PAPER UNDER REVIEW APRIL 8, 2021

Six solutions for more reliable infant research

Krista Byers-Heinlein (Concordia University; k.byers@concordia.ca; ORCID: 0000-0002-7040-2510)

Christina Bergmann (Max Planck Institute for Psycholinguistics; Christina.Bergmann@mpi.nl)

Victoria Savalei (University of British Columbia; v.savalei@ubc.ca)

Data, materials, and code availability statement: All data and code used to create the tables, figures, and examples in this manuscript are available via the Open Science Framework at https://osf.io/e7j9k/

Funding statement: This work was supported by a grant to KBH from the Natural Sciences and Engineering Research Council of Canada (grant number 2018-04390).

Conflict of interest disclosure: The authors declare no conflicts of interest.

Acknowledgments: Thank you to the members of the Concordia Infant Research Lab as well as Angeline Sin Mei Tsui for their comments on earlier versions of this paper.

Abstract

Infant research is often underpowered, undermining the robustness and replicability of our findings. Improving the reliability of infant measures offers a solution for increasing statistical power independent of sample size. Here, we discuss two senses of the term reliability in the context of infant research: reliable (large) effects and reliable measures. We examine the circumstances under which effects are strongest and measures are most reliable, and provide simulations to illustrate the relationship between effect size, measurement reliability, and statistical power. We then present six concrete solutions for improving measurement in infant research: (1) routinely estimating and reporting the effect size and measurement reliability of infant tasks, (2) selecting the best measurement tool, (3) developing better infant paradigms, (4) collecting more data points per infant, (5) excluding unreliable data from analysis, and (6) conducting more sophisticated data analyses. Deeper consideration of measurement in infant research will improve our ability to study infant development.

Keywords: reliability, replicability, effect size, measurement, infancy, methodology

Highlights

- Reliable measures are those with large effect sizes (group-level studies) and/or with good measurement reliability (individual differences studies)
- Measurement reliability in infant research is seldom reported, and low in cases where it is assessed; observed effect sizes and resulting power are typically low
- Low reliability has concerning implications for conducting robust studies and drawing reliable conclusions
- We present six solutions for increasing effect sizes and measurement reliability, which can boost statistical power independent of sample size

Six solutions for more reliable infant research

Studying human behavior is difficult, particularly when those humans are tiny, squirmy, and don’t follow instructions (i.e., infants). Since the 1950, infant researchers have developed innovative measurement instruments that capitalize on infants’ natural repertoire of behaviors such as looking, reaching, and sucking. These have provided important insights into infant development (Aslin, 2014). Yet, infant researchers seldom consider the measurement properties of our research tools, even though the importance of accurate measurement has been understood by psychometricians for more than 100 years (Spearman, 1904). To be able to draw robust conclusions about infant development – including when theorizing, modelling, or designing studies and interventions – we need to account for our ability to measure it. This paper will overview the role of measurement in infant behavioral research – focusing on effect size and measurement reliability – and provide practical solutions for improving it.

Measurement in infant research

Infant researchers use carefully-designed experimental tasks to study constructs as diverse as attention, word learning, and theory of mind. The process of creating a number that represents a variable under study is called measurement (Flake & Fried, 2019). Making accurate measurements is hard. For example, in order to detect elusive subatomic particles, billions of dollars were spent to build the Large Hadron Collider (CERN, 2021). Although budgets are not as large, and the object of study is not as tiny, infant research too faces measurement challenges. Any measurement – be it of an infant or of a particle – is affected by measurement error. Measurement error is the difference between a true value and the measured value. Measurement error is assumed to be random, such that a measured value fluctuates around the true value. No measure can be totally precise, and different measures have different degrees of precision. The more precise the measure, the easier it is to detect the phenomenon of interest.

Measurement error reduces statistical power, defined as the probability of detecting a true effect (see Button et al., 2013). For example, imagine that a researcher wishes to measure children’s height, but the only instrument available involves stacking and counting homemade chocolate chip cookies. Cookie-based height measurement is likely to have substantial measurement error. Thus, it could be difficult to observe that 10-year-olds (on average 68 cookies tall) tend to be shorter than 11-year-olds (on average 69 cookies tall), or that the 10-year-olds that are the tallest in the class today would likely be the tallest in the class next year. Such results would not be impossible to observe even with this suboptimal measurement instrument, but the researcher would have to include a very large sample of children to detect such relatively small effects. If interest was in a larger effect, say whether 10-year-olds are taller than 1-year-olds (on average 38 cookies tall), or that individual children grow between the ages of 1 and 10 years, the researcher would easily have sufficient power even with a small sample. As these examples illustrate, there is an important relationship between measurement error, statistical power, and sample size.

Infant studies are often underpowered (Bergmann et al., 2018; Oakes, 2017), and the field has primarily focused on increasing sample size as a way to increase statistical power. Some examples are innovative recruitment methods for lab-based studies (Brand et al., 2019; Brouillard & Byers-Heinlein, 2019), testing infants in alternate settings such as in museums or online (Callanan, 2012; Scott & Shultz, 2017; Scott et al., 2017; Sheskin et al., 2020), and conducting large-scale multi-lab collaborations (ManyBabies Consortium, 2020; Byers-Heinlein et al., 2020). However, holding sample size constant, statistical power can also be increased by decreasing measurement error. Decreasing measurement error increases observed effect sizes and boosts measurement reliability, two key constructs that are discussed more fully in the next section. While infant researchers increasingly consider the role of effect size in experimental design and interpretation, much less attention has been paid to measurement reliability. We therefore set out to diagnose and suggest possible improvements to amend this state of affairs.

Reliable (large) effects versus reliable measures

The term “reliable” is sometimes used in a casual way, to describe a method that works well to answer a research question. However, what makes a measure good for answering a research question depends on whether the research takes a correlational versus an experimental approach (Hedge et al., 2018; Pérez-Edgar et al., 2020). The *correlational* approach is interested in individual differences, for example whether infants’ performance on two different tasks is related. In correlational research, the term “reliability” refers to measurement reliability, defined as the precision or the consistency of a measurement instrument when a measurement is repeated (Hedge et al., 2018). This corresponds to the sense of the word reliability used in the methodological and psychometrics literature. By contrast, the *experimental* approach asks questions at the group level, for example whether infants at a particular age have a particular ability. The methodological and psychometrics literature uses the term “effect size” to refer to this meaning, rather than the word “reliable”. As we will discuss further in this section, paradigms that have large effects do not necessarily have high measurement reliability. For consistency, the rest of this paper will use the terms “measurement reliability” and “effect size” to refer to these two distinct aspects of measurement. Note that issues of measurement are closely intertwined with recent discussions of replicability, although those discussions have largely focused on whether the underlying effect being measured is real and accurately described (e.g., Davis-Kean & Ellis, 2019).

