Archives of Sexual Behavior (2021) 50:773–777
https://doi.org/10.1007/s10508-021-01929-6

COMMENTARY

Broad Agreement, But Notes of Caution With the Implications of Sakaluk's (2020) Critique of Wisman and Shrira (2020)

Randy J. McCarthy[1] · Jennifer M. Erickson[1] · Xinyu Hu[1] · Joy S. Pawirosetiko[1] · Hannah L. Tarleton[1] · Courtney L. Thomas[1] · Morgan G. Tillery[1] · Brad J. Sagarin[1]

Received: 8 January 2021 / Revised: 17 January 2021 / Accepted: 19 January 2021 / Published online: 16 February 2021

Wisman and Shrira (2020) presented three studies from which they concluded men could detect olfactory signals of women's sexual arousal. Subsequently, upon close inspection of the reported statistics, Sakaluk (2020) found several apparent statistical errors and improbable findings that left him "critical and ultimately cynical of the effects the authors describe" (Sakaluk, 2020, p. 2743). We mostly agree with Sakaluk's critiques. Most fundamentally, we agree that if the statistics in Wisman and Shrira were not internally consistent, then we cannot make confident claims based on the evidence presented within the article. We also agree that even if the statistics were internally consistent, the totality of the other critiques was still discouraging.

Our agreement with Sakaluk's (2020) critiques is not wholehearted though. We feel that some of the "heuristic tests of evidentiary value" are aimed at features of the studies that we do not see as inherently problematic. Namely, and to use Sakaluk's terminology, we have concerns with the conclusions drawn from the "intraocular trauma test of sample sizes" and the "intraocular trauma test of effect size plausibility." In short, Sakaluk argued that Wisman and Shrira's (2020) sample sizes were so blatantly small and effect sizes were so blatantly large that readers could (or should) be suspicious about the credibility of the findings.

Our concern is that readers may infer that these heuristics are *always* appropriate for critiquing the evidence presented within a study. Heuristics are useful, but they can be blunt. As the saying goes, the Devil is in the details. Our goal herein is to spell out the details we believe were glossed over. We hope our discussion will provide readers with additional context in evaluating whether Sakaluk's (2020) critiques were apt and enable readers to apply these heuristics more accurately in evaluating research and in planning their own studies.

We have three points we want to make:

1. Small sample sizes cannot be used ipso facto to claim a statistical test is underpowered;
2. There are several ways to compute effect sizes, and these different ways may not allow for direct apples-to-apples comparisons with heuristically useful benchmarks; and
3. The magnitude of an effect for a phenomenon observed in an artificial laboratory setting is unlikely to represent the magnitude of the effect for that same theoretical phenomenon in the "real world," and this lack of generalizability does not invalidate the worth of the lab study.

Small Samples Need Not Always Cause Intraocular Trauma

Wisman and Shrira (2020) reported three studies with samples of 24, 32, and 35 male students. On its face, these are not impressively large samples. However, each of the focal hypothesis tests compared the same men's ratings of sweat samples from sexually aroused female donors to their ratings of sweat samples from non-sexually aroused female donors. As Sakaluk (2020) acknowledged, this repeated-measures design would in theory increase statistical power, although, also as noted by Sakaluk, because Wisman and Shrira did not present the correlation between the repeated-measures or the intra-class correlation from the ratings that were averaged together, we cannot calculate to what extent statistical power was actually increased. Nevertheless, the implication of the critique is that Wisman and Shrira's data were underpowered because it is unlikely that such small samples would repeatedly produce statistically significant findings (see also Button et al., 2013; Fraley & Vazire, 2014).

✉ Randy J. McCarthy
rmccarthy3@niu.edu

[1] Department of Psychology, Northern Illinois University, DeKalb, IL 60115, USA

Even though Sakaluk's (2020) wording is careful, we worry that some readers may infer that "small sample size" is interchangeable with "low statistical power." This would be an especially unfortunate misreading because large samples are extremely difficult to routinely and affordably collect in some areas of sex research. For example, Singer, Crooks, Johnson, Lutnick, and Matthews (2020) interviewed 21 sex workers regarding their experience during the COVID-19 pandemic and Rieger, Watts-Overall, Holmes, and Gruia (2020) evaluated video recordings of 33 sets of adult identical twins when investigating sexual orientation concordance and gender nonconformity. Readers who infer that small samples are inherently underpowered may be discouraged from conducting such studies where large samples would be prohibitive.

