Is Reliability of Cognitive Measures in Children Dependent on Participant Age? A Case Study with Two Large-Scale Datasets

Okko Räsänen (okko.rasanen@tuni.fi)
Unit of Computing Sciences, Tampere University
P.O. Box 553, FI-33101, Finland

María Andrea Cruz Blandón (maria.cruzblandon@tuni.fi)
Unit of Computing Sciences, Tampere University
P.O. Box 553, FI-33101, Finland

Jukka Leppänen (jukka.leppanen@utu.fi)
Department of Psychology and Speech-Language Pathology, Faculty of Social Sciences,
20014 University of Turku, Finland

Abstract

When assessing children in laboratory experiments, the measured responses also contain task-irrelevant participant-level variability ("noise") and measurement noise. Since experimental data are used to make inferences of development of cognitive capabilities with age, it is important to know if reliability of the used measurements depends on child age. Any systematic age-dependent changes in reliability could result in misleading developmental trajectories, as lower reliability will necessarily result in smaller effect sizes. This paper examines age-dependency of task-independent measurement variability in early childhood (3–40 months) by analyzing two large-scale datasets of participant-level experimental responses: the ManyBabies infant-directed speech preference (MB-IDS) dataset and a saccadic reaction time (SRT) dataset collected from rural South Africa. Analysis of participant- and study-level data reveals that MB-IDS shows comparable reliability across the included age range. In contrast, SRTs reflect systematically increasing measurement consistency with increasing age. Potential reasons and implications of this divergence are briefly discussed.

Keywords: child development; empirical data; statistical analysis; data reliability; eye tracking; large-scale data

Introduction

Controlled behavioral experiments are one of the basic tools for scientific study of human cognition and its development. By exposing participants to a series of carefully designed stimuli and measuring their responses, one can make inferences regarding the underlying mechanisms responsible for participants' learning and information processing. In this context, coming up with good experimental designs is far from trivial and requires suitable lab environments and technological tools to capture the phenomena of interest. Yet, even with the best designs and tools, the collected data is never a pure image of the phenomenon of interest. Instead, participant responses are affected by various sources of variability (Faisal, Selen & Wolpert, 2008), such as participant factors (e.g., general vigilance, attentiveness, comprehension of task instructions, increasing task fatigue), as well as neural (e.g., sensory, cellular, and synaptic) and measurement noise (e.g., finite measurement resolution, human observer effects, or quantization of originally non-discrete behaviors).

When measuring children, especially young infants, experimenters face a series of additional challenges: young infants cannot comprehend and follow explicit task instructions. Therefore, spontaneous behaviors, such as looking times of the infants, are often used to make inferences of stimulus processing (e.g., through eye tracking or monitoring head orientation towards spatial location of the stimuli). Moreover, babies have limited attention span, general executive function, and are yet to develop skills or can alternate between exhibiting or not exhibiting the skill even within the same assessment session (see Adolph, Hoch & Cole, 2014). Hence, experimental research with babies is complicated, and the collected data can potentially include a notable amount of additional variability that is not dependent on the experimental conditions per se.

Notably, measurement variability could be expected to change as a function of infant development and maturation (e.g., the standard deviation of infant oculomotor response decreases by age, Rose et al., 2002). This age-dependency of variability ("noise") becomes a potential problem when age-dependent change in cognitive skills is of interest. Such cases include derivation of developmental trajectories of language skills from meta-analytic models that use age as a moderator in the statistical model (e.g., Bergmann et al., 2018; Gasparini, Tsuji & Bergmann, 2022; Lewis et al., 2016) and use of the developmental trajectories as a reference in evaluation of computational models of language development (Cruz Blandón, Cristia & Räsänen, 2022). Even though meta-analytic models do consider different sources of variation, the standard approach for developmental trajectory estimation does not reveal whether the potentially observed age-dependent effect size changes are driven by changes in measurement uncertainty or by developmental changes in the language capability of interest. Thereby, it would be important to understand what kind of impact participant age and the associated experimental designs (if they co-vary with age) have on reported effect sizes. This is also related to the ongoing debate on how to interpret empirical findings across

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

several studies, given the uncertainties associated with each study (see, e.g., Kvarven, Strømland & Johannesson, 2020, and Lewis et al., 2022, for a recent discussion).

