Limited evidence of test-retest reliability in infant-directed speech preference in a large pre-registered infant sample

Melanie S. Schreiner[1,2], Christina Bergmann[3], Michael C. Frank[4], Tom Fritzsche[5], Nayeli Gonzalez-Gomez[6], Kiley Hamlin[7], Natalia Kartushina[8], Danielle J. Kellier[4], Nivedita Mani[1], Julien Mayor[8], Jenny Saffran[9], Melanie Soderstrom[10], Mohinish Shukla[11], Priya Silverstein[12,13], Martin Zettersten[9,14], & Matthias Lippold[1]

[1] University of Goettingen

[2] Leibniz Science Campus PrimateCognition

[3] Max Planck Insitute for Psycholinguistics

[4] Stanford University

[5] University of Potsdam

[6] Oxford Brookes University

[7] University of British Columbia

[8] University of Oslo

[9] University of Wisconsin-Madison

[10] University of Manitoba

[11] Università di Padova

[12] Lancaster University

[13] Center for Open Science

[14] Princeton University

Author Note

Acknowledgements. This work was supported in part by Leibniz ScienceCampus Primate Cognition seed funds awarded to MS and ML, and grants from XXXX awarded to XXXX and from the NSF awarded to MZ (NSF DGE-1747503).

Open Practices Statement. All code for reproducing the paper is available at https://github.com/msschreiner/MB1T. Data and materials are available on OSF (https://osf.io/zeqka/).

CRediT author statement. Outside of the position of the first and the last author, authorship position was determined by sorting authors' last names alphabetical order. An overview of authorship contributions following the CRediT taxonomy can be viewed here: https://docs.google.com/spreadsheets/d/1jDvb0xL1U6YbXrpPZ1UyfyQ7yYK9aXo002UaArqy35U/edit?usp=sharing.

Correspondence concerning this article should be addressed to Melanie S. Schreiner, Gosslerstr. 14, 37073 Göttingen. E-mail: melanie.schreiner@psych.uni-goettingen.de

Abstract

The ManyBabies1 collaborative research project (hereafter, MB1; Frank et al., 2017; ManyBabies Consortium, 2020) explores the reproducibility of the well-studied and robust phenomenon of infants' preference of infant-directed speech (hereafter, IDS) over adult-directed speech (hereafter, ADS; Cooper & Aslin, 1990). The current study is a follow-on project aiming at further investigating the test-retest reliability of infant speech preference measures. In particular, labs of the original study were asked to bring in tested babies for a second appointment retesting infants on their IDS preference. This allows us to estimate test-retest reliability for the three different methods used to investigate preferential listening in infancy: the head-turn preference procedure, central fixation, and eye-tracking. Our results suggest that the test-retest reliability of infants' speech preference measures is rather low. While increasing the number of trials that infants needed to contribute for inclusion in the analysis from 2 to 8 trial pairs revealed a descriptive growth in test-retest reliability, it also considerably reduced the study's effective sample size. Therefore, future research on infant development should take into account that not all experimental measures might be appropriate to assess individual differences between infants, and hence, the interpretation of findings needs to be treated with caution.

Keywords: language acquisition; speech perception; infant-directed speech; adult-directed speech, test-retest reliability

Word count: 4506

Limited evidence of test-retest reliability in infant-directed speech preference in a large pre-registered infant sample

Obtaining a quantitative measure of young infant and children's cognitive abilities is an extraordinarily difficult endeavor; the most frequent way to assess what they know or can do is done via tracking overt behavior. However, the attention span of this specific group of participants is rather short, they do not follow instructions, their mood can change instantly and their behavior can be best described as unstable and volatile. Therefore, most measurements are noisy and the typical sample size of a study is often just around 20 infants per group resulting in low power (Oakes, 2017). In addition, there is individual and environmental variation that may add even more noise to the data (e.g., as discussed by Johnson & Zamuner, 2010). Despite these demanding conditions, infancy researchers, nevertheless, need to have reliable and robust ways to assess infants' behavior.

In order to address these challenges in infant research, the ManyBabies collaborative research project was launched (Frank et al., 2017). Within a collaborative framework, the aim of this initiative is to conduct large-scale conceptual, consensus-based replications of seminal findings to identify sources of variability and establish best practices for experimental studies in infancy.