We agree with the spirit of Sakaluk's (2020) critique; high statistical power is desirable. And when an article provides insufficient information to calculate precise statistical power, readers might only be able to check the sample size as a rough guess. But it is critical that researchers use the sample size heuristic judiciously. We should call out studies with low statistical power, not studies with small samples.

Large Effect Sizes Need Not Always Cause Intraocular Trauma: Decisions in Calculating Effect Sizes Matter

Wisman and Shrira (2020) conducted a mini-meta-analysis of their three studies. As described in their article, Wisman and Shrira first converted the eta-squared effect size from each study's repeated-measures ANOVA into a *d* effect size, which they then converted into the *r* effect size for their meta-analysis.[1]

Focusing on the *d* effect sizes is illustrative, so we focus on these for our discussion. The *d* effect sizes reported in Wisman and Shrira (2020) represent the differences between men's ratings of the scents of sexually aroused women and of non-sexually aroused women (i.e., average rating for the sexually aroused trials minus average rating for the non-sexually aroused trials) in the units of standard deviations. Unfortunately, Wisman and Shrira did not describe which standard deviations were used to compute their *d* effect sizes, which makes it difficult to interpret and to judge whether the effects were implausibly large.

However, before we talk about the magnitude of the effect sizes, we want to point out what we believe to be an error. From what we can gather, Wisman and Shrira (2020) computed their *d* effect sizes from their reported η^2 effect sizes incorrectly. We believe they used the following formula to convert η^2 to *d*:

$$d \approx 2\sqrt{\frac{\eta^2}{1-\eta^2}}$$

Applying the η^2 effect sizes from Wisman and Shrira (2020) to this formula gives *d* effect sizes of 1.19, 0.77, and 0.70 for their Study 1, 2, and 3, respectively. This is nearly identical to their reported effect sizes of 1.18, 0.77, and 0.70 for those same studies. However, that conversion formula is only valid for between-participants designs (e.g., Brysbaert, 2019). In repeated-measure designs such as those used by Wisman and Shrira, the multiplication by 2 is inappropriate because the repeated observations provided by participants are not independent. The correct conversion for repeated-measures designs is the following formula:

$$d \approx \sqrt{\frac{\eta^2}{1-\eta^2}}$$

Applying the η^2 effect sizes from Wisman and Shrira (2020) to this (correct) formula gives *d* effect sizes of 0.59, 0.39, and 0.35 for their Study 1, 2, and 3, respectively.[2] These *d* effect sizes represent the mean differences in the unit of standard deviation of the difference scores.[3] Thus, we believe that Wisman and Shrira overestimated their *d* effect sizes by a magnitude of 2. As can be seen, these recomputed effect sizes are more modest in magnitude.

These conversion errors are important, but are tangential to our main point; namely, there are several ways to compute a standardized mean difference for repeated-measures designs (e.g., Westfall, 2016) and it is important to explore these differences before declaring effect sizes prima facie implausibly large. Alternative ways to compute *d* effect sizes

[1] The same result can be obtained simply by taking the square root of the reported eta-squared effect sizes. Specifically, Wisman and Shrira report effect sizes of $\eta^2 = .26$, $\eta^2 = .13$, and $\eta^2 = .11$ for Studies 1, 2, and 3, respectively. The square root of these η^2 effect sizes corresponds to Wisman and Shrira's reported *r* effect sizes in their Table 1: Study 1, $r = \sqrt{.26} = .51$, Study 2, $r = \sqrt{.13} = .36$, and Study 3, $r = \sqrt{.11} = .33$. Additionally, Sakaluk noted that the meta-analysis reported by Wisman and Shrira was not reproducible when the *F*-values were entered into an online *p*-checker app. We were able to reproduce the results of Wisman and Shrira's meta-analysis using these *r* effect sizes (see our R code here: https://osf.io/prqsv/).

[2] To illustrate how easy it is to make this error, we only discovered this error because we initially made the same missteps as Wisman and Shrira.

[3] These same *d* effect sizes also can be computed with formula $d = \sqrt{F}/\sqrt{n}$, which gives effect sizes of 0.59, 0.38, and 0.34, for Studies 1, 2, and 3, respectively. The slight differences between these *d* effect sizes and the *d* effect sizes computed by converting η^2 was due to the rounding of η^2 by Wisman and Shrira. If one computes the exact η^2 from the *F* ratio using the formula $F{*}df_{\text{between}}/(F{*}df_{\text{between}} + df_{\text{within}})$, the resulting *d* effect sizes are equivalent.

