Received: 8 March 2022 | Revised: 25 April 2022 | Accepted: 29 April 2022

DOI: 10.1002/icd.2338

COMMENTARY

WILEY

Reducing measurement error with ecologically valid testing methods

Kelley E. Gunther | Berenice Anaya | Koraly Pérez-Edgar

Penn State University, University Park, PA, USA

Funding information
National Institute of Neurological Disorders and Stroke, Grant/Award Number: 1F99NS120411-01; National Science Foundation, Grant/Award Number: DGE1255832

Handling Editor: Syed Moin

Abstract

'Six solutions for more reliable infant research' outlines important steps in improving research with the youngest of participants. Here, we suggest that a complimentary avenue for further improving reliability can be found in employing more ecologically valid tasks. Because laboratory tasks are often designed as simple, heavily controlled representations of a construct, they can be devoid of the necessary context in which the construct naturally unfolds. We posit that additional error is generated in translating laboratory task metrics back to the construct of interest. Thus, it may be necessary to supplement more traditional laboratory tasks with iterations that more closely match the research construct of interest and how it may unfold in the 'real world'. We make suggestions for adjusting to the 'noise' incurred by moving research out of traditional laboratory environments, while also recognizing how these adjustments may improve the sociodemographic profiles within research studies.

KEYWORDS
ecological validity, infancy, measurement error

1 | REDUCING MEASUREMENT ERROR WITH ECOLOGICALLY VALID TESTING METHODS

Infant research, particularly when examining sensory or cognitive processes, has often relied on highly controlled experimental paradigms situated within a laboratory. These paradigms have been designed, in part, to minimize measurement noise. With precisely timed trials and testing forums mostly devoid of additional environmental influences, these methods aim to carefully control, or account for, all possible aspects of the task beyond the behaviour in question. Protocols often aim for a replicable laboratory experience across all infants, so any differences in behaviour can

https://doi.org/10.1002/icd.2338

be attributed to the specific experimental manipulation or developmental mechanism of interest. This is one important avenue for boosting signal:noise ratio to conduct research in infant populations that is reliable and robust. Indeed, in their article, Byers-Heinlein et al. (2021) emphasize the importance of measurement reliability, the ability of a tool to precisely and/or consistently measure a construct. One way to achieve reliable measurement is to hold constant the exact environment in which the measurement takes place and tightly control the parameters of a task.

However, an outsized emphasis on experimental control often leads researchers to rely on tasks that lack strong ecological validity. That is, paradigms may not capture behaviours or processes in a way that readily translates to the way they unfold outside of a laboratory. For example, a baby's attention to pictures of toys on a computer screen may be different from how they deploy attention to the toy when it is presented in the form of a three-dimensional object to visually process or when they can physically hold the toy. Furthermore, their attention to a picture of a single toy may differ from when the toy is in their home, surrounded by other salient objects and competing activities. While both 'tasks' may be measuring attention to the same toy, the physical and emotional context in which these attention patterns unfold are radically different and may significantly alter the observed attention patterns. While we must employ the most consistent, precise measure (i.e., high reliability) of a construct to ensure reliability, measurement error is still a concern if our operationalization for the phenomenon of interest is misaligned with or narrowly scoped to only certain aspects of the true phenomenon (i.e., low validity). That is, measurement error arises not only from noisy or hard to reproduce measures of behaviour, but also in a mismatch between the measure and the broader construct of interest.

Reliability is concerned with consistency and precision, while validity focuses on the accuracy of the measurement. Although the two are separate constructs, they also compliment each other in reducing measurement error. Vazire et al. (2022) assert that improving the reliability of measures also calls for improving the validity of these metrics. We posit that bolstering validity by way of maximizing ecological validity will in turn improve reliability. Specifically, measuring a behaviour in a way most similar to how it unfolds in the 'real world' reduces error introduced when trying to tie fairly constrained laboratory tasks to more complex psychological processes.

