

Editorial

Direct Replication in Psychological Assessment Research

Mark S. Allen¹, Dragos Iliescu², and Samuel Greiff³

¹School of Psychology and Therapeutic Studies, Leeds Trinity University, Holesforth, Leeds, United Kingdom

²Faculty of Psychology and Educational Sciences, University of Bucharest, Romania

³Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg

Direct replication is the best (and possibly only) believable evidence for the reliability of an effect

(Simons, 2014, p. 76)

Independent verification through replication is the foundation of science. Observation of an effect becomes an established and accepted finding only when it has been duplicated many times by different research teams. Direct replications are scientific studies that test the repeatability of a previous finding by duplicating the design and methods as closely as possible. Direct replications should not be confused with conceptual replications that look for the same effect using different methods and under different conditions (see Derksen & Morawski, 2022). The value of performing direct replications is that they help to establish the generalizability of effects and help identify false positive results (Nosek et al., 2022). This is critical for progress in science as single-study findings are far from reliable. For example, the Open Science Collaboration (2015) directly replicated 100 psychology experiments and found that only 36% could be replicated, with effect sizes about half as large as the original research. This surprisingly low repeatability rate is thought to be a product of sloppy research practices (e.g., low statistical power) in combination with publication bias – the finding that studies that fail to demonstrate positive results are less likely to be published – and its associated effects (i.e., questionable research practices, such as selective reporting, HARKing, and *p*-hacking; see Kerr, 1998; Simmons et al., 2011). Because psychological assessment research is also susceptible to sloppy research practices and publication bias (e.g., studies that fail to confirm the validity or reliability of a measure are less likely to be published), psychological assessment tools that lack validity or reliability might remain in use for much longer than they should. Direct replication studies can provide valuable information regarding the reliability of research testing the reliability of psychological assessment tools.

A Look Back to 2011

The year 2011 is considered a turning point for psychological science and marked the beginning of what became known as the replication crisis in psychology. For a long time, the academic culture was such that little emphasis was placed on repeating the methods of other researchers, and if researchers wanted to look for the same effect they needed to do it in a different (novel) way. However, the publication of some seemingly implausible findings in 2011, in combination with journal editors' unwillingness to publish a refutation of those findings, resulted in a backlash from the scientific community and widespread discussion about the problems inherent in academic publishing (see Chambers, 2018). Central issues of concern were publication bias, data sharing, open access, statistical power, and questionable research practices (*p*-hacking in particular). These issues were thrown into the public light and the reputation of psychology as a whole was in crisis.

Throughout the 2010s, an increasing amount of research was published that highlighted problems in modern academic publishing and the importance of reproducibility. Researchers became unafraid to highlight poor editorial practices when they emerged, and journal editors started to recognize dated editorial policies and implement key changes to improve the quality (i.e., among others, reproducibility) of work featured in their journals. This replication crisis in psychological science might better be described as a renaissance period, with some noting that the scientific practices of psychologists have improved dramatically (Nelson et al., 2018). Many journals now explicitly encourage submissions of direct replication studies (e.g., *Journal of Personality and Social Psychology*, *Evolutionary Behavioral Sciences*), and researchers started going to great lengths to ensure that important findings in psychological science did in fact replicate (see e.g., Bouwmeester et al., 2017; Klein et al., 2018; Vohs et al., 2021; Wagenmakers

et al., 2016). Fuelled by this new recognition that replication is illustrative of excellent research practices (and perhaps by the severe reputational damage that journals might incur from refusing to publish direct replications of research featured in their own journal), the indifference and hostility toward direct replication appeared to be a thing of the past.

Where Do We Stand in 2023

To give a little perspective on how psychology has progressed, it is worth providing a few examples of comments that we have received from journal editors and reviewers when attempting to publish direct replication studies to this past year. In one example, combined with a recommendation not to publish, a reviewer commented: “Direct replication is indeed a powerful tool, but in itself, it is not generally publishable. To contribute to the field, replication should be used with extension efforts.” Just as discouraging are the following editorial comments: “A direct replication is nice but typically not the purview of the journal and the mere fact that your sample sizes are larger is unconvincing a case to be accepted ... and we have to prioritize studies that provide novel findings ...” and “we cannot go back and un-do past publications but instead, we, like science must move forward and hope/pray that over time bad ideas will die ... perhaps you can modify your paper into a brief report ... this would make the minor contribution easier to swallow for the journal.”