This is not to say that age-dependency of measurement reliability is completely ignored. For instance, the proportion of (in)valid trials per participant is often employed as a proxy of data reliability. For example, ManyBabies consortium (2020), focusing large-scale study of infant-directed speech (IDS) preference, found participant age as a significant (positive) predictor of missing data. However, the authors did not make strong inferences regarding the impact of data quality on the observed developmental pattern.

In this paper, we try to dig deeper into the measurement reliability of infant behavior as a function of infant age by explicitly focusing on confidence intervals of effect sizes estimated from infants and studies consisting of groups of infants. We use the large-scale MB-IDS data to study whether age-dependency of IDS preference is potentially affected by age-dependent trial-level data reliability. In addition to the MB-IDS data, we conduct a similar analysis on a large-scale dataset of saccadic reaction times (SRTs) measured from babies from a non-WEIRD environment to see if a similar pattern of data reliability as a function of age emerges. These two datasets were chosen as they both contain responses from hundreds (SRT) or thousands (MB-IDS) of infants with tens of thousands of responses from the same well-established experimental design, and with well-documented primary findings from the respective studies.

Throughout the remaining paper, we will use the term *variability* to refer to the subject- or study-level dependent measure variability that is not explained by experimental or other common factors shared by study participants. Moreover, we do not address *different sources* of within-subject variability, but focus on the overall contribution and age-dependency of variability on the measures of interest.

Why Effect Sizes Shrink with More Variability?

The basic problem with task-independent measurement uncertainty (variability) is that, on average, the resulting effect size estimates will be lower than those of noise-free measurement. Our present aim is to estimate the amount of variability in infant responses as a function of infant age, i.e., to estimate the reliability of the data (see also DeBolt, Rhemtulla & Oakes, 2020).

To approach this formally, a simplistic model of infant behavioral response *a*(*x*) for a stimulus *x* in a single test trial can be written as

$$a(x) = f(x) + N(0, \sigma_b^2), \quad (1)$$

where *f*(*x*) is stimulus-related cognitive processing (e.g., exogenous attention driven by the stimulus) and $N(0, \sigma_b^2)$ is normally distributed stimulus-independent variability in the responses. Variance $\sigma_b^2 > 0$ reflects the total contribution of all sources of internal variability at the participant level (e.g., fussiness, attentiveness, neural noise etc.), and differs across participants. In addition, the datum recorded for the given trial is not the infant response *a*(*x*) as such, but some external observation *r*(*x*) = *g*(*a*(*x*)) together with measurement noise:

$$r(x) = g\big(a(x)\big) + N(\mu_m, \sigma_m^2) \quad (2)$$

Even if we don't know the form of *f*() or *g*(), we know that larger σ_m^2 and σ_b^2 result in a smaller effect size (on average). This is since the effect size is inversely proportional to the variance across the measurements:

$$d \sim \frac{1}{\sigma^2} \sim \frac{1}{\sigma_s^2 + \sigma_m^2 + g\{\sigma_b^2\}} \quad (3)$$

where $g\{\sigma_b^2\}$ denotes variance resulting from application of g() to $N(0, \sigma_b^2)$, and σ_s^2 is *task-dependent* across-subject variability in the coded responses. In the extreme case, poor measurements or a complete lack of task engagement results in stimulus-independent and hence experimental condition independent responses with zero effect.

Usually, the aim of an experiment is to measure ES related to *f*(x) while minimizing the impact of σ_b^2 (e.g., by using engaging stimuli, limiting experiment length, controlling for infant vigilance) and the impact of σ_m^2 (e.g., by using sensitive experimental paradigm, calibrating the measurement system, avoiding coding bias etc.). However, in practice, these "noise terms" always exist due to individual variation and through finite measurement fidelity. This means that, on average, *the measured effect sizes are smaller than what actual differences in infants' stimulus-dependent processing between experimental conditions would entail* (statistically speaking; but see, e.g., Oakes, 2017, or Bergmann et al., 2018, for discussion on the opposite effect of publication bias on reported effect sizes).

