When to use a cookie, and when to use a ruler: A response to Byers-Heinlein, Bergmann and Savelei's "Six solutions for more reliable infant research"

Abstract

Here we provide a response of broad agreement to the authors' six suggestions for improving the reliability of infant research. We also draw attention to three points we feel are important additions. We discuss the importance of considering measurement validity alongside measurement reliability as both contribute to overall measurement error, as well as emphasizing the selection of methods that are theory driven and balancing this with the need for precise or accurate measures. We also briefly discuss how differences in training, access to resources and other factors can influence choices of method and analysis. Sometimes simple is good enough.

When to use a cookie, and when to use a ruler: A response to Byers-Heinlein, Bergmann and Savelei's "Six solutions for more reliable infant research"

The particular challenges of infant research are well documented, and a desire to understand infant development has driven the design of ingenious paradigms with which to investigate such development in many areas. We broadly agree with the central claims and suggestions made by the authors, but seek to further this important discussion by drawing attention to three points we feel require more emphasis. Our first point is that measurement of mental processes relies on inference from behaviors and that this measurement validity forms part of the measurement error the authors discuss. Secondly that theory should always be the primary driver of methodological choices, and finally that there are some pragmatic limitations on implementing more complex methods and analysis which should be considered.

Measurement Error Includes Measurement Validity

As the authors acknowledge, some effects are more difficult to measure than others, and some are larger than others. But we would go even further than this, and say that given our current state of knowledge, some are not possible to measure at all. For example, we cannot *directly* measure attention, perception, or understanding. Instead, we rely on innovative experimental designs as proxy measures for phenomena like these. For example, the high amplitude sucking procedure has been used to determine whether infants can detect change in auditory stimulus (e.g., Eimas et al., 1971) and infant looking times have been used to detect categorical perception (Skelton et al. 2017). Using the authors' own example, this is rather like using cookies to measure height from a photograph, rather than *directly measuring* the person. Whilst innovation may improve the methods we have in the future, it is likely that certain aspects of infant development will never be directly accessible and so will always be estimated at some

degree of distance from the phenomena in question, and thus rely on a number of assumptions. This means many aspects of human research rely on effective inference. We typically refer to how successfully a measure reflects the underlying phenomena as measurement validity. Measurement error as presented by Byers-Heinlein and colleagues focused solely on how well a particular measure can capture a behaviour, but not how well the behaviour measured actually reflects the underlying phenomena we seek to understand. This is important because both of these are part of the overall "net" measurement error.

The example of eye tracking is an interesting example of this, not least because the authors found such interesting effects on effect size with tighter inclusion criteria. If you are interested in measuring where an infant is looking and for how long, eye tracking, even with infants, has very high spatial resolution, and is therefore a fairly accurate and precise measure of where an infant's eyes look. It is when you use this looking behaviour to make inference about phenomena such as infant learning or attention that this measure might become a less accurate and precise reflection of the underlying phenomena (e.g., understanding, perception). Eye movements effectively include a lot of other data about individuals (e.g., interests, fatigue levels, previous experience) and other competing environmental factors which can mean that even at the level of the individual, test-retest reliability would be low. Therefore, for this particular method, the lack of precision or accuracy is arguably due more to issues of validity than which may not be fixed by increasing the number of data points collected. Whilst the authors are correct that the net effect may be referred to as measurement error, it might be helpful for individual researchers to consider where the variability in measurements comes from and whether it can be reduced before considering alternative methods, perhaps relying on triangulation of a variety of measures and methods instead (LoBue et al., 2020).

Theory Should Drive Method Choice

We agree whole heartedly that infant researchers should aspire to use stimuli and paradigms which are ‘better’. But, we see ‘better’ as being driven primarily by theory, rather than better per se. To use the authors’ cookie-height measurement analogy, often the cookie method will be sufficient for answering at least some of the questions we may have about childrens’ height, provided our hypotheses have been built on strong theory, and the conclusions we are drawing from the data are appropriate. That is to say, prioritising the validity of our methods and conclusions will build stronger theories than relying on reliability alone.

To illustrate this, let's consider two research questions and the most appropriate way to answer them. Firstly, can infants see colour? Secondly, what is the perceptual experience of colour for infants? Teller et al (1978) contributed evidence for the first question using preferential looking to coloured vs achromatic lights. As measurements of colour vision go, this is a cookie. The data can’t be used to give us much precision in answering the question of perceptual experience, but it is sufficient to answer the question of whether infants have any colour vision at all. It's ability to do so is thanks in part to a meticulous selection of stimuli which force a single conclusion to a specified hypothesis (Aslin, 2000), and not just the specificity of the method or analysis used.

