*, 49, 2761–2763. <https://doi.org/10.1007/s10508-020-01851-3>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147–163. <https://doi.org/10.1037/1082-989X.9.2.147>
- McCarthy, R. J., Erickson, J. M., Hu, X., Pawirosetiko, J. S., Tarleton, H. L., Thomas, C. L., ... Sagarin, B. J. (2021). Broad agreement, but notes of caution with the implications of Sakaluk's (2020) critique of Wisman and Shrira (2020) [Commentary]. *Archives of Sexual Behavior*, 50, 773–777. <https://doi.org/10.1007/s10508-021-01929-6>
- McKinsey & Company. (2021). *Achieving an inclusive US economic recovery*. Retrieved from <https://www.mckinsey.com/industries/public-and-social-sector/our-insights/achieving-an-inclusive-us-economic-recovery>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, 69, 511–534. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348, 1422–1425. <https://doi.org/10.1126/science.aab2374>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Almenberg, A. D., ... Vazire, S. (2021). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*. <https://doi.org/10.31234/osf.io/kfsfvq>
- Sakaluk, J. K. (2016). Promoting replicable sexual science: A methodological review and call for metascience. *Canadian Journal of Human Sexuality*, 25, 1–8. <https://doi.org/10.3138/cjhs.251-CO1>
- Sakaluk, J. K. (2020). Getting serious about the assessment and promotion of replicable sexual science: A commentary on Wisman and Shrira (2020) and Lorenz (2020) [Commentary]. *Archives of Sexual Behavior*, 49, 2743–2754. <https://doi.org/10.1007/s10508-020-01795-8>
- Sakaluk, J. K., & Graham, C. A. (2018). Promoting transparent reporting of conflict of interests and statistical analyses. *Journal of Sex Research*, 55, 1–6. <https://doi.org/10.1080/00224499.2017.1395387>
- Sakaluk, J. K., Williams, A. J., Kilshaw, R. E., & Rhyner, K. T. (2019). Evaluating the evidential value of empirically supported psychological treatments (ESTs): A meta-scientific review. *Journal of Abnormal Psychology*, 128, 500–509. <https://doi.org/10.1037/abn0000421>
- Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17, 551–566. <https://doi.org/10.1037/a0029487>
- Seto, M. (2019). SA editorial about next phase of more open science. *Sexual Abuse*, 31, 371–373. <https://doi.org/10.1177/1079063219838430>
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Nosek, & B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1, 337–356. <https://doi.org/10.1177/2515245917747646>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12(6), 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Simonsohn, U. (2014). *We cannot afford to study effect size in the lab*. Retrieved from <http://datacolada.org/20>
- Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, 144, 1146–1152. <https://doi.org/10.1037/xge0000104>
- Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, 10, 886–899. <https://doi.org/10.1177/1745691615609918>
- Steltenpohl, C. N., Montilla Doble, L. J., Basnight-Brown, D., Dutra, N. B., Belsky, A., Kung, C., ... Darda, K. M. (2021). *Society for the Improvement of Psychological Science Global Engagement Task Force Report*. <https://doi.org/10.31234/osf.io/4upqd>
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association*, 54, 30–34. <https://doi.org/10.1080/01621459.1959.10501497>
- Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13, 411–417. <https://doi.org/10.1177/1745691617751884>
- Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q., Finley, A. J., ... Wagenmakers, E.-J., & Albarracín, D. (in press). A multi-site preregistered paradigmatic test of the ego depletion effect. *Psychological Science*.
- Wisman, A., & Shrira, I. (2020). Sexual chemosignals: Evidence that men process olfactory signals of women's sexual arousal. *Archives of Sexual Behavior*, 49, 1505–1516. <https://doi.org/10.1007/s10508-019-01588-8>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.