The fact that there is a substantial amount of literature on infants' preference of infant-directed speech (hereafter, IDS) over adult-directed speech (hereafter, ADS, Cooper & Aslin, 1990) led to the decision of exploring the reproducibility of this well-studied phenomenon within the first ManyBabies collaborative research project (hereafter, MB1, ManyBabies Consortium, 2020). Across many different cultures, infants are commonly addressed in IDS, which typically is characterized by higher pitch, greater pitch range, and shorter utterances, compared to the language used between interacting adults (Fernald et al., 1989). A large body of behavioral studies finds that across ages and methods, infants show increased looking times when hearing IDS compared to ADS stimuli (Cooper & Aslin, 1990; see Dunst, Gorman, & Hamby, 2012 for a meta-analysis). This attentional enhancement is also documented in neurophysiological studies showing increased neural activation during IDS compared to ADS exposure (Naoi et al., 2012; Zangl & Mills, 2007). In addition to the heightened attention, IDS has also been identified as facilitating early word learning. In particular, infants' word segmentation abilities Thiessen, Hill, & Saffran (2005) and their learning of word-object associations (Graf Estes & Hurley, 2013; Ma, Golinkoff, Houston, & Hirsh-Pasek, 2011) seems to be enhanced in the context of IDS. In sum, IDS seems to be beneficial for early language development.

Within MB1, altogether 67 labs contributed data of 2,329 infants showing that babies generally prefer to listen to IDS over ADS. Nevertheless, the overall effect size $d = 0.35$ was much smaller than the meta-analytic effect size of $d = 0.67$ reported by Dunst et al. (2012). The results revealed a number of additional factors that influenced the effect size. First, there was a developmental change with older infants showing larger preferences. Second, the stimulus language was also linked to the speech preference with North American English learning infants being more likely to prefer IDS over ADS than infants learning other languages. Third, comparing the different methods employed, the head-turn preference procedure yielded the highest effect size, followed by the central fixation paradigm, with eye-tracking revealing the smallest effect size. Finally, exploratory analyses assessed the effect of different inclusion criteria. Across methods, using stricter inclusion criteria led to an increase in effect sizes despite the larger proportion of excluded participants (see also Byers-Heinlein, Bergmann, & Savalei, 2021, for the relation between sample size, effect size, and exclusion criteria).

However, there is a difference between a result being reliable in a large sample of infants and the individual measure of an individual infant being reliable. In studies tracking individual differences, the measured behavior during an experimental setting is often used to predict a cognitive function or specific skill later in life. A precondition for this link to be observable is that inter-individual differences between infants do exist and can be detected with a high reliability at these earlier stages. However, how reliable are the measures used in infancy research?

A salient example of this line of individual differences research is the literature showing that infants' behavior in speech perception tasks can be linked to later language development (see Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014 for a meta-analysis), potentially identifying infants at risk for later language delays or disorders. This finding thus has substantial implications for theoretical and applied work. However, to be able to make such claims robustly would require a high reliability within infants across multiple testing sessions. Previous attempts to address the reliability of measurements are either limited to adult populations investigating various tasks (Hedge, Powell, & Sumner, 2018), or have been conducted with very small sample sizes (e.g., Houston, Horn, Qi, Ting, & Gao, 2007).

In one example of such a study, Colombo, Mitchell, and Horowitz (1988) used a paired comparison task, in which infants were familiarized with a stimulus and for the test trials presented with the familiarized and a novel stimulus side-by-side. Results indicated that infants' novelty preference was extremely variable from task to task. Assessing infants' performance from one week to another revealed that infants' attention measures were moderately reliable. However, reliability seemed to increase with the number of tasks infants were able to complete in the younger age group suggesting that reliability is influenced by the number of assessments. In addition, infants' performance from 4 to 7 months was longitudinally stable but somewhat smaller than the week-to-week reliability.

Cristia, Seidl, Singh, and Houston (2016) also retested infant populations by independently conducting 12 different experiments on infant speech perception at three different labs with different implementations of the individual studies. Hence, it was only after completed data collection that the data was pooled together by the different labs revealing potential confounds. Nevertheless, the results showed that reliability was extremely variable across the different experiments and labs and overall low.

Against this background, the current study investigates test-retest reliability of infants' performance in an auditory preference task. Within MB1, a multi-lab collaboration, we examine whether infants' preferential listening behavior to IDS and ADS is reliable across two different test sessions. We also aim to address whether time between test and retest or infants' language background influences the reliability of the preference measure.

Our study was faced with a critical design choice: what stimulus to use to assess test-retest reliability. A first constraint of our study was that, since it was a follow-on to MB1, any stimulus we used would always be presented after the MB1 stimuli. One option would be simply to bring back infants and have them hear exactly the same stimulus materials. A weakness of this design would be the potential for stimulus familiarity effects, however, since infants would have heard the materials before. Further complicating matters, infants might show a preference for or against a familiar stimulus depending on their age (Hunter & Ames, 1988). The ideal solution then would be to create a brand new stimulus set with the same characteristics. Unfortunately, because of the process how MB1 stimuli were created, we did not have enough normed raw recordings available to make brand new stimulus items that conformed to the same standards as the MB1 stimuli.