Many readers will no doubt have had similar experiences. There is a fair bit to unpack here, but the old fallacies appear to be alive and well, including the quest for novelty (rather than truth), disinterest in the replicability of results, and content for (potential) false positive findings to remain unchallenged. The first thing to point out is that direct replications do not represent a minor or inferior contribution, but rather, have the same level of scientific value as the original studies deemed publishable by the journal. Given the problems outlined earlier (publication bias, researcher bias), direct replication studies should be a priority for journals, rather than something that is reluctantly accepted. The second thing to note is that direct replications should not typically be extended variations of the original design. Even for simple cross-sectional research, adding more variables changes the study (e.g., boredom effects due to a longer questionnaire) and the study no longer qualifies as a direct replication. The third thing that warrants mention is the convenience of declaring that replication is not the purpose of a journal. Since independent verification through replication is the foundation of science, then all scientific journals should, in principle, be publishing replication studies. At *EJPA*, we very much consider ourselves a scientific journal and encourage authors to submit direct replications of studies previously published in the journal.

What Is and What Is not Direct Replication?

It is not possible to conduct an exact replication since there will always be countless minor differences between any two studies (Nosek et al., 2022) – the new sample is slightly older, the laboratory wall is painted blue rather than red, different pencils were used to complete questionnaires, and so on. At first glance, this might appear to be a problem for direct replication because the results of the original study might be dependent on these unknown factors. However, these differences actually benefit progress in science because findings that emerge only under highly specific conditions that will never occur again are not particularly useful to theory development or progress in science. Rather, replication has been described as a theoretical commitment: “A study is a replication when the innumerable differences from the original study are believed to be irrelevant for obtaining the evidence about the same finding” (Nosek et al., 2022, p. 722). For direct replications in psychological assessment research, it is important to consider the factors that are believed to be irrelevant. In particular, new assessment tools are often developed for a particular population and are not explicitly designed for all geographical regions. Therefore, replicating the study in a new world region (even if identical methods are used) would not qualify as direct replication, but rather, would fall under the umbrella of a cross-cultural validation study. For example, a questionnaire measure of “sense of humor” developed in one world region might have little value in other world regions given the strong connection between culture and humor. However, if a questionnaire is developed with the explicit intention of being valid across multiple world regions, then direct replication would be appropriate in any such region.

Direct replications differ from generalizability studies, sometimes termed “conceptual replication,” which look for the same effect using different methods and under different conditions (Derksen & Morawski, 2022). The term conceptual replication has received some criticism for being misleading (see Chambers, 2018) since methods are not duplicated and therefore these studies do not really qualify as replication. This is not to imply that conceptual replication is not also extremely valuable. Determining whether findings emerge under different conditions and using different methods is essential to progress in science. However, ideally, it should first be established whether a finding is robust, before exploring its effects across settings and populations. Both direct and conceptual replications provide information about generalizability. In direct replication, if the original finding is not replicated then this provides information that something about the original setting might have affected results and those results cannot be

generalized beyond the very specific conditions of the original study (or that results might represent a false positive finding). In conceptual replication, if findings do not emerge in the new conditions, then this provides information that findings of the original setting cannot be generalized to the new one (or that the original finding is not robust). Both are valuable, but direct replication appears to be much rarer.

Direct Replication in Psychological Assessment

In a previous editorial, we pointed out that inferential analysis in psychological assessment often differs from experimental research (that typically uses a dichotomy of rejecting or failing to reject a null hypothesis), and the implications this has for the registered report publication option at *EJPA* (Greiff & Allen, 2018). That is, psychological assessment research might conclude that the validity of a new measure is excellent, good, acceptable, questionable, poor, or unacceptable. In other words, psychological assessment research often uses a continuum (rather than a dichotomy) to formulate conclusions about assessment measures and this has implications for direct replication. Let us say a new questionnaire is developed and the authors use confirmatory factor analysis to demonstrate construct validity, presenting a series of model fit indexes such as: RMSEA = .08, CFI = .93, NNFI = .92, SRMR = .06. An independent researcher takes a look at the questionnaire and (on face value) the items appear somewhat confusing or unclear and so decides to replicate the study in order to find out whether the new questionnaire is in fact valid. The new study provides the following estimates: RMSEA = .11, CFI = .88, NNFI = .89, and SRMR = .10. In this example, the researcher might conclude that the study is a successful replication given that values are similar, but they could also argue that it is a failed replication given that the new study reports more imprecise estimates.