In the present study, we are interested in the impact of task-independent noise factors to the observed effect sizes in behavioral studies, and how they may change as a function of infants' age. Both the measurement noise σ_m^2 (e.g., due to different experimental paradigms or their suitability to infants of a particular age) and subject-level task-independent noise σ_b^2 can, and most likely will, change with infant development. If this is the case, then any effect size -based developmental trajectory estimates or age group comparisons should take the effects of noise into account when interpreting age-dependent changes in language capabilities.

Why Younger Baby Data Could be Less Reliable?

Many researchers working with infant development can probably relate to the anecdotal notion that measuring young infants in a lab is more complicated than that of older children or adults. This implicitly suggests that data collected from older participants could also be more reliable. Besides the effort needed for participant recruitment and running the practicalities at the lab, the greater variability associated with younger infants might be reflected in the collected data in at least two ways: 1) a higher proportion of failed trials or excluded participants, as determined by the exclusion criteria of the study, and 2) larger variance in trial-by-trial responses of the infants due to a larger role of σ_b^2 compared to task-dependent processing in the control of observable action.

Note that for 2) to hold, the observed variability should be decorrelated across the infants and not dependent on experimental or population-level variables, such as stimulus identity, presentation order, or native language. By properly controlling for different potential explanatory factors, one could try to estimate how the σ_b^2 change with infant age, other things equal. That is something we try to measure in this work by analyzing trial-level response data from a large participant population of different-aged babies.

However, the overall picture is much more complicated than the above one. The same experimental paradigm or same set of stimuli might not be ideal for babies of different ages, and therefore σ_m^2 may also vary with age (see also ManyBabies Consortium, 2020). In addition, the older the babies, the more likely they are to reflect individual variation in developmental trajectories and stages. This can lead to, e.g., additional variability across different stimuli of the same experimental condition, whereas the experimenter assumes the stimuli to be equally representative of the phenomenon of interest. This is a limitation we acknowledge, and hence our results should be subject to careful interpretation.

Case study 1: ManyBabies IDS Preference

Dataset

As our first dataset, we use the ManyBabies (MB) infant-directed speech (IDS) preference dataset publicly available at https://osf.io/re95x/ from MB study by ManyBabies Consortium (2020). The MB study consisted of IDS preference experiment conducted separately at 67 different labs around the world with a total of 2329 infants using the same study design and the same set of IDS and ADS stimuli spoken in North American English. The only differences between the studies in different labs were 1) whether infant responses were either collected using eye tracking, head-turn preference procedure (hpp), or single-screen central fixation method, 2) whether the infants were native English listeners, 3) and the infant age group(s) tested at each lab.

In the MB experiment, infants were exposed to 8 IDS and 8 ADS audio clips in North American English while their looking times (LTs) to targets were measured. The dependent variable was derived as the looking-time difference $\text{LTD}_i = \text{LT}_{\text{IDS},i}–\text{LT}_{\text{ADS},i}$ from pre-defined IDS and ADS stimulus pairs (from now on referred to as trials *i*), reflecting the attentional preference towards IDS over ADS stimuli. The MB response dataset (*03_data_diff_main.csv* in the OSF repository) consists of these trial-level LTDs for each infant from each participating lab together with info on infant age and native language. In addition, invalid trials (e.g., shorter than 2-s looking time; fussiness etc.; see ManyBabies Consortium, 2020 for detailed criteria) are separately marked in the data.