Table 1 Different ways of computing effect sizes

				Different effect size computations			
	n	Sexually aroused *M*(*SD*)	Non-sexually aroused *M*(*SD*)	M_{diff}	d_z	d_{control}	d_{average}
Study 1	24	3.60(0.66)	3.33(0.51)	0.27	0.59	0.53	0.46
Study 2	32	2.44(0.83)	2.18(0.76)	0.26	0.39	0.34	0.33
Study 3	35	3.02(0.96)	2.84(0.91)	0.18	0.35	0.20	0.19

Note. d_z uses the standard deviation of difference scores as the standardizer. d_{control} uses the standard deviation of the non-sexually aroused ratings scores as the standardizer. d_{average} uses the pooled standard deviation of the sexually aroused ratings and the non-sexually aroused ratings scores as the standardizer

for within-participants designs would be to express the mean differences using the standard deviation of the "control" condition (which would be the non-sexually aroused trials in Wisman and Shrira's, 2020, studies) or the average standard deviations of both trial types (e.g., Cumming, 2014; Lakens, 2013). The advantage of these approaches would be that the "standardizer" is in the units of the original scales and can be more directly compared to *d* effect sizes from between-participants designs (such as the effect sizes used as comparisons by Sakaluk, 2020). Alternatively, we can express Wisman and Shrira's (2020) effect sizes as simple mean differences. This is perhaps the most directly interpretable option because the reader can evaluate the magnitude of the effect merely by looking at the scale on which participants provided their responses.

We recomputed the effect sizes from Wisman and Shrira (2020) using these alternative approaches (see Table 1). For instance, in Study 1, Wisman and Shrira report an effect size of $d = 1.18$, which we believe should be $d = 0.59$ for the reasons stated above, when using the standard deviation of difference scores as the standardizer. However, this same effect is "only" a mean difference of 0.27 points on a scale ranging from *1 = not at all sexy* to *7 = very sexy*, which is "only" about a half a standard deviation of the ratings. An effect size of about $d = 0.5$ is comparable to many of the effect sizes that Sakaluk (2020) used as points of comparison in his critique. Importantly though, the change in the denominator from the standard deviation of difference scores to the standard deviation of the original scale matters not just quantitatively, but also qualitatively: The latter are directly comparable to effect sizes from a between-subjects design, the former are not.

Even with alternative ways of computing the effect size, it is possible that the effects reported in Wisman and Shrira (2020) were still implausibly large. We do not know. But these alternative ways of computing the effect sizes look less alarming and seem to be within the range of effects that can be readily produced in social science research.

Large Effect Sizes Need Not Always Cause Intraocular Trauma: Generalizability Is Not Always the Goal of a Study

The goal of Wisman and Shrira (2020) was to test a theoretical proposition: Do participants rate the scents of sexually aroused women as sexier than the scents of non-sexually aroused women? To accomplish this, they designed a study to have high internal validity perhaps at the expense of ecological validity.[4] To us, this is a fine tradeoff (see Anderson & Bushman, 1997; Mook, 1983, for detailed arguments). Studies with high internal validity are an indispensable tool in the research repertoire even if that means the results are not directly exportable to the "real world." Indeed, Sakaluk (2020) raised these ideas in his footnote 2, but we want to underscore them more forcefully.

Because the goal of Wisman and Shrira (2020) was to have high internal validity, we should evaluate their studies on whether they accomplished that goal. Namely, did their studies allow readers to draw strong inferences about their theoretical proposition? This means evaluating their study on whether they controlled for extraneous variables even if this creates a situation that has no real world counterpart and whether they isolated the theoretically important factors even if those factors are never this cleanly isolated in the real world. Notably, these methodological features—i.e., controlling for extraneous variables and isolating theory-relevant factors—probably create a maximally favorable condition for an effect to emerge and, consequently, may maximize the magnitude of the effect that could be detected in such an artificial situation. If an effect is found in such an artificial and contrived situation, then perhaps this effect also would be found in more natural settings, but probably to a lesser magnitude. The point is that testing whether an effect exists is a different question than testing whether an effect generalizes.

[4] We do not deny that the articles intended for general audiences that are cited by Sakaluk (2020) claim this effect has "real world" relevance. But we believe Wisman and Shrira (2020) stayed fairly close to the data.

We are not denying that there is some upper limit to the effects that can be produced in an ideal laboratory setting. Of course, at some magnitude, a peculiarly large effect size should raise an alarm. However, it would be a mistake to infer that the effect sizes in artificial laboratory settings, which are often (and intentionally) not concerned with ecological validity, can be directly compared to effects that might be found in the real world. Further, estimating the "real world" magnitude of an effect is not always the goal of research. Studies should not be condemned for failing to achieve what they intentionally did not seek to achieve.

