UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Rocha, S., Ní Choisdealbha, Á., Attaheri, A., Mead, N., Olawole-Scott, H., Grey, C., Williams, I., Gibbon, S., Boutris, P., Brusini, P., Brough, C., Alfaro e Oliveira, M., & Goswami, U. (2024). Language Acquisition in the Longitudinal Cambridge UK BabyRhythm Cohort. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.92998

Developmental Psychology

Language Acquisition in the Longitudinal Cambridge UK BabyRhythm Cohort

Sinead Rocha[1,2][a], Áine Ní Choisdealbha[1], Adam Attaheri[1], Natasha Mead[1], Helen Olawole-Scott[1], Christina Grey[1], Isabel Williams[1], Samuel Gibbon[1], Panagiotis Boutris[1], Perrine Brusini[1], Carmel Brough[1], Maria Alfaro e Oliveira[1], Usha Goswami[1]

[1] University of Cambridge, Cambridge, UK, [2] Anglia Ruskin University, Cambridge, UK

Keywords: infant, language development, longitudinal, temporal sampling framework, brain-behaviour analyses

https://doi.org/10.1525/collabra.92998

Collabra: Psychology

Vol. 10, Issue 1, 2024

The Cambridge UK BabyRhythm project is a study of 122 infants as they age from 2 – 30 months, investigating cortical tracking and sensorimotor synchronisation to acoustic and visual rhythm in relation to language acquisition. As there are few standardised language tasks appropriate for this age range, the BabyRhythm project adapted a range of parent-report and infant-led experimental measures that could be used within a home testing environment. Here we present a rich description of infant performance on tasks intended to sample 5 linguistic domains: semantics, phonology, grammar, rhythmic timing and gesture. For each task we describe infant performance (mean, median, range), and we also report performance by sex (N female = 57) and by monolingual (N = 91) versus multilingual (N = 31) home environments. We report relations between measures. We share our unique longitudinal database (all data available on OSF), and 'lessons learned' on adapting language assessments for very young children. Critically, we identify the language tasks that will be utilised in our longitudinal brain-behaviour analyses, providing the benchmark upon which future neural and behavioural markers will be measured.

The Cambridge UK BabyRhythm project is a longitudinal study of language acquisition in typically-developing infants that has been designed to test the Temporal Sampling theory of individual differences in language acquisition (Goswami, 2011, p. 2020). Temporal Sampling (TS) theory is a sensory-neural theory concerning the role of neural rhythmic oscillatory alignment (neural entrainment) and motor rhythmic synchronisation to speech in building a nascent language system (Goswami, 2011, p. 2020). Rhythm has long been described as a universal precursor of language acquisition (Mehler et al., 1988). TS theory provides a mechanistic framework regarding how neural rhythms and sensory processing of rhythm (acoustic, visual and motor) may contribute to language development. TS theory integrates the neural multi-sensory resolution processing theory of adult speech processing (e.g. Poeppel, 2003) with known developmental features of language acquisition (e.g. Kuhl, 2004). Neural multi-sensory resolution processing theory proposes that parallel encoding of multi-modal speech information at a range of timescales with millisecond accuracy by cortical oscillations is necessary for efficient speech comprehension (Giraud & Poeppel, 2012). TS theory argues that infants may differ intrinsically in the accuracy with which this automatic temporal (speech-brain) alignment is achieved, at one or more timescales, thereby affecting the developmental trajectory for language acquisition. To test this proposal, the BabyRhythm project has gathered a series of neural entrainment and motor synchronisation measures during the first 2 – 11 months of life for 122 infants, and has also administered a series of language outcome tasks from 8 months until age 2.5 years. We have already described brain (Attaheri et al., 2022) and behavioural (Rocha et al., 2021) indices of rhythm perception and production. Here we describe the language tasks used in the project by the 24 months test point. We provide a rich, open dataset of typical language development over the first two years of life, and critically, select those tasks that appear most robust as a basis for longitudinal analyses of brain-behaviour relationships.

Classically, environmental contributions to individual differences in language acquisition have received more research attention than neural differences, largely because they are easier to measure. It is now well-established that individual differences in both the quantity of language that an infant hears, and in the quality of that language (amount of infant-directed speech, IDS), make significant contri-

[a] Corresponding author: Sinead Rocha. Email: sineadrocha@gmail.com

butions to language acquisition (Ramírez-Esparza et al., 2014; Weisleder & Fernald, 2013). Although few infants produce much language before their first birthday, longitudinal studies based on the MacArthur-Bates Child Development Inventory (CDI, Fenson et al., 2007) show that the earliest age for producing a first word is approximately 9 months (Fenson et al., 1994). The CDI was designed around the first few hundred words and phrases typically acquired by American-English-learning children. By 16 months of age, median spoken vocabulary size for the original cohort was 55 words and by 23 months, it was 225 words. The CDI also helped to establish the important mediating role of communication gestures such as pointing and enactive gestures (Zinober & Martlew, 1985). By age 6, the average American child has a spoken vocabulary of around 6000 words and a comprehension vocabulary of around 14,000 words (Dollaghan, 1994). Clearly, some powerful learning mechanisms are at work.

There are very few longitudinal studies of individual differences between infants in sensory or neural processes that predict later language development (but see Kalashnikova et al., 2019; Kuhl et al., 2007; Ortiz-Mantilla & Benasich, 2013). However, cross-sectional perceptual learning studies in infants suggest many candidate learning mechanisms. These include the discrimination of acoustic features that specify phonetic categories in the native language (e.g. Eimas et al., 1971; Werker & Tees, 1984), the perception of prosodic information (e.g., Jusczyk et al., 1987), the exploitation of phonotactic information (e.g., Jusczyk & Aslin, 1995), and statistical learning mechanisms such as learning conditional probabilities between syllables in words (e.g., Saffran et al., 1996).

By the end of their first year, experimental work shows that typically-developing infants are already building a lexical phonological system that enables the distinction of minimal pairs (words differing by a single phoneme; Vihman et al., 2004), and by the end of their second year, most infants are beginning to produce grammatically-accurate utterances (e.g., Tomasello, 2000, 2014). The language tasks selected for the BabyRhythm project hence aimed to measure semantic development, development of gestures, phonological development and grammatical development. As TS theory is based on rhythmic timing, infant motor rhythmic synchronization to speech rhythm was also of interest; accordingly tasks measuring toddler rhythmic timing were created.

To measure semantic development, we utilized the official UK adaptation of the MacArthur Bates CDI (UK-CDI; Alcock et al., 2020). The CDI was administered at 10, 12, 15, 18 and 24 months. Two experimental tasks, a word recognition measure based on Bergelson and Swingley (2012), and a toddler-controlled receptive vocabulary task based on Friend and Keplinger (2008) were also administered, at 8 months of age and at 18 and 24 months of age respectively. Both experimental tasks present infants with two images, one image is named, and then looking time (at 8 months) or active image selection (at 18 or 24 months) provides the dependent measure. Bergelson and Swingley (2012) provided evidence for semantic understanding in US infants as young as 6 months of age with their looking task (though see Kartushina & Mayor, 2019 and Steil et al., 2021, for contrasting evidence with Norwegian and German speaking infants). Friend and her colleagues developed the Computerized Comprehension Task (CCT) to measure lexical knowledge directly from older infants (Friend & Keplinger, 2008). Toddlers view pictures on a touch screen while one image is named, and select the corresponding image. Good convergent validity between the CCT and the CDI is reported (Friend et al., 2012). Vocabulary was expected to increase with age for both the CDI and the CCT. Vocabulary can be smaller in multilingual learners (Poulin-Dubois et al., 2013; Hurtado et al., 2014). Approximately a quarter of our sample were from multilingual families. Accordingly, we explore linguistic status as a variable in the analyses presented here.

To measure phonological development, we utilized two distinct literatures, the literature on phonological awareness (Goswami & Bryant, 2016), and the literature on nonword repetition (Adams & Gathercole, 1995; Gathercole & Baddeley, 1989). Phonological awareness is largely studied in relation to reading development, and children with poorer awareness of phonology at all linguistic levels (stressed syllable, syllable, rhyme, phoneme) are known to experience difficulties in learning to read (Ziegler & Goswami, 2005). A longitudinal causal relationship has been established using rhyme awareness tasks with preschoolers such as the rhyme oddity task (here children select the odd word out [the non-rhyme] in oral word triples like *cat, fit, pat*) and nursery rhyme knowledge (Bradley & Bryant, 1983; Bryant et al., 1989). For the current project, we created a touchscreen game intended to simulate the rhyme oddity task using families of toys. Nonword repetition tasks require children to repeat accurate novel phonological forms that usually contain multiple syllables. A longitudinal relationship has been established between nonword repetition and later language development (Gathercole, 2006), and nonword repetition can also identify children with developmental language disorder (Bishop et al., 1996). For the BabyRhythm project, we adapted a nonword repetition task developed by Hoff, Core and Bridges (2008) for children aged under 2 years. Finally, we adapted a study of nursery rhyme knowledge from Bryant and colleagues (1989), previously demonstrated as a strong predictor of phonological awareness.

