Test-retest reliability in infant speech perception tasks

Alejandrina Cristia[1]

Amanda Seidl[2]

Leher Singh[3]

Derek Houston[4]

[1]Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS)

Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University

29, rue d'Ulm, 75005, Paris, France; alecristia@gmail.com; +33144322623

[2]Purdue University

3142 Lyles-Porter Hall, W. Lafayette, IN 47907, USA; aseidl@purdue.edu; +17654963863

[3]National University of Singapore

9 Arts Link, AS5 04-30, Singapore 117570, Singapore; psyls@nus.edu.sg; +6565167750

[4]Indiana University Medical School and The Ohio State University

915 Olentangy River Rd, Columbus, OH 43212 USA; houston.200@osu.edu; +13172744923

Abstract

A long line of research investigates how infants learn the sounds and words in their ambient language over the first year of life, through behavioral tasks involving discrimination and recognition. More recently, individual performance in such tasks has been used to predict later language development. Does this mean that dependent measures in such tasks are reliable and can stably measure speech perception skills over short time spans? Our three labs independently tested infants with a given task, and retested them within 0-18 days. Together, we can report data from 12 new experiments (total number of paired observations N=409), ranging from vowel and consonant discrimination to recognition of phrasal units. Results reveal that reliability is extremely variable across experiments. We discuss possible causes and implications of this variability, as well as the main effects revealed by this work. Additionally, we offer suggestions for the field of infant speech perception to improve the reliability of its methodologies through data repositories and crowd-sourcing.

Word count: Abot 8,000

Test-retest reliability in infant speech perception tasks

Hundreds of experiments have probed infants' abilities to discriminate sounds and recognize words in continuous speech. Results of these studies have delineated the development and changes in speech perception skills during the first two years of life. Furthermore, work has accumulated, particularly in the last ten years, suggesting that measures of speech perception gathered in infancy (including both sound discrimination and word segmentation) can predict later language (see a recent meta-analysis in Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014). The potential predictive value of infant speech perception measures gathered thus far strongly suggests that such skills could be a fundamental landmark in early language acquisition. But can laboratory-based measures of infant speech perception be used as part of a diagnostic battery for language delays or disorders? This goal would require that these measures be reliable within individual infants across testing occasions, a feature that has very seldom been explored within the domain of language processing. The aim of this article is to provide a joint report on the work of three different labs, each of which independently collected test-retest data on infant speech perception tasks. Together, these data provide a solid starting point in the search for reliable infant speech perception tasks by suggesting promising avenues and highlighting potential challenges.

While such work is rare in the field of early language acquisition, test-retest reliability questions have been more widely researched for other areas of developmental cognition. In a seminal study, Fantz (1964) reported that infants' attention towards a visual stimulus tended to decrease with repeated exposure; he found that infants reliably habituated, based on a sample of 28 infants between 6 and 25 weeks of age. In the 50 years that followed, literally thousands of children were tested with similar materials to document the stability, reliability, and predictive value of measures derived from this general finding, and to understand the constructs underlying the power of measures gathered with such a task (to name a very few recent examples, Colombo, Shaddy, Richman, Maikranz, & Blaga, 2004; Fagan, Holland, & Wheeler, 2007; Rose, Feldman, & Jankowski, 2003). The scale of this work confirms the usefulness of these information processing measures (for example, Visual Recognition Memory), which are more sensitive to risk and have a better predictive value with respect to childhood outcomes than the best established battery of standardized tests (Rose et al., 2003). However, this certainty has only been attained through the investigation of numerous implementations of the general habituation task, and through the accumulation of hundreds of studies, since individual experiments sometimes yielded spuriously large correlations (see

discussions in Bornstein & Sigman, 1986 and McCall & Carriger, 1993).

Few measures thought to be specific to language processing have been put to as rigorous trials as the aforementioned information processing ones. To our knowledge, the only study assessing the potential of a speech perception task to measure individual variation reliably is Houston, Horn, Qi, Ting, and Gao (2007). In this study 9-month-old infants were habituated to an audio-video presentation of a talker saying the non-word 'seepug'; once they habituated, they were presented with two types of trials. In one, the familiarized audio-video alternated with a novel 'seepug' video; in the other, it alternated with a video of the same talker saying the novel word 'boodup'. Results from 10 infants retested between 1 and 3 days after the first testing occasion revealed a robust preference for the trials containing the novel word on both testing days; and a significant correlation $r > .65$ between performance on the first and the second days.

In only one other experiment were infants retested on a similar task with the purpose of assessing changes in sound perception over time. Cardillo (2010) tested and retested 20 English-learning infants at 7 and 11 months of age on the vowel contrast /u-y/, using a variant of the Conditioned Head Turn paradigm. Her results reveal a (non-significant) correlation $r = .19$ between d' (a measure of discrimination) measured at each visit. It is unclear whether the lower correlation coefficient here reveals non-stability over longer time spans, or rather could be attributed to one of the methodological or conceptual differences across these two experiments. Indeed, at least one of the sounds used by Cardillo (2010) was non-prototypical in the infants' native language. Thus, it could have been predicted that infants would, as a group, become overall less sensitive to the contrast as they age and learn more about their ambient language.[1]

Given the importance of establishing whether infant speech perception measures can indeed be reliable and the scarcity of this key information,[2] we sought to contribute some much needed evidence on their possible psychometric value. Experimental design and data collection was done independently in three of our labs, which we will call "labA", "labB", and "labC". As a result, the experiments we report on

[1] It should be noted that this experiment reports a significant *in*crease in performance with age, although this might be related to the fact that Conditioned Head Turn is a rather challenging task cognitively, and children may fare better with it as they age.