In our analyses, we started with all data from all participants and labs. If one lab had tested multiple age groups (in bins of 3–6, 6–9, 9–12, and 12–15 months) and/or using multiple methods, each age-group/method combination was treated as a separate study. Babies with less than 5 valid trials were excluded from bootstrap analyses (next Section), resulting in valid data from 1433 infants between 3.0 and 15.0 months of age. For study-level bootstrap statistics, a study was included if there were valid data (≥ 5 trials) from at least 10 babies. This resulted in a dataset of 62 studies with 1155 infants. For the specific methods, the number of valid studies (babies) corresponded to 29 (618) for hpp, 26 (407) for central fixation, and 7 (130) for eye tracking. All trials of all babies were used for counting the proportion of valid trials.

Measuring Age-Dependency of Data Reliability

To investigate the potential age-dependency of measurement reliability, we used two complementary methods to estimate the amount of noise in infant responses: 1) the proportion of invalid trials, and 2) 95% confidence intervals (CI95) of effect sizes (ES), as obtained from bootstrapping. For IDS, the first one was already reported in ManyBabies Consortium (2020), but we replicate the analysis for completeness. For invalid trial proportion, we simply count the number of invalid trials for each baby and then report the average proportions for the different age groups in the two datasets.

For ES CI95 estimation, we conduct standard empirical bootstrapping by resampling trial-level responses of each infant with replacement. On MB data, for each bootstrap sample, we calculate 1) IDS preference ES (Cohen's *d)* for *each infant* using the bootstrap sample of LTDs of valid trials of that baby, and 2) IDS preference ES for *each study* using the means of participant-level bootstrap samples of LTDs. We perform bootstrap resampling 10,000 times for each baby, and then use 2.5% and 97.5% percentiles of the resulting ES to define the CI95 for individual babies (CI95_B) and studies (CI95_S). For all CI95 results, we report the width of the confidence interval (max–min) and study how it changes with age (Pearson correlation between age and CI).

Note that while derivation of participant-level effect sizes is not normally meaningful (corresponding to a study with *N*=1), it allows us to test the stability of participant-level responses as a function of age. We expect more noise to result in higher LT variability across trials, and hence also increasing the CI95_B captured by the bootstrap analysis.

In the main analyses, we pool the MB-IDS data from all three testing methods (eye-tracking, hpp, central fixation), as pooling is also often employed in developmental trajectory estimation, but also report results for the separate methods.

To ensure that the bootstrapping is capturing variance that is independent of cross-subject factors, the MB-IDS data was normalized by subtracting the LTD predictions of the following linear model:

$$LTD \sim trial + method + NA + stimID + trial * age + method * age \qquad (4)$$

from the original LTDs, where *NA* = North American English as infant L1, *stimID* = stimulus identity (categorical) and *trial* = trial number (1–8). After the normalization, CI95 estimates from the bootstrapping reflect uncertainty associated with within-subject residual variance, and we are interested if this depends on the age of the infants. Intuitively, magnitude of the resulting CI95$_B$ reflects the randomness in infant responses across trials, as measured separately for each baby, while CI95$_S$ reflects how this participant-level uncertainty propagates to study-level effect size statistics.

Results for MB-IDS

Top row of Fig. 1 shows the effect sizes for MB-IDS data at participant (left) and study levels (right). As reported by ManyBabies Consortium (2020), ES representing infant IDS preference increases with infant age ($r = 0.31$; $p = 0.0151$) with a positive mean effect. As for the *reliability* of the data, Fig. 1 middle row shows the confidence interval estimates of the bootstrapping (left and center) and the proportion of invalid trials (right) in the data. Bottom row shows the corresponding results for the three experimental methods.

As seen from Fig. 1, the CI95s of participant- and study-level ES do not have a statistically significant change as a function of age when data is pooled across the three methods. Concerning the individual methods, central fixation has a slight increase in CI95$_B$ with age ($r = 0.118, p = 0.011$) while the other two methods do not show age dependency. As noted by ManyBabies consortium (2020), the proportion of invalid trials increases with infant age ($r = 0.194$, $p < 0.001$). The original authors provided faster habituation of older infants as a likely explanation to this. Due to this, valid sample size per study (after exclusion criteria) was also slightly negatively correlated with age ($r = –0.158, p = 0.001$).