To measure the development of gesture and grammar, we created two experimental measures, a pointing task and a grammar elicitation task. The pointing task was based on a game with puppets, designed to elicit joint attention behaviours. It has been reported that children who show earlier pointing to elicit shared attention also show better language development as measured by the CDI (Carpenter et al., 1998). The grammar elicitation task was based on prior tasks intended to measure knowledge of plurals, the present imperfect tense, and the past tense (Berko, 1958; Yuan & Fisher, 2009). Yuan and Fisher (2009) showed that novel grammatical forms such as "blicking" and "blicked" could be recognised by toddlers aged 2 years, while Berko (1958) worked with 5-year-olds and elicited novel grammatical constructions. Yuan and Fisher (2009) used short videos of novel actions being performed by human agents, while Berko used pictures of unfamiliar cartoon characters (Berko, 1958). The younger children were asked to choose which video showed "blicking", while the older children were asked elicitation questions, such as "See this picture? This is a wug! Now there is another one! There are two of them. There are two -?" [*wugs*]. We combined these approaches and created an interactive game involving novel toys and actions, designed to elicit spontaneous production of plurals, the present continuous tense, and simple past tense from our sample.

Finally, given our theoretical interest in temporal rhythmic parameters, we sought to investigate whether individual differences in the ability to time motor production of speech (single words) or gestures (clapping) when singing nursery rhymes would be predictive of later language acquisition. Many of the linguistic routines of early childhood involve the temporal matching of perception and production. For example, English nursery rhymes like "Row, row, row your boat, gently down the stream" and "If you're happy and you know it, clap your hands" are accompanied by rhythmic actions such as rocking the body to the beat or clapping to the rhythm, and during knee-bouncing songs with infants such as "Horsie Horsie don't you stop" the infant experiences rhythmic movement in time with the words of the song. The rhythmic timing task utilised nursery rhymes such as "If you're happy", which were sung along with the toddler, and then gaps were left for the child to clap to the beat or produce missing words of the song to the rhythm.

The aim of current paper is to present the emerging language skills of the BabyRhythm cohort on our diverse range of language tests. To this end we describe infant performance on each of our dependent variables, and the relationship between dependent variables within linguistic domains. We provide evidence for similarities and differences between males and females, and monolingual and multilingual language learners. Finally, we select our most robust measures of language development to be used in longitudinal brain-behaviour analyses.

Methods and Results

Participants

122 infants (65 male, 91 monolingual) were recruited to the longitudinal Cambridge UK BabyRhythm project. 108 infants were recruited prior to the first brain recording (2-month) visit, and 14 infants were recruited prior to the second brain recording (4-month) visit. Families were recruited from the local area via flyers and online advertisements, forming a sample of convenience. The project spanned eight brain recording visits to our laboratory over the first year of life (2-11 months), when infants took part in a battery of EEG (Attaheri et al., 2022) and motion capture (see Rocha et al., 2021 pre-print) tasks, followed by home visits at 12, 15, 18, 24 and 30 months (the latter data are still being analysed). Attrition from the sample was low; 10 families withdrew during the laboratory visits, two withdrew following the laboratory visits. A further four families relocated abroad and continued to provide questionnaire data but not behavioural data. The study was approved by the University of Cambridge ethics committee. The study was re-approved by the same committee to facilitate remote testing during the COVID-19 pandemic. The caregiver gave written, informed consent concerning the experimental procedure. Infants received a certificate and small age-appropriate gift as a thank you for participation in some of the laboratory and home visits (e.g. book, toy), and any travel expenses incurred were refunded to the caregiver. Following remote visits, families were sent a £5 book voucher.

Information about infants' language exposure was primarily collected using a language exposure questionnaire (based on Bosch & Sebastián-Gallés, 2001; Molnar et al., 2013) when the infants were 18 and 30 months old. In our sample, infants were categorised as multilingual if they were exposed to a language or combination of languages other than English at least 30% of the time. These data are available for N = 102 infants. Where the in depth-questionnaire data were not available, infants were classified using data from the families' registration questionnaire (N = 20; question "Does your child regularly hear a language that is not English?"). Our sample consists primarily of monolingual English infants (N=91, females = 44). Within the multilingual group (N = 31, females = 13), a total of 21 additional languages were reported by parents (Afrikaans, American Sign Language, Catalan, Chinese, Danish, Dutch, Finnish, French, German, Hindi, Italian, Japanese, Lithuanian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Taiwanese, Urdu). Multilingual infants were each exposed to a maximum of three different languages (M=2.2). The majority were dominantly exposed to English (N=15/23 with in-depth data) with varying degrees of exposure (M=57.9%, Min=36.9%, Max=68.5%). Multilingual infants with dominant exposure to another language (N=8/23 with in-depth data) were all exposed to English to some extent (M=34.2%, Min=1.0%, Max=47.6%). Note that the minimum here relates to one infant who had their 30-month-visit including language exposure questionnaire (i.e. after the data collection reported in this manuscript) in April 2020, in the peak of the first UK lockdown. This resulted in a reduction in English exposure to 1% at 30-months of age, from 13.1% at 18-months. The English exposure for non-English dominant multilinguals is higher when this exceptional circumstance is excluded (M=38.9%, Min=10.7%, Max=47.6%).

Procedure

The current manuscript describes the different language assessments administered between 8 – 24 months over a series of laboratory visits (until 11 months), home based (12 – 24 months), and (from March 2020) remote home visits, the latter conducted over video communication platform Zoom. Remote visits were a necessary adaptation to the restriction of in-person interaction during the COVID-19 pandemic. Home visits involved one or two experimenters visiting the family's home and conducting interactive tasks along with parent questionnaires. These visits were always recorded via a Canon Powershot Sx620 Hs video camera for offline coding. Five 18-month and two 24-month sessions were cancelled and not rescheduled during the transition to remote testing in the first UK COVID-19 lockdown. Subsequent appointments used adapted protocols for remote administration. Remote visits were conducted by one experimenter via Zoom, who guided the parents in parental administration of the tasks. Physical copies of parent instructions were sent to the families in advance of their remote visit, in addition to a reward chart and stickers for the child. Parents also had access to YouTube videos explaining task administration, and containing video stimuli where appropriate. Experimenters recorded the Zoom call, but the primary data analysed were from parents' own recordings of the session, taken on their personal device and uploaded via secure transfer to the University of Cambridge. Once coded, all data were uploaded to REDCap (Research Electronic Data Capture; Harris et al., 2009, p. 2019). In the following sections, each of the language tasks are described individually. Full standard operating procedures and coding schemes for each task, in addition to data and analysis scripts, can be found on the OSF platform (https://osf.io/ftejv/). Information on number of infants participating by task and missing data for each task is available in the supplementary materials.

A. Semantic Tasks

Communicative Development Inventory (CDI): Receptive and Expressive Vocabulary (Parental Estimation)

CDI Task Description

The UK-CDI is a parent-completed questionnaire on which parents fill in a large checklist of their child's communicative behaviours. The UK-CDI Words and Gestures is used for children aged 8-18 months and is the official UK adaptation of the MacArthur Bates CDI: Words and Gestures. A shortened version was given to children in this cohort at 10 and 12 months containing 350 items, then the standard version was given at 15 and 18 months, containing 395 items. The Lincoln Toddler CDI, a longer version containing 689 items, adapted for older children (18-30 months) from the original US CDI: Words and Sentences, was given at 24 months.

CDI Procedure

Caregivers were given a paper copy of the UK-CDI (Words and Gestures) questionnaire when their infants were 10, 12, 15 and 18 months old. At 24 months parents were given the Lincoln Toddler CDI. As our sample were primarily English-learning infants, but included those exposed to multiple additional languages, caregivers were asked to mark for each word on the CDI if the child understood the word or if they could understand and say the word, and in how many languages the child could understand and/or say each word. This is not best practice for use of CDIs, where rather than translation (here provided by the parents), *adaptation* is critical to encompass cultural relevance and word frequencies across languages (Pena, 2007). The current approach was the result of a pragmatic decision due to the complexities of our sample (see Participants section, above, for description of 21 additional languages). Additional sections on gestural production and grammar were collected but are not analysed in the current paper. Parents completed the form at home and returned it using a prepaid stamped addressed envelope. Due to COVID-19 restrictions some participants were asked to fill in their CDI online.

CDI Data Processing

Number of words understood (Comprehension) and words produced (Production) were calculated. Infants received a score of 1 for each item that they could comprehend or produce, regardless of how many languages and in which language they knew or spoke that word, giving a score of 'concept vocabulary'. As there is no clear consensus in the literature regarding whether to use total or concept vocabulary for multilingual children (see Weisleder et al., (preprint), for a scoping review and recommendations for best practice), a comparison for our data is included in the supplementary materials.

Word Recognition (Receptive Vocabulary, Infant Looking Procedure)

Word Recognition Task Description

The task was adapted from Bergelson and Swingley (2012). In each trial, two images appeared onscreen. In turn, the images made an attention-getting movement for one second, accompanied by an attention-getting sound. As this happened, the parent heard a pre-recorded English sentence via headphones. The parent repeated the sentence out loud using infant directed speech. For example, "Where's the apple? Look at the apple". Parents could repeat the sentence in their child's preferred language (see Supplementary Materials). The child was then presented with two pictures on the screen: one target, and one distractor. Both static pictures were presented side-by-side for 6 seconds, including 1.867 seconds for the parent to prompt and the infant to saccade, and a 3.5 second analysis window. Infants participated in 32 trials testing knowledge of 8 food items and 8 body parts in one of two pseudo-randomised trial orders. Pictures were approximately 16.9 x 12.7cm and displayed to the left and right sides of the screen, with side of presentation counterbalanced. Pictures were approximately matched for size and colour.