[2] It is possible that other test-retest data exist but have not been rendered public. A public database on individual variation in infant speech perception was created recently (Cristia et al., 2014, invarinf.acristia.org), and submissions (including of unpublished effects) have been invited through posts in the main mailing lists of the field. Inspection of this database in August 2015 revealed no test-retest studies, other than the two we already discuss.

are more diverse than if a single person had designed them all, but all the experiments were carried out with similar behavioral methods, and they all aimed to assess the reliability in individual infants' speech discrimination or recognition scores. Given the similarity in procedures and objectives, we pooled our efforts to provide a single, maximally informative report. We reasoned that a joint report would both allow us to statistically address the robustness of the underlying reliability effects and provide a comprehensive overview for readers, performing the task of synthesizing results that would have been necessary if separate reports had been rendered public.

In putting together the present report, we have followed recent recommendations for a more accurate psychological science, including the incorporation of previous data into a cumulative meta-analysis (Braver, Thoemmes, & Rosenthal, 2014)[3], using confidence intervals rather than null-hypothesis testing (Cumming, 2014), preferring completeness over "neatness" of the results (Nosek, Spies, & Motyl, 2012), and reducing false positives by clearly separating main and exploratory analyses (Simmons, Nelson, & Simonsohn, 2011). Only the main planned analyses are reported here. The data and scripts used for the analyses reported here, as well as numerous additional figures, summaries, and exploratory analyses are accessible in a public scientific repository (https://osf.io/62nrk/).

Motivation of tasks and ages

All of these data have been gathered under the conceptual goal of assessing reliability of infant performance measures when performing speech perception tasks. We all favored behavioral tasks, either the central fixation or the headturn preference procedure, for several reasons: they have been widely used in the past; they place few motor and cognitive requirements on the child; and they are relatively inexpensive to set up. As a result, they could eventually be used by researchers and clinicians who are working with populations at risk. We all also focused on stimuli that infants (as a group) should be able to process at the ages tested, based on prior data. Finally, we all used short test-retest intervals, not only because that is what previous work had used (Houston et al., 2007's infants had been retested within three days), but also because there is a great deal of research showing that infant speech processing changes rapidly within the first year of life (see specific explanations in the individual experiments' reports, https://osf.io/rbhy3/).

[3]We have not, however, included Cardillo's data given the many differences in procedure and inter-test delay evident between all of the present work and hers.

While the conceptual goal was the same, our implementations were somewhat different because we did not become aware of each other's data until after it was collected. As a result, our three labs used different criteria when deciding on the tasks and ages.

LabA aimed to build directly on Houston et al. (2007), and thus used only discrimination tasks. One set of experiments explored an audio-only version of Houston's earlier study ('seepug' versus 'boodup'), and expanded the ages tested to a younger and an older group. Another set of experiments attempted to hone in on more specific aspects of speech processing. Since discrimination of those two dissimilar words could reflect consonant sensitivity, vowel sensitivity, or even sensitivity to the holistic form of the sound file, two subsequent studies focused on minimal pairs, recruiting infants at about 9 months of age for comparability with Houston et al. (2007). One of the experiments focused on /sa-ʃa/, a consonant contrast that has separately been found to reflect meaningful individual variation (Cristia, 2011). The other focused on the vowel contrast /i-u/.

LabB aimed to test infants in the cusp of developing specific linguistic skills. Therefore, infants were tested on their discrimination of a minimal vowel contrast embedded in /ʃip-ʃɪp/ at around 6 months in view of previous work stating that vowel perception becomes language-specific at around this age (Tsuji & Cristia, 2014). Consonant discrimination has been thought to become specialized to the native language only by 10-12 months (Werker & Tees, 1984), and thus 11-month-olds were tested on /sip-ʃip/ (the choice of the contrast independently inspired by Cristia, 2011's results). Two other benchmarks were explored, namely prosodic processing and word recognition. For the prosodic processing task, 6-month-olds were tested on their preference for well-formed intonational phrases, following previous work showing language-specific processing of acoustic cues to prosodic boundaries at about this age (see Johnson & Seidl, 2008 and citations therein). Finally, wordform recognition could have been studied at a range of ages, but given that two tasks (one on vowels, one on phrases) had been carried out with 6-month-olds, this same age group was chosen. As a consequence, an age-appropriate paradigm was adopted by (1) aligning the target word with a phrase edge; and (2) using passages for the familiarization, and bare words for the test phase (Seidl & Johnson, 2006).

The study in LabC was inspired by a different strand of literature, documenting the predictive value of word segmentation (Newman, Ratner, Jusczyk, Jusczyk, & Dow, 2006). Following this work, the headturn preference procedure was employed, first familiarizing infants with a pair of words until they

reached a threshold, and then testing them on their recognition of the same words embedded in passages at test. In view of the interest in prediction, and unlike all of the other experiments we report on, the design was longitudinal, meaning that the same infants were tested and retested at three ages, for a total of 3 paired observations for each child. These three ages were selected on the basis of key benchmarks of word processing that have been described previously: 7.5 months (clear signs of word segmentation); 9 months (word segmentation despite pitch changes); and 11 months (flexible word segmentation; see Singh, White, & Morgan, 2008; Singh, 2008; Singh, Steven Reznick, & Xuehua, 2012).