Ironically, translating task performance from the laboratory to predict behaviour outside of the laboratory, laboratory measures generate an additional source of noise and error. That is, attempts to fully account for metrics that ensure measurement reliability in a highly controlled setting may inadvertently produce a task less attuned to natural variation in performance, harming both validity and generalizability. Consequently, the additional noise (i.e., error) incurred with these practices limits how robust a measurement can be. Therefore, we propose that developing a more robust infant science requires coupling improved reliability from our repertoire of tightly-controlled laboratory tasks with approaches that capture behaviours in their most true-to-life form.

To mitigate this translation error, we suggest supplementing traditional tasks with re-imagined infant paradigms that better approach ecological validity, testing constructs in ways that are more true-to-life. Advances in ambulatory data collection have made it easier than ever to collect precision measures, including functional near infrared spectroscopy (fNIRS; Lloyd-Fox et al., 2014), vocalization data (Richards et al., 2017) and mobile eye-tracking (Franchak & Yu, 2022), while a participant completes paradigms in nontraditional testing environments. These can include a laboratory-managed playroom, their own home, or an outside location, like a school or museum. As devices become smaller, they have become even more infant-friendly and increasingly wearable for even the smallest of participants. As a result, infants are able to move and play as they normally would, in any space that a child may comfortably occupy, while accurate physiological, behavioural and/or neural data are collected.

Ecologically valid methods are not simply interesting, but may be necessary to accurately understand the limits and scope of our constructs of interest. Indeed, a burgeoning body of work based on wearable technologies suggests that findings from laboratory studies, particularly if computer-based, need not match their more ecologically valid counterparts for all individuals. This discrepancy has been highlighted in studies examining social attention, where static faces on a computer screen are quite disparate from 'real-life' people, whose faces are dynamic and contingently responsive (Risko et al., 2016). Hirsch et al. (2017) found unique patterns of brain activity via fNIRS when an individual gazed at a real-life social partner versus a picture of a face. They found greater activation in brain regions

associated with communication, including the pars opercularis, pre-motor cortex, supplementary motor cortex, and subcentral area for the face-to-face condition versus a face on a computer screen. Fu et al. (2019) found that behaviourally inhibited (BI) children, a temperamental profile at elevated risk for social anxiety (Chronis-Tuscano et al., 2009), were not differentiated from their non-BI counterparts on a computer task assessing attention biases to threatening faces. However, children high in BI had fewer gaze shifts to a stranger than non-BI children during a live interaction using mobile eye-tracking. Additionally, patterns of gaze on the computer task predicted the number of gaze shifts to the stranger only for BI children.

In infants, despite the traditional notion that attention is biased to human faces from infancy (Slater et al., 2010), Franchak et al. (2016) found that infants engaged in a naturalistic interaction actually spend very little time looking at their caregiver's face. The authors postulate that this discrepancy may be in part due to differences in posture in naturalistic settings versus laboratory tasks. In natural interactions, there may be a relatively high motor cost when a prone infant looks at their mother's face directly (Franchak et al., 2016). Laboratory paradigms, with standardized highchairs and ideal visual angles, often remove this cost. These discrepancies may also extend to non-social domains. Foulsham and Kingstone (2017) compared patterns of fixations to a naturalistic scene presented on a computer screen to patterns of fixations while adults walked through that same location with mobile eye-tracking glasses. Gaze patterns to the static scene on the screen did not predict gaze patterns during the walk. Rather, the best predictor of gaze during the walk was a model that accounted for where the eyes were positioned in an individual's head.

Inevitably, transitioning research into more naturalistic settings also introduces more noise into data collection. Moving testing into more naturalistic environments, like a participant's home, means that a litany of variables are no longer under the experimenter's control. For example, extraneous activity in the testing venue may distract the participant or interfere with the presentation of the task. Opportunities to minimize this noise include using 'lab-controlled' naturalistic paradigms in which an experimenter guides a game or interaction with a participant, so it is relatively replicable and consistent across all participants (e.g., Fu et al., 2019; Gunther, Brown, et al., 2021; Gunther, Fu, et al., 2021, Gunther et al., 2021). Traditional approaches of trial-by-trial data cleaning can also be employed post-hoc, cleaning data segments that are outliers in a measurement of interest to partially homogenize tasks across participants. For example, in our own work we examined eye-blink rate during wait periods of a modified Jenga game. Before analysis, we removed wait period 'trials' that were more than 2 standard deviations above or below the mean to standardize analysable data across participants as much as possible (Gunther, Fu, et al., 2021).