Word Recognition Procedure

Testing took place during the 8-month laboratory visit. Infants were sat on their parent's lap in the sound attenuated EEG booth, approximately 1 m from the presentation screen and speakers. Parents were given a set of blacked-out glasses and headphones to wear, through which the pre-recorded sentences were played. Eye tracking data were collected using a Tobii TX300 eye tracking camera (sampling rate 300Hz) located and fixed at the base of the presentation screen (23" TFT monitor). The eyetracker was calibrated using a 5-point calibration to an animated circle with accompanying sound.

Word Recognition Data Processing

Looking times were calculated using the eye-tracking data in the first instance. In case eye-tracking data was not saved or was otherwise not useable, a trained coder video-coded where the infant was looking. Among the 55 infants who provided sufficient data for analysis (see OSF for details), data from 46 infants were coded from eye-tracking data and data from nine infants were manually coded from video. For the full sample including infants whose data were subsequently omitted, these numbers were 68 and 21 respectively. Nineteen of the video coded datasets were double coded, with good inter-rater reliability (ICC(2,1) = .862, p < .001). Although the degree of temporal precision will have differed between manually coded and eye-tracking data, the paired analysis approach means that any systematic biases will have applied to both the target and distractor trials. For each trial looking 'left', 'right', 'neither', or 'away' was calculated for the analysis window. Trials were included in analysis when total looking time across the 367 to 3,867 ms analysis window was longer than 1000 ms. Following the original paper, a score of word recognition was calculated by observing the difference in fixation for paired pictures. The fixation to picture A compared to B when A was the target word was compared to when picture A was the distractor and B was the target. A positive difference is assumed to reflect the infant's knowledge of the target word.

Computerised Comprehension Task (CCT; Receptive Vocabulary, Infant-Controlled Procedure)

CCT Task Description

The CCT allows infants to demonstrate word recognition by pointing to an image that matches a target word, and was adapted from Friend and Keplinger (2003), with some images updated to fit UK cultural norms (e.g. typical British bus and fire engine). The experimenter presented the task on a touchscreen tablet. The caregiver was provided with a printed list of 41 items and instructions on the exact cue to give the child (e.g. 'Can you touch the ball?'). Caregivers could use the infants' preferred language, language used for each item was not recorded. Testing stopped after five consecutive incorrect or missing responses.

CCT Procedure

The experiment was run on a Lenovo Miix 510 tablet, using Windows 10, via OpenSesame 3.1 with Python 2.7 inline script, during in-person visits only.

CCT Data Processing

Responses were video coded offline, as infant touch was found to be unreliably monitored by the tablet, and some infants pointed without making contact with the screen. As in Friend and Keplinger (2003), no response or responses to the edges or centre of the screen were taken as incorrect responses. Proportion of infants' first touch/point to the target item was taken as the dependent variable. Double coding of infant responses was conducted for 9 infants at 18-months (ICC(2,1) = 0.89, *p* < .001), and 11 infants at 24-months (ICC(2,1) = 0.96, *p* < .001).

Semantic Task Results

Infant performance on the CDI, Word Recognition and CCT are detailed in Table 1. All analyses were performed in RStudio (R Core Team, 2021)), or JASP (JASP Team, 2022).

Infant scores on the CDI increase steadily with age, as would be expected, and are always greater for comprehension than for production, as would also be expected. Regarding the Word Recognition task, a positive difference is assumed to reflect knowledge of the target word. As will be recalled, the dependent variable was the difference between fixation to picture A rather than B when A was the target word, compared to when picture A was the distractor and B was the target. Infants (N = 55) showed a difference score close to zero (i.e. chance; M = .021, SD = .127). Only 54% of infants showed an overall positive looking score (v = 946, *p* = .1145), equating to a 2.13% increase in looking (SE = 1.69%). Other preferential looking studies testing European language learners in this age range show a similar pattern of results to ours (Kartushina & Mayor, 2019; Steil et al., 2021). The touchscreen CCT provides an even more active receptive vocabulary score. At 18-months, infants (N = 99) correctly identified ~one third of the items presented (M = .349, SD = .224). At 24-months, infants (N = 60) generally identified over two thirds of the same items (M = .694, SD = .186). Raw data for the Word Recognition and CCT tasks, broken down by participant sex and mono/multilingual status, are shown in Figure 1.

Figure 1 appears to show monolingual infants identifying more correct words on the CCT. To test for group differences, Bayesian t-tests with the default priors, or where data are not normally distributed, Bayesian Mann-Whitney U tests based on five chains of 1000 iterations, provide the strength of evidence supporting the alternate hypothesis over the null hypothesis. Bayes Factors > 3 demonstrate moderate support for the alternate hypothesis, and Bayes Factors < .3 indicate moderate support for the null hypothesis of no difference between groups. Across semantic tasks, we do not find evidence for any differences between males and females or monolinguals and multilinguals, with some moderate evidence to support the null hypothesis of no difference between groups (see supplementary materials for details).

We also analysed the relationships between our semantic measures. In the correlograms in Figure 2, pairwise Pearson correlation values are reported in the top right panels. Pearson's correlations over successive timepoints also give an index of the measurement reliability of each repeated task (see Byers-Heinlein et al., 2022). Bayes Factors (BFs) were calculated to establish the relative support for the hypothesis that each pair of variables correlate, compared to the null hypothesis of no relationship. BFs are represented visually by the colour of the top right squares. Scatterplots of the raw data are shown in the bottom left

Table 1. Descriptive statistics for Semantic Measures

Variable	*n*	*mean*	*sd*	*median*	*min*	*max*	*range*	*se*
CDI Comprehension (10m)	99	29.83	37.00	15.00	0.00	248.00	248.00	3.72
CDI Production (10m)	99	1.47	2.26	0.00	0.00	10.00	10.00	0.23
CDI Comprehension (12m)	109	66.94	59.97	47.00	0.00	308.00	308.00	5.74
CDI Production (12m)	109	3.80	5.07	2.00	0.00	28.00	28.00	0.49
CDI Comprehension (15m)	106	123.88	83.94	98.50	1.00	371.00	370.00	8.15
CDI Production (15m)	106	14.99	20.59	9.50	0.00	153.00	153.00	2.00
CDI Comprehension (18m)	106	196.78	87.37	194.00	31.00	384.00	353.00	8.49
CDI Production (18m)	106	49.65	51.91	34.50	1.00	296.00	295.00	5.04
CDI Comprehension (24m)	95	430.12	145.63	465.00	86.00	688.00	602.00	14.94
CDI Production (24m)	95	270.76	158.79	265.00	0.00	609.00	609.00	16.29
Word Recognition (8m)	56	0.02	0.13	0.02	-0.32	0.30	0.62	0.02
CCT (18m)	99	0.35	0.22	0.32	0.00	0.85	0.85	0.02
CCT (24)	60	0.69	0.19	0.73	0.00	0.98	0.98	0.02

NB: Maximum score for 10- and 12-month CDI is 350 items. Maximum score for 15- and 18-month CDI is 395 items. Maximum score for 24-month CDI is 689 items. Word Recognition scores are difference scores, scores close to zero indicate chance level performance. CCT scores are proportion of correct responses, scores range from 0-1.

