Received: 31 March 2022 | Accepted: 14 April 2022

DOI: 10.1002/icd.2326

COMMENTARY

WILEY

Valid points and looks: Reliability and validity go hand-in-hand when improving infant methods

Martin Zettersten[1] | Ron Pomper[2] | Jenny Saffran[3]

[1]Department of Psychology, Princeton University, Princeton, New Jersey, USA

[2]Center for Childhood Deafness, Language and Learning, Boys Town National Research Hospital, Omaha, Nebraska, USA

[3]Department of Psychology, University of Wisconsin-Madison, Madison, Wisconsin, USA

Correspondence
Jenny Saffran, Department of Psychology, University of Wisconsin-Madison, 1202 W Johnson St, Madison, WI 53706, USA.
Email: jenny.saffran@wisc.edu

Handling Editor: Moin Syed

Abstract

In this commentary, we suggest that infancy researchers should carefully consider validity when taking steps to improve reliability. Zooming in to focus on looking-time methods, we argue that limitations in validity represent perhaps an even more fundamental issue than reliability. At the same time, focusing single-mindedly on reliability comes with two possible pitfalls: maximizing reliability at the expense of construct validity, and overvaluing parental report measures compared to direct measures of infant behaviour. Finally, we articulate several promising avenues for improving validity in infant research: experimental and modelling efforts to characterize the functional relationship between measures such as looking time and infant cognition, using multiple measures to establish convergent validity, and improving our understanding of how measures vary across a broader set of stimulus characteristics.

KEYWORDS
infant methods, looking-time methods, reliability, validity, word recognition

Byers-Heinlein et al. (2021) do commendable work drawing attention to the importance of measurement reliability. In our commentary, we broaden the focus to include issues concerning validity. The goals of optimizing reliability and validity are intertwined—and may sometimes even conflict. While we focus on looking-time methods, many similar questions arise when considering other infant methodologies.

1 | THE CHALLENGE OF ESTABLISHING RELIABLE AND VALID INFANT MEASURES

Reliability and validity are closely connected concepts. In psychological research, validity—whether or not conclusions drawn from a measurement are well-founded—can be difficult to ascertain, because most constructs are latent or unobservable and therefore cannot be measured directly. Reliability is often thought to be *necessary*, but not *sufficient*, to establish a measure's validity (Flake et al., 2017). Although a measure without reliability will fail to consistently measure the intended construct, improvements in reliability cannot guarantee validity. Even a perfectly reliable measure does not necessarily index the construct that researchers are hoping to tap. Recent reviews have highlighted the absence of consistent construct validation practices in psychological science (Flake & Fried, 2020). These concerns are particularly acute in infant research (Kominsky et al., 2020).

To illustrate the problem of validity, consider studies in which infants are exposed to some to-be-learned material (such as a sequence of sounds) and preferential looking times to novel versus familiar stimuli serve as the index of learning (Saffran et al., 1996). Most analytic approaches treat looking time as a continuous metric, implicitly assuming that there is a meaningful relationship between the magnitude of individual infants' looking-time preferences and learning—for example, that larger novelty preferences indicate greater learning (for a recent example, see Hoareau et al., 2019). However, our understanding of factors that determine the functional relationship between looking-time differences and underlying learning processes are limited (Aslin, 2007; Bergmann et al., 2013). Has an infant who shows a 1000 ms novelty preference learned the familiarization material better than—or perhaps even twice as well as—an infant who shows a 500 ms novelty preference? Note that this problem remains unsolved even if we assume (for argument's sake) that we can measure individual infants' preferences with perfect reliability. That is, even if we know with certainty that infant A will consistently show a larger novelty preference than infant B, it is not clear that this means that infant A is learning faster/more than infant B. Even if measurement reliability is vastly improved, fundamental questions about the interpretation and validity of infant looking methods will remain.

2 | POSSIBLE DANGERS OF FOCUSING TOO HEAVILY ON RELIABILITY

In principle, the twin goals of optimizing reliability and validity do not necessarily conflict. In practice, however, efforts to improve either of these measurement properties may lead to tradeoffs. We identify two possible pitfalls if researchers overzealously maximize reliability without careful consideration of validity (as also cautioned by Byers-Heinlein et al., 2021).