To sum up, we report data from twelve experiments, that we identify by combining the procedure, the linguistic level targeted, the age group when relevant, and the lab:

- h-vowel-LabA (habituation and dishabituation using two different vowels),
- h-vowel-LabB (habituation and dishabituation using two one-syllable words differing only in their vowels),
- h-fric-LabA (habituation and dishabituation using two one-syllable words differing only in the initial fricative),
- h-fric-LabB (habituation and dishabituation using two one-syllable words differing only in the initial fricative),
- h-word1 through 3-LabA (habituation and dishabituation using two different wordforms, and this at 3 different ages),
- f-word-LabB (familiarization and test of a word across isolation and passage contexts),
- f-word 1 through 3-LabC (familiarization and test of a word across isolation and passage contexts, and this at 3 different ages), and
- f-phrase-LabB (familiarization and test of a prosodically well- versus ill-formed phrase).

Methods

In this section, we provide general methodological aspects that are common or diverse across our experiments. Further information can be found in reports focusing on each experiment individually (https://osf.io/rbhy3/).

Stimuli

The visual stimuli (be it images or blinking lights) was not central to the studies' goals but rather served to provide the child with a fixation point.

As for auditory stimuli, all of the experiments focused on speech. Most of them used natural speech, the only exceptions being the vowel and sibilant experiments in LabA. As mentioned previously, the different experiments varied in terms of the linguistic level targeted, and the stimuli varied accordingly. Thus, all of the experiments that focused on vowel or sibilant processing used monosyllabic isolated words. In contrast, all of the word experiments except for one (h-word-LabA) employed to a certain extent connected speech, as did the phrase experiment.

In all cases, irrelevant dimensions (such as amplitude and interstimulus intervals) were matched across different types of stimuli used in the same experiment.

Procedure

The procedure was essentially the same for all experiments in labs A and B, which used the Central Fixation paradigm. The infant sat on a caregiver's lap, viewed neutral visual stimuli presented on a screen and heard auditory stimuli from speakers placed in the same approximate location as the screen. Throughout all phases, the infant's attention was directed to the screen before each trial using an attention-getter visual stimulus, such as a silent video of a laughing baby. When the infant fixated the screen, the trial began, and continued until the infant looked away for longer than 1 s or the maximum trial duration had been reached. If the infant looked away and back in less than that criterion duration, the trial did not stop but this time was not counted as looking.

In LabC, the headturn preference procedure was used. The infant sat on a caregiver's lap and their attention was attracted by one front and two side lights. When looking at a side light, the infant heard auditory stimuli from speakers placed in the same approximate location as the light itself. Throughout all phases, the infant's attention was directed to the central light before each trial, and when the infant looked a side light was turned on instead. When the infant fixated on it, the trial began, and continued until the infant looked away for longer than 2 s or the maximum trial duration had been reached. As before, looks away shorter than the criterion duration were not counted towards the overall looking time for that trial.

All experiments had two key phases: exposure, and test. Depending on how the exposure phase was

terminated, experiments can be classified in two types, “habituation" and “familiarization". In the former, the exposure phase was interrupted when a habituation criterion was met through an average looking time decrement, whereas in the latter it was based on total exposure length.

In all studies, the test phase contained somewhat familiar and somewhat novel stimuli, although the precise implementation of these two types varied. For instance, in all habituation experiments, the test phase consisted in 14 pseudo-randomized trials: 10 old trials, when the habituation stimuli was presented again, and 4 $novel$ trials, when a novel and the habituation auditory stimuli were presented in alternation. In contrast, the test phase in LabC experiments presented infants with 3 times 4 passages, some of which contained the word that had been presented during familiarization, and others did not.

It should be noted that looking time data was sometimes available for other trial types, such as pretest trials in the habituation experiments, and an intermediate familiarity category at test in LabC. To avoid an explosion in terms of the dependent measures considered, and for comparability across all experiments, we have focused on one derived measure which best captures the linguistic process of interest, namely performance during the test phase when comparing the familiar and novel trials.

Participants

All experiments have been approved by appropriate institutional review boards, and written parental consent was obtained for each participating infant. Key characteristics of the final samples of infants included in each experiment are shown in Table 1. In all experiments, infants were full-term, normally-hearing, and exposed to English at least 90% of the time (with the exception of 4 bilinguals in the f-word-LabC series). Ages of infants in the studies ranged from 5 to 12 months of age; the precise distribution for each experiment is shown in Table 1.

Although the stopping rule was not decided in advance, stopping did not occur because of our inspection of statistical results. Sample sizes depended mainly on the availability of resources to recruit and test participants. We point out that three of the experiments have Ns that are below current recommendations (e.g., 20 according to Simmons et al., 2011). Nonetheless, we have included these data because they are not smaller than previously published comparable data (Houston et al., 2007) and because our meta-analytic approach allows us to integrate these results down-weighting them in accordance to their small sample size.

The separation between testing occasions was most often 1-3 days, but it varied depending upon family availability, with slightly longer separations when the family could not come back within that period (maximum 18 days). There were two exceptions to this general pattern. First, 3 children in h-fric-LabA and 1 in h-word-LabA were tested twice on the same day. Nonetheless, we will speak of performance on day 1 and day 2 for *a*ll experiments. The second exception occurred in the h-vowel-LabA experiment, where infants were randomly assigned to one of three inter-test separations: 1-2 days, 3-7 days or 13-15 days.[4]

Finally, 69 infants participated in more than one experiment. Since this cannot be taken into account in the meta-analytic analyses, we have complemented them with a mixed linear regression model, which yielded the same results. This is not reported in detail for conciseness, but interested readers can download data and sample code from the section of our Supplementary Materials that is focused on the analyses (https://osf.io/gen2f/).