Given this situation, we chose an intermediate path: we reversed the ordering of MB1 stimuli. Average looking times in MB1 were always lower than 9s per trial, even for the youngest children on the earliest trials (the group who looked the longest on average). So most children in MB1 did not hear the second half of most trials. Thus, by reversing the order, we had a perfectly matched stimulus set that was relatively unfamiliar to most infants. The disadvantage of this design was that infants who looked longer might be more likely to hear a familiar clip that they had heard in the previous study. If infants then showed a familiarity preference – an assumption which might not be true – the end result could be to inflate our estimates of test-retest reliability slightly, since longer lookers would on average look a bit longer at retest due to their familiarity preference. We view this risk as relatively low, but do note that it is a limitation of our design.

The MB1 study was also interested in effects of experimental method on infant-directed speech preference. Using central fixation (hereafter, CF), eye-tracking, and the head-turn preference procedure (hereafter, HPP), the current study also explores whether there are any differences in test-retest reliability between the three widely used methods. Exploring differences in CF, eye-tracking, and HPP, Junge et al. (2020) provide some evidence in favor of using the HPP in speech segmentation tasks. In addition, their re-analysis of the meta-analysis on infant speech segmentation conducted by Bergmann and Cristia (2016) also suggested numerically larger effect sizes for HPP than for eye-tracking. Similarly, the MB1 project reported an increase in the effect size for HPP compared to CF and eye-tracking (ManyBabies Consortium, 2020). HPP requires gross motor movements relative to other methods, such as CF and eye-tracking paradigms, for which subtle eye movements towards a monitor located in front of the child are sufficient. One possible explanation for the stronger effects with HPP may be a higher sensitivity to the contingency of the presentation of auditory stimuli and infants' head turns away from the typical forward-facing position. While these findings suggest that HPP may be a more sensitive index of infant preference, they do not necessarily imply higher reliability for individual infants' performance using the different methods. Hence, it remains an open question whether the same measures that produce larger effect sizes at the group-level also have higher test-retest reliability for individual infants (Byers-Heinlein, Bergmann, et al., 2021). Therefore, assessing the test-retest reliability of the different preference measures is crucial, so that researchers can make informed decisions about the appropriate methods for their particular research question. Critically, only measures with high test-retest reliability should be used for studies of individual differences.

Preregistration. Prior to the start of data collection, we preregistered the current study on the Open Science Framework (OSF; see https://osf.io/v5f8t). We also had secondary planned analyses which were not run in the end as we had fewer participants than originally expected.

Method

A call was issued to all labs participating in the original MB1 study on January 24th, 2018 (ManyBabies Consortium, 2020). The collection of retest session data was initially set to end on May 31st, 2018, one month after the end date of the original MB1 project. Due to the fact that the original MB1 project extended the time frame for data collection and the late start of data collection for the MB1 test-retest study, we also allowed participating labs to continue data collection past the scheduled end date.

Participants

Contributing labs were asked to re-recruit their monolingual participants between the ages of 6 to 12 months who had already participated in the MB1 project. If participating labs had not committed to testing either of these age groups, they were also allowed to re-recruit participants from the youngest age group of 3- to 6-month-olds and/or the oldest age group of 12- to 15-month-olds. Labs were asked to contribute half (n=16) or full sample (n=32), however, a lab's data was included in the study regardless of the number of included infants. The study was approved by each lab's respective ethics committee and parental consent was obtained for each infant prior to participation in the study.

Our final sample consisted of 158 monolingual infants from 7 different labs. In order to be included in the study, infants needed a minimum of 90% first language exposure, to be born full term with no known developmental disorders, and normal hearing and vision. We excluded 11 participants due to session errors and 11 participants who did not have at least one valid trial per condition (IDS and ADS) at their first or second session. The mean age of infants included in the study was 245 days (range: 108 – 373 days). Further information on labs and participants are provided in Table 1.

Materials

Visual stimuli. The visual stimuli were identical to MB1. For the central fixation paradigm and eye-tracking, labs were asked to use a multicolored static checkerboard as the fixation stimulus as well as a multicolored moving circle with a ringing sound as an attention getter to reorient infants toward the screen in between trials. Labs using the HPP method were instructed to use their lab's standard procedure, as in MB1.

Speech stimuli. We used the identical training stimuli of piano music from MB1. A second set of naturalistic IDS and ADS recordings of mothers either talking to their infant or to an experimenter was created for the retest session by reversing the order of clips within each sequence of the original study. This resulted in eight new sequences of natural IDS and eight new sequences of natural ADS with a length of 18 seconds each. This was in order to prevent infants who still remembered the stimuli from their first test session from easily getting bored.