Overall, the results show that data collected from younger infants do not show larger trial-by-trial variability in responses. This suggests that the amount of noise in the data, as captured with the present methodology, should not substantially bias age-dependent effect size fits to the study-level data for developmental trajectory estimation (e.g., ManyBabies Consortium, 2020). More specifically, the observed increase in IDS preference with aging does not appear to be simply a result of more task-independent randomness in the data of younger babies.

Case study 2: Saccadic Reaction Times from Eye Tracking

Dataset

As the second dataset, we utilize SRT data collected from Greater Tzaneen area within Mopani District, Limpopo Province, South Africa, as originally described in Leppänen et al. (2023). The data consists of responses from an SRT experiment administered to infants at 7-, 17-, and 36-month checkpoints in a longitudinal manner. Participant caregiver-infant dyads were a subsample of dyads taking part in broader study investigating the impact of a package of early childhood interventions in the area.

SRTs were collected when the dyads visited a lab where the children were administered EEG and eye tracking assessments on the same visit. Eye tracking measurements were collected in a quiet room using Tobii X3-120 equipment. Each test consisted of two 3–4 min sessions. One session consisted of calibration targets, videos depicting short (5–45 s) social scenes (data not used here), and visual saccadic target sequences for SRT measurement. The three target types were sequentially presented in the given order several times across the session. In total, 6–18 calibration targets and 40 saccade targets were presented to each child across the two sessions. The targets consisted of colored animated cartoon drawings of objects (e.g., bird, fish, face, soccer ball; size ~ 5.7° x 5.7°), starting from center and then with pseudo-random 10° shifts between subsequent targets.

For the SRT data, a saccadic trial was considered as valid if the starting position of the saccade was placed on the previous target (discarding the first target), there were no missing samples exceeding 100-ms between target onset and saccade registration, gaze entry to target was not preceded by a missing sample, and SRT fell within expected target range of 100–1000 ms from target onset. In addition, outlier SRTs (2.5 SDs from grand average SRT) were discarded.

In addition to SRTs, the dataset contains estimates of household wealth of the individual children, a proxy for relative socioeconomic status of the children. Household wealth was estimated based on a checklist of 29 assets that a household might own, subjected to principal component analysis as described by Filmer and Pritchett (2001). The reader should see Leppänen et al. (2023) for complete details of the experimental setup, participants, and collected data.

In our experiments, we analyzed data from all the participants who had at least 10 valid trials per visit and at least one visit. This left us with 357 out of 386 original participants and with 717 unique participant/age-bin combinations. This corresponded to a total of 28,520 SRT trials and 15,835 valid trials across the three age bins.

Measuring Age-Dependency of Data Reliability

The basic procedure for CI95 estimation on the Tzaneen SRT data was similar to the MB-IDS data. However, instead of measuring ES, we directly estimate CI95s of SRTs of individual participants and test if they depend on age, as the SRTs are the primary measure of interest in this case. In addition, as an aggregate "study-level" phenomenon, we measure correlations and their CI95s between SES scores of infant households and infant SRTs for the different age bins. This is since earlier research has reported a reduction in SRTs with increasing household wealth on the same data (Leppänen et al., 2023).

For individual SRT CI95 analysis, the model used for residual variance calculation was

$$SRT \sim stimID + SES + trial + trial * age + SES * age \quad . \tag{5}$$

The same model but without the SES term was used for the correlation $r\{\text{SES, SRT}\}$ CI95 estimation. As with the MB-IDS data, residual variance after applying the above model was subjected to 10,000 bootstrap simulations to estimate the 95% CIs for the participant SRTs (CI95$_B$) and the correlation between SES and SRT.

