Developmental psychologists should care about measurement precision

Shane Lindsay & Emily Mather
Department of Psychology, University of Hull

February 2022

This is the peer reviewed version of the following article: Lindsay, S., & Mather, E. (2022). Developmental psychologists should care about measurement precision. Infant and Child Development, e2321, which has been published in final form at doi.org/10.1002/icd.2321. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for self-archiving.

Developmental psychologists should care about measurement precision

In a wide ranging article in this journal, Byers-Heinlein et al. (2022) make a persuasive case for paying close attention to reliability in developmental research. They focus on measurement reliability, which indexes how well individuals can be ranked across repeated measurements. We highlight the importance of measurement precision as a complement to understanding measurement reliability. We describe how this more absolute type of reliability can be quantified by the within-participant standard deviation across trials (SD_W) also known as the Standard Error of Measurement (SEM). We outline how reporting and understanding measurement precision can play a key role in achieving the 'six solutions' for improving reliability in developmental research offered by Byers-Heinlein et al.

Developmental psychologists should care about measurement precision

The replication crisis in psychology has led to claims of a host of other crises in psychology. One such crisis is the "measurement crisis" (e.g., Lilienfeld & Strother, 2020). The article by Byers-Heinlein et al. (2022) can be seen in this context as a practical response to address historical neglect of measurement across psychology, which some have argued as a major contributing factor to the replication crisis (e.g., Loken & Gelman, 2017). Byers-Heinlein et al.'s article is important and timely in raising awareness of measurement issues within development psychology and should be welcomed, especially for experimental psychologists who might not be as familiar with measurement concepts compared with other research traditions.

Byers-Heinlein et al. provide a helpful introduction to reliability, given that there are different types of reliability and terminology where meanings can vary in different fields. Two approaches to reliability can be distinguished in an attempt to further unmuddy the waters. One approach stems from psychometrics, with a focus on individual differences research and the use of correlational designs. Here a key goal is to distinguish individuals, where reliability can be defined as the ability to consistently rank or distinguish individuals in repeated measurements, and hence is a form of relative reliability. This is what Byers-Heinlein et al. refer to as measurement reliability. As the first of their six proposed solutions to improve reliability in infant research, Byers-Heinlein et al. suggest that researchers should routinely report the Intraclass Correlation Coefficient (ICC). In such an approach the ICC is a useful and appropriate statistic, and we agree that routine reporting of ICC would be beneficial to the appropriate use of measures in individual differences and correlational research.

A contrasting approach, often more associated with the physical sciences (where reporting an ICC would be unusual), sees the term reliability as concerned with accuracy (being close to a true value) and consistency (getting similar values), where the goal is to minimise error in multiple measurements. Byers-Heinlein et al. refer to this as measurement error or measurement precision. For the experimental psychologist, this sense of absolute reliability is likely to be of more primary interest if they want to understand and maximise the effect sizes they obtain. As Byers-Heinlein et al. describe, measurement precision has direct bearing on effect sizes, and they provide a nice demonstration of how increased measurement error can decrease Cohen's *d* (see their Figure 1).

However, for assessing the measurement precision of an instrument the ICC is not an appropriate tool since its value depends on the heterogeneity of the population it is used on. We would be unlikely to want to use a method to assess the accuracy of measuring children if that method gave different answers depending if we were measuring the heights of newborns or 5 year olds, where height has a much greater absolute range. Therefore if we are concerned with trying to better understand measurement in order to optimise our research designs we need a better tool for the job.

Such a tool is provided through a calculation very familiar to the psychologist: the standard deviation. Researchers need no compunction to report inter-participant variation in their studies, as usual practice is to report the standard deviation of the mean scores between participants in a study (here SD_B for *B*etween-participants). But whenever there are multiple trials per participant the measurement error can be estimated by calculating the standard deviation across the trials for each participant. Following Bland and Altman (1996), we refer to this measure of intra-participant variation as the within-participants standard deviation, SD_{Wi} (the subscript $_i$ denotes the score for an individual participant).