2.1 | Reliable, but still valid?

In some cases, increasing reliability could come at the expense of deriving meaningful measures. To make this concern concrete, consider the target article's important recommendation to increase trial numbers. While doing so will typically increase reliability, it also makes the task more taxing. In addition to the construct of interest, many infant behavioural measures capture how readily infants continue to engage with the task over time. Taken to its extreme, a task with many, many trials may become a highly reliable measure of individual differences in infants' willingness to continue to stick with the task, rather than the targeted cognitive ability (see DeBolt et al., 2020 for practical guidance on trial numbers).

Another example comes from recent work on improving reliability in looking-while-listening measures of infant word recognition (Zettersten et al., 2021). These tasks track eye movements to quantify accuracy in fixating on a target referent after it is named (e.g., *Find the book*). Recognition accuracy has typically been measured during a short time window immediately after the onset of the target word (e.g., 300–1800 ms), based on the rationale that

fixations later in the trial are less likely to be driven by the spoken target word (Fernald et al., 2008). A recent reanalysis of a large number of datasets demonstrated that longer time windows (e.g., extending up to 3000–4000 ms) may instead serve as better default windows of analysis, because they tend to maximize inter-item reliability. However, this approach also raises important questions about construct validity: to what extent do accuracies computed from longer windows still capture target word recognition? By including looking behaviour several seconds after the onset of the target label, dependent variables will measure not only infants' lexical processing, but also their tendency to sustain attention on a single picture (and visual disengagement in general; see Venker, 2017, for an example). More generally, taking steps to increase reliability must be carefully weighed to ensure that they do not come at the expense of measuring the desired construct of interest.

2.2 | Throwing the baby behaviour measures out with the bathwater

Another concern is that focusing primarily on reliability may lead us to overvalue certain types of measurements—namely, ones with the highest reliability—even if they lack validity as measures of our construct of interest. In particular, developmental measures that rely on adults reporting on child behaviours may come to be valued at the expense of measures of children's behaviour itself. Adult-based measures will typically exhibit superior reliability properties relative to measures of young children's behaviour. For example, parental report measures (such as the MacArthur-Bates Communicative Development Inventories) have high rates of test–retest reliability ($r > 0.8$ for multiple measurements within a retest range of several months; Frank et al., 2021) that will be difficult to attain for measures based on children's behaviour, even if measurement practices for child behaviour are improved. However, in some cases, parental report may provide a less valid measure of children's abilities. For example, parental report underestimates infants' comprehension of specific words as revealed by more sensitive looking-time measures (Bergelson & Swingley, 2012; Houston-Price et al., 2007; Venker et al., 2016). For many research purposes, we would not want to consider parental report to be superior to infant behavioural measures on the basis of reliability properties alone.

3 | PATHS TO IMPROVING VALIDITY

One step towards improving both reliability and validity of gaze-based measures is to deepen our understanding of the factors that systematically predict looking behaviour. Several current lines of work illustrate promising avenues towards this goal. Recent projects have begun studying whether the strength of learning shapes the magnitude of looking-time differences (Zettersten et al., 2020). Improved modelling techniques have begun to parse the components that shape infants' looking behaviour in a more fine-grained manner (McMurray, 2020; Piantadosi et al., 2014; Poli et al., 2020). Efforts to combine pre-existing data into large-scale databases give us sufficient power to systematize our understanding of the measurement properties of looking-time data (Zettersten et al., 2021; Zettersten, Yurovsky, et al., 2022). Finally, large-scale multi-site efforts such as the ongoing ManyBabies5 project can help us understand the degree to which variance in looking time is predicted by properties of participants (e.g., age) and properties of the task design (e.g., familiarization time). Together, efforts of this kind promise to substantially improve our understanding of the link between looking-time measures and infants' cognitive processes.

Another key step towards solidifying the construct validity of infant measures is developing converging evidence across multiple measures (Havron, 2022; LoBue et al., 2020). When multiple measures targeting the same construct converge on similar conclusions, it increases confidence that we are measuring what we intend to measure. Inspiring examples using multiple measures include efforts combining looking-time methods with manual exploration to study curiosity (Perez & Feigenson, 2022; Stahl & Feigenson, 2015) and measuring infants' attentional profile using both looking time and heart-rate variability (Richards, 1987; Wass et al., 2018).