Procedure. Infants were retested using the identical procedure as during the first testing day: central fixation, HPP, or eye-tracking. Participating labs were asked to ideally schedule test and retest session 7 days apart with a minimum number of 1 day and a maximum number of 31 days. Three infants whose time between test and retest exceeded 31 days were also included in the analyses. The mean number of days between test and retest was 10 (range: 1 to 49 days).

A total of 18 trials, including two training, eight IDS, and eight ADS trials, were presented in one of four pseudo-randomized orders. Trial length was either infant-controlled or fixed depending on the lab's standard procedure, that is a trial stopped either if the infant looked away for 2 seconds or after the total trial duration of 18 seconds. The online coding experimenter and the parent listened to music masked with the stimuli of the study via noise-cancelling headphones. If the experimenter was in an adjacent room separate from the testing location, listening to masking music was optional for the experimenter.

Data exclusion. A child was excluded if they had a session error, i.e., an experimenter error (e.g., inaccurate coding, or presentation of retest stimuli on the first test session), or equipment failure (visual stimuli continued to play after the end of a trial). Trials were excluded if they were marked as trial errors, i.e., if the infant was reported as fussy, an experimental or equipment error occurred, or there was parental interference during the task (e.g., if the parent spoke with the infant during the trial). Trials were also excluded if the minimum looking time of 2 s was not met. If a participant was unable to contribute at least one IDS and one ADS trial for either test or retest, all data of that participant was excluded from the test-retest analyses.

Results

IDS preference

First, we examined infants' preferences for IDS in both sessions. Two two-samples t-tests revealed that the children in Session 1, $t(157) = 6.47$, $p < .001$, and in Session 2, $t(157) = 4.19$, $p < .001$, showed a preference of IDS over ADS (see Table 2 for the looking times in each session). In the first session, 68.35 percent of the children showed a preference for IDS, and in the second session, 63.29 percent of the children showed a numerical preference for IDS. In other words, we replicated the previous finding from the main MB1 study. There was no difference in the strength of the preference effect, as a multilevel analysis with a random slope and random intercept for session on the lab level revealed no significant impact of session on infants' preference, with an estimate of zero and very small variance, β=-0.26, SE=0.31, p=.424.

Table 1

Statistics of the included labs. n refers to the number of infants included in the final analysis.

Lab	Method	Language	Mean age (days)	N
babylab-potsdam	HPP	German	227	22
babyling-oslo	Eyetracking	Norwegian	249	10
brookes-babylab	central fixation	English	267	18
InfantCog-UBC	central fixation	English	147	7
infantll-madison	HPP	English	230	30
lancslab	Eyetracking	English	236	16
wsi-goettingen	central fixation	German	280	39
wsi-goettingen	HPP	German	242	16

Table 2

Looking times in s for each session and condition

Trial type	Session 1 $Mean$	Session 1 SD	Session 2 $Mean$	Session 2 SD
ADS	7.72	2.77	6.96	2.92
IDS	8.76	2.85	7.75	2.75

Table 3

Coefficient estimates from a linear mixed effects model predicting IDS preference in Session 2.

	Estimate	SE	t	p
Intercept	0.874	0.456	1.920	0.102
Session One	0.035	0.085	0.414	0.679

Reliability

We assessed test-retest reliability in two ways. First, we conducted a multilevel analysis, with Lab as random intercept, predicting the IDS preference in Session 2 based on the IDS preference in Session 1. The results revealed that we could not predict the preference score in Session 2 based on Session 1 (see Table 2). Second, we calculated the Pearson correlation coefficient. While a simple correlation coefficient might overestimate the test-retest reliability in our sample because it does not control for the differences between different labs and methods (HPP, CF, and eye-tracking), we felt it was important to also conduct a Pearson correlation as it is commonly used to assess reliability. Again, the size of the correlation coefficient was not statistically different from zero and the estimate was, in fact, approaching nil, $r = .09$, 95% CI $[-.06, .25]$, $t(156) = 1.19$, $p = .237$. Furthermore, we calculated the percentage of preference reversal between test and retest counting the number of participants for whom the preference changed between test and retest, and dividing it by the number of all participants. The results revealed that 41.77 percent of the infants had a preference reversal from test to retest session. Taken together, our results lead us to conclude that there is no overall test-retest reliability for the three infant preference measures used within the current study.

To test whether the results were different for a specific method, we calculated the

Table 4

Coefficient estimates from a linear mixed effects model predicting IDS preference in Session 2 for each method separately.