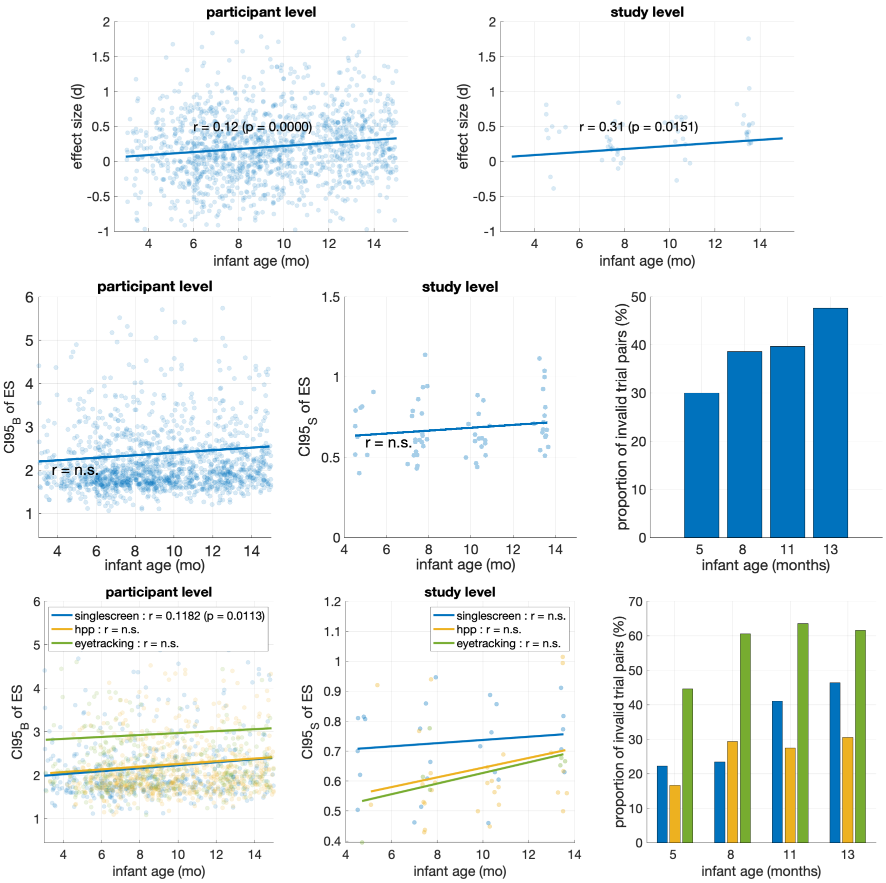

Figure 1: Top row: IDS preference effect sizes for individual participants (left) and studies (right) as a function of infant age. Individual dots denote individual infants (left) or studies (right) and solid line shows least-squares fit to the data together with Pearson correlation between effect sizes and infant ages. Middle row: participant (left) and study level (middle) 95% confidence interval widths (CI95) for the effect sizes together with the proportion of invalid trials (right) as a function of infant age. Bottom row: CI95s and invalid trial results shown separately for the three used experimental paradigms (singlescreen = central fixation, hpp = head-turn preference paradigm).

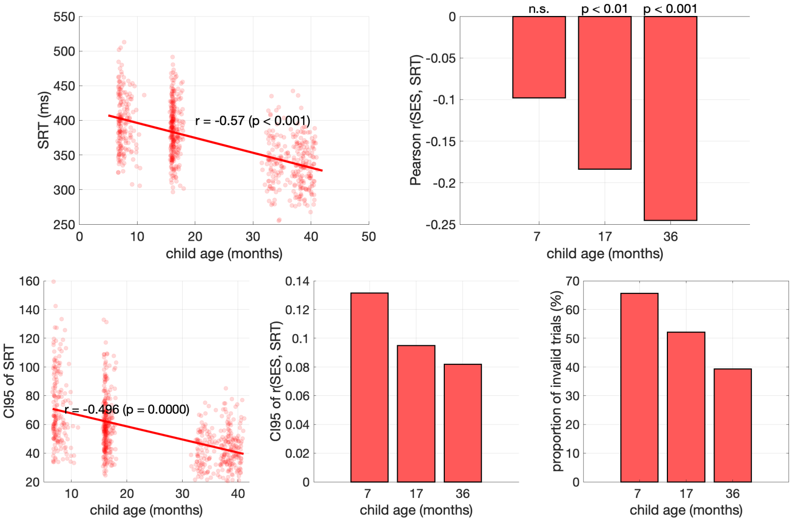

Figure 2: Top left: participant SRTs as a function of their age. Top right: Pearson correlation between family SES and infant SRT for different age groups. Bottom left: 95% confidence interval width (CI95) for SRT data. Bottom middle: CI95s for the correlation between family SES and infant SRT. Bottom right: proportion of invalid trials for the different age groups. Straight lines on the left plots denote least squares fits to the data.