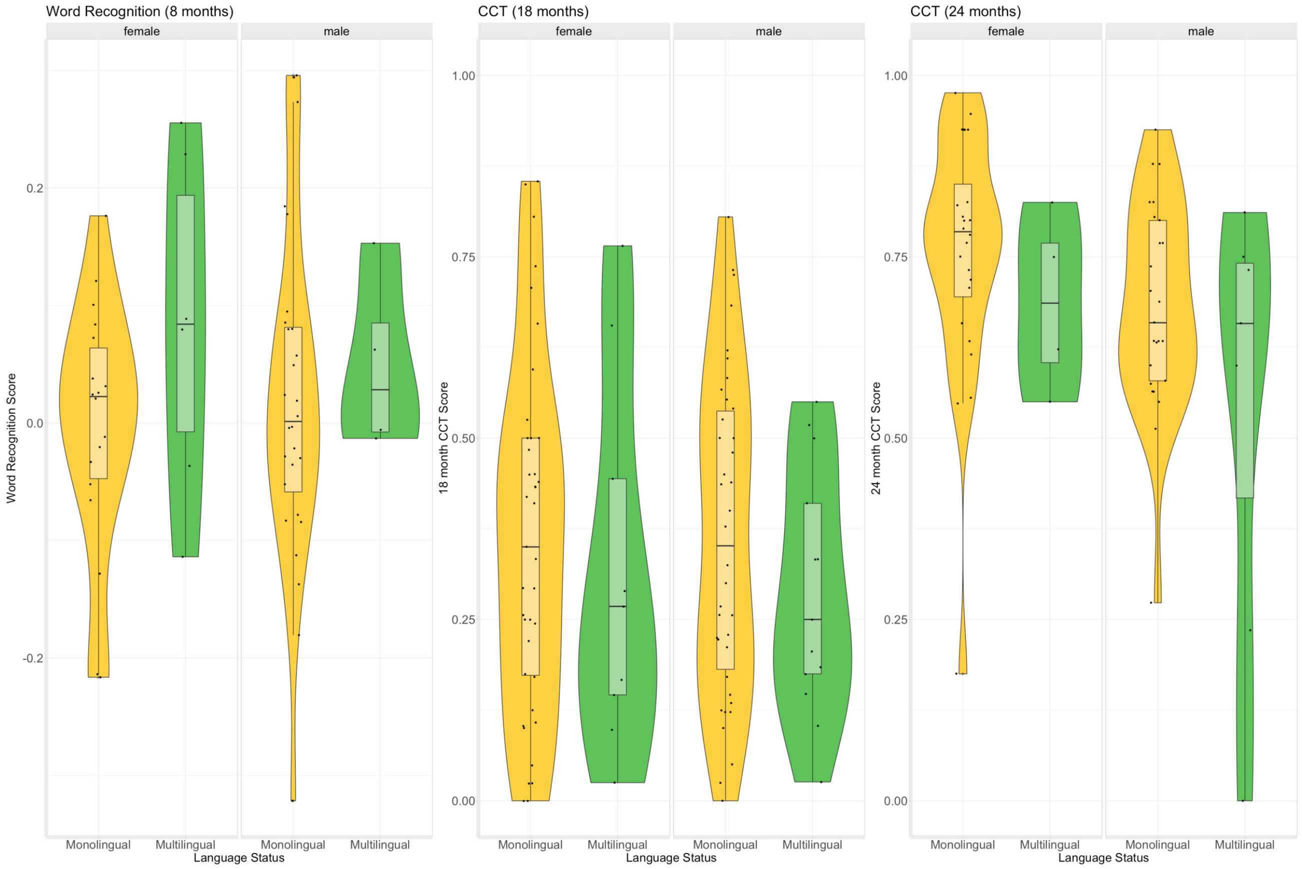

Figure 1. Descriptive statistics for experimental Semantic measures.

Violin plots presented for Word Recognition and CCT, overlaid with jittered raw data and shaded box plots. Performance on each measure is shown separately by infant sex (male/female) and linguistic status (mono/multilingual)

panels, and the distribution of raw data is shown through the centre panels. CDI measures of comprehension and production were positively correlated with each other across ages. 15-, 18- and 24-month CDI scores were all highly correlated, (all BF_{10} > 100, see Figure 2). Whilst 10- and 12-month CDI scores were highly correlated with each other (r range = .33 to .71, BF_{10} range = 28.61 to > 1000), correlations are weaker between early and late CDI measures, with evidence for the null hypothesis of no correlation between 10-month CDI measures and 24-month CDI

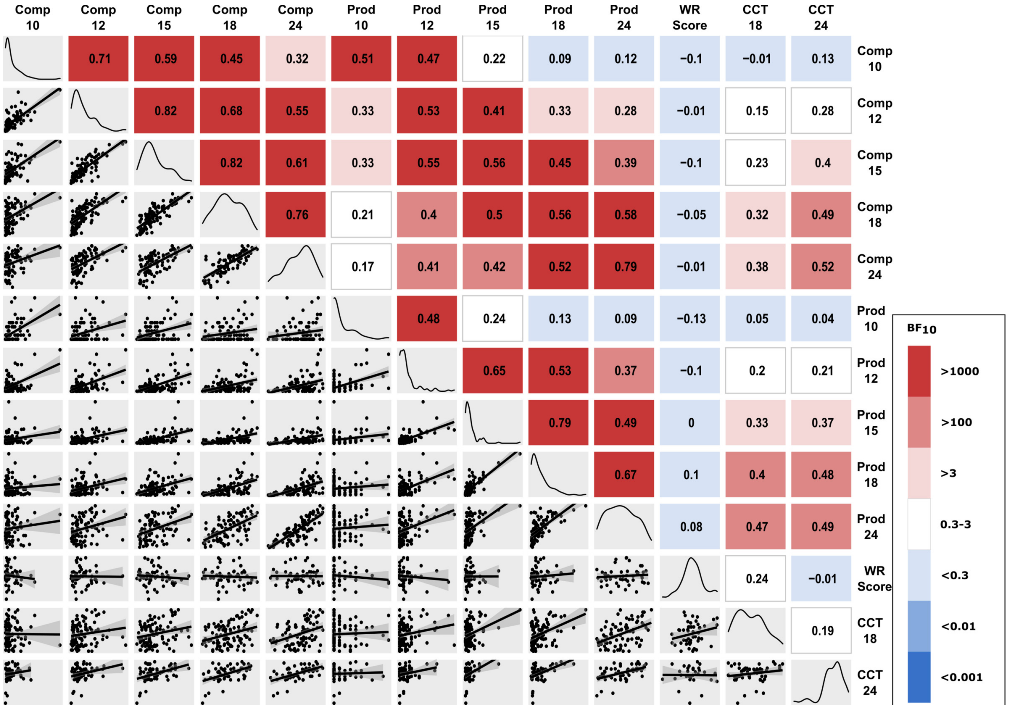

Figure 2. Correlogram showing relationships between semantic tasks.

Comp = CDI Comprehension, Prod = CDI Production, WR Score = Word Recognition score (8 months), CCT = Computerised Comprehension Task. Numbers in variable names refer to the age of testing, where tasks were administered over multiple timepoints.

measures (BF_{10} < .3). All CDI measures are positively correlated with 18 and 24-month CCT measures. There is good BF evidence for correlations between CCT and CDI from 15-months onwards (15-month production and 24-month CCT BF_{10} = 20.322, all CDI measures from 18 months and CCT measures BF_{10} > 3, range = 8.394 to > 1000). Word recognition scores did not correlate with other measures of word learning, with all BF_{10} < .3 (except with the correlation between word recognition and 18-month CCT, BF_{10} = 1.100). Given the chance performance in this task, this is not surprising.

B. Gesture

As noted above, the CDI sections on gestural production are not analysed in the current paper. An experimental pointing task intended to measure joint attention and communicative intent was administered and is analysed here.

Pointing

Pointing Task Description

The task was based on a fox puppet who would sometimes appear during a game (see supplementary Figure S1), to which it was expected that the infant would point and draw adult attention. The initial design of the experiment (adapted from Liszkowski et al., 2007) was intended to elicit referential and declarative pointing. Infants therefore participated in six 20 second trials divided into three conditions: Attend referent (puppet appeared, experimenter looked towards puppet, see Figure S1A), attend nothing (puppet appeared, experimenter looked away from puppet, see Figure S1B), and attend baby (experimenter talking directly to the infant, puppet not present, see Figure S1C). On trials where the puppet was present, the experimenter would turn to one side, protrude a puppet through one of the holes, and begin reciting a script (see OSF repository) which included questions like “Did you see Mr Fox? Did you see him dancing?”, while moving the puppet for the duration of the script. Each testing session began with the attend referent condition, but whether the puppet first appeared either to the left or to the right was counterbalanced across infants.

Pointing Procedure

The experimenter sat behind a customised folding display board, from which a fox puppet attached to a bamboo stick could appear on one side or the other. Parents sat on the floor approximately 1.5m from the experimenter with their child on their lap. Parents were instructed to try to keep their child on their lap, but otherwise remain passive during testing.

Pointing Data Processing

Videos were coded in ELAN (2019, Version 5.7, Nijmegen: Max Planck Institute for Psycholinguistics) and annotated for index finger points, whole hand reaches and any associated vocalisations that occurred during points and reaches. The number of index points, whole hand reaches and vocalisations were annotated for each condition. 10% of videos were second coded, and interrater reliability was good, κ = .817, p < .001). Many infants did not point during the experiment, the Median number of points made by in-