Method	estimate	SE	pvalue	cor	pvalue2
HPP	0.151	0.137	0.276	0.134	0.276
Eyetracking	0.034	0.162	0.835	0.021	0.919
central fixation	-0.195	0.125	0.125	0.080	0.530

Pearson correlation coefficients and the multilevel analyses for the three different methods, HPP, central fixation and eye-tracking, separately (see Table 4). Splitting the data per method also did not lead to different results. Neither the Pearson correlation coefficients nor the coefficients of the multilevel analysis were significant, all p-values $> .286$. We also tested for the possibility that the Time between sessions might have an impact on the reliability. The subsequent multilevel analysis, with Lab as random intercept, predicting the IDS preference in Session 2 based on the IDS preference in Session 1, the number of days between Session 1 and Session 2 and the interaction of these two variables, did not indicate that Time between sessions had an effect. Neither the main effect of Time between sessions, β=0.00, SE=0.03, p=.916, nor the interaction term, β=-0.01, SE=0.02, p=.479, showed significant effects. As NAE-learning infants showed greater IDS preferences than their non-NAE counterparts in the original study, we also assessed if test-retest reliability interacted with children's native language. A multilevel analysis with Lab as random intercept, predicting the IDS preference in Session 2 based on the IDS preference in Session 1, NAE and the interaction of these two variables, revealed no interaction, β=0.29, SE=0.18, p=.115 (see Figure 1).

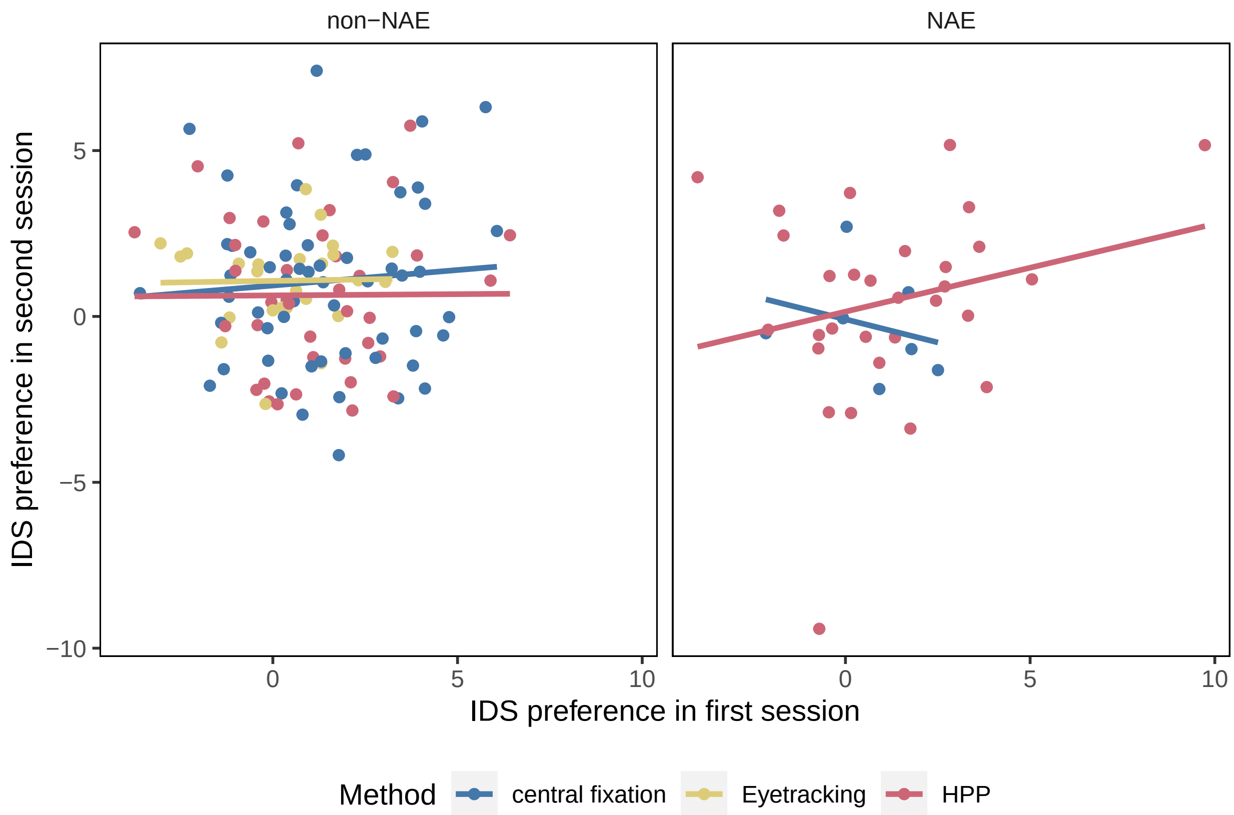

Figure 1. Infants' preference in Session 1 and Session 2 with individual data points and regression lines color-coded by method (central fixation, eye-tracking, or HPP). Results are plotted separately for North American English-learning infants (right panel) and infants learning other languages and dialects (right panel).