Large effects. A large effect is one that is “shown by most participants in any study” and that “produces consistent effect sizes” (Hedge et al., 2018). Effect size can be quantified via standardized effect size metrics such as Cohen’s *d*. For a within-subjects design, this is calculated as the ratio of the mean difference to the standard deviation of the mean difference:

$$d = M/SD$$

A higher Cohen's *d* corresponds to a stronger effect such that, all else being equal, a higher Cohen's *d* will result in greater statistical power. Effect sizes can be computed based on group-level information usually reported in papers (i.e., sample size, together with either means/standard deviations or test statistics such as *t*-values). Note that differences in observed effect size stem from two possible sources of variation, which both contribute to the observed standard deviation (the denominator of the formula): differences in the underlying true effect and differences in measurement error (or both). It is not possible to determine which source of variation is greater from the observed effect size alone.

A recent meta-meta-analysis of a variety of topics in infant research found wide variability among effect sizes, ranging from .12 – 1.24 with a median of .45 (Bergmann et al., 2018). Thus, the effect size of infant research methods varies significantly by domain, meaning that some group-level phenomena can be detected more readily (i.e., with more statistical power) than others.

Reliable measures. A reliable measure is defined as one that "consistently ranks individuals" (Hedge et al., 2018). Measurement reliability is quantified by metrics such as r_{xx}, defined as the ratio of true score variance to observed score variance (which is the sum of true score variance and measurement error):

$$r_{xx} = \mathrm{var}_T / \mathrm{var}_O$$

Measurement reliability can only be assessed when infants contribute two or more measures of the same construct, either during the same testing session (e.g., multiple trials of the same type) or during different testing sessions. Unlike effect size, it is not possible to estimate reliability if individuals only contribute a single score.

Unfortunately, measurement reliability statistics are seldom reported in the infant literature – an important point we will return to in a later section. Extant investigations have reported measurement reliability that varied widely across studies and tasks (speech perception tasks; Cristia et al., 2014), or measurement reliability close to zero (visual preference procedures; DeBolt et al., 2020; Nighbor et al., 2017), although some earlier work reported moderate measurement reliability (infant attentional measurements; Colombo et al., 1988). Note that in nearly all cases, reported measurement reliability was less than .5, which is generally considered poor (Koo & Mae, 2016).

Generalizability of effect sizes and measurement reliability

Note that estimates of effect size and measurement reliability relate to the measurement of a particular sample under particular circumstances. Thus, we would not expect values to be identical for infants of different backgrounds or ages, or those tested in different contexts (e.g., in the lab versus online at home), even when tested in the same apparent task. For example, we might expect infants of different ages to show different effect sizes on the same task, for example older infants to perform better than younger infants. Similarly, a sample of 18-month-old infants might have very similar abilities (true scores) on a task, whereas a sample of 9- to 18-month-old infants might have a wide range of abilities. That is, assuming that measurement error remains constant across age, estimated reliability will be higher for the group with the wide age range than the group with the narrow age range, as the measure can more consistently rank the infants with more varying abilities than infants with more similar abilities. The more similar two studies are with respect to their methods and the population tested, the more similar we expect their effect sizes and measurement reliabilities to be.

A simulation study of effect size and measurement reliability

Perhaps surprisingly, paradigms that produce the largest effect sizes are not necessarily the ones with the highest measurement reliability (Hedge et al., 2018). This is because observed effect sizes are largest when true between-participant variability is low (i.e., infants all have the same underlying ability), while measures are most reliable when true between-participant variability is high (i.e., infants have a range of different underlying abilities). For infant research, this means that the methods that are optimal for producing the largest group-level effects may be different from the ones that are optimal for detecting individual differences.

We illustrate this apparent paradox via four simulated datasets, which we could imagine arising from a set of different studies analyzing infants' looking time difference scores (e.g., looking to experimental trials minus looking to control trials). Recall that variability in individual scores arises from two distinct sources: true score variability (i.e., real underlying differences between infants) and measurement error (which is by definition random, meaning it does not systematically bias scores and averages out to zero). The datasets cross these two sources of variability, such that true score variability and measurement error were either low (SD = .5) or high (SD = 1). The mean of participants' true difference scores was set at 1 for all datasets. Based on these parameters, we calculated observed effect sizes (Cohen's *d*) and measurement reliability (r_{xx}) for each dataset[1]. The code used to generate all Tables and Figures presented in this article is available via the Open Science Framework at https://osf.io/e7j9k/

To make this example more concrete, we can imagine that infants have either been tested in a quiet laboratory (small measurement error) or in a noisy community center full of distractions (large measurement error). Moreover, we compare two types of samples: infants sampled within a narrow age range (low true variability), and infants sampled across a wide age range (high true variability). We observe that each group has an average one-second looking time difference to an experimental stimulus compared to a control stimulus.

Figure 1 plots infants' true scores (left side of each panel) and their observed scores which include measurement error (right side of each panel) for a simulated 50 infants per

[1] Specifically, we first calculated the total variance by summing true score variance and measurement variance (themselves calculated by squaring their respective standard deviations). Taking the square root of this value, we arrived at the observed standard deviation, which we used to calculate Cohen's *d* = mean/SD. r_{xx} is calculated by dividing the true score variance by the total variance. The reported values reflect these calculations, rather than the values from the plotted infants, which are used for illustrative purposes only.

group. Observed means, standard deviations, observed effect size (Cohen's *d*) and measurement reliability (r_{xx}) are indicated at the bottom of each panel. Note that the true (latent) effect size is *d* = 1 in panels 1A and 1B and *d* = 2 in panels 1C and 1D.