In order to summarise the within-participants standard deviations across participants, the common practice (for mathematical reasons relating to the calculation of variances) is not to take the arithmetic mean but the root mean square of the SD_{Wi} scores (SD_W). In the psychometrics literature this value is termed the Standard Error of Measurement (*SEM*) and sometimes known as the typical error (not to be confused with the standard error of the mean; see Weir, 2005, for further discussion and comparison with the ICC). SD_W can also be calculated from a reliability score (i.e., SD_B * sqrt(1- ICC)), or preferably (since it then does not depend on the type of ICC used) from the square root of the residual error variance in an ANOVA or linear model (Weir, 2005; see *r* package *SimplyAgree* for one example of a package that provides both ICC and SD_W).

Unlike the ICC, the SD_W is not a standardised unit. This could be seen as a comparative disadvantage, but we take the opposite view. When the goal is to improve research tools through a better understanding of measurement precision, we consider it important to get your hands dirty with measurement properties, rather than abstracting away from them. A similar argument can be made for effect sizes (e.g., Baguley, 2009), as standardised measures such as Cohen's *d* also obscure the nature of an effect compared with examining raw effect sizes and their variation separately. It is currently common practice to report the constituents of an effect size (the mean and SD_B) along with an effect size (i.e., Cohen's *d*). Likewise, along with the standardised measure we suggest its constituents be reported alongside with it, including both SD_B and SD_W. They have both inherent value as informative measures of variability in the scale of the dependent variable, and they aid interpretation of an ICC. Furthermore, the SD_W is not dependent on the number of trials used, unlike the ICC value Byers-Heinlein et al. recommend reporting, which reduces as trials increase. This makes it a less design dependent measure of a measurement tool.

Along with their Solution 1 to improve reporting practices, Byers-Heinlein et al. provide five additional solutions to improve reliability in infant research. We will briefly describe how each of these would benefit from increased focus on measurement precision provided by the reporting and informed utilisation of SD_W values from different paradigms.

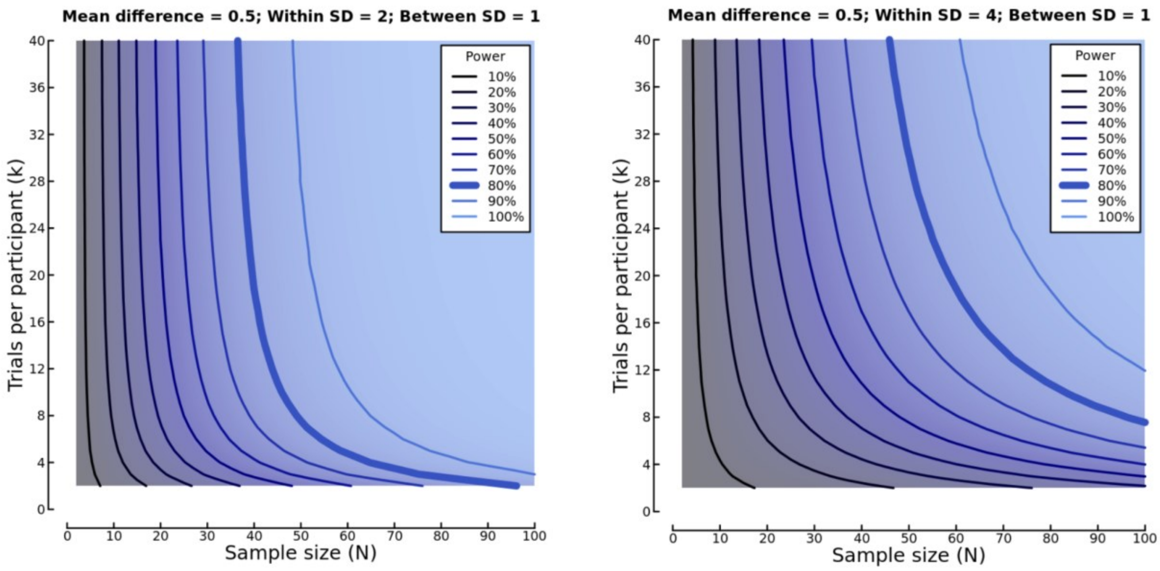
Solution 2: "Select the best measurement tool" and Solution 3:" Develop better infant paradigms" are related. In both cases SD_W can be used as a direct measure of the quality of an existing or new measurement tool. This is irrespective of an interest in group or individual differences, whereas ICC would primarily only be of interest to the latter audience.