Dependent measures and analyses

All analyses were carried out in R (R Core Team, 2015). In all experiments, a preference quotient (PQ) was estimated for each day separately on the basis of looking times to the two main types of trials presented during the test phase. In habituation experiments, the PQ was (LT_n - LT_o)/(LT_n + LT_o), where LT_n is the average looking time during new trials and LT_o is the looking time during old trials.[5] Our PQ measure can take a value between -1 and 1, with negative numbers indicating preference for familiarity and positive ones preference for novelty. Given the habituation-dishabituation design, positive outcomes (novelty preferences) are expected.

In familiarization experiments, the PQ was (LT_u - LT_f)/(LT_u + LT_f), where LT_u is the average looking time during mismatch or novel trials and LT_f is the looking time during match or familiar trials. It has the same properties mentioned before, except that since habituation has not been ensured, novelty is not a necessary outcome. On the contrary, previous recognition work often reveals familiarity preferences,

[4]In exploratory analyses available from https://osf.io/brge5/, we show that the separation between test and retest significantly modulates the association strength in performance across days, with stronger correlations for short delays (0-9 days) than longer ones.

[5]This implementation follows work on lateralization of brain function (e.g., Chlebus et al., 2007) rather than Houston et al. (2007). The denominator in the 2007 study was the average looking time across test trials. This definition does not fit in with our intuition of novelty preference because it compares a difference looking times against an overall average where old and novel trials are not represented equally (the overall average contains 10 old and 4 novel trials).

which would be evidenced in negative PQs.

To investigate stability in individual performance, we calculated Pearson's correlations on the individual day 1 and day 2 scores, for each experiment separately. The weighted mean correlation was then estimated across all experiments using the package metafor (Viechtbauer, 2010). Given the relatively small samples of most of the experiments, correlations were combined using the Fisher's z transformation. Since the experiments differed on key design characteristics, a random effects model was fit (as the experiments could not be viewed as strict replications of each other), using the DerSimonian-Laird estimate (see page 47 of Schwarzer, 2007). A test for moderators was carried out with the same metafor package.

Since correlations in individual performance varied significantly across the different experiments, we also report performance in each experiment. To convey group effects, we calculated effect sizes as the mean PQ divided by the standard deviation across children for each day separately. Finally, we carried out a number of exploratory analyses, which are available in the online supplementary materials (https://osf.io/brge5/).

Summary of similarities and differences across experiments

As noted above, all of these data were gathered with the same goal: Establishing the short-term test-retest reliability of individual infants' performance in speech perception tasks. All of the studies have been gathered using the same general procedure, whereby the child controls auditory stimuli presentation via their fixation on a visual stimulus that is unrelated to the auditory ones. Infants were tested and retested with the same experiment, and the two testing occasions were separated by relatively short intervals.

The most salient differences across the twelve experiments involve the ages of the participants, the linguistic level targeted, and the lab that designed and gathered the data. There are other methodological differences across two or more of the experiments, but to facilitate the discussion, we relegate the report of these experiment-specific characteristics to the individual reports in our online supplementary materials (https://osf.io/rbhy3/).

Before carrying out a meta-analysis, one must decide whether the studies included are similar enough to avoid "mixing apples and oranges". In any case, it should be clear from this description that the experiments are not straightforward replications of one another, but they are extremely similar nonetheless. The diversity found here is clearly no greater than that in other excellent meta-analyses. To take a classical

example, Glass and Smith (1979) looked at the relationship between number of students in a class and academic achievement “mixing" together older and younger students, poorer and richer schools, those that followed a child-centered or a centralized curriculum. All of these classrooms were undoubtedly unique; yet the meta-analytic method allows us to study an effect of theoretical and practical interest despite this variation.

We take into account variance across experiments through two analysis choices. First, we fit a random effects, rather than a fixed effects, model, to signal that studies are not strict replications of each other. Second, studies are weighted based on their sample size, with more importance given studies with larger Ns. The diversity in implementations present in our data is an asset rather than a liability, because we are interested in finding out, and communicating to our readers, a realistic estimate of the reliability effect, e.g., to what extent they themselves can expect to see test-retest reliability if they share our same conceptual.

Results

We report next on the main effects within each experiment and day, first, and on test-retest correlations, second. Our online supplementary materials contain additional analyses for each experiment separately (https://osf.io/rbhy3/) as well as some further cross-lab analyses (https://osf.io/brge5/). Interested readers are strongly encouraged to download the data and analysis scripts for further exploration (available from https://osf.io/gen2f/).

Main effects within each experiment and day

Table 2 shows effect sizes separated by day for each experiment, as well as other relevant information. Overall, there is considerable variation across experiments, justifying our choice of a random effects model.

Absolute effect sizes in habituation experiments tend to be larger, on average, than those in familiarization experiments. Although p-values cannot be directly compared across studies here, given the large diversity in sample size, we point out that most habituation experiments elicited a significant preference at test on both days, with the exceptions affecting the studies focused on the consonantal contrast /s-ʃ/. In contrast, results in group performance were more variable among familiarization studies.

It may seem that absolute effect sizes are somewhat larger on day 2 than on day 1, but the only significant difference in individual's PQs occurs in h-fric-LabA (and probably relates to the effect size reversing sign on day 2).