From Figure 1, we can make several observations about the interplay between effect size and measurement reliability. First, measurement reliability is highest when true variability is high and measurement error is low (panel 1B). By contrast, the observed effect size is largest when both true variability and measurement error are low (panel 1D). Thus, although reducing measurement error boosts both effect sizes and measurement reliability, greater true score variability yields higher measurement reliability but smaller effect sizes. Finally, observed effect size is affected by total variability, but agnostic to whether this variability is due to true score variability or measurement error (e.g., panels 1B and 1C have identical values of *d*).

A key take-home message is that a method can produce a large effect size but have low measurement reliability, or conversely produce a small effect size but have high measurement reliability. Those methods that yield large effects will be powerful for detecting group differences, while those with high reliability will be powerful for detecting individual differences. Without measuring *both* the effect size and reliability of our measures, we cannot know which infant measures are suited to which research purposes.

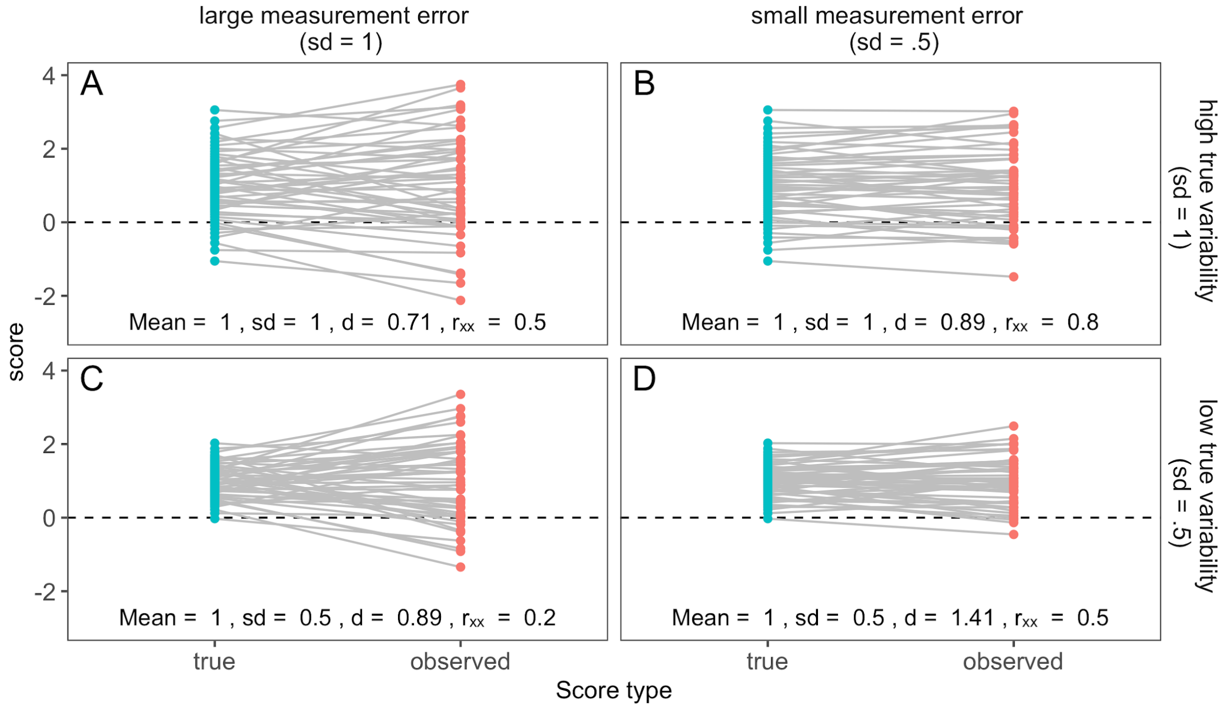

Figure 1: True and observed scores for four hypothetical datasets, under conditions of high/low true variability, and large/small measurement error. N = 50 points from simulated datasets are plotted to illustrate. Expected means, standard deviations (*sd*), observed Cohen's *d*, and measurement reliability (r_{xx}) are shown. True (latent) Cohen's *d* is 1 in panels 1A and 1B, and 2 in panels 1C and 1D.

The problem with small effect sizes and low measurement reliability

At this point, most infant researchers are aware that in experimental studies, using tasks that produce small effect sizes will result in low statistical power at typical sample sizes. Table 1 illustrates the relationship between sample size and effect size to achieve 80% power using a two-tailed paired t-test, α = .05. Values were calculated using the *pwr* package in R (Champley, 2020). Larger effect sizes sharply reduce the sample size needed to achieve sufficient statistical power. In fact, when measuring large effects, required sample sizes are quite reasonable for a paired samples *t*-test. Note that well-powered samples will differ for other statistical tests, and will typically be larger for example in between-group comparisons or for interactions.

Table 1: Relationship between observed effect size (Cohen's *d*) and sample size (N) to achieve 80% power in a two-tailed, paired samples *t*-test, α = .05.

Effect size (Cohen's d)	**N**
1.0	17
.8	26
.6	45
.4	99
.2	393

What is often less understood by infant researchers is that low measurement reliability leads to low statistical power in correlational studies. This is because in correlational studies, statistical power depends on the measurement reliabilities of both of the constructs being measured (Trafimow, 2005). Researchers often think about the true correlation that they expect between their variables. However, due to a statistical phenomenon called attenuation of correlation, the observed correlation will always be weaker than the true correlation unless measurement reliability is perfect[2] (Spearman, 1904). Table 2 (adapted from Hedge et al., 2018) illustrates the relationship between measurement reliability **(**r_{xx}**)**, the true correlation between two variables (true *r*), the observed correlation between two variables (observable *r*), and the sample size necessary to achieve 80% power in a Pearson's correlation test. For simplicity, we assume that the two variables that will be correlated have the same measurement reliability (r_{xx}). As before,

[2] Mathematically, the observed correlation is the true correlation times the product of the square root of each measure's reliability, following the formula $r_{observed} = r_{true}$ sqrt(r_{xx} * r_{yy}) (Spearman, 1904).

improving measurement reliability can decrease the sample size necessary to achieve a particular level of statistical power, or can improve power at an identical sample size. Conversely, combinations of low measurement reliability, low true correlation, and/or small sample size will result in low statistical power.

Table 2. Relationship between measurement reliability, true correlation, and observable correlation between two variables with 80% power (alpha = .05). Adapted from Hedge, Powell, & Sumner (2018), Table 5.