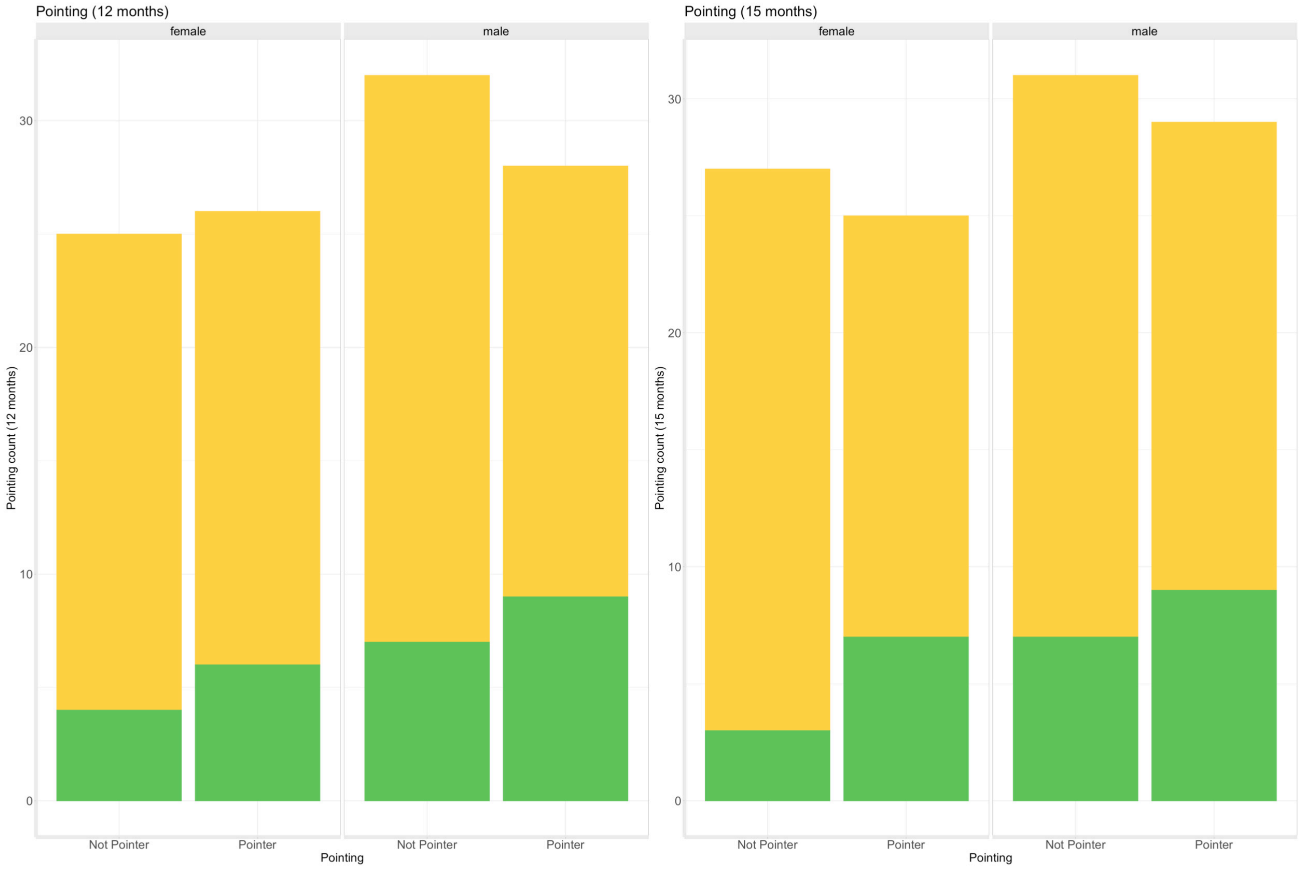

Figure 3. Descriptive statistics for Pointing.

Stacked bar charts show the number of infants who did and did not point at 12-months (left) and 15-months (right). Monolingual infants are depicted in yellow, multilingual infants in green.

fants who did point was 3 (SE = .267). As there was evidence for no difference between the Attend nothing and Attend referent conditions at 12 (BF_{10} = 0.207) or 15 (BF_{10} = .105) months, for our final analyses we dichotomised our sample into two groups; infants who pointed and infants who did not point.

Gesture Results

Infant pointing was measured at 12-months (N = 111) and 15-months (N = 112). At 12-months, 57 of 111 infants pointed during the experiment. At 15-months, 58 of 112 infants pointed, see Figure 3.

Figure 3 indicates minimal difference in whether infants pointed by demographic features. Bayesian contingency table tests provide anecdotal to moderate evidence for no difference between males/females and mono/multilinguals, see supplementary materials.

C. Phonology

Two measures of phonological development were originally devised, a nonword repetition measure and a rhyming measure. Due to COVID-19, the rhyming measure was later substituted for a strong predictor of phonological awareness assessing infants' knowledge of well-known nursery rhymes, described below.

Nonword Repetition (NWR)

NWR Task Description

This measure used toys as stimuli and was adapted from Hoff, Core and Bridges (2008). Items signifying six real words (fish, pig, puppy, monkey, banana, tomato) and six matched non-words (kish, dap, eppy, punky, tanina, kamito) were presented to the infant in one of two set orders. Nonwords always followed the matched real word, and the real words became progressively longer (four monosyllabic, four bisyllabic, four trisyllabic).

NWR Procedure

During home visits, the experimenter took each item from a bag, verbally labelled the item, and asked the infant to repeat the label back to them. The toddlers were encouraged to repeat the target word/non-word, which was modelled up to four times by either the experimenter or the parent (e.g. "can you say fish?"). Infants were allowed to handle the objects, and congratulated for correct responses. During remote visits, the structure of the task remained the same except infants watched a video of the items being presented. They first heard the experimenter label the item, then the parent was instructed to prompt their child up to three times. The experimenter intervened to prompt the child if the parent failed to do so.

NWR Data Processing

The recordings were annotated offline using ELAN. The first four target-related infant responses occurring for each item? were transcribed using IPA and following the conventions of broad transcription. Three dependent variables were taken: number of consonants produced correctly, number of syllables produced correctly, and accurate placement of primary syllable stress (primary stress pattern correct). These measures were expressed as proportions as some stimuli had 2 syllables and some had 3 syllables. The primary stress measure was intended to reflect emergent prosodic accuracy, and was independent of the number of syllables correct. For example a child who said "NA-na" or "BA-na" for the trisyllable "banana" would score 100% for primary stress pattern correct but 67% for syllables correct. Scoring was consistent across home and remote visits. Data from 10% of infants were double coded for consistency, there was strong agreement between the two coders' judgements (ICC(2,1) = .791, p < .001). Administration of this task altered from in-person to a virtual presentation. Bayesian analyses did not reveal a systematic advantage for either procedure, except for strong evidence for a greater proportion of correct stress placement in the remote administration (see supplementary materials for full details).

Rhyme Oddity (in person testing)

Rhyme Oddity Task Description

An experimental game about cartoon mummies and babies was devised, based on the rhyme oddity task used with pre-school children (Bradley & Bryant, 1983). Using a touch screen tablet, infants were first trained to touch the screen via a bubble-popping game. Following this warmup activity, infants received two demonstration trials. A 'mummy' animal was presented in the top half of the screen and three 'babies' presented at the bottom of the screen. The infants were encouraged to touch the three babies. They were then instructed that the babies would try to copy what their mummy said. In the first demonstration, the mummy said 'ba' and the babies said 'ga', 'ta' and 'wa', respectively. The experimenter then said 'good babies! They sound the same!'. In the second demonstration, the mummy said 'ba' and the babies said 'ga', 'ta', and 'wit'. The experimenter gave feedback that one of the babies said something different, and asked the infant to point to the different baby. Infants were given two practice trials in which they were encouraged to point to the 'odd baby out', followed by 10 experimental trials, where they received up to three prompts to touch the screen. In each trial, real words were used. Rhyming and non-rhyming words for each trial were matched for lexical neighbourhood density.

Rhyme Oddity Procedure

The experiment was run on a Lenovo Miix 510 tablet, using Windows 10, via OpenSesame 3.1 with Python 2.7.12 in-line scripts. The order and location of the odd trials were counterbalanced across two order sets. Infants received feedback at the end of each trial, incorrect trials were replayed with the experimenter showing the infant the correct response.

Rhyme Oddity Data processing

Data were video-coded. Responses were coded as correct if the infant touched, pointed to, or described the correct baby, or repeated the target non-rhyming word correctly. Only the infant's first response was coded, and feedback was only given at the end of the trial. The dependent variable was the proportion of valid, completed trials in which the infant indicated the correct baby.

Rhyme Knowledge (remote testing)

Rhyme Knowledge Task Description

This task was adapted from Bryant et al. (1989), as a notable predictor of phonological awareness over longitudinal studies. Parents were asked to prompt their infant to produce the first two lines of five English nursery rhymes (Humpty Dumpty, Baa Baa Black Sheep, Hickory Dickory, Jack and Jill, and Twinkle Twinkle). Parents first asked if the child could sing the rhyme alone. If the child remained silent the parent prompted them by singing the start of the rhyme. If the child still did not respond, the parent sang the rhyme themselves, omitting the final word of each line for the child to deliver. Note that parents were not asked to rate their child's familiarity with the rhymes included, which were chosen to replicate Bryant et al (1989).

Rhyme Knowledge Procedure

The nursery rhymes were presented successively, with graduated prompts presented for the parent to read via screenshare on Zoom. Parents were asked to give the child time to respond between prompts.

Rhyme Knowledge Data Processing

A revised version of the coding scheme used in Bryant et al (1989) was devised using a 5-point rating scale. Children could score between 0 (completely incorrect or no response) and 5 (child recited full two lines of the rhyme accurately and with no further prompt than the nursery rhyme name; full coding scheme on OSF). Scores from the five nursery rhymes were added together for a maximum score of 25, and adjusted proportionately in case of a missed or invalid trial. Four videos (10% of the sample) were double coded and showed good interrater reliability (r = 0.88 across each nursery rhyme; weighted κ = 0.96).

Phonology Results

Infants took part in two phonology tasks, Nonword Repetition (NWR; 18- and 24-months) and Rhyme Oddity (24-months, in-person only) As the touchscreen Rhyme Oddity task was not suitable for use in remote visits following COVID-19, a Rhyme Knowledge task was utilised instead, as a potential strong predictor of phonology. The NWR task yielded three dependent variables: proportion of consonants correct, proportion of syllables correct, and proportion of attempts with correct primary stress pattern-

Table 2. Descriptive Statistics for Phonology Tasks

variable	*n*	*mean*	*sd*	*median*	*min*	*max*	*range*	*se*
NWR Consonants (18m)	105	0.130	0.173	0.069	0	0.810	0.810	0.017
NWR Consonants (24m)	101	0.529	0.303	0.621	0	1.000	1.000	0.030
NWR Syllables (18m)	105	0.137	0.158	0.083	0	0.583	0.583	0.015
NWR Syllables (24m)	101	0.453	0.301	0.500	0	1.000	1.000	0.030
NWR Stress (18m)	105	0.108	0.147	0.000	0	0.500	0.500	0.014
NWR Stress (24m)	101	0.420	0.282	0.459	0	1.000	1.000	0.028
Rhyme Oddity (24m)	43	0.259	0.196	0.222	0	0.600	0.600	0.030
Rhyme Knowledge (24m)	37	6.561	6.148	5.000	0	20.000	20.000	1.011

NB: NWR and Rhyme Oddity variable scores are proportions of correct responses, ranging from 0 - 1. Rhyme Knowledge scores are normalised totals out of a maximum of 25.

ing. Descriptive statistics for each variable are given in Table 2.