Results for SRTs

Top row of Fig. 2 shows the basic results for SRT analysis with age-dependency of SRTs on the left and SES-SRT correlation on the right. As expected (e.g., Alahyane et al., 2016), the SRTs systematically decrease with age (r = –0.57, p < 0.001). In addition, household wealth (SES) and SRTs are inversely correlated in the 17- and 36-month age groups, as reported by Leppänen et al. (2023).

As for the respective CI95s (Fig. 2, bottom row), there is now also a decreasing trend with age for both the subject-level SRTs (r = -0.496, p < 0.0001) and age-group SES-SRT correlations, and for the proportion of invalid trials. This shows that not only that the SRTs become lower for older infants, but also the reliability of individual saccadic trials becomes higher with age. This is in contrast with the MB-IDS results, where no notable reliability differences were observed between the age groups. As a posthoc analysis, we also verified that a significant age-dependent decline in the CI95s remains when the CIs are adjusted for the mean SRT of each participant (r = -0.297, p < 0.0001).

Finally, we checked the complementarity of invalid trial counts and CI95s from the bootstrapping by measuring their correlation at participant level using the SRT data. The analysis reveals that they are related but still complementary (r = -0.6531; p < 0.0001). This indicates that participant-level CI95s estimated from residual variance can provide additional information on data reliability.

Discussion and Conclusions

The study set out to investigate whether reliability of cognitive measures in children depends on the child age, as measured by confidence intervals of subject- and study-level behavioral responses. For the MB-IDS data, we did not find evidence of higher across-trial variability and hence lower reliability of the measurements for younger children. In contrast, there was a very slight increase in uncertainty of the measurement of participant-level ES as a function of child age, but reliability of full studies did not have an age-dependent trend. The result suggests that comparison of IDS-preference effect sizes between age groups is not critically biased by age-dependent noise in the empirical data, hence supporting the idea of developmental trajectory estimation with age-dependent statistical models fit to large-scale data (cf., Bergmann et al., 2018; Cruz Blandón et al., 2022).

In contrast, results with SRTs from eye tracking showed that the trial-by-trial variability of the SRTs decreased by age, being consistent with Rose et al. (2002). Note that no such decrease in variability was observed even for eye tracking in the case of MB-IDS data. The higher variability at younger ages may also explain why the associations to other measures, such as SES here, become stronger with age, as the "signal-to-noise" ratio of the SRT measurements might be worse with younger infants.

The contradictory findings from the two datasets are puzzling. If the MB-IDS task were to be more suitable for younger compared to older children, that may counterweight otherwise potentially higher response variability at a younger age. However, the IDS preference effect increases with age, hence the experimental setup is at least somewhat applicable across the studied age groups. Also, the used normalization of the LTD in Eq. (4) (esp. *trial*age*) attenuates systematic effects of faster habituation at older ages (see ManyBabies Consortium, 2020, for a discussion) that could otherwise cause higher across-trial variability with increasing age and thereby counterbalance age-dependent effects of variability. Also, the age ranges in the two tasks overlap only partially, which may hide non-linearly changing variability factors.

As another potential factor, MB-IDS data come from a multimodal task where auditory processing is measured via visual looking behavior, and where the measured LTDs (attentional preference) are a result of relatively advanced cognitive analysis of prosodic and linguistic differences in stimulus characteristics. In contrast, SRT only involves relatively simple sensory-motor transformation of visual information needed for visual target pursuit. Given the gradual maturation of cortical connectivity and development of higher-level cognitive processes (see also Blumberg & Adolph, 2023), sensorimotor processing required for the MB-IDS task may simply be cognitively more demanding than saccadic reactions, which could be seen as a more "elementary" cognitive process. The age differences between the MB-IDS and SRT participant populations can also increase the relative contribution of the task complexity on the measured responses.