Even in the absence of imposed structure, we suggest that the signal:noise ratio is often still strong enough to robustly and reliably measure behaviour in well-designed studies that match the 'effect size' of the construct of interest to the selected environment and sample. A shift to naturalistic paradigms allows us to use a task that more closely resembles the construct of interest within the system it naturally operates, such as a familiar game for the infant to play or a parent–child interaction. We can then examine this scenario directly. With a shift to more naturalistic tasks, one can mitigate the error that may be accumulated in extrapolating the findings of a more sterile lab task to the broader construct and/or environment of interest.

Additionally, by using an environment that is more familiar or friendly to the infant, an experimenter is more likely to better capture the infant's typical behaviour, and less likely to capture the child's reaction to the novelty of the testing environment. This in turn, may bolster test–retest reliability. Finally, flexibility in testing space may enhance our ability to include historically underrepresented and excluded groups in research. By bringing research to a family's home, rather than asking a family to travel to a laboratory, we can mitigate barriers to participation such as time and cost, creating samples that are more representative of the general population. Prior work has been very successful with moving data collection out of the lab and into less traditional testing environments, coding parenting behaviours during video of a child's bedtime routine from cameras installed in the home as well as patterns of the infant's natural sleep using actigraphy (e.g., Jian & Teti, 2016), collecting data on an infant's home linguistic environment via wearable devices (Warlaumont et al., 2022), and visits to the home to video record infants' natural activity for subsequent behavioural coding (e.g., Herzberg et al., 2021), to name just a few examples.

The shift to more naturalistic tasks is not simple and may also involve redefining our notion of 'clean' data. Infant research has often utilized adapted thresholds for 'clean' data to take into account that infants may not always be the most patient research participants. While making these concessions for increased ecological validity, researchers may worry about additional noise imposed when paradigms are less structured. However, we believe that with any potential increased noise comes a boosted signal, and thus a stronger measurement of the construct of interest. Toggling back-and-forth between tight control and greater ecological validity will help us determine the relation between potential, or ideal, behaviour profiles, and the presence and form of that behaviour 'in the wild' (Pérez-Edgar & Hastings, 2018). This shift in view will make way for more accurate and reliable characterizations of human behaviour from early infancy, laying groundwork for a more thorough understanding of human development.

AUTHOR CONTRIBUTIONS

Kelley Gunther: Conceptualization; writing – original draft. **Berenice Anaya:** Conceptualization; writing – review and editing. **Koraly Pérez-Edgar:** Conceptualization; writing – review and editing.

FUNDING INFORMATION

The study was supported by grants from the National Science Foundation Graduate Research Fellowship under Grant No. DGE1255832 (to K. E. G.) and the National Institute of Neurological Disorders and Stroke under Grant No. 1F99NS120411-01 (to B. A.).

CONFLICT OF INTEREST

Authors have no conflicts of interest to report.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/icd.2338.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Kelley E. Gunther https://orcid.org/0000-0002-9964-7896

REFERENCES

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

Chronis-Tuscano, A., Degnan, K. A., Pine, D. S., Pérez-Edgar, K., Henderson, H. A., Diaz, Y., Raggi, V. L., & Fox, N. A. (2009). Stable early maternal report of behavioral inhibition predicts lifetime social anxiety disorder in adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*(9), 928–935. https://doi.org/10.1097/CHI.0b013e3181ae09df

Foulsham, T., & Kingstone, A. (2017). Are fixations in static natural scenes a useful predictor of attention in the real world? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *71*(2), 172–181. https://doi.org/10.1037/cep0000125

Franchak, J. M., Kretch, K. S., & Adolph, K. E. (2016). See and be seen: Infant-caregiver social looking during locomotor free play. *Developmental Science*, *21*(4), e12626. https://doi.org/10.1111/desc.12626

Franchak, J. M., & Yu, C. (2022). Beyond screen time: Using head-mounted eye tracking to study natural behavior. *Advances in child development and behavior*, *62*, 61–91. https://doi.org/10.1016/bs.acdb.2021.11.001