At 18-months, infants (N = 105) typically scored poorly across all NWR measures (consonants correct M = .130, SD = .172, syllables correct M = .137, SD = .303, typical stress M = .107, SD = .147). By 24-months, infant performance (N = 101) showed substantial improvement (consonants correct M = .529 SD = .172, syllables correct M = .453, SD = .300, typical stress M = .420, SD = .282). For rhyming, the in-person Rhyme Oddity task yields one dependent variable, the proportion of 'odd one out' targets detected. Performance was low at 24-months (N = 43, M = .259, SD = .196). Finally, the remote participation Rhyme Knowledge task at 24-months also showed low performance, with infant (N = 37) ability to recite the first two lines of well-known nursery rhymes scoring on average 6.561 (SD = 6.148) of a maximum 25 points (hence 26.2%, similar to the 25.9% average for Rhyme Oddity). The breakdown by linguistic status and sex is shown in Figure 4.

Inspection of Figure 4 suggests a possible multilingual disadvantage across all the phonology tasks. It is particularly possible that in the Rhyme Knowledge task, infant performance was affected by their familiarity with the English nursery rhymes. However, this pattern is not statistically supported, with anecdotal evidence for no difference between mono/multilinguals (all BF_{10} < 1). There is moderate evidence for no difference between sexes in all NWR measures (all BF_{10} < .3), and anecdotal evidence for no sex differences in the rhyme oddity and rhyme knowledge measures (all BF_{10} < 1), see supplementary materials.

Correlograms for the phonology tasks are shown in Figure 5. There are strong correlations between the three NWR dependent variables within timepoints (r range = .71 to .89, all BF_{10} > 1000) and moderate correlations between timepoints (r range = .24 to .49, all BF_{10} > 1000), with the latter suggesting good measurement reliability. The exception is for producing typical primary stress patterns at 24 months and 18-months, where evidence is anecdotal to moderate (BF_{10} = 2.088 - 4.717). It is possible this is influenced by changes in administration, with stronger performance in the stress metric for the remote testing version of the task. Whilst NWR performance is positively correlated with Rhyme Oddity performance (r range = .14 to .36), the relationship appears anecdotal (Rhyme Oddity and 18-month NWR variables BF_{10} = .999 to BF_{10} = 2.805, 24-month NWR variables (BF_{10} .898 to BF_{10} .287)), see Figure 5. Regarding the remote participation nursery rhyme knowledge measure, there is moderate evidence for infant performance on the Rhyme Knowledge task at 24-months correlating with proportion consonants correct and typical primary stress production measured at 18-months (consonants BF_{10} = 5.269, stress BF_{10} = 3.789), but not at 24-months (consonants BF_{10} = .893, stress BF_{10} = .856). Bayes Factors are also around 1 for correlations between Rhyme Knowledge and proportion syllables correct, as measured at both ages (18-month BF_{10} = 2.274, 24-month BF_{10} = .636).

D. Grammar

Grammar Task (in-person testing)

Grammar Task Description

Games with novel objects were devised to try and elicit production of 3 grammatical forms: plurals, the present continuous tense, and simple past tense. Plurals were always attempted first, followed by counterbalanced present versus past tense, with four trials per condition. For plural trials, experimenters labelled a novel object with a target nonword, followed by a second identical object. The prompt to elicit a plural from the child was 'This is one *wug*, and here is another *wug*! There are two...?'. Prompts were repeated up to three times, for each of four object pairs. During present continuous and simple past trials, one of two hand puppets demonstrated novel actions involving jumping, scooting, and bowing. Different puppets, actions and stem non-words were used for the continuous present and simple past tense blocks. In both blocks, children were shown an action and given its label twice (e.g. "Sooty knows how to blick". Puppet blicks. "Let's see him blick". Puppet blicks). In present continuous trials, prompts like "What is Sooty doing? Sooty is..." were accompanied by continuous repeated presentation of the action. In simple past tense trials, a single iteration of the action was followed by a prompt like "What did Sooty do?".

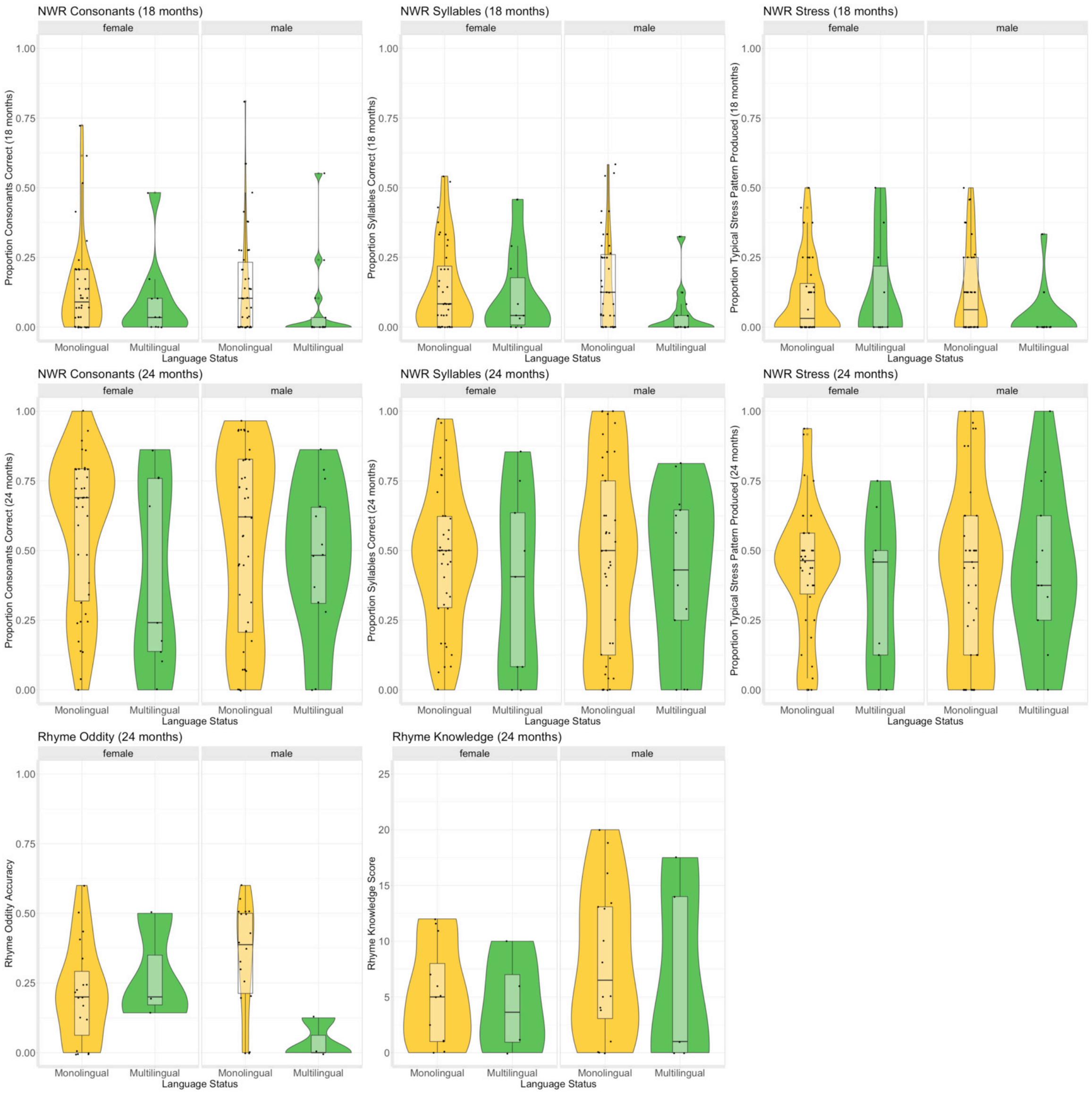

Figure 4. Descriptive statistics for Phonological measures.

Violin plots are presented for the Non Word Repetition, Rhyme Oddity and Rhyme Knowledge tasks, overlaid with jittered raw data and shaded box plots. Performance on each dependent variable is shown separately by infant sex (male/female) and linguistic status (mono/multilingual).

Grammar Procedure

The experimenter sat behind a display board and operated the objects being used in the game while delivering the relevant script (see Figure S1). Parents sat on the floor approximately 1.5m from the experimenter with their child on their lap. Parents were instructed to try to keep their child seated, but otherwise remain passive during testing. Note that a remote-testing alternative of this task was trialled, and the results are available in the supplementary materials.

Grammar Data processing

Recordings were annotated using ELAN. Infants were given binary scores of either ‘produced correct grammar at least once’ or ‘did not produce correct grammar’ for each of the three conditions.

Grammar Results

In-person home visits tested 24-month-old infant (N = 45) ability to apply conventional grammatical endings to novel stem words. Most infants tested were not able to produce any correct grammar (39/45, 86.66%). No infants gained the maximum score of 3 for manipulating the novel stem for appropriate plural, present continuous and past tense endings, see Figure 6.

Figure 6 does not show any apparent differences in emergent grammar knowledge by sex, and illustrates that only a small number of multilinguals participated in this task. Bayesian contingency tables show anecdotal evidence for no difference between sexes (BF_{10} = .650), and good evidence for no difference between monolinguals and mulitlinguals (BF_{10} = .191). Infants performed largely at floor; emergent grammar at this early age may be better picked up by larger sample sizes and a less demanding task.