Finally, MD-IDS relies on a difference of two noisy constituent measures whereas SRT does not. As a result, the "signal-to-noise" ratio in the two datasets is very different with trial-level SRTs having relatively modest variability around the mean participant SRTs (CI95s are approx. 1/8th of the means). In contrast, participant-level CI95s in MB-IDS are approx. tenfold to the means. Hence, it is possible that even if age-dependent factors to data reliability exist for MB-IDS, they are swamped by other sources of variability in the given experimental setup.

In general, the present study highlights how reliability of child data as a function of child age can vary from an experimental setting to another. In addition, the intuition of getting less reliable data from younger children does not always necessarily hold—at least to the extent that the present methodology can dig into this issue. Given that only two qualitatively distinct datasets were investigated, it is difficult to draw broader conclusions or predictions on measurement reliability with participant age. Instead, additional analyses with different experimental paradigms and cognitive phenomena should be conducted for cases where suitable large-scale data are available.

Acknowledgment

O.R and M.A.C.B. were supported by Academy of Finland grants no. 314602 and 345365 and by Kone Foundation grant L-SCALE. The study in Tzaneen was funded by the South African Medical Research Council (www.samrc.ac.za) through a grant to the Health Economics and Epidemiology Research Office at the University of the Witwatersrand. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Adolph, K. E., Hoch, J. E., & Cole, W. G. (2018). Development (of walking): 15 Suggestions. *Trends in Cognitive Sciences*, *22*(8), 699–711.

Alahyane, N., Lemoine-Lardennois, C., Tailhefer, C., Collins, T., Fagard, J., & Doré-Mazars, K. (2016). Development and learning of saccadic eye movements in 7- to 42-month-old children. *Journal of Vision*, 16(1):6.

Bergmann, C., Tsuji, S., Piccini, 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.

Blumberg, M. S., & Adolph, K. E. (2023). Protracted development of motor cortex constraints rich interpretations of infant cognitions *Trends in Cognitive Sciences*, 27, 232–245.

Cruz Blandón, M. A., Cristia, A., & Räsänen, O. (2022). Evaluation of computational models of infant language development against robust empirical data from meta-analyses: what, why and how? *OSF preprint:* https://doi.org/10.31234/osf.io/yjz5a

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, 393–419.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, 9(4), 292–303.

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data-or tears: an application to educational enrollments in states of India. *Demography*, 38(1), 115–132.

Gasparini, L., Tsuji, S. & Bergmann, C. (2022). Ten easy steps to conducting transparent, reproducible meta-analyses for infant researchers. *Infancy*, 27, 736–764.

Kvarven, A., Strømøland, E, & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behavior*, 4(4), 423–434.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, article 863.

Leppänen, J. M., Tarullo, A., Evans, D., Coetzee, L., Fink, G., Yousafzai, A. K., Hamer, D. H. & Rockers, P. C. (2023). Socioeconomic gradients in children's eye movement behaviors. *OSF preprint*: https://doi.org/10.31219/osf.io/pjhk4.

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P., Cristia, A., & Frank, M. C. (2016). A quantitative synthesis of early language acquisition using meta-analysis. *Preprint:* https://doi.org/10.31234/osf.io/htsjm.

Lewis, M., Mathur, M. B., VanderWeele, T. J., & Frank, M. C. (2022). The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science*. 9: 211499.

ManyBabies Consortium: Frank, M. C., Alcock K. J., Arias-Trejo N., et al. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, 3(1), 24–52.

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, 22(4), 436–496.

Rose, S. A., Feldman, J. F., Jankowski, J. J., & Caro, D. M. (2002). A longitudinal study of visual expectation and reaction time in the first year of life. *Child Development*, 73(1), 47–61.