Psychological Medicine

cambridge.org/psm

Correspondence

Cite this article: Luo Y, Furukawa TA (2023). Effect size calculation needs to be specified with details: comment on Ying *et al.* *Psychological Medicine* **53**, 4300–4301. https://doi.org/10.1017/S0033291722002057

Received: 19 May 2022
Revised: 2 June 2022
Accepted: 15 June 2022
First published online: 8 July 2022

Key words:
Cognitive-behavioral therapy; depression; effect size; result interpretation; standardized mean difference

Author for correspondence:
Yan Luo,
E-mail: lilacluo@gmail.com,
luo.yan.2u@kyoto-u.ac.jp

© The Author(s), 2022. Published by Cambridge University Press

Effect size calculation needs to be specified with details: comment on Ying *et al.*

Yan Luo and Toshi A. Furukawa

Department of Health Promotion and Human Behavior, School of Public Health in the Graduate School of Medicine, Kyoto University, Kyoto, Japan

To the editor

We read with great interest the article by Ying et al. (Ying et al. 2022), who reported on a well-conducted randomized controlled trial of cognitive-behavioral therapy (CBT) in alleviating depressive symptoms among Chinese patients with subthreshold depression. The results indicated that the internet-based CBT (ICBT) was significantly superior not only to the waiting list but also to face-to-face CBT. In interpreting the results from clinical trials, effect sizes are critically informative. However, we have some concerns about the effect sizes reported in this study.

In general, the standardized mean difference (S.M.D.) is widely used in clinical trials when the outcomes are continuous. The S.M.D. is standardized by dividing the mean difference (M.D.) by the standard deviation (S.D.), and allows comparison between studies which use different measuring instruments. However, there are various methods to calculate the S.M.D.: the S.M.D. can be calculated from different M.D.s (e.g. M.D. of endpoint scores, M.D. of change scores from baseline, M.D. from a model where the baseline score is adjusted) and S.D.s (e.g. pooled S.D. of endpoint scores, pooled S.D. of change scores, pooled S.D. of baseline scores, or the S.D. converted from model statistics). The S.M.D.s estimated by these methods are substantially different from one another, raising potential problems regarding reproducibility, selective reporting, and proper interpretation of how large the effect is (Luo et al., 2022). For instance, without prespecifying the calculation method, researchers may compute S.M.D.s using different methods, and select the largest one to report. Additionally, Cohen's rule of thumb, often used as a reference to interpret the effect size in clinical research, could be hard to apply if different calculation methods produce different S.M.D. values for the same outcome of the study.

In Ying et al.'s study, the method to calculate the effect size was explained in the **Method** section as, '*Within-group and between-group effect size* (*Cohen's d*) *were based on the method suggested for mixed model analysis* (Feingold, 2009; Morris & DeShon, 2002; Thorsell et al., 2011)'. However, it is still unclear which M.D. and which S.D. they used to calculate the effect sizes. Both Feingold's and Morris & DeShon's papers suggest using the pooled baseline S.D. for the M.D. estimated from a mixed model (Feingold, 2009; Morris & DeShon, 2002). On the other hand, in Thorsell et al.'s study, the square root of the variance estimate from the mixed model was used to calculate the effect size (Thorsell et al., 2011).

Ying et al.'s reported S.M.D.s do not seem to follow these methods. Let's take for example the between-group effect size of CES-D at post-intervention for ICBT *v.* face-to-face CBT, which was reported to be 0.06 (95% confidence interval: 0.02–0.09) in their Table 4. Using the values reported in Table 3 (the M.D. at endpoint was 1.6, the pooled baseline S.D. was 3.75, and the pooled endpoint S.D. was 3.80), the S.M.D. would be calculated as 0.43 using the baseline S.D. or 0.42 using the endpoint S.D.. The reported S.M.D. could have been calculated from other S.D.s that were not reported in the paper, but for an M.D. of 1.6 to generate an S.M.D. of 0.06, the S.D. would need to be approximately 27. It would be very difficult to imagine a population that has such a large variability in CES-D scores whose score ranges between 0–60. There are similar discrepancies for the other between-group effect sizes reported in their Table 4.

Because the S.M.D. values that we calculated and that was reported by the authors were very different, the interpretation of how large the effect of the intervention was could have substantively different clinical interpretations. The authors stated in the article that Cohen's rule of thumb was used as an aid for interpretation. Applying this rule, an S.M.D. of 0.4 would be moderate, while an S.M.D. of 0.06 would be less than a small effect. We are wondering how the effect of ICBT and CBT should be properly interpreted, and whether the effect sizes estimated in this article could be appropriately compared to previous studies, which might have used different S.M.D. calculation methods.

In summary, because the S.M.D.s can be calculated by different M.D.s and S.D.s and these S.M.D. estimates can vary substantially, researchers should be careful in reporting them and readers should be mindful how the reported S.M.D.s were calculated. As it is still hard to recommend a single calculation method that should be used universally for now (Luo et al., 2022), it is desirable for researchers to report their calculation methods in detail to increase transparency and reproducibility.

Prespecifying the method beforehand may help to avoid selective reporting bias. Meanwhile, future methodological studies are warranted that elucidate which S.M.D. calculation methods are recommendable.

Conflict of interest. None.

References

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods*, *14*(1), 43–53. doi: 10.1037/a0014699

Luo, Y., Funada, S., Yoshida, K., Noma, H., Sahker, E., & Furukawa, T. A. (2022). Large variation existed in standardized mean difference estimates using different calculation methods in clinical trials. *Journal of Clinical Epidemiology*, *149*, 89–97. doi: 10.1016/j.jclinepi.2022.05.023

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. doi: 10.1037/1082-989x.7.1.105

Thorsell, J., Finnes, A., Dahl, J., Lundgren, T., Gybrant, M., Gordh, T., & Buhrman, M. (2011). A comparative study of 2 manual-based self-help interventions, acceptance and commitment therapy and applied relaxation, for persons with chronic pain. *The Clinical Journal of Pain*, *27*(8), 716–723. doi: 10.1097/AJP.0b013e318219a933

Ying, Y., Ji, Y., Kong, F., Wang, M., Chen, Q., Wang, L., … Ruan, L. (2022). Efficacy of an internet-based cognitive behavioral therapy for subthreshold depression among Chinese adults: A randomized controlled trial. *Psychological Medicine*, 1–11. doi: 10.1017/S0033291722000599

https://doi.org/10.1017/S0033291722002057 Published online by Cambridge University Press