Reliability and Replicability in Infant Research:

A commentary on Byers-Heinlein et al. (2021)

Francesco Margoni

Department of Psychology, University of Oslo, Norway

Correspondence

Forskningsveien 3A, 0373 Oslo, Norway

Email: francesco.margoni@psykologi.uio.no

Acknowledgements: I wish to thank Erik Kjos Fonn, Joakim Haugane Zahl, and Lotte Thomsen for their useful comments on a previous version of this article.

Abstract

Byers-Heinlein et al. (2021) argue that a greater focus on reliability in infant research is much needed and desirable for the field. Moreover, they offer a number of possible solutions to increase reliability in infant research. Here, I do not discuss the suitability of the offered solutions. Instead, I further develop the authors' argument by considering the worryingly downstream consequences of low reliability for the very possibility of drawing meaningful inferences from the results of replication studies. I argue that the sources of error related to sampling and measurement greatly affect the estimates, resulting in confidence intervals being often very wide. Consequently, interpreting the differences between small-scale original and replication studies becomes a meaningless task, which, I maintain, could be replaced by a meta-analytic mindset and the habit of incorporating uncertainty directly in the research output or estimate.

Keywords: replicability, measurement error, methodology, meta-analysis, infancy

Word count: 1490

A major problem affecting infant research is certainly the use of small sample sizes and tests with low statistical power. In the target article, Byers-Heinlein and colleagues (2021) convincingly showcase how the statistical power of a test can be increased independently of sample numerosity by improving reliability. In doing so, they are calling our attention to an aspect that they diagnose to be often neglected by infant researchers: *measurement reliability*. Because, I suspect, especially in the study of infant cognition much focus has been placed on inferring whether at a certain point in the child's development a certain ability is present or absent, and less focus has been placed on investigating individual differences or correlations between different abilities, currently infant researchers seldom report indexes of the reliability of their measurements. Inevitably, it often simply remains unconsidered to what extent the results of a study are due to measurement error.

Low reliability of measures can easily lead to estimates far from the true one. Moreover, and interestingly enough, high measurement reliability in correlational and individual differences studies seem to increase the statistical power of the tests independently of sample size. The solutions offered by authors are therefore ways of improving reliability. First, given the current state of affairs in infant research, it would be crucial to start calculating and reporting estimates of reliability of the measures. This is something that authors conducting and reporting the study themselves should do, and something which would in the long run be extremely helpful not only in assessing the empirical results but also in evaluating the suitability of the measurement tools at disposal to investigate the phenomenon at hand. Second, as showcased by DeBolt et al. (2020), a promising partial solution to reduce measurement error would be to design studies in a way that would allow researchers to collect more data points per infant.

This commentary will not focus on assessing the adequacy or general feasibility of the six solutions offered in the target article. Instead, my aim is to build further on authors' main concerns for the current state of affairs in infant research, by expanding on a point that authors themselves recognize as important, that is, issues of measurement are closely intertwined with issues of replicability. If the diagnosis is correct, and indeed infant research suffers from having low measurement reliability in (correlational) studies that really aim at detecting true variability, then the great uncertainty around the findings necessarily affects not only the interpretation of the original results but also the interpretation of any single small-scale attempt to *replicate* that original result. And I will argue that it affects the interpretation of the findings to a point that it becomes meaningless to draw any conclusion with respect to the confirmation or disconfirmation of a given hypothesis or theory from either a failed or a successful attempt to replicate.

To see why, please note that across studies results can vary because of two main unavoidable sources of error: *sampling error* and *measurement error*. These sources of error and uncertainty are fundamental to explain why two or more studies investigating the very same phenomenon are unlikely to produce the same output. Especially when sample sizes are small, the single estimate reported in a single study

can and should be understood as only one possible outcome among a very wide range of possible estimates. These intervals and the possibility to find large differences between estimates across studies increase as a function of the amount of error in each study.

In my prior work, I have considered how the literature on infants' preference for prosociality over antisociality, spearheaded by Kiley Hamlin and colleagues, and the debate about whether the original results suggestive of such a preference are or are not trustworthy in light of the failed attempts to replicate conducted by independent laboratories, can be reinterpreted by considering how sampling error alone can affect the estimates (Margoni & Shepperd, 2020). For instance, it is possible to demonstrate how the result reported by Salvadori et al. (2015), i.e. 12/24 infants preferring prosociality ($P = 0.50$), which was interpreted as suggesting a failed replication of the original result by Hamlin and Wynn (2011), i.e. 12/16 successes ($P = 0.75$), falls indeed within the prediction interval, calculated to be $PI = (0.43, 0.96)$. In other words, it is possible to show that this large difference between the studies can be expected and explained as due to chance or sampling error *alone*. And please also consider that the calculations do not substantially change if we assume that the true underlying effect is $P = 0.64$ (i.e. taking a best guess with respect to where the effect truly lies, based on a meta-analysis, see Margoni & Surian, 2018) (see Tables 2-3 in Margoni & Shepperd, 2020).