Results with different inclusion criteria

To this point, all analyses were performed on data with the inclusion criteria from MB1. For this, infants needed only 1 out of 8 valid trial pairs (i.e., any combination of an IDS and ADS trial) to be included in the analyses. Given that the use of more stringent inclusion criteria yielded larger effects sizes within the original MB1 study, we also assessed test-retest reliability by applying stricter inclusion criteria and thereby increasing test length to 2, 4, 6, and 8 included test trial pairs per condition. Applying a stricter criterion - and thereby increasing test length - increased reliability numerically from r = 0.07 to r = 0.34 (Figure 2). However, in part likely due to the decrease in sample size, only one of these correlations (when requiring a minimum of 6 trial pairs) was statistically significant: 2 valid trial pairs, $t(152) = 0.90$, $p = .367$; 4 valid trial pairs, $t(143) = 1.03$, $p = .306$; 6 valid trial pairs, $t(98) = 2.23$, $p = .028$; 8 valid trial pairs - all trials on both testing days - $t(22) = 1.68$, $p = .108$. The analyses provide tentative evidence that stricter inclusion criteria might lead to higher test-retest reliability but at the same time comes with tremendous decreases in sample size.

General Discussion

The current study set out to explore the test-retest reliability of the infant speech preference of IDS over ADS. Infants of the original MB1 project were retested on a reversed order of stimuli in order to assess if their listening pattern would be similar to that of their initial assessment. While we replicated the original effect of infants' speech preference for IDS over ADS in the current MB1 follow-up study for both test and retest session on the group-level using the same MB1 protocol, we found that infants' speech preference measures had no test-retest reliability. In other words, we were unable to detect any stable individual differences of infants' speech preference. This finding is in line with other research indicating a rather low test-reliability for different developmental paradigms (Cristia et al., 2016). Given that most experimental procedures conducted in developmental research are interested in the comparison of groups, individual differences between participants within a specific condition are usually minimized by the experimental procedure while differences between conditions are maximized. Therefore, the infant preference measure may be a good approach to capture universal phenomena but does not seem to be appropriate for examining factors that may lead to individual differences in development.

Consistent with general psychometric theory (e.g., DeBolt, Rhemtulla, & Oakes, 2020) a larger number of included test trials was associated with higher reliability. However, in our dataset, this association was based on exploratory analyses and was only