This is where the registered report option can be extremely helpful. Researchers can submit a Stage 1 registered report that pre-specifies the values required to have successfully replicated the result. For example, researchers could specify that the original study reported values that fall in the “good” range (RMSEA = .06–.08, CFI = .90–.95, NNFI = .90–.95, and SRMR = .06 to .08). The researchers might then specify that the RMSEA and CFI are the most critical estimates for formulating conclusions about construct validity and should either one fall below this pre-specified range then the study will be considered a failed replication. The key point is that values can be specified and agreed on prior to conducting the study, making the research conclusions

much less ambiguous. For this reason, we strongly encourage those conducting direct replications in psychological assessment to adopt the registered report format.

Conclusion

Independent verification through replication is the cornerstone of science, and it is a shame that directly replicating others’ work is often taken as a hostile act rather than a desirable (and even flattering) part of the research process (Nosek & Errington, 2020). Here at *EJPA*, we encourage researchers to conduct direct replication studies in psychological assessment, and these can be submitted in the registered report format (preferred) or as a regular article. Researchers can submit direct replications of previous studies that have been published in *EJPA* but can also submit direct replications of assessment research published elsewhere. Direct replications of research published in *EJPA* have demonstrated sufficient quality to warrant publication in the journal (as evidenced by the publication of the original study), but direct replications of research published elsewhere will need to demonstrate that the research is of sufficient quality to feature in the journal. That is, the study being replicated might have several weaknesses and not meet quality standards at *EJPA*. Nevertheless, we will entertain all replication studies and particularly encourage those that aim to replicate original studies of assessment tools that have been widely used for many years. In fact, we have made several attempts to encourage the submission of replication studies over the past few years (e.g., editorials, editorial presentations, and “meet the editors” open meetings at conferences). However, despite this explicit openness toward replication studies, it is striking that we have received so few of them. We hope this editorial encourages authors to conduct replication studies in psychological assessment research, and we look forward to receiving such submissions in the future.

References

- Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., Chmura, T. G. H., Cornelissen, G., Døssing, F. S., Espín, A. M., Evans, A. M., Ferreira-Santos, F., Fiedler, S., Flegr, J., Ghaffari, M., Glöckner, A., ... Wollbrant, C. E. (2017). Registered replication report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, 12(3), 527–542. <https://doi.org/10.1177/1745691617693624>
- Chambers, C. (2018). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Derksen, M., & Morawski, J. (2022). Kinds of replication: Examining the meanings of “conceptual replication” and “direct replication”. *Perspectives on Psychological Science*, 17(5), 1490–1505. <https://doi.org/10.1177/1745691621104116>

- Greiff, S., & Allen, M. S. (2018). *EJPA* introduces registered reports as new submission format. *European Journal of Psychological Assessment*, 34(4), 217–219. <https://doi.org/10.1027/1015-5759/a000492>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443–490. <https://doi.org/10.1177/2515245918810225>
- 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., & Errington, T. M. (2020). The best time to argue about what a replication means? Before you do it. *Nature*, 583, 518–520. <https://doi.org/10.1038/d41586-020-02142-6>
- Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. S., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719–748. <https://doi.org/10.1146/annurev-psych-020821-114157>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 253–267. <https://doi.org/10.1126/science.aac4716>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1), 76–80. <https://doi.org/10.1177/1745691613514755>
- Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., Alquist, J. L., Baker, M. D., Brizi, A., Bunyi, A., Butschek, G. J., Campbell, C., Capaldi, J., Cau, C., Chambers, H., Chatzisarantis, N. L. D., Christensen, W. J., Clay, S. L., Curtis, J., ... Albarracín, D. (2021). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, 32(10), 1566–1581. <https://doi.org/10.1177/0956797621989733>
- Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., ... Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928. <https://doi.org/10.1177/1745691616674458>

Published online January 16, 2023

Mark S. Allen

Mark S. Allen
School of Psychology and Therapeutic Studies
Leeds Trinity University
Horsforth
Leeds, LS18 5HD
United Kingdom
m.allen@leedstrinity.ac.uk