A third avenue is improving our understanding of how infant behaviour varies across stimuli. Infant designs often suffer not only from a limited number of trials but also from small item sets, limiting our ability to derive generalizable conclusions (Yarkoni, 2022). For example, the vast majority of looking-while-listening studies test infants' ability to recognize a small set of familiar words (e.g., dog) given prototypical exemplars of each word (e.g., Golden Retrievers). Manipulating stimulus properties parametrically (Pomper et al., 2021; Pomper & Saffran, 2019; Zettersten, Weaver, & Saffran, 2022) and establishing greater variety in stimuli through large-scale data collection (Visser et al., 2022; Zettersten et al., 2021) can help establish effects of cross-stimulus variability and improve our understanding of the factors driving infant behaviour.

4 | CONCLUSION

To some, the path ahead for infant measures may appear daunting. Improving the psychometric properties of infant methods is hard. However, this is a particularly exciting time in infant research. With large-scale empirical investigations and continual improvements in analytic techniques, the field has never been so well-positioned to tackle the questions that have challenged us for over 60 years—what's in an infant's look? (Aslin, 2007; Fantz, 1964; Haith, 1998). We hope that researchers will continue to embrace the difficulty of reliable and valid measurement in infant behaviour and cognition.

AUTHOR CONTRIBUTIONS

Martin Zettersten: Writing – original draft; writing – review and editing. **Ron Pomper:** Writing – original draft; writing – review and editing. **Jenny Saffran:** Writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

We thank Jessica Kosie and Casey Lew-Williams for feedback on an earlier version of this manuscript, and members of the UW-Madison Infant Learning Lab (past and present) for many useful conversations about these issues.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/icd.2326.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Martin Zettersten https://orcid.org/0000-0002-0444-7059
Ron Pomper https://orcid.org/0000-0001-5595-4192
Jenny Saffran https://orcid.org/0000-0003-3749-8773

REFERENCES

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Bergmann, C., Ten Bosch, L., Fikkert, P., & Boves, L. (2013). A computational model to investigate assumptions in the headturn preference procedure. *Frontiers in Psychology*, *4*, 676. https://doi.org/10.3389/fpsyg.2013.00676

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development*, e2296. https://doi.org/10.1002/icd.2296

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*(4), 393–419. https://doi.org/10.1111/infa.12337

Fantz, R. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668–670. https://doi.org/10.1126/science.146.3644.668

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. In I. A. Sekerina, E. M. Fernández, & H. Clahsen (Eds.), *Developmental psycholinguistics: On-line methods in children's language processing* (pp. 97–135). John Benjamins Publishing Company.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press.

Haith, M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior and Development*, *21*(2), 167–179. https://doi.org/10.1016/S0163-6383(98)90001-7

Havron, N. (2022). Why not both? Using multiple measures to improve reliability in infant studies. PsyArXiv. https://doi.org/10.31234/osf.io/nu465

Hoareau, M., Yeung, H. H., & Nazzi, T. (2019). Infants' statistical word segmentation in an artificial language is linked to both parental speech input and reported production abilities. *Developmental Science*, *22*, e12803. https://doi.org/10.1111/desc.12803

Houston-Price, C., Mather, E., & Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. *Journal of Child Language*, *34*(4), 701–724. https://doi.org/10.1017/S0305000907008124

Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, K. (2020). Simplicity and validity in infant research. PsyArXiv. https://doi.org/10.31234/osf.io/6j9p3

LoBue, V., Reider, L. B., Kim, E., Burris, J. L., Oleas, D. S., Buss, K. A., Pérez-Edgar, K., & Field, A. P. (2020). The importance of using multiple outcome measures in infant research. *Infancy*, *25*(4), 420–437. https://doi.org/10.1111/infa.12339

McMurray, B. (2020). I'm not sure the curve means what you think it means: Toward a [more] realistic understanding of the role of eye-movement generation in Visual World Paradigm. PsyArXiv. https://doi.org/10.31234/osf.io/pb2c6