To attempt to answer the question of perceptual experience, we need a ruler. Knoblauch et al (2001) combined preferential looking with psychophysical measurements of colour intensity thresholds in infants and children. In theory, we *could* use this data to answer the question of whether infants have colour vision, however, that approach would not be theoretically appropriate because it relies on assumptions on the mechanisms and structure of infants colour vision. Knoblauch et al did have access to this data thanks to the ‘cookie’ studies such Teller et

al. The consequences of attempting such precise measurements before accuracy could be over-interpretation/fitting of our data (being inaccurate), or a set of data which cannot reliably rule out an alternative hypothesis (being imprecise). Comparison of how these studies answer our two research questions show our inference is constrained by the appropriateness of each method given the existing theory, and not the method or analysis itself (i.e. the validity of these methods is critical in driving understanding and theory forward). In areas where there is little existing theory present to guide method choices, researchers should aim to systematically understand which parameters and experimental conditions an effect holds under (Steinle, 1997), which likely requires testing using a diverse range of methods. Quantifying measurement error has a clear role to play in providing evidence for the strength of the data.

It is important that we do not lose sight of the aim of building general theories of development, which requires both accuracy and precision, while in the pursuit of the 'best' single experimental paradigm or analysis.

Some Pragmatic Constraints

There is little benefit in being precise at analysis and measurement of data which is not reflective of the ground truth. There are pragmatic constraints on how this might arise, for example as a result of the narrow sociodemographic sample of participants, or from inadvertently excluding evidence from studies using ‘simpler’ methods or analyses when defining our hypotheses. More complex methods and analysis often require a skills-diverse and well-resourced baby lab (e.g., time, specialist lab space, investment in training, access to equipment, software etc), and as a result are not equally accessible to all researchers. Cross-lab collaborations and open science practices with which the authors are of course most familiar, can help close this resource gap across labs. There are reasonably accessible methods (e.g., the

practice of reporting a Bayes Factor with every p-value, Dienes 2014), that researchers can use to interpret and present experimental data, and clearly, much of what is suggested by Byers-Heinlein et al can also become accessible. Nonetheless, we raise this issue to stress that theoretically grounded research whose conclusions are appropriate for the data are valuable *regardless* of the method used, and that who is measuring and who is being measured is as important as how we measure them and what with.

Conclusion

We share the concerns of the authors, and believe their recommendations form an excellent basis for some rule of thumb guides to thinking about research designs when embarking on new studies and we hope to further the discussion of how these recommendations can be refined. We believe it is important to think about inference error as a separate form of measurement error because in some cases this may change the evaluation of *some* methods for answering *some* questions. We would argue that it may be more important to be thinking about novel ways of getting closer to measuring the real objects of our interest (e.g., attention or comprehension) than refining our measurement of the behaviours we believe reflect them (although of course this may never be possible). We also believe that a greater emphasis should be placed on the value of theory development as the central driver of the methods selected, and that the value of applying many diverse methods to a problem should not be under-rated. Considering what we are trying to measure should be at least as important as what we use to take the measure- we should prioritise high validity in our methods and conclusions. Finally, we considered some of the practical problems for those from less well-funded areas of academia who may be unwittingly excluded from contributing on the basis of accessing resources or highly specific training.

References

Aslin, R. N. (2000). Why take the cog out of infant cognition?. *Infancy*, *1*(4), 463-470 https://doi.org/10.1207/S15327078IN0104_6

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(781) doi: 10.3389/fpsyg.2014.00781

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science, 171*(3968), 303-306.

Knoblauch, K., Vital-Durand, F., & Barbur, J. L. (2001). Variation of chromatic sensitivity across the life span. *Vision research*, *41*(1), 23-36 .https://doi.org/10.1016/S0042-6989(00)00205-4

LoBue, V, Reider, LB, Kim, E, Burris, J. L., Oleas, D. S., Buss, K. A., ... & Field, A. P. (2020). The importance of using multiple outcome measures in infant research. *Infancy 25*, 420–437. https://doi.org/10.1111/infa.12339

Skelton, A. E., Catchpole, G., Abbott, J. T., Bosten, J. M., & Franklin, A. (2017). Biological origins of color categorization. *Proceedings of the National Academy of Sciences, 114*(21), 5545-5550. https://doi.org/10.1073/pnas.1612881114

Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science*, *64*, S65-S74. https://doi.org/10.1086/392587

Teller, D. Y., Peeples, D. R., & Sekel, M. (1978). Discrimination of chromatic from white light by two-month-old human infants. *Vision Research, 18*(1), 41-48. https://doi.org/10.1016/0042-6989(78)90075-5