Reliability of measurement (*r*xx)	**True *r***	**Observable *r***	**N**
1.0	.7	.7	13
1.0	.3	.3	84
.6	.7	.42	41
.6	.3	.18	239
.2	.7	.14	397
.2	.3	.06	2117

Given that, in many cases, the measurement reliability of infant behavioral experimental tasks may be low, it is crucial to consider how observed correlations in infant research should be interpreted. For example, imagine a researcher tests infants' performance on a task at age 18 months and again at age 24 months. Following Table 2, even if the true correlation of infants' abilities at the two timepoints is .7 (so performance is reasonably stable), if the measurement reliability of the task is low (.2), the sample must include 397 infants to detect the observable correlation with 80% power. In fact, due to attenuation of correlation, the observable correlation is *r* = .14, even though the true correlation is *r* = .7. The same issue arises for correlations computed on concurrent measurements, for example correlating task performance and vocabulary size. When measurement reliability is low, observing a small correlation could be due to a small true correlation, and/or low measurement reliability of one or both measures. When a measure is unreliable and a large correlation is observed, then this is likely due to chance, rather than reflecting the true underlying relationship.

In sum, low measurement reliability makes it difficult to detect true effects in correlational studies, just as small effect sizes make it difficult to detect true effects in experimental studies. In the next section, we review six practical solutions for reducing measurement error in infant research, thereby increasing effect size and measurement reliability.

Solutions for increasing effect size and measurement reliability of infant research

Solution 1: Routinely report effect size and measurement reliability

To improve the robustness of our research, infant researchers must begin by determining the effect sizes and measurement reliability of existing methods. Fortunately, effect size estimates are largely available from the infant literature, either because they have been included in published reports (which is increasingly standard practice) or because they can be readily computed from available information (i.e., means and standard deviations or exact test statistics). However, similar to other fields that use behavioral tasks (Parsons et al., 2019), measurement reliability estimates are seldom reported in infant research. Moreover, it is usually impossible to estimate measurement reliability from the information reported in published papers.

There are multiple approaches to estimating measurement reliability that might be appropriate for infant research, and here we provide a brief overview. Measurement reliability can be estimated any time infants provide two or more data points for the same measure. Note that much of the psychometric literature on reliability discusses reliability across different raters/judges. In infant research, the raters/judges can be thought of as different trials of the same type within a single testing session (e.g., several different preference trials from the same experimental condition, or multiple difference scores across trial pairs), or different testing sessions using the same task (i.e., test–retest reliability).

To compute measurement reliability from two data points (e.g., from two different testing sessions), researchers can simply compute Pearson's *r* using infants' scores across the two sessions. Simple correlations can be computed using the *cor.test* function in R, which is available by default (R Core Team, 2020). Mathematically, correlations can take values from -1 to +1, although when computing measurement reliability, we expect values from 0 (no reliability) to 1 (perfect reliability). Negative values imply that individuals who did better on one assessment did worse on the other, and are usually observed due to low measurement reliability coupled with sampling error.

To compute measurement reliability from multiple data points (e.g., two or more different trials of the same type), researchers can compute the Intraclass Correlation Coefficient (ICC), for example using the *psych* package in R (Revelle, 2018; see also Parsons et al., 2019). The ICC ranges from 0 to 1, with higher values representing better measurement reliability. The ICC has several different variants, and researchers will need to take four considerations into account in selecting the most appropriate one (Koo & Mae, 2016). The first is whether all participants encountered the same items, which will most often be the case in infant behavioral research. The second is whether the researcher wishes to generalize beyond the specific items tested (i.e., fixed versus random effects), which will usually be the case in infant experiments. The third is whether the researcher is interested in consistency (i.e., the degree to which participants are in the same rank order across timepoints) or in absolute agreement (i.e., the degree to which participants have the same exact scores across timepoints). The fourth is related to how the measurement will take place in the future, and usually depends on whether researchers are comparing across multiple testing sessions (single rater type), or across trials within a single testing session (multiple raters type).

For the bulk of cases, where researchers use the same materials for all participants and wish to generalize beyond their particular stimulus set, the ICC should be calculated using a two-way random effects model. Infant researchers will often be more interested in consistency than absolute agreement, given that absolute scores often vary due to uninteresting factors such as item salience, practice effects, fatigue, etc. In this most common case, infant researchers should use the single measures variant (ICC3 in the *psych* package) when computing ICC across multiple testing sessions, and should use the multiple measures variant (ICC3k in the *psych* package)[3] when computing ICC within the same testing session. While this recommendation will be appropriate in many or most cases, the choice of which ICC variant(s) to report must be informed by the researcher's specific experimental design and research goals. We refer readers to Koo and Mae (2016) and Parsons et. al (2019) for more detailed guidance.

As an example of how to compute the ICC using the *psych* package, we provide sample code that analyzes open data from ManyBabies 1 (ManyBabies Consortium, 2020) available at https://osf.io/e7j9k/. The obtained ICC value might be reported as follows:

> Reliability of the looking time difference to the infant-directed speech (IDS) stimuli versus the adult-directed speech (ADS) stimuli across the 8 trial pairs was estimated with an intraclass correlation coefficient (ICC), based on a mean-rating (k = 8), consistency, 2-way random-effects model (ICC3k) using the *psych* package in R (Reville 2018). The estimated consistency was .14, 95% CI = [.09, .18].

Even after carefully considering which ICC variant to compute, researchers will need to be thoughtful in interpreting estimates of measurement reliability. First, it is important to note that the meaning of measurement reliability estimated from infant data will depend on the timeframe across which the different measurements were taken. For infants, we might not expect a lot of change in infants' true abilities if the measures are taken within the same testing session or only a week apart, but we might expect a large change if the two measures are taken a year apart. In the first case, the measurement reliability estimate would reflect the dependability of the measures, and in the second case it would reflect the stability of the trait (Hussey & Hughes, 2018).

Second, as illustrated in the previous section, measures can have large effect sizes without having reliable measurements, and vice-versa. Group-level studies should not be criticized on the basis of low measurement reliability, just as individual differences studies should not be criticized on the basis of small effect sizes. However, in the context of studies where both group-level and individual differences are examined (e.g., a researcher compares groups of infants from different backgrounds, and also tests whether performance is correlated with vocabulary size), a careful examination of both effect sizes and measurement reliability is necessary in interpreting the observed pattern of results.

[3] Note that, with complete data, ICC3k is equivalent to Cronbach's alpha, which can also be computed using the *alpha* function in the *psych* package in R (Revelle, 2018).