Meta-analysis of test-retest reliability

We fit a random effects meta-analysis on z-transformed correlations using sample size weighting. The weighted mean was 0.065, with a 95% confidence interval of [-0.121; 0.251], which was not different from zero ($z = 0.686$, $p = 0.49$). Results also indicated that there was significant heterogeneity among the datapoints ($Q(12) = 33.957$, $p = 0.0007$) with about 64.66% of the total variance being attributable to diversity across the studies. Indeed, the forest plot shown in Figure 1 illustrates that some of the confidence intervals for individual experiments fail to overlap.[6]

Test-retest reliability within each experiment

Given the significant heterogeneity found in the meta-analysis and the observed variation across experiments, we describe results from each experiment in more detail. Correlations and confidence intervals are provided in Table 3. In all, the confidence intervals for five experiments did not include zero, and surprisingly in only three of the cases was the correlation significantly positive (h-vowel-LabA, f-word-LabB, and the published Houston2007).

The other two experiments with significant correlations, but in the unexpected negative direction, were h-sib-LabA and f-word2-LabC. Regarding the former, given the considerable sample size, we are inclined to believe that this is a valid result, particularly since the other experiment bearing on the same consonantal contrast also yielded a correlation coefficient that was negative even though it had been carried out with different stimuli and in a different lab (h-sib-LabB). The same cannot be said of f-word2-LabC, which is exactly like f-word1 and 3 from the same LabC, in experimental terms, and can be conceptually likened to f-word-LabB, which yielded a significant positive correlation.

Data were insensitive for the other experiments, because the confidence intervals included zero and spanned a large range of both positive and negative potential correlation coefficients. For the LabB studies,

[6] We also carried out an analysis declaring study type (habituation versus familiarization) as moderator, which was not significant $QM(1) = 0.474$, $p = 0.491$. Similarly, the test for moderator was not significant in an analysis declaring stimulus type (vowel, sibilant, word, phrase) [$QM(3) = 2.942$, $p = 0.401$] nor was it in an analysis declaring lab [$QM(3) = 4.536$, $p = 0.209$].

this may be due to the small sample size and low power, but the same explanation cannot be extended to h-word1 through 3-LabA, which are similar to the published experiment (Houston et al., 2007), except that discrimination was purely auditory and the sample size had been increased 2- to 3-fold.[7] Finally, one may wonder whether there is some relationship between effect size and stability in performance. Comparison of r values and t values in Table 2 suggest no clear association.

Discussion

The measurement of individual variation and its stability over infancy and early childhood has proved fruitful in other cognitive domains, as it contributes to the theoretical goal of revealing the stability and emergence of constructs like attention and memory, as well as the applied goal of predicting cognitive and educational outcomes. In contrast, the investigation of infant speech perception measures lags behind, with only a handful of studies beginning to explore the potential stability and predictive value of infant perception landmarks. Within this line of work, we are the first to document the reliability of a variety of measures and testing procedures across short test-retest periods. Overall, our results underline the diversity in stability that can be obtained.

Interpreting variability in reliability

Some readers may wonder if the studies differ in quality, reasoning that if a correlation is not significant (or it is significant, but in the “wrong" direction - a question to which we return below), then perhaps the experimental setup or the stimuli were somehow bad. It is very common to hear statements by infant researchers about, for example, direction of preference not mattering, and only its significance. Indeed, significance is viewed as a badge for quality of the data.

However, such viewpoints are increasingly under attack in other scientific traditions. This is clearest when one considers medical research, where Ioannidis (2005) and others have quite clearly laid out the dangers of confusing significance with quality and ignoring null results, thus misguiding decisions that

[7] Notice that this should not be viewed as a failed replication, given both the differences in procedure and the fact that the correlation observed in the current experiment is included in the confidence interval of the previous studyâĂŹs correlation estimate. The issue of how to decide whether a study has been replicated successfully or not is one that is receiving a great deal of attention in the psychological sciences at present and is leading to a refinement of the criteria for replication: see for instance Klein et al. (2015).

have an impact on both health and public spending. In a nutshell, if we only select significant results, then the resulting literature is intrinsically biased, and it will contain many more false positives than if we valued experiments based on their methods (Sterling, 1959) – and, some years ago, this bias was found to be even greater in psychological than medical journals (Sterling, Rosenbaum, & Weinkam, 1995). It is important to bear in mind that significant results can be “duds" (being, in fact, false positives) and null results can be true (indicating, in fact, that there is no difference between conditions, or no correlation between two variables).

The problem of bias takes center stage today in several major journals, causing a move towards the evaluation of manuscripts (and data more generally) on the basis of methods and not results. In the article explaining *Psychological Science*'s shift away from traditional practices based on p-values, Cumming (2014) explains that “To ensure [that published literature be complete, coherent, and trustworthy], we must report all research conducted to a reasonable standard, and reporting must be full and accurate." Using quality metrics, it would have been impossible for us to predict which one of our experiments would have stronger (or positive) correlations before actually seeing the results. Indeed, all of our labs are productive, and we have well-honed recruitment and testing procedures. Moreover, in addition to this fact that vouches for the general quality of our work, we even have some “twin" studies in the present dataset that yielded very dissimilar results, such as h-word-LabA and h-fric-LabA, which differ only in the stimulus used, and yield significant results in *opposite* directions.

To a certain extent, this variability could be due to different practices in our different labs, and to variance intrinsic in the infant population we seek to study – but if so, it is crucial to take such variation into account explicitly since it could directly impact replicability. So how can our field move forward?