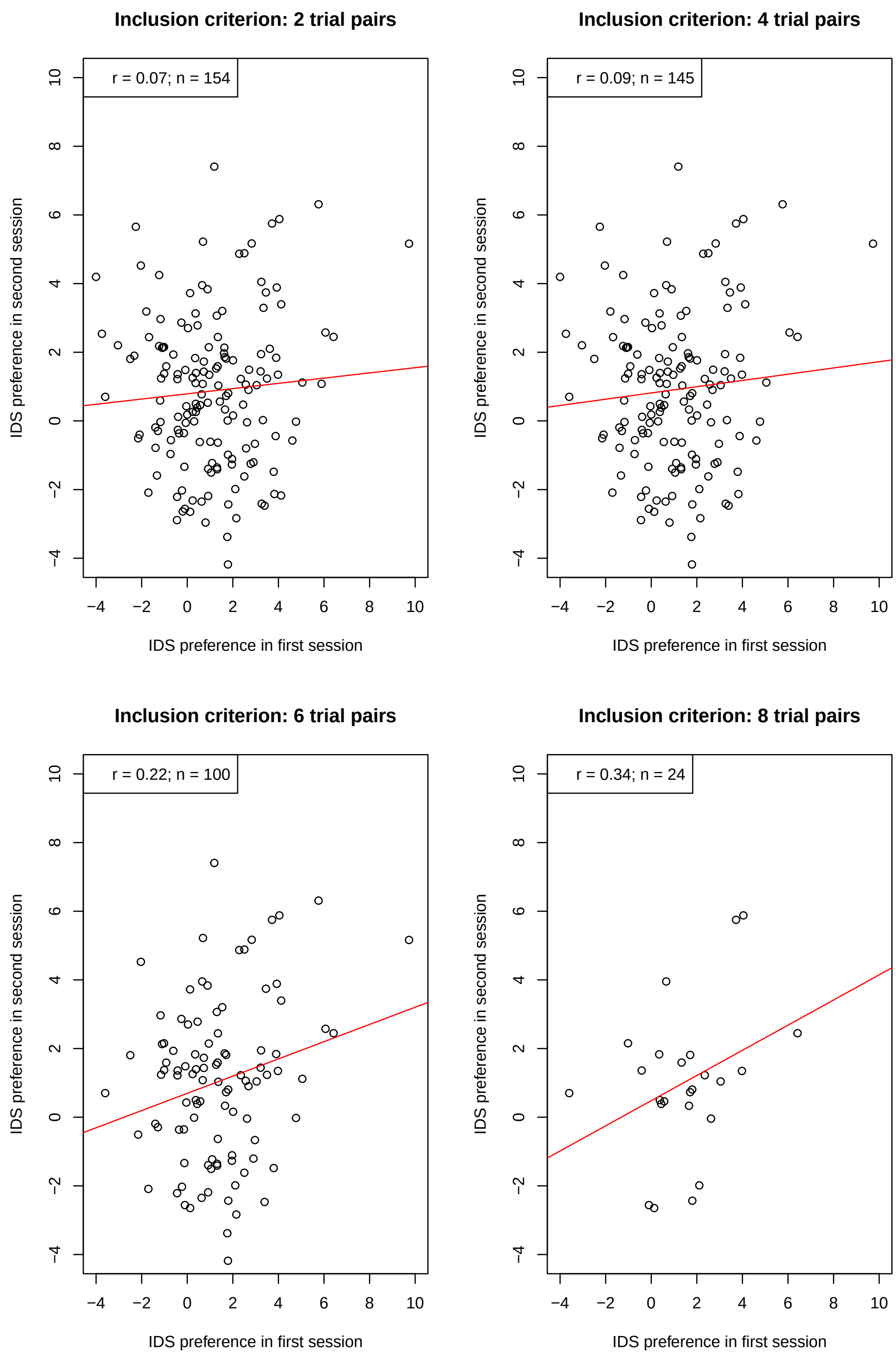

Figure 2. IDS preferences of both sessions plotted against each other for each inclusion criterion. n indicates the number of included infants, r is the Pearson correlation coefficient as the indicator for reliability.

found descriptively, hence, a replication is warranted. A similar effect on the group-level was found in the MB1 project, where a stricter inclusion criterion led to bigger effect sizes (ManyBabies Consortium, 2020).

In our study, as in the MB1 original study, higher reliability through strict exclusions came at a high cost. In particular, with the strictest criterion, only a small portion of the original sample size, that is 24 out of 158 infants, could be included in the final sample for this particular analysis. In other words, applying a stricter criterion leads to a higher drop out rate and reduces the actual sample size enormously. In the case of studies in the field of developmental science, where there are many practical restrictions in collecting large samples of infants (e.g., birth rate in the area, restricted lab capacities, budget restrictions), a strict drop out criterion might not be easy - if even possible at all - to implement. Note that studies in developmental science already have above average drop out rates (Miller, 2017). In addition, drop out may not be random, and so having high drop out rates can further limit the generalisability of a study. Particularly in the context of turning individual differences measures into diagnostic tools, high drop-out rates have an additional limitation of not being broadly usable.

An alternative approach to increase the number of valid trials might be to also increase the number of collected trials. In this case, a participant can have a high number/proportion of invalid trials and still be included into the final sample as the absolute number of trials is high and thereby decreasing trial-to-trial variability (DeBolt et al., 2020; see Silverstein, Feng, Westermann, Parise, & Twomey, 2021 for an example). While this approach might sound promising, it must be seen if this is realistic, because the attention span of a typical participant of a developmental study is rather short. Therefore, prolonging the experimental procedure to maximize the absolute number of trials might also be practically challenging. Further potential attempts in obtaining higher numbers of valid trials may include changes in the procedure (e.g., Egger, Rowland, & Bergmann, 2020) or implementing multi-day test sessions (Fernald & Marchman, 2012). As our results are only based on the particular phenomenon of IDS preference (albeit, with three widely used methods: HPP, central fixation; eye-tracking) it is essential to further assess the underlying reliability of these measures within other areas of speech perception. While most infants prefer IDS over ADS (Dunst et al., 2012), predicting a pattern of preference, for instance, within speech segmentation tasks, i.e. familiar versus novel words, seem not that straightforward (Bergmann & Cristia, 2016). Especially in the context of relating a direction of preference to later language development, there seem to be controversial findings. That is, both familiarity and novelty responses have been suggested to be predictive of infants' later linguistic abilities (DePaolis, Vihman, & Keren-Portnoy, 2014; R. S. Newman, Rowe, & Ratner, 2016; R. Newman, Ratner, Jusczyk, Jusczyk, & Dow, 2006). In light of findings from the current study, researchers conducting longitudinal studies with experimental data from young infants predicting future outcomes should be cautious as there may be inter-individual variability affecting their preferences.

Limitations

While we had an above average sample size for a study in developmental research, we were unable to reach the number of participants collected within the original MB1 study. In addition to a delayed call, the extra effort of having to schedule a second lab visit for each participant and the fact that there were already other collaborative studies taking place simultaneously (MB1B, Byers-Heinlein, Tsui, Bergmann, et al., 2021; MB1G, Byers-Heinlein, Tsui, Van Renswoude, et al., 2021), might have contributed to the rather low turnout. A higher sample size and a larger number of participating labs from different countries might have enabled us to test for possible differences of the test-retest reliability of the different methods (HPP, central fixation, eye-tracking) and NAE versus non-NAE language backgrounds. Further, a larger sample size might have enabled us to conduct meaningful tests of moderators such as age of the child on the test-retest reliability.