Because reporting measurement reliability has not been standard in infant research, many infant researchers lack the necessary training on how to determine measurement reliability, or examples of how to report such information in their papers (for a recent example of an infant study that did report measurement reliability see Egger et al., 2020). We hope that the information provided here will help infant researchers to embrace a standard practice of computing and reporting the measurement reliability of infant measures. Even when less relevant to a particular study's goals, reporting measurement reliability is useful for guiding the design of future studies (e.g., to determine whether an experimental paradigm is suitable for studying individual differences). Sharing trial-level data is also beneficial, as it enables other researchers to compute different metrics of measurement reliability. In studies where it is not possible to estimate reliability (for example, in habituation tasks where there is a single critical test trial), researchers can simply state that, to their knowledge, there is no procedure to estimate the reliability of the measure (Parsons et al., 2019). Researchers planning longitudinal studies should consider including the same measure at multiple timepoints in order to estimate the test–retest reliability of their measures, especially in tasks where it is not possible to compute measurement reliability from a single testing session. Routine reporting of both effect sizes and measurement reliability will go a long way to improving the robustness of infant research.

Solution 2: Select the best measurement tool

Researchers in many branches of psychology routinely aim for measurement tools with high validity and measurement reliability, while balancing other concerns such as ease of administration. By contrast, infant researchers often make methodological decisions based on historical convention, rules of thumb, and standard laboratory practices, rather than based on known psychometric properties of our methods (DeBolt et al., 2020; Eason, Hamlin, & Sommerville, 2017; Oakes, 2017). Although there are many factors that have contributed to this state of affairs (the difficulty of testing infant participants being especially salient), one important factor is that researchers do not have the necessary information about effect sizes or measurement reliability, as these have not always been measured or reported in the literature.

Recent efforts have begun to systematically gather information about the effect sizes of different infant tasks, making this information much more accessible than before. For example, MetaLab (https://metalab.stanford.edu) is an aggregation platform for meta-analyses of infant cognitive and language research. At the time of writing the database contained information from 30 meta-analyses, many of which are coded for moderators such as age and methodological factors. Thus far, MetaLab focuses on measures of observed effect size (e.g. Cohen's d) because this information can typically be extracted from published papers. Researchers can look up the expected average effect size of commonly used infant paradigms in the published literature, keeping in mind that estimates may be inflated due to publication bias, and that moderators such as age and methodologies are not randomly assigned.

Unfortunately, MetaLab does not (yet) include information about the measurement reliability of infant methods, as this would require papers to either report reliability statistics (which is extremely rare), or to provide trial-level data (which is often unavailable, although is becoming more common). Until measurement reliability estimates are available in a central repository, researchers will need to compute the internal consistency of comparable measures in existing datasets from their own or other labs (for an example of this approach, see DeBolt et al., 2020), using the approaches described in the previous section.

Large-scale collaborations are also beginning to provide information about the effect size and measurement reliability of infant paradigms. For example, labs participating in ManyBabies 1 (which tested infants' preference for infant-directed speech over adult-directed speech in a looking time paradigm) were free to use one of three common infant methods (ManyBabies Consortium, 2020). The observed effect size was larger for labs that tested using headturn-preference than those that used central fixation or eye-tracking, even controlling for factors such as infants' language background and age (see Table 3). An in-progress pre-registered study is examining the measurement reliability of the ManyBabies 1 task using a test–retest approach (Schreiner et al., 2020). In line with other reports (e.g. Cristia et al., 2016), overall measurement reliability was low, although reliability was higher when the analysis was limited to infants who contributed more valid test trials.

Without adequate effect sizes (for group-level studies) or measurement reliability (for individual-differences studies), infant research is "bound to fail" (Rouder et al., 2019). Where it is available, it is crucial that researchers use information about effect size and measurement reliability in guiding infant study design and interpretation. Where it is not available, the field may wish to devote resources towards measuring the reliability of common paradigms. Nonetheless, once both effect size and measurement reliability estimates are more regularly reported in the literature, they will provide important guidance for researchers designing studies.

Solution 3: Develop better infant paradigms

As the field begins to understand the measurement properties of our current paradigms, we may find that some areas of research lack paradigms with acceptable effect sizes for group-level studies, and/or measurement reliability for individual differences studies. It is also possible that some paradigms produce stable individual differences but weak group-level results, or vice-versa. A fuller understanding of the measurement properties of our current methods can serve as a guide for research areas ripe for methodological innovation, pointing to where the field most needs to develop better infant paradigms. Paradigms with large effect sizes and strong measurement reliability will both be needed.

As an example Houston et al. (2007) sought to develop a task that would allow reliable assessment of speech discrimination in individual infants, which could be useful for clinical diagnosis. However, as Houston et al. pointed out, existing tasks had been developed to maximize effect size, rather than measurement reliability. Houston et al. developed three variants of a visual habituation procedure, and used a test–retest approach to assess measurement reliability, and identified one particular variant that had higher reliability than the others. Note that only 10 infants were tested per variant, making these specific results highly preliminary (see also de Klerk et al., 2019, for a replication study that reported a much reduced effect size, and Schott et al., 2019, for a discussion of why results from small-scale pilot studies can be misleading). Nonetheless, this paper provides a nice example of how infant researchers can think about the development of infant procedures with better measurement reliability.

Work directly aimed at improving infant behavioral methods is complemented by other methodologically-related research. For example Santolin et al. (2020) recently reported evidence that infants' experience with a paradigm is related to the direction of preference they show (i.e. whether infants look more to novel or familiar stimuli). As another example, ManyBabies 5 is conducting a large-scale collaborative study aimed at understanding the processes that underlie looking time, which could be beneficial for designing looking time experiments with larger effect sizes and/or better measurement reliability. In general, research that addresses methodological questions directly could yield a large return on investment, as such results could be used to inform many subsequent studies and potential clinical assessments.

Solution 4: Collect more data points per infant

For more than a century, psychometricians have known that, in most cases, a "longer" test (one with more items) will produce a more reliable score (Symonds, 1928; Spearman, 1910; Brown, 1910). This relatively simple solution – collecting more data points per infant – has the potential to reduce measurement error, thereby increasing effect size and measurement reliability.