Conclusion

Given the heated exchanges that critical commentaries can provoke, we appreciate Sakaluk's (2020) respectful tone and thorough analysis in his critique of Wisman and Shrira (2020). Scrutiny is an indispensable scientific activity, and it works best when done tactfully. We also share Sakaluk's long-term optimism in the field. The fact that Sakaluk's critique and responses to the critique (e.g., Imhoff, 2020) were published strengthens our optimism.

In the spirit of not wanting readers to walk away with incorrect take-home messages, we want to emphasize what we hope readers take from this Commentary. First, the current Commentary is not intended to redeem Wisman and Shrira (2020). Although we were able to replicate Wisman and Shrira's meta-analysis and our corrected calculations of Wisman and Shrira's effect sizes place them in a more plausible range, the remaining anomalies in Wisman and Shrira's data are troubling. And, even if we disagree with some of Sakaluk's (2020) points, we find the totality of Sakaluk's arguments to be persuasive. As for the critiques that are specific to Wisman and Shrira's data, the original authors have access to the data on which those analyses are based and are in the best position to confirm or rebut those criticisms (and we hope the authors take the opportunity to do so).

Second, we are not arguing that small samples and large effect sizes are never problematic. Sometimes they are. And when they are, we should not be timid in pointing that out. However, small sample sizes and large effect sizes are not inherently problematic. Determining whether they are in a particular instance requires closely inspecting the study for whether the study had sufficient statistical power, whether the design and methods of the study were appropriate, whether the researchers computed the effect size in a meaningful way, and whether the researchers intended for their observed effect size to generalize to a "real world" phenomenon.

Third, we want to publicly endorse the recommended changes Sakaluk (2020) discusses in his "What's Next?" section (see also Lorenz, 2020). Many of Sakaluk's suggestions are practical and, if implemented, would be constructive.

We hope researchers are judicious in using sample sizes and effect sizes to heuristically critique research. We want our critiques to be accurate, because that is when they are most productive. It would be a loss to our field if the fear of these critiques discouraged researchers from conducting potentially valuable and informative studies.

Acknowledgment The second through seventh authors are listed alphabetically.

Funding Not applicable.

Data Availability Not applicable.

Compliance with Ethical Standards

Conflicts of interest We have no conflicts of interest to disclose.

Code availability https://osf.io/prqsv/

References

Anderson, C. A., & Bushman, B. J. (1997). External validity of "trivial" experiments: The case of laboratory aggression. *Review of General Psychology, 1*(1), 19–41. https://doi.org/10.1037/1089-2680.1.1.19.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 1–38. https://doi.org/10.5334/joc.72.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. https://doi.org/10.1038/nrn3475.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966.

Fraley, R. C., & Vazire, S. (2014). The N-Pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019.

Imhoff, R. (2020). Assessment of evidential value requires more than a single data point [Commentary]. *Archives of Sexual Behavior, 49*(8), 2755–2759. https://doi.org/10.1007/s10508-020-01836-2.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4,* Article 863. https://doi.org/10.3389/fpsyg.2013.00863.

Lorenz, T. K. (2020). Reproducibility and registration in sexuality research [Commentary]. *Archives of Sexual Behavior, 49*(2), 367–372. https://doi.org/10.1007/s10508-020-01650-w.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*(4), 379–387. https://doi.org/10.1037/0003-066X.38.4.379

Rieger, G., Watts-Overall, T. M., Holmes, L., & Gruia, D. C. (2020). Gender nonconformity of identical twins with discordant sexual orientations: Evidence from video recordings. *Archives of Sexual Behavior, 49,* 2469–2479. https://doi.org/10.1037/dev0000461.

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.

Singer, R., Crooks, N., Johnson, A. K., Lutnick, A., & Matthews, A. (2020). COVID-19 prevention and protecting sex workers: A call to action [Letter to the Editor]. *Archives of Sexual Behavior, 49*(8), 2739–2741. https://doi.org/10.1007/s10508-020-01849-x.

Westfall, J. (2016). Five different "Cohen's *d*" statistics for within-subject designs. *Cookie Scientist*. http://jakewestfall.org/blog/index.php/2016/03/25/five-different-cohens-d-statistics-for-within-subject-designs/

Wisman, A., & Shrira, I. (2020). Sexual chemosignals: Evidence that men process olfactory signals of women's sexual arousal. *Archives of Sexual Behavior, 49*(5), 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.