But even before considering the influence of sampling error, we can see how measurement error alone can determine a great deal of variability in the estimates. Stanley and Spence (2014) offers a simple case to illustrate the point. Imagine that a first original study reported a correlation between age and well-being of $r = 0.30$. The case is simple enough because (a) age is an objective measure, and it is therefore not subject to measurement error; (b) we set the reliability associated to the well-being measurement at 0.70 (a perhaps somewhat optimistic assumption if we are to take this as a representative example of an infant study); (c) we assume omniscience and set the value of the true correlation at 0.30; and (d) we do not take into account sampling error, therefore we imagine that all the replication studies are conducted using the same materials and the very same participants, who also do not have any memory of the previous testing sessions. Even under these favorable conditions, by running a simulation of (100,000) replication studies, it is possible to calculate that results would range between $r = 0.03$ and $r = 0.46$, and distribute normally. This interval will then vary depending on the reliability of the measures, the sample sizes of the studies, the true unfortunately always unknown effect size, and the amount of sampling error. Just to give an example, Stanley and Spence (2014) shows that once sampling error is also considered and the true correlation is set at $r = 0.10$, then one can expect to legitimately find a negative correlation of $r = -0.33$ in a follow-up replication study. That is, one could expect these dissimilarities between findings in the original and the replication study just because of sampling and measurement error.

Thus, the downstream consequence of having low reliability, or even only a large sampling error, is that it becomes very difficult (and, as a result, meaningless) to interpret a failed (or successful) attempt to replicate. This is because what ostensibly is for instance a failed attempt to replicate often really is an estimate that falls between a replication or prediction interval. Still, often clear differences between the original and the replication studies emerge. However, no one seems to have a clear or definitive answer to the question of how much results can differ before they can be considered as mutually inconsistent. A simple answer would be to state that once results in the replication attempt are below the threshold of statistical significance, they can be considered inconsistent with those reported in the original study. However, we have seen how sampling and measurement errors can alone provide an alternative explanation of the variability across the results. In other words, because the confidence intervals surrounding the estimates are often very wide, it is extremely difficult to assess where the true estimate of the population parameter lies, and whether a small-scale replication study should be trusted more than an original small-scale study.

Asking whether a replication study confirms or disconfirms an original study often totally misses the point. What is really left to researchers is the opportunity to endorse a meta-analytic mindset and pool together all the findings at hand. As many have already suggested (Cumming, 2014; Halsey, 2019; Wilson et al., 2020), setting aside the logic of confirmation/disconfirmation under the Null Hypothesis Significance Testing approach in favor of the effort to best estimate effect sizes would be greatly beneficial and to some extent the only meaningful option left to researchers. Certainly, because meta-analyses estimates could be greatly biased and their accuracy ultimately depends on the quality of the single studies, increasing sample sizes and measurement reliability will be of paramount importance. Reducing measurement error would be a desirable goal even independently of the effort to increase sample numerosity, and it will have a positive effect on the accuracy of the estimates and the subsequent capacity to draw inferences based on the results. One very last point to conclude: Meanwhile infant researchers work toward these important goals, a viable and promising option would be to start developing the habit of incorporating uncertainty (directly) in the research output.

References

- Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development*, e2296.
- DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy*, 25, 393-419.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29.
- Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters*, 15, 20190174.
- Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26, 30-39.
- Margoni, F., & Shepperd, M. (2020). Changing the logic of replication: A case from infant studies. *Infant Behavior and Development*, 61, 101483.
- Margoni, F., & Surian, L. (2018). Infants' evaluation of prosocial and antisocial agents: A meta-analysis. *Developmental Psychology*, 54, 1445-1455.
- Salvadori, E., Blazsekova, T., Volein, A., Karap, Z., Tatone, D., Mascaro, O., & Csibra, G. (2015). Probing the strength of infants' preference for helpers over hinderers: Two replication attempts of Hamlin and Wynn (2011). *PloS one*, 10, e0140570.
- Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, 9, 305-318.
- Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*, 117, 5559–5567.