To examine how presenting infants with more trials could affect statistical power, DeBolt et al. (2020) conducted a series of simulations based on data from studies that used preference procedures, where infants' relative looking time to two images was measured. Across five datasets, they observed variability in the effect size of the tasks, but near-zero measurement reliability. Their simulation demonstrated that, in such cases, adding new trials from the same infants can increase power for detecting group-level effects just as much as increasing sample size. Moreover, the quality of the data did not appear to decrease over time. In another example, Houston et al. (2007) reported higher measurement reliability from a paradigm that presented infants with more test trials for analysis than in paradigms that presented fewer trials, which increased the power to detect individual differences (for additional discussion, see Cristia, 2016; de Klerk et al., 2019).

There are other approaches that could increase the number of analyzed trials per infant without increasing the number of trials that infants encounter. For example, Egger et al. (2020) created a gaze-triggered looking-while-listening paradigm where the target (i.e. the object that was labeled on a particular trial) depended on infants' fixation. This approach provided more trials from which to derive a reaction time score, which in this paradigm crucially depends on whether the infant was looking to the distractor at the moment of hearing the target word. Another approach could be to adapt experiments in ways that enable infants to complete more trials, for example by using varied attention getters between trials, short filler movies, pauses, and so on. The feasibility of these different strategies will depend on the study type and might warrant their own line of research to be able to make an informed choice as to how to increase experiment duration without compromising data quality.

Certainly, not every type of research question or experimental design will be amenable to increasing the number of analyzed trials per infant. Moreover, there may be limits to this approach as infants become overly fatigued or fussy. However, in many cases adding additional trials or adapting experiments so that more existing trials can be analyzed is a low-effort option for increasing measurement reliability and experimental power.

Solution 5: Exclude low quality data from analysis

Infant researchers have a long history of systematically excluding subsets of their data that are considered to be of low quality, for example excluding trials with very short looking times, or infants who only contribute a small number of trials. The intuition is that doing so removes data where infants are “off-task”.

What is the relationship between infant exclusions, effect size, and measurement reliability? ManyBabies 1 addressed this question by applying different infant-level exclusion criteria to their data in a set of exploratory analyses. Infants participated in up to 16 experimental trials, and effect sizes were calculated when including infants who contributed 2 or more, 4 or more, or 8 or more useable trials. As shown in Table 3, stricter exclusion criteria yielded larger effect sizes. For example, in eye tracking (the method that showed the most striking pattern), including infants with as few as two trials (one per condition) yielded an effect size of *d* = .24, while a stricter criterion of including infants with at least 8 trials nearly doubled the effect size *d* = .41. At the same time, stricter criteria decreased the effective sample size, as more infants were excluded from analysis. Again, eye tracking showed the most striking pattern, with 85% of infants included with the loosest criterion, but only 36% of infants included with the strictest criterion. While ManyBabies 1 focused on the role of infant-level exclusion on effect sizes, future explorations of this dataset can also investigate the effect of data exclusions on measurement reliability within the same session. For example, the ongoing assessment of test–retest reliability by Schreiner and colleagues (2020) also suggests that limiting the analysis to infants that contributed more trials could substantially improve measurement reliability.

Table 3. ManyBabies 1 effect sizes (d), percentage of included participants (% included), number of participants needed to test prior to exclusions (N needed - tested), and the number ultimately analyzed (N needed - analyzed) to yield 80% power under a single-samples *t*-test applying different exclusion criteria (Min # trials). Table adapted from Table 6 in ManyBabies Consortium (2020).

Min # trials	Effect size (Cohen's d)	% included	N needed - tested	N needed - analyzed
Central Fixation				
2	.29	98	191	188
4	.34	88	155	137
8	.40	73	136	99
Eyetracking				
2	.24	85	322	273
4	.33	59	246	145
8	.41	36	262	94
Headturn Preference Procedure				
2	.51	98	63	61
4	.53	92	62	57
8	.63	78	52	41

Infant exclusions can increase effect size (which increases power), but they also decrease the size of the sample available for analysis (which decreases power). What is the tradeoff between these two factors? Is it better to use a stricter criterion with a smaller analyzed sample, or use a looser criterion with a larger analyzed sample? We again used the data from ManyBabies 1 to explore this question, by calculating how many total infants would need to be tested to achieve 80% power using the different exclusion criteria. Table 3 indicates the number of infants that would need to be tested, and the number that would be analyzed under different exclusion criteria.

For the headturn preference procedure, the optimal strategy would be to use the strictest criterion: only 52 infants would need to be tested (of which 41 would be analyzed) compared to the loosest criterion whereby 63 infants would need to be tested (of which nearly all – 61 – would be analyzed). Similarly, for central fixation, the optimal strategy is to use the strictest criterion, which necessitates testing 136 infants to analyze 99. In contrast, for eye-tracking, the intermediate strategy of a 4 trial minimum appears optimal, requiring testing 246 infants to include 145 infants in the final analysis, as opposed to the strictest criterion which would require testing 262 infants to include 94 infants in the final analysis.

Overall, this example demonstrates an interesting interplay between inclusion criterion and experimental power, due to different effect sizes. Different strategies might be optimal for different paradigms, depending on the tradeoff between gains in effect size versus loss of participant numbers when stricter inclusion criteria are implemented (see also Pour Iliaei et al., 2020, for another example where stricter inclusion criterion yielded more robust results). Note that for previous studies on infant-directed speech, when stated, inclusion criteria were much stricter than the strictest criterion assessed in ManyBabies 1 (50%), often requiring infants to complete 100% of trials (e.g. Fernald, 1987, Inoue et al, 2011). It is an open question whether additional gains can be made in applying even stricter inclusion criteria than those explored here.

In sum, optimizing approaches to data exclusion can increase observed effect size, without necessarily requiring testing more infants in total to achieve appropriate statistical power. Note that to avoid *p*-hacking, plans for data exclusion should be pre-registered (see Havron et al., 2020). At the same time, transparent exploration of the effects of different exclusion criteria, even if not pre-registered, could provide researchers with guidance in developing data exclusion plans for future studies.

Solution 6: Conduct more sophisticated statistical analyses

Infant behavioral research has historically relied on analytic techniques such as *t*-tests and ANOVAs, which collapse responses across time and across trials to yield one or two data points per infant for analysis. Indeed, the examples in this paper so far have been within this framework. However, in its raw state, infant data is considerably richer than what is often analyzed, with infants contributing data points on multiple experimental trials, and/or fine-grained data within trials such as looking patterns over time.