Given the importance of the topic, the cost to the individual labs, and the fact that multi-site data would speak to the robustness and validity of effects, we propose that the next step should be towards crowd-sourcing reliability experiments. For example, the first “Many Labs initiative" aimed to assess the replicability of 13 influential effects in the social sciences (Klein et al., 2015). Researchers in 36 sites volunteered, collecting data in a coordinated manner, and eventually jointly contributing data for a total N of over 6000 adult participants. This effort not only allowed to determine which, of 13 key results, were in fact replicable, but it also provided researchers with a more realistic estimate of the effect size they might obtain *once site and participant variation is taken into account*. This is crucial if we hope, at any point, to develop individual measures of performance for use as diagnostic.

An alternative to organized crowd-sourcing is that individual researchers systematically contribute their results to repositories. For instance, for individual variation (including test-retest) there exists an appropriate public repository, which accepts both published and unpublished submissions (invarinf.acristia.org, see Cristia et al., 2014 for further details). We believe it is possible that there are other test-retest studies that have been carried out elsewhere and that have yielded varied results, as ours did. If so, it is entirely plausible that the lead researchers have refrained from submitting their data for publication. And yet it would be incredibly informative for those of us who do venture in this difficult territory to know exactly what has been done, so that we avoid the methodological choices, ages, or tasks that are not conducive to reliability. This is precisely the biggest risk that Sterling identified, over 50 years ago, whereby under-reporting of null results lead to squandering of resources (since the same study is unknowingly conducted by other researchers, until a false positive is encountered and published as “true"; Sterling, 1959). In the long run, the short time investment required to report unpublished work in a repository would save a great deal of resources to the community, as well as potentially highlight promising and challenging avenues for the important goal of establishing reliable infant speech perception tasks.

What do our strongest results tell us?

In this subsection, we would like to speculate as to the specific factors that lead to strong positive test-retest correlations based on the insights from the 13 test-retest experiments we have discussed. Before doing so, we would like to point out once more that the experiments were carried out in different labs, on different age groups, and with slightly different methods. This variation was, in part, not parametric because we were unaware of each other’s work prior to the data being fully collected. Therefore, some factors are at present confounded, and further work is needed to tease them apart.

Nonetheless, we believe the present set of experiments already provide us with interesting clues about which directions to explore further and which to avoid in the quest for large positive test-retest correlations. We would like to argue that, by and large, the central fixation setup using a habituation procedure is to be preferred because it leads to more consistent and larger preferences at the group level. In addition, as mentioned above, this procedure places low motoric and cognitive demands on the child and it is inexpensive to set up, thus lending itself to more widespread use than, for example, the headturn preference procedure.

As for the stimuli and age group to be targeted, our results certainly show that this is a crucial dimension, which can make or break a study (compare h-vowel-LabA and h-fric-LabA, differing only on the stimuli, yet yielding correlations in opposite directions). In a very tentative manner, we would like to advance the speculation that test-retest positive correlations are more likely if infants are asked to discriminate more phonologically distinct items. This description would encompass the whole-word discrimination studies (Houston et al., 2007 and h-word1 through 3-LabA - which were also positive, albeit non-significant) as well as the vowel discrimination study using the distinct vowels /i/ and /u/ (h-vowel-LabA).[8] Our description is aimed to exclude the experiments focusing on sibilants and the /i-ɪ/, admittedly a subtle vocalic contrast.

Bear in mind that much current work on infant predictors focuses on very subtle contrasts in the hope of tapping language-specific processing (cf. the meta-analysis in Cristia et al., 2014). If our interpretation of factors leading to higher reliability is correct, then speech tasks as they are being used today would provide a suboptimal diagnostic tool. This is just a speculation, and it is an empirical question that requires more research than that presented here, in order to determine whether there are implementations of discrimination tests that (a) are robust in terms of test-retest reliability, (b) measure emergent linguistic knowledge (rather than, e.g., general lab performance); and (c) are strong predictors of later language, thus making a strong and specific diagnostic tool.

In fact, a comparison with Visual Recognition Memory (VRM; e.g., Rose et al., 2003) is useful at this point. This cognitive measure has remained a standard and powerful measure for several decades. The regular VRM battery has 9 distinct 'problems'. In each one, the child is familiarized with a visual stimulus and subsequently presented with this same stimulus paired with a new one, and a novelty preference score is calculated for each problem and then averaged across all 9 problems. Each problem uses a different stimulus, and the stimuli are drawn from different categories (e.g., faces, geometric shapes); and different familiarization and test times are employed. Clearly, this is totally unlike our habituation-dishabituation paradigm, particularly that which uses only two syllables presented dozens of times within a couple of minutes, or our word segmentation experiments, where the same infants heard the same passages many times.

[8] All of these studies were carried out with infants about 9 months of age – but this is probably not a key factor, since several other studies with lower and negative correlations also targeted this age group, most saliently h-fric-LabA.

An avenue of research we hope future work explores involves compounding the most promising measures into a $single$ study, perhaps combining both audiovisual word discrimination and /i-u/ discrimination. The idea of compounding the tasks, along the lines of the VRM battery, entails presenting infants with streamlined familiarization times (which have been independently varied to assess their effects on group performance) and fewer test trials in order to produce an overall shorter test.