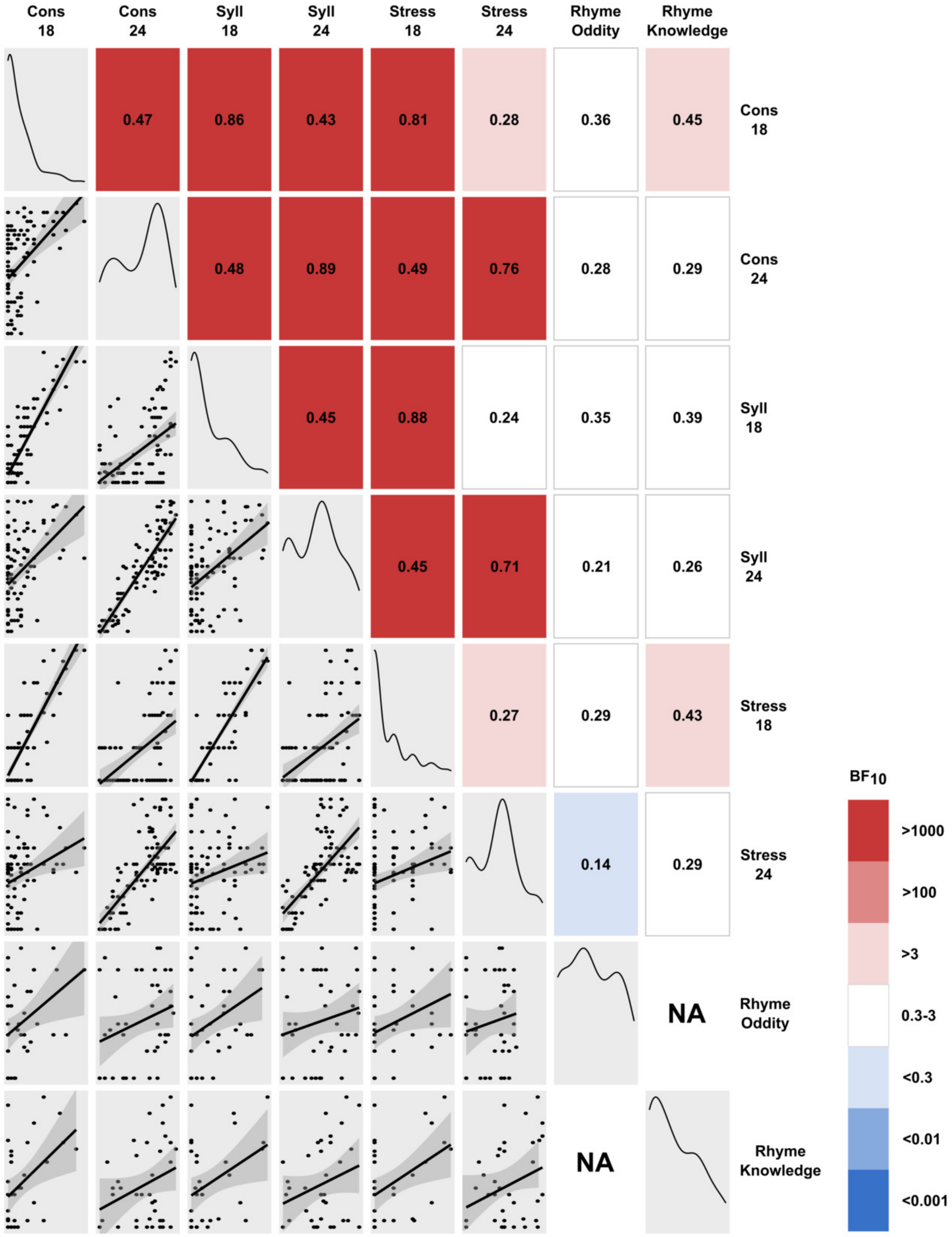

Figure 5. Correlogram showing relationships between phonology tasks.

Cons = Consonants correct, Syll = Syllables correct, Stress = primary stress pattern correct (all taken from NWR). Numbers in variable names refer to age of testing. NA values reported as infants could not take part in both Rhyme Oddity (home testing) and Nursery Rhyme Knowledge (remote testing).

E. Rhythmic Timing

TS theory proposes that rhythmic ability in perception and production is related to language development. To measure rhythmic timing in our toddlers, we devised a nursery rhyme completion game. Note that this was in use before COVID-19, and that it uses different English nursery rhymes to the Rhyme Knowledge task described above.

Nursery Rhyme Completion (NRC)

NRC Task Description

Infants were presented with three familiar British nursery rhymes - ‘The wheels on the bus’, ‘Row your boat’ and ‘If you’re happy and you know it’. Infants heard demonstration rhymes (whole rhyme sung) and test rhymes (identical recordings with target words/actions omitted, see OSF repository), and were encouraged to fill in the missing words and actions themselves. Prior to the testing session, parents were sent the demonstration recordings of the

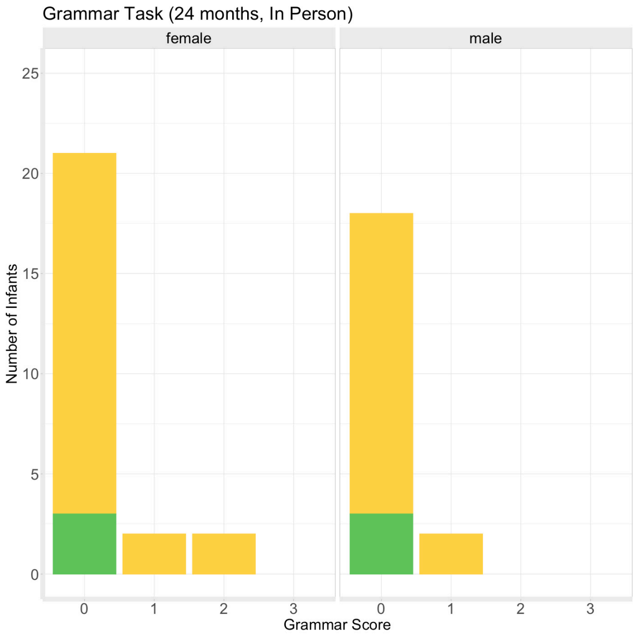

Figure 6. Descriptive statistics for Grammar measures.

Stacked bar charts show the number of infants who achieved each of the possible scores (0-3). Monolingual performance is shown in yellow, and multilingual is shown in green.

rhymes with instructions to play them while singing along so the child would be familiar with them.

NRC Procedure

During home visits, demonstration rhymes were played to the child twice. First, both experimenter and parent sang along, then, only the experimenter sang along. Following this familiarisation, the test version of the rhyme was played twice. The infant was prompted that some words would be missing and asked to sing along and fill in the gaps. Missing words and actions were always at the end of the musical phrase (e.g. 'The wheels on the bus go round and [...]'). During remote visits, the protocol followed the same structure but using a Sing-along style YouTube video (OSF repository) in place of the live experimenter. The YouTube video was recorded while the experimenter listened to the original stimuli through earphones to replicate the pace of the rhymes. The video also gave instructions to the parents about when they were meant to sing along (demo rhymes) and when only the child was meant to sing (test rhymes). The omitted target words and actions remained the same.

NRC Data Processing

Praat textgrids showing the correct timings of target utterances and actions were overlaid onto video recordings taken from the testing session. Any utterances and claps produced by the infants were marked onto the textgrids. The average utterance/clap mismatch and average clap inter-onset interval (IOI) were calculated from the textgrid in ELAN. Clap IOI is defined as the interval (onset to onset) between claps produced during the full length of the trial. Mismatch is defined as the time difference in milliseconds between when the target utterance/clap should have been produced and when the infant produced the action. Mismatch values close to zero therefore indicate that the infant was more accurate, with negative values showing that the child was early and positive values that the child was late (early or late in comparison to the target timing of the utterance/clap). Infant responses outside of the response window were excluded (window length varied by trial, see SOP on OSF for exact timings). Infant responses were averaged across all trials and rhymes. Double-coding of 10 infants showed good inter-rater reliability (Clap Mismatch ICC(2,1) = .930, p < .001, Clap IOI ICC(2,1) = .770, p < .001). Home and remote visit data were coded identically, except that the timings for target utterances and claps were slightly different in the YouTube video, and mismatch values were therefore adjusted.

Rhythmic Timing Results

Infant ability to time both speech output and motor responses to a rhythm were assessed using the dependent variable of mismatch to the correct timing in ms. As can be seen from comparing the minimum and maximum scores, infant performance was extremely variable. Descriptive statistics are shown in Table 3.

Of the infants who participated (N = 92), 37 made an attempt to fill in at least one gap in the rhyme during test trials. The average mismatch between the target timing and infant production was 715 ms (SD = 440 ms). 37 infants also attempted to clap during the gap in 'If you're happy and you know it clap your hands' (M mismatch = 511 ms, SD = 244 ms). Finally, 36 infants produced two or more claps during 'Happy' trials, and the IOI of their clapping was on average 187 ms (SD = 94 ms) away from the target tempo of 580 ms. The raw data and the breakdown by sex and linguistic status are shown in Figure 7. A Bayesian independent samples t-test shows strong evidence for females performing more accurately in the speech mismatch measure (BF_{10} = 60.139). Mann-Whitney U tests show anecdotal evidence for females being more accurate in the timing of their first clap (BF_{10} = 1.781), but anecdotal evidence of no difference between males and females on their rate of clapping (BF_{10} = .347). There is anecdotal evidence for no difference between mono/multilinguals across all three measures (see supplementary materials).