In traditional analytic techniques, variation over time or trials ends up lumped together, and is attributed to measurement error. However, infant researchers are aware, at least implicitly, of systematic sources of variance hidden within these data. For example in many studies, infants tend to habituate over time, such that overall attention decreases across trials. As another example, some test items might be more difficult for infants than others.

Using more sophisticated analytic techniques, it is possible to directly model these known sources of variance, so that the focal sources of variance (e.g., experimental manipulations, age effects) can be more precisely quantified (Gelman, 2006). Approaches such as mixed-effects models can take into account individual differences across participants or items (random effects), as well as fixed effects such as linear increases or decreases in performance across trials. With more accessible software packages, better computing power, and advanced statistical techniques, it may be possible to do more with the data we have. Below, we illustrate with three examples.

As a first example, a recent paper (Pour Iliaei et al., 2020) investigated cognitive differences between 7-month-old monolingual and bilingual infants based on the seminal work of Kovács & Mehler (2009). On nine training trials, infants saw a central cue followed by a reward on the left side of the screen (counterbalanced) and on nine test trials the reward switched sides such that it appeared on the right side of the screen. In the original paper, data had been averaged across 3-trial blocks to yield up to six data points per infant. However, Pour Iliaei et al. developed a new analytic technique that modeled the change in performance on trials over time, as well as the slope of infants' looking within trials, to yield up to 80 data points per infant. Comparing the old and new techniques using original data as well as several other open datasets, the paper reported robust differences between monolinguals and bilinguals only with the more sensitive analysis, and not with the traditional one (see also Humphrey & Swingley (2018) for a simulation of a similar approach with infant data).

As a second example, de Klerk et al. (2019) tested infants in a discrimination task. Infant saw a series of alternating (fap-fep) and non-alternating (fap-fap) trials. At the group level, infants at 6, 8, and 10 months old clearly discriminated the contrast. The authors wished to determine which infants discriminated the contrast at an individual level. Following the individual-level regression procedure developed by Houston et al. (2007), they found very limited evidence for individual-level discrimination. However, using a Bayesian Hierarchical modeling approach, which incorporated information from each age group to inform the model for each individual infant, they found evidence for individual-level discrimination in 77% of 10-month-olds, 53% of 6-month-olds, and 27% of 8-month-olds.

As a final example, van Renswoude et al. (2017) noted that typical eye-tracking software detects fixations and saccades using algorithms that are optimized for adults, which do not consider individual differences in eye movements. They developed a software tool called "GazePath" that takes individual behavior at the trial level into account, and interpolates missing data. Across several different infant and adult datasets, the researchers demonstrated the efficacy of their method for picking up on small eye movements that traditional algorithms missed, as well as for processing noisy infant data.

These three examples illustrate the diversity of ways that advanced statistical and computational techniques, particularly ones that model data at a fine level of granularity, can in some cases better separate signal from noise, thus making data from extant infant paradigms more informative. The evidence presented here about these particular analytic approaches is anecdotal, and methodological research will be needed to better understand which statistical approaches maximize statistical power in the context of infant research. Nonetheless, we hope that particularly with the rise of open data, analyses such as the ones outlined here can further showcase the use of specific modelling techniques for infant data and allow researchers to build on previous work when planning and pre-registering their own analyses.

Conclusion

Infant research can benefit from carefully considering the properties of our measures. This paper has distinguished two important types of reliability: whether an *effect* is reliable (shows a large group-level effect size) and whether a *measure* is reliable (consistently measures individuals). Effect size and measurement reliability may be optimized under different conditions, and researchers should be aware of which one is most relevant to their research question. Effect size is important for studies looking at group-level effects, whereas measurement reliability is important for studies looking at individual differences. Here, we have illustrated six ways that infant researchers can improve measurement at each step of the research process: routinely reporting effect size and reliability statistics, selecting the best measurement tool, developing improved paradigms, collecting more data points per infant, excluding low quality data from analysis, and applying more sophisticated analytic techniques.

There are multiple considerations as we embark on this work. First, improving effect sizes and measurement reliability of infant research must go hand-in-hand with careful consideration of measurement and ecological validity (Kominsky et al., 2020). Second, developmental changes over time, as well as cross-population differences, could impact both the effect sizes and measurement reliabilities of our paradigms. Finally, we must guard against undisclosed flexibility in research, which can undermine our best efforts (Davis-Kean & Ellis, 2019). A more concerted consideration of measurement in infant research has the potential to increase experimental power independently of increases in sample size, and will ultimately yield a more robust and replicable science of infant development.

References

Aslin, R. N. (2014). Infant learning: Historical, conceptual, and methodological challenges. Infancy, 19(1), 2-27. https://doi.org/10.1111/infa.12036

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. Child Development, 89(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Brand, R. J., Gans, R. T., Himes, M. M., & Libster, N. R. (2019). Playdates: A win-win-win strategy for recruitment of infant participants. Infancy, 24(1), 110-115. https://doi.org/10.1111/infa.12269

Brouillard, M., & Byers-Heinlein, K. (2019). Recruiting hard-to-find participants using Facebook sponsored posts. Retrieved from https://osf.io/9bckn/

Brown, W. (1910). Some experimental results in the correlation of mental abilities. British Journal of Psychology, 3, 296–322.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365-376. https://doi.org/10.1038/nrn3475

Byers-Heinlein, K., Tsui, R. K. Y., van Renswoude, D., Black, A. K., Barr, R., Brown, A., ... & Singh, L. (2020). The development of gaze following in monolingual and bilingual infants: A multi-laboratory study. Infancy. https://doi.org/10.1111/infa.12360

Callanan, M. A. (2012). Conducting cognitive developmental research in museums: Theoretical issues and practical considerations. Journal of Cognition and Development, 13(2), 137-151. https://doi.org/10.1080/15248372.2012.666730

Champely, S. (2020). pwr: Basic Functions for Power Analysis. R package version 1.3-0. https://CRAN.R-project.org/package=pwr

CERN. Facts and figures about the LHC. https://home.cern/resources/faqs/facts-and-figures-about-lhc Retrived April 1, 2021.