We are not, at present, suggesting that a study like f-word-LabB, which revealed a significant positive correlation, be included in such a battery because this was a familiarization-style study which targeted word segmentation. Yet two out of the three word segmentation studies carried out in LabC yielded a $negative$ correlation, and the last one was very close to zero. At the very least, this indicates that word segmentation studies can yield very variable results across laboratories, and are therefore a risky choice in the quest for test-retest reliability at present.

Turning now to our strongest results on the opposite side of the spectrum, we also found five $negative$ correlations in performance, two of which were significant. We do not have a single interpretation to provide that could encompass all of the data, since the experiments vary in both procedure (3 using central fixation, 2 the headturn preference procedure) and stimuli (2 on sibilants, 2 on word segmentation, and 1 on phrase well-formedness). Conceptually, one could imagine that negative correlations emerge if the infant remembers the preceding test phase items and, having listened longer to the novel than the habituated category, now turns attention to the habituated category. This interpretation is only valid if the habituation and novel stimuli are kept the same between test and retest – but for the two negative correlations that are significant, the order was switched at retest.[9] Although interpreting such negative correlations is challenging at present, we think it is crucial to present them in the public light, to avoid misleading the community into thinking that they do not exist.

A final question to address when discussing promising avenues for test-retest reliability relates to what can be expected in terms of group performance. Before analyzing our data, some of us believed that it was desirable to have large group effects, because this would indicate that the task as a whole was clear to the children, whereas small effects might indicate random performance. Others among us worried that very large effects would indicate ceiling performance, leading to reduced inter-individual variability. Neither of

[9]For further discussion, see individual reports available from https://osf.io/rbhy3/, and targeted investigation of mediation through habituation times available from https://osf.io/brge5/.

these interpretations were supported when we attempted to relate group effect sizes and test-retest correlations. To say it more clearly, it is not the case that studies where the main effect is significant yield larger test-retest correlations. It is also not the case that studies with a great deal of internal variability lead to larger test-retest correlations. Thus, test-retest reliability needs to be studied on its own right, as it cannot be guessed at from group-level effects.

Conclusions

The present paper summarizes the results of 12 experiments, carried out independently in three labs. The main finding emerging from this work is that it is not the case that one can simply take any well-described experimental task and expect individual performance across two test days to correlate positively. Additionally, we have discussed some promising avenues for future research aiming at developing infant speech perception tasks that lead to stable performance using a procedure that does not place great demands on the infant, and is low cost and easy to implement. In this quest, we believe that cross-lab collaboration, and potentially even large-scale crowd-sourcing, is the most promising way of being able to pin down the characteristics of reliable tasks that can be successfully used in different labs and environments.

References

Bornstein, M. H., & Sigman, M. D. (1986). Continuity in mental development. *Child Development*, *57*, 251-274.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342.

Cardillo, G. C. (2010). *Predicting the predictors*. University of Washington.

Chlebus, P., Mikl, M., Brázdil, M., Pažourková, M., Krupa, P., & Rektor, I. (2007). fMRI evaluation of hemispheric language dominance using various methods of laterality index calculation. *Experimental brain research*, *179*(3), 365–374.

Colombo, J., Shaddy, D. J., Richman, W. A., Maikranz, J. M., & Blaga, O. M. (2004). The developmental course of habituation in infancy and preschool outcome. *Infancy*, *5*, 1-38.

Cristia, A. (2011). Fine-grained variation in caregivers' speech predicts their infants' discrimination. *Journal of the Acoustical Society of America*, *129*, 3271-3280.

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.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Fagan, J. F., Holland, C. R., & Wheeler, K. (2007). The prediction, from infancy, of adult IQ and achievement. *Intelligence*, *35*, 225-231.

Fantz, R. L. (1964). Visual experience in infants. *Science*, *146*, 668-670.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational evaluation and policy analysis*, *1*(1), 2–16.

Houston, D., Horn, D. L., Qi, R., Ting, J., & Gao, S. (2007). Assessing speech discrimination in individual infants. *Infancy*, *12*, 119–145.

Ioannidis, J. P. (2005). Why most published research findings are false. *Chance*, *18*(4), 40–47.

Johnson, E. K., & Seidl, A. (2008). Clause segmentation by 6-month-old infants: A crosslinguistic perspective. *Infancy*, *13*(5), 440–455.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., … others (2015). Investigating variation in replicability. *Social Psychology*.

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development*, *64*, 57-79.

Newman, R. S., 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*, 643–655.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org`

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2003). The building blocks of cognition. *Journal of Pediatrics*, *143*, S54-S61.

Schwarzer, G. (2007). Meta: An R package for meta-analysis. *R News*, *7*, 40–45.

Seidl, A., & Johnson, E. (2006). Infant word segmentation revisited. *Developmental Science*, *9 (6)*, 565-573.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359–1366.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*(2), 833–870.

Singh, L., Steven Reznick, J., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development. *Developmental science*, *15*(4), 482–495.

Singh, L., White, K. S., & Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input. *Language Learning and Development*, *4*(2), 157–178.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American statistical association*, *54*(285), 30–34.

Sterling, T. D., Rosenbaum, W., & Weinkam, J. (1995). Publication decisions revisited. *The American Statistician*, *49*(1), 108–112.

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels. *Developmental Psychobiology*, *56*(2), 179–191.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Werker, J. F., & Tees, R. (1984). Cross-language speech perception. *Infant Behavior and Development*, *7*, 49–63.

Table 1

Background information on included participants, broken down by experiment. Age refers to the age in months at the first test. Separation indicates the number of days separating the two visits. NA indicates that this information is not available.