A further limitation concerns the stimuli. While the order of the clips within trials presented to the participating children in the second session was different than in the first session, the exact same stimulus material as in MB1 was used in both sessions. In particular, all children heard the exact same voices in Session 1 and in Session 2. From a practical point of view, it was the easiest solution. However, familiarity effects might have influenced infants' looking behavior. Assuming that only infants with longer looking times in Session 1 might have had the chance to recognize the voices in Session 2 from their session a week ago as familiar clips would only be towards the end of trials, infants with shorter looking times might not have had the opportunity to listen to the voices from their first session. Therefore, for some children, familiarity with the stimulus material might have led to artificially lowering test-retest reliability.

Conclusion

Following the MB1 protocol, the current study could not detect test-retest reliability of infants' preference measures for IDS over ADS. Subsequent analyses showed that a stricter criterion for the inclusion of data points may enhance the test-retest reliability at the cost of high drop out rates. Developmental studies which rely on stable individual differences of their participants need to consider the underlying reliability of their measures, and we recommend a broader assessment of test-retest reliability in infant research.

Data and materials availability statement

The data and materials that support the findings of the current study are openly available on OSF at https://osf.io/ZEQKA/.

References

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science*, *19*(6), 901–917.

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development*, e2296.

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., et al.others. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920974622.

Byers-Heinlein, K., Tsui, R. K.-Y., Van Renswoude, D., Black, A. K., Barr, R., Brown, A., et al.others. (2021). The development of gaze following in monolingual and bilingual infants: A multi-laboratory study. *Infancy*, *26*(1), 4–38.

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*, 1198–1210.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595.

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.

Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy*, *21*(5), 648–667.

DeBolt, M. C., Rhemtulla, M., & Oakes, L. M. (2020). Robust data and power in infant research: A case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy*, *25*(4), 393–419.

DePaolis, R. A., Vihman, M. M., & Keren-Portnoy, T. (2014). When do infants begin recognizing familiar words in sentences? *Journal of Child Language*, *41*(1), 226–239.

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1), 1–13. Retrieved from http://www.earlyliteracylearning.org/cellreviews/cellreviews_v5_n1.pdf

Egger, J., Rowland, C. F., & Bergmann, C. (2020). Improving the robustness of infant lexical processing speed measures. *Behavior Research Methods*, *52*(5), 2188–2201.

Fernald, A., & Marchman, V. A. (2012). Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child Development*, *83*(1), 203–222.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B. de, & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., … Vihman, M. (2016). British english infants segment words only with exaggerated infant-directed speech stimuli. *Cognition*, *148*, 1–9.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., … Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

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.

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.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research*, *5*, 69–95.

Johnson, E., & Zamuner, T. (2010). *Using infant and toddler testing methods in language acquisition research.*

Junge, C., Everaert, E., Porto, L., Fikkert, P., Klerk, M. de, Keij, B., & Benders, T. (2020). Contrasting behavioral looking procedures: A case study on infant speech segmentation. *Infant Behavior and Development*, *60*, 101448.

Ma, W., Golinkoff, R. M., Houston, D. M., & Hirsh-Pasek, K. (2011). Word learning in infant-and adult-directed speech. *Language Learning and Development*, *7*(3), 185–201.

ManyBabies Consortium. (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.

Miller, S. A. (2017). *Developmental research methods.* Sage publications.

Naoi, N., Minagawa-Kawai, Y., Kobayashi, A., Takeuchi, K., Nakamura, K., Yamamoto, J., & Shozo, K. (2012). Cerebral responses to infant-directed speech and the effect of talker familiarity. *Neuroimage*, *59*(2), 1735–1744.

Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173.

Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006).

Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology*, *42*(4), 643.

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469.

Schreiner, M. S., & Mani, N. (2017). Listen up! Developmental differences in the impact of IDS on speech segmentation. *Cognition*, *160*, 98–102.

Silverstein, P., Feng, J., Westermann, G., Parise, E., & Twomey, K. E. (2021). Infants learn to follow gaze in stages: Evidence confirming a robotic prediction. *Open Mind*, 1–15.

Singh, L., Nestor, S., Parikh, C., & Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy*, *14*(6), 654–666.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71. https://doi.org/10.1207/s15327078in0701_5

Zangl, R., & Mills, D. L. (2007). Increased brain activity to infant-directed speech in 6-and 13-month-old infants. *Infancy*, *11*(1), 31–62. https://doi.org/10.1207/s15327078in1101_2