Correlational analyses suggested that the speech and motor timing within the NRC task were not correlated (see Figure 8), with evidence to support the null hypothesis (all BF_{10} < .3).

Identification of appropriate outcome variables

A principle aim of the BabyRhythm project is to relate individual differences in early brain recording and motor rhythm production to later language outcomes. Given the wide range of tasks utilised in the BabyRhythm project, it is important to select a priori the most suitable language tasks for longitudinal brain-behaviour analyses. Tasks were considered suitable if data were available for the majority of our sample, maximising our chances of sufficient statistical power for longitudinal analyses (note that this is

Table 3. Descriptive Statistics for Timing Measures in ms

variable	*n*	*mean*	*sd*	*median*	*min*	*max*	*range*	*se*
Speech Mismatch (24m)	37	715.960	440.205	621.600	43.000	1830	1787.000	72.369
Clap Mismatch (24m)	37	510.771	244.163	496.500	211.000	1680	1469.000	40.140
Clap IOI Mismatch (24m)	36	186.975	94.008	174.487	42.600	433	390.400	15.668

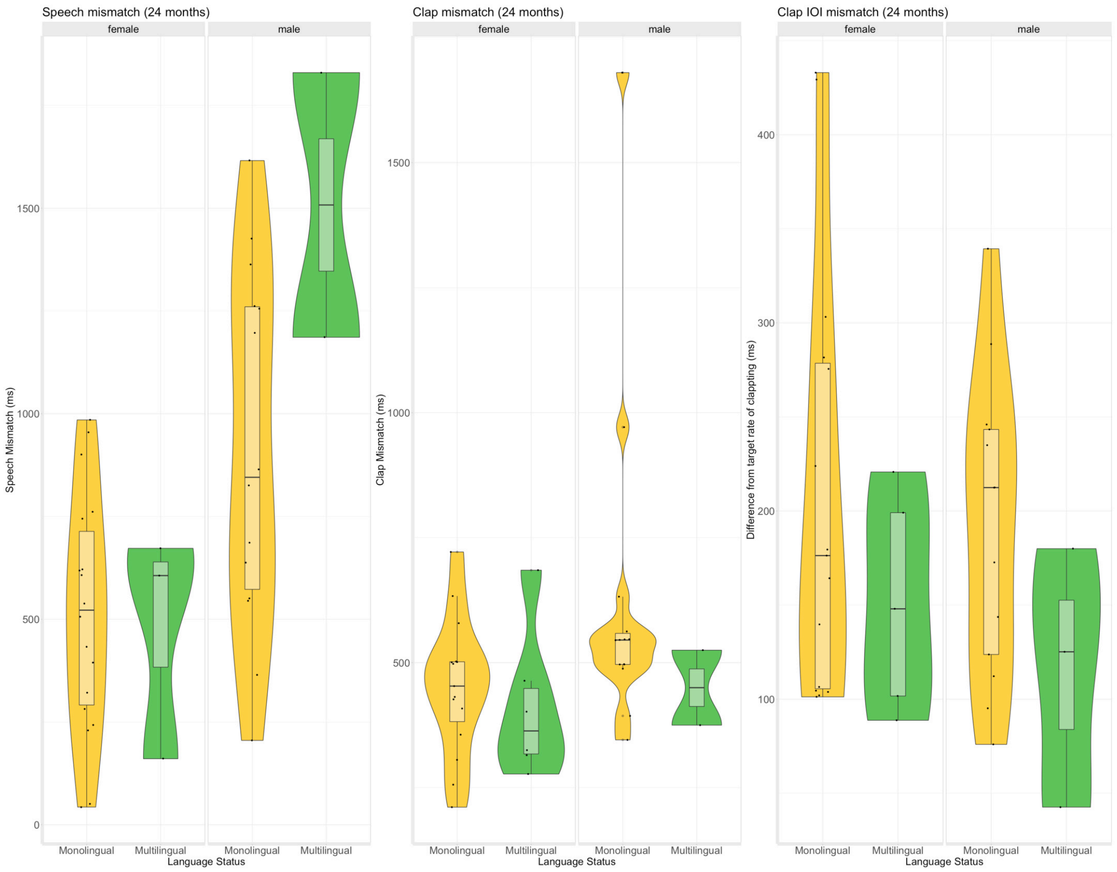

Figure 7. Descriptive statistics for Timing measures.

Violin plots presented for Nursery Rhyme Completion variables, overlaid with jittered raw data and shaded box plots. Performance on each measure is shown separately by infant sex (male/female) and linguistic status (mono/multilingual). Note that lower values indicate more accurate performance.

particularly important as although we had very good participant attendance at our early testing sessions from age 2 – 11 months, our neural measures are inherently more susceptible to data attrition due to infant fussiness, movement artefacts, technical issues, etc.). Further, the project attempted to assess an extremely broad range of language skills very early in life, and some measures attempted were revealed to be overly difficult for the infants, especially where advanced production was involved. Exclusion was therefore also made on the basis that infants, as a group, were performing at floor (i.e. performance was not above chance level, or the majority of infants did not attempt a response and remained silent during the task). Table 4 sets out which of variables will and will not be included in brain-behaviour analyses. Variables that are not selected here may be used for exploratory analyses in future work.

Three experimental measures were retained for longitudinal brain-behaviour analyses, namely pointing at 12-months, vocabulary knowledge (CCT at 18-months), and nonword repetition at 24-months. In addition, we have robust data from the parent-report CDI, which will also be retained for brain-behaviour analyses. Accordingly, longitudinal analyses up to age 24 months will use data from within three of our five domains of interest, semantic development (CDI, CCT), phonology (NWR) and gesture (pointing).

Table 4. Description of each of the experimental variables, with justification for exclusions from brain-behaviour analyses.

DV	Domain	Age	Select?	Justification for exclusion
Word Recognition	Semantics	8	No	Performance around chance
CCT	**Semantics**	**18**	**Yes**	
CCT	Semantics	24	No	Low sample size, not possible to collect full sample due to COVID-19 move to remote testing
Pointing	**Gesture**	**12**	**Yes**	
Pointing	Gesture	15	No	Low reliability with parent-report of pointing at 15-months, see supplementary materials
NWR Consonants	Phonology	18	No	Many infants performing at floor – majority of infants did not generate a response
NWR Syllables	Phonology	18	No	Many infants performing at floor – majority of infants did not generate a response
NWR Stress	Phonology	18	No	Many infants performing at floor – majority of infants did not generate a response
NWR Consonants	**Phonology**	**24**	**Yes**	
NWR Syllables	**Phonology**	**24**	**Yes**	
NWR Stress	**Phonology**	**24**	**Yes**	
Rhyme Oddity	Phonology	24	No	Small sample size, not possible to collect full sample due to COVID-19 move to remote testing
Nursery Rhyme Knowledge	Phonology	24	No	Small sample size, task introduced only after move to remote testing
Grammar	Grammar	24	No	Many infants performing at floor – majority of infants did not generate a response
NRC Speech mismatch	Timing	24	No	Many infants performing at floor – majority of infants did not generate a response
NRC Clap mismatch	Timing	24	No	Many infants performing at floor – majority of infants did not generate a response
NRC Clap tempo mismatch	Timing	24	No	Many infants performing at floor – majority of infants did not generate a response

Variables selected for future brain-behaviour analyses are displayed in bold.

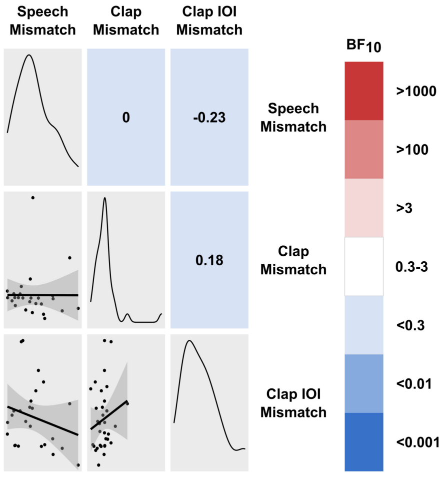

Figure 8. Correlogram showing relationships between timing measures at 24-months.

Discussion

The Cambridge UK BabyRhythm project has gathered a unique longitudinal dataset of measures of language development that may be of great interest to other researchers in the field. The current report offers a methodological contribution to the infant literature, presenting an overview of the different tasks used in the project. Empirically, we use robust Bayesian analyses to test how performance in each task reflects the linguistic status of the infant (monolingual versus multilingual) and whether the infant was male or female. Finally, we make a theoretical contribution to the literature, framing our results within Temporal Sampling theory and assessing which tasks are most robust regarding our planned brain-behaviour analyses. Of the five linguistic domains that were measured, namely semantic development, phonological development, grammatical development, gesture and rhythmic timing, only three domains provided sufficiently robust experimental data (CDI data aside) between the ages of 8 and 24 months. These were semantic development (CCT at 18 months), phonological development (NWR at 24 months) and gesture (pointing at 12 months). Infant performance in some of the other tasks used in the project was still largely at floor at 24 months, however data for these tasks (rhyming, grammatical development and rhythmic timing) were also collected at 30 months of age. These data are still being coded, leaving open the possibility that future work will be able to assess brain-behaviour relations for these domains also.

As noted