Perez, J., & Feigenson, L. (2022). Violations of expectation trigger infants to search for explanations. *Cognition*, *218*, 104942. https://doi.org/10.1016/j.cognition.2021.104942

Piantadosi, S. T., Kidd, C., & Aslin, R. (2014). Rich analysis and rational models: Inferring individual behavior from infant looking data. *Developmental Science*, *17*(3), 321–337. https://doi.org/10.1111/desc.12083

Poli, F., Serino, G., Mars, R. B., & Hunnius, S. (2020). Infants tailor their attention to maximize learning. *Science Advances*, *6*(39), eabb5053. https://doi.org/10.1126/sciadv.abb5053

Pomper, R., Kaushanskaya, M., & Saffran, J. (2021). Change is hard: Individual differences in children's lexical processing and executive functions after a shift in dimensions. *Language Learning and Development*. https://doi.org/10.1080/15475441.2021.1947289

Pomper, R., & Saffran, J. R. (2019). Familiar object salience affects novel word learning. *Child Development*, *90*(2), e246–e262. https://doi.org/10.1111/cdev.13053

Richards, J. E. (1987). Infant visual sustained attention and respiratory sinus arrhythmia. *Child Development*, *58*(2), 488–496. https://doi.org/10.2307/1130525

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94. https://doi.org/10.1126/science.aaa3799

Venker, C. E. (2017). Spoken word recognition in children with autism spectrum disorder: The role of visual disengagement. *Autism*, *21*(7), 821–829. https://doi.org/10.1177/1362361316653230

Venker, C. E., Haebig, E., Edwards, J., Saffran, J. R., & Ellis Weismer, S. (2016). Brief report: Early lexical comprehension in young children with ASD: Comparing eye-gaze methodology and parent report. *Journal of Autism and Developmental Disorders*, *46*(6), 2260–2266. https://doi.org/10.1007/s10803-016-2747-z

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., Kaldy, Z., Kulke, L., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., … Zettersten, M. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. *Behavioral and Brain Sciences*, *45*, E35. https://doi.org/10.1017/S0140525X21000455

Wass, S. V., de Barbaro, K., Clackson, K., & Leong, V. (2018). New meanings of thin-skinned: The contrasting attentional profiles of typical 12-month-olds who show high, and low, stress reactivity. *Developmental Psychology*, *54*(5), 816–828. https://doi.org/10.1037/dev0000428

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Zettersten, M., Bergey, C., Bhatt, N., Boyce, V., Braginsky, M., Carstensen, A., de Mayo, B., Kachergis, G., Lewis, M., Long B., MacDonald, K., Mankewitz, J., Meylan, S., Saleh, A., Schneider, R. M., Tsui, A., Uner, S., Xu, T. L., Yurovsky, D., & Frank, M. C. (2021). Peekbank: Exploring children's word recognition through an open, large-scale repository for developmental eye-tracking data. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

Zettersten, M., Black, A., Bergmann, C., Bacon, D., Weaver, H., & Saffran, J. (2020, October 22–-23). *Investigating the relationship between infant learning and measured effect size in preferential looking paradigms* [poster presentation]. Many paths to language, Nijmegen, The Netherlands, Max Planck Institute.

Zettersten, M., Weaver, H., & Saffran, J. (2022). Becoming word meaning experts: Infants' processing of familiar words in the context of typical and atypical exemplars. PsyArXiv. https://doi.org/10.31234/osf.io/njh38

Zettersten, M., Yurovsky, D., Xu, T., Uner, S., Tsui, A., Schneider, R. M., Saleh, A. N., Meylan, S., Marchman, V., Mankewitz, J., MacDonald, K., Long, B., Lewis, M., Kachergis, G., Handa, K., de Mayo, B., Carstensen, A., Braginsky, M., Boyce, V., Bhatt, N. S., Bergey, C. Frank, M. C. (2022). Peekbank: An open, large-scale repository for developmental eye-tracking data of children's word recognition. PsyArXiv. https://doi.org/10.31234/osf.io/tgnzv

How to cite this article: Zettersten, M., Pomper, R., & Saffran, J. (2022). Valid points and looks: Reliability and validity go hand-in-hand when improving infant methods. *Infant and Child Development*, e2326. https://doi.org/10.1002/icd.2326