Fu, X., Nelson, E. E., Borge, M., Buss, K. A., & Pérez-Edgar, K. (2019). Stationary and ambulatory attention patterns are differentially associated with early temperamental risk for socioemotional problems: Preliminary evidence from a multimodal eye-tracking investigation. *Development and Psychopathology*, *31*(3), 971–988. https://doi.org/10.1017/S0954579419000427

Gunther, K. E., Brown, K. M., Fu, X., MacNeill, L., Jones, M., Ermanni, B., & Pérez-Edgar, K. (2021). Mobile eye tracking captures changes in attention over time during a naturalistic threat paradigm in behaviorally inhibited children. *Affective Science*, *2*, 495–505. https://doi.org/10.1007/s42761-021-00077-3

Gunther, K. E., Fu, X., MacNeill, L., Vallorani, A., Ermanni, B., & Pérez-Edgar, K. (2021). Profiles of naturalistic attentional trajectories associated with internalizing behaviors in school-age children: A mobile eye tracking study. *Research on Child and Adolescent Psychopathology*, *50*, 637–648. https://doi.org/10.1007/s10802-021-00881-2

Gunther, K. E., Fu, X., MacNeill, L. A., Jones, M., Ermanni, B., & Perez-Edgar, K. (2021). Now it's your turn!: Eye blink rate in a Jenga task modulated by interaction of task wait times, effortful control, and internalizing behaviors. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/nfq53

Herzberg, O., Fletcher, K. K., Schatz, J. L., Adolph, K. E., & Tamis-LeMonda, C. S. (2021). Infant exuberant object play at home: Immense amounts of time-distributed, variable practice. *Child Development*, *93*, 150–164. https://doi.org/10.1111/cdev.13669

Hirsch, J., Zhang, X., Noah, J. A., & Ono, Y. (2017). Frontal temporal and parietal systems synchronize within and across brains during live eye-to-eye contact. *NeuroImage*, *157*, 314–330. https://doi.org/10.1016/j.neuroimage.2017.06.018

Jian, N., & Teti, D. M. (2016). Emotional availability at bedtime, infant temperament, and infant sleep development from one to six months. *Sleep Medicine*, *23*, 49–58. https://doi.org/10.1016/j.sleep.2016.07.001

Lloyd-Fox, S., Papademetriou, M., Darboe, M. K., Everdell, N. L., Wegmuller, R., Prentice, A. M., Moore, S. E., & Elwell, C. E. (2014). Functional near infrared spectroscopy (fNIRS) to assess cognitive function in infants in rural Africa. *Scientific Reports*, *4*(1), 1–8. https://doi.org/10.1038/srep04740

Pérez-Edgar, K., & Hastings, P. D. (2018). Emotion development from an experimental and individual differences lens. In J. T. Wixted (Ed.), *The Stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 4, 4th ed., pp. 289–321). Wiley.

Richards, J. A., Xu, D., Gilkerson, J., Yapanel, U., Gray, S., & Paul, T. (2017). Automated assessment of child vocalization development using LENA. *Journal of Speech, Language, and Hearing Research*, *60*(7), 2047–2063.

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the fourth wall of cognitive science: Real-world social attention and the dual function of gaze. *Current Directions in Psychological Science*, *25*(1), 70–74. https://doi.org/10.1177/0963721415617806

Slater, A., Quinn, P. C., Kelly, D. J., Lee, K., Longmore, C. A., McDonald, P. R., & Pascalis, O. (2010). The shaping of the face space in early infancy: Becoming a native face processor. *Child Development Perspectives*, *4*(3), 205–211. https://doi.org/10.1111/j.1750-8606.2010.00147.x

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Warlaumont, A. S., Sobowale, K., & Fausey, C. M. (2022). Daylong mobile audio recordings reveal multitimescale dynamics in infants' vocal productions and auditory experiences. *Current Directions in Psychological Science*, *31*(1), 12–19. https://doi.org/10.1177/09637214211058166

How to cite this article: Gunther, K. E., Anaya, B., & Pérez-Edgar, K. (2022). Reducing measurement error with ecologically valid testing methods. *Infant and Child Development*, e2338. https://doi.org/10.1002/icd.2338