		Age (months)		Separation (days)	
Study	N	Mean	SD	Mean	SD
h-vowel-LabA	58	9.64	1.04	7.03	5.11
h-vowel-LabB	17	5.89	0.38	3.82	3.97
h-sib-LabA	89	8.72	1.48	4.37	3.13
h-sib-LabB	10	11.08	0.32	5.2	3.99
h-word1-LabA	30	6.08	0.48	2.1	2.52
h-word2-LabA	18	9.03	0.39	1.89	0.9
h-word3-LabA	17	11.58	0.21	2.12	0.86
f-word-LabB	30	5.98	0.33	3.37	2.93
f-phrase-LabB	10	6.2	0.2	3.2	2.3
f-word1-LabC	40	7.56	NA	3.67	1.83
f-word2-LabC	40	10.03	NA	1.85	1.31
f-word3-LabC	40	11.05	NA	1.8	1.18

Table 2

Effect sizes (ES), t from a t-test against chance preference (zero) and its p-value, for the first and second tests (day). Across days: Pearson correlations in individual performance and its p-value; t from a paired t-test across days and its p-value. All statistics are rounded to two decimals.

	Day 1			Day 2			Across days			
Experiment	ES	t	p	ES	t	p	r	p	t	p
Habituation experiments										
h-vowel-LabA	0.791	6.03	0	1.127	8.58	0	0.31	0.02	1.59	0.12
h-vowel-LabB	0.814	3.36	0	1.338	5.52	0	0.06	0.82	1.34	0.2
h-sib-LabA	0.11	1.03	0.3	-0.221	-2.09	0.04	-0.23	0.03	-2.04	0.04
h-sib-LabB	0.308	0.98	0.35	0.366	1.16	0.28	-0.32	0.37	0.26	0.8
h-word1-LabA	0.868	4.76	0	1.28	7.01	0	0.28	0.14	1.44	0.16
h-word2-LabA	0.942	4	0	0.861	3.65	0	-0.17	0.5	-0.33	0.75
h-word3-LabA	0.976	4.02	0	0.913	3.76	0	0.36	0.16	-0.09	0.93
Familiarization experiments										
f-word-LabB	-0.281	-1.54	0.14	-0.448	-2.45	0.02	0.43	0.02	-0.64	0.53
f-word1-LabC	-0.427	-2.7	0.01	-0.798	-5.05	0	-0.02	0.92	-0.43	0.67
f-word2-LabC	-0.135	-0.85	0.4	-0.719	-4.55	0	-0.41	0.01	-1.19	0.24
f-word3-LabC	-0.389	-2.46	0.02	-0.66	-4.17	0	0.05	0.75	-0.93	0.36
f-phrase-LabB	-0.607	-1.92	0.09	0.175	0.55	0.59	-0.17	0.63	1.64	0.13
Previous study (habituation)										
Houston2007	2.665	8.43	0	2.453	7.76	0	0.7	0.02	0.36	0.73

Table 3

Key information for the meta-analysis integration: rz indicates z-transformed correlation coefficient; 95%-CI the 95 percent confidence intervals. The asterisk facilitates the detection of experiments where confidence intervals did not overlap with zero.

Experiment	rz	95%-CI
h-vowel-LabA	0.32*	[0.06; 0.58]
h-vowel-LabB	0.06	[-0.46; 0.58]
h-sib-LabA	-0.23*	[-0.45; -0.02]
h-sib-LabB	-0.33	[-1.07; 0.41]
h-word1-LabA	0.29	[-0.09; 0.66]
h-word2-LabA	0.17	[-0.68; 0.33]
h-word3-LabA	0.38	[-0.15; 0.90]
f-word-LabB	0.46*	[0.08; 0.84]
f-word1-LabC	-0.02	[-0.34 , 0.30]
f-word2-LabC	-0.44*	[-0.76 ,-0.11]
f-word3-LabC	0.05	[-0.27 , 0.37]
f-phrase-LabB	-0.17	[-0.91 , 0.57]
Houston2007	0.87*	[0.13; 1.61]

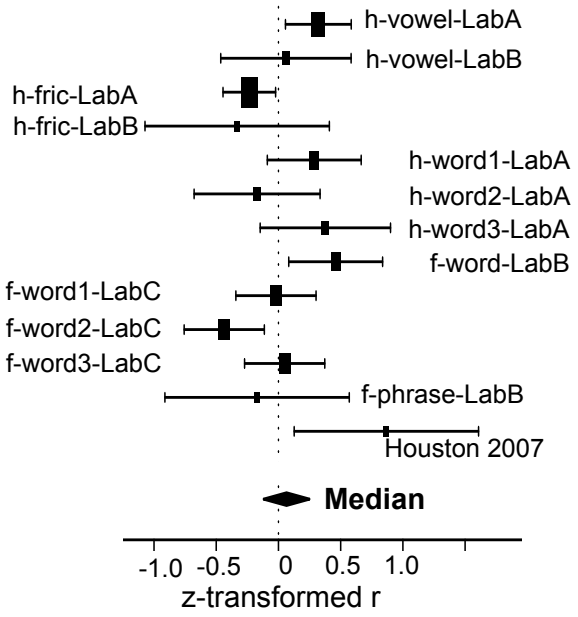

Figure 1. Forest plot of test-retest reliability effect sizes. Each line is an experiment, for which we show its effect size and 95% confidence interval (black simple lines) and its weight on the regression (through the size of the square). The black diamond indicates the weighted median effect size, with the edges of the diamond marking the 95% confidence interval.