For Solution 4: "Collect more data points per infant", SD_W can help in determining when this is most beneficial, as the higher the SD_W, the greater the benefit in statistical power from increased trials. Baker et al. (2020) provide an online tool (https://shiny.york.ac.uk/powercontours) that depends on the input of SD_B and SD_W that allows power calculations taking into consideration both prospective participant and trial numbers. We illustrate with an example shown in Figure 1, where the effect size has Cohen's $d = .5$. On the left, SD_B is half that of the size of SD_W and power of 80% is achieved with 8 trials and ~50 participants. More than 8-12 trials has limited effect on the contours and the primary value comes from increasing sample size. However, the right plot shows where SD_B is a quarter of SD_W, and here trial numbers have a more marked effect on power, where power of

80% with 50 participants is only achieved with ~40 trials. Using just 8 trials would necessitate running 100 participants to achieve the same level of power.

Figure 1.

Power contours for designs with low (left) and high (right) within-participant variability



For Solution 5: "Exclude low quality data from analysis" SD_{Wi} values can be used as a measure of data quality at an individual level (in comparison to the typical value SD_W), and potentially lead to removal of participants with very high values (for a discussion of this approach in an EEG research context, see Luck et al., 2021).

Finally, it remains to be seen how often Solution 6: "Conduct more sophisticated statistical analyses" will increase power while balancing Type 1 error rates. But we agree that further use of models such as hierarchical linear or Bayesian mixed-effect models is important. This is because they allow inclusion of item specific variance and trial level variance and hence measurement error to be more clearly quantified, in contrast to approaches such as ANOVA where measurement error can be obscured (see Singmann et al., 2021, for the perils of standardisation and aggregating away from trial variance). This should encourage awareness of the importance of considering measurement error and we would expect better inferences from such models and more realistic power analyses than where measurement precision is assumed to be perfect (see Debruine and Barr, 2021 for a guide to using simulations for power analyses with mixed-effects models that incorporates SD_W).

In conclusion, we believe that infant research needs to be both more reliable and to have higher measurement precision: we should care about both relative and absolute reliability, and we should report and use measures of both. While Byers-Heinlein et al. have a focus on improvements to statistical power from increasing our effect sizes, we would highlight that our ultimate goal should be to advance our theoretical understanding of the effects and the variability of their components. To achieve this goal we need to see our effects with more resolution in order to understand the different sources of variation that generate infant behaviour.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported?. *British Journal of Psychology, 100*(3), 603-617.

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2020). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods, 26*(30). 295-314.

Bland, J. M., & Altman, D. G. (1996). Statistics notes: measurement error. *British Medical Journal, 312*(7047), 1654.

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development.*

DeBruine, L. M., & Barr, D. J. (2021). Understanding mixed-effects models through data simulation. *Advances in Methods and Practices in Psychological Science, 4*(1).

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584-585.

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 281.

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology, 58*(6), e13793.

Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S., Davis-Stober, C., Dunn, J. C., … Shiffrin, R. (2021, June 20). Statistics in the Service of Science: Don't let the Tail Wag the Dog. https://doi.org/10.31234/osf.io/kxhfu

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *The Journal of Strength & Conditioning Research, 19*(1), 231-240.