Colombo, J., Mitchell, D. W., & Horowitz, F. D. (1988). Infant visual attention in the paired-comparison paradigm: Test-retest and attention-performance relations. Child Development, 59(5), 1198-1210. https://doi.org/10.2307/1130483

Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. Child Development, 85(4), 1330-1345. https://doi.org/10.1111/cdev.12193

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. Infancy, 21 (5), 648-667. https://doi.org/0.1111/infa.12127

Davis-Kean, P. E., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. Infant Behavior and Development, 57, 101339. https://doi.org/10.1016/j.infbeh.2019.101339

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. https://doi.org/10.1111/infa.12337

Eason, A. E., Hamlin, J. K., & Sommerville, J. A. (2017). A survey of common practices in infancy research: Description of policies, consistency across and within labs, and suggestions for improvements. Infancy, 22(4), 470-491. https://doi.org/10.1111/infa.12183

Egger, J., Rowland, C.F. & Bergmann, C. (2020). Improving the robustness of infant lexical processing speed measures. Behavior Research Methods, 52, 2188–2201 (2020). https://doi.org/10.3758/s13428-020-01385-5

Fernald, A., & Kuhl, P. K. (1987). Acoustic determinants of infant preference for motherese speech. Infant Behavior & Development, 10(3), 279–293. https://doi.org/10.1016/0163-6383(87)90017-8

Flake, J. K., & Fried, E. I. (2019, January 17). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. https://doi.org/10.31234/osf.io/hs7wm

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do Technometrics, 48(3), 432-435. https://doi.org/10.1198/004017005000000661

Havron, N., Bergmann, C., & Tsuji, S. (2020). Preregistration in infant research—A primer. Infancy, 25(5), 734-754.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. Behavior Research Methods, 50(3), 1166-1186. https://doi.org/10.1177/0013164410384856

Houston, D. M., Horn, D. L., Qi, R., Ting, J. Y., & Gao, S. (2007). Assessing speech discrimination in individual infants. Infancy, 12(2), 119–145. https://doi.org/10.1111/j.1532-7078.2007.tb00237.x

Humphrey, C., & Swingley, D. (2018). Regression analysis of proportion outcomes with random effects. arXiv preprint arXiv:1805.08670.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. Advances in Methods and Practices in Psychological Science, 3(2), 166-184. https://doi.org/10.1177/2515245919882903

de Klerk, M., Veen, D., Wijnen, F., & de Bree, E. (2019). A step forward: Bayesian hierarchical modelling as a tool in assessment of individual discrimination performance. Infant Behavior and Development, 57, 101345. https://doi.org/10.1016/j.infbeh.2019.101345

Inoue, T., Nakagawa, R., Kondou, M., Koga, T., & Shinohara, K. (2011). Discrimination between mothers' infant-and adult-directed speech using hidden Markov models. Neuroscience research, 70(1), 62-70. https://doi.org/10.1016/j.neures.2011.01.010

Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, K. (2020). Simplicity and validity in infant research.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163. https://10.1016/j.jcm.2017.10.001

Kovács, Á. M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. Proceedings of the National Academy of Sciences, 106(16), 6556-6560. https://doi.org/10.1073/pnas.0811323106

ManyBabies Consortium. (2020). Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference. Advances in Methods and Practices in Psychological Science, 24–52. https://doi.org/10.1177/2515245919900809

Nighbor, T., Kohn, C., Normand, M., & Schlinger, H. (2017). Stability of infants' preference for prosocial others: Implications for research based on single-choice paradigms. PloS one, 12(6), e0178818. https://doi.org/10.1371/journal.pone.0178818

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. Infancy, 22(4), 436-469. https://doi.org/10.1111/infa.12186

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. Advances in Methods and Practices in Psychological Science, 2(4), 378-395. https://doi.org/10.1177/2515245919879695

Pérez-Edgar, K., Vallorani, A., Buss, K. A., & LoBue, V. (2020). Individual differences in infancy research: Letting the baby stand out from the crowd. Infancy, 25(4), 438-457. https://doi.org/10.1111/infa.12338

Pour Iliaei, S., Killam, H., Dal Ben, R., & Byers-Heinlein, K. (2020). Bilingualism affects infant cognition: Insights from new and open data. https://doi.org/10.31234/osf.io/ex76a

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle W (2021). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.1.3, https://CRAN.R-project.org/package=psych.

Rouder, J.N., Haaf, J.M. (2019). A psychometrics of individual differences in experimental tasks. Psychononmic Bulletin & Review 26, 452–467. https://doi.org/10.3758/s13423-018-1558-y

Santolin, C., Garcia-Castro, G., Zettersten, M., Sebastian-Galles, N., & Saffran, J. R. (2020). Experience with research paradigms relates to infants' direction of preference. Infancy. https://doi.org/10.1111/infa.12372

Schreiner, M. S., Lippold, M. (2020). Assessing test-retest reliability of the infant preference measures. Poster presented at the Virtual International Congress of Infant Studies (vICIS 2020), Glasgow, UK. Pre-registration retrieved from: https://osf.io/s5xqn/

Schott, E., Rhemtulla, M., & Byers-Heinlein, K. (2019). Should I test more babies? Solutions for transparent data peeking. Infant Behavior & Development, 54, 166–176. https://doi.org/10.1016/j.infbeh.2018.09.010

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. Open Mind, 1(1), 4-14.https://doi.org/10.1162/OPMI_a_00002

Scott, K., Chu, J., & Schulz, L. (2017). Lookit (Part 2): Assessing the viability of online developmental research, results from three case studies. Open Mind, 1(1), 15-29. https://doi.org/10.1162/OPMI_a_00001

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., ... & Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. Trends in Cognitive Sciences, 24(9), 675-678. https://doi.org/10.1016/j.tics.2020.06.004

Spearman, C. (1904). The proof and measurement of association between two things. The American Journal of Psychology, 15(1), 72–101. https://doi.org/10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Symonds, P. M. (1928). Factors influencing test reliability. *Journal of Educational Psychology, 19*(2), 73–87. https://doi.org/10.1037/h0071867

Trafimow D. The ubiquitous Laplacian assumption: Reply to Lee and Wagenmakers (2005) Psychological Review. 112: 669-674. DOI: 10.1037/0033-295X.112.3.669

van Renswoude, D.R., Raijmakers, M.E.J., Koornneef, A. et al. Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. Behav Res 50, 834–852 (2018). https://doi.org/10.3758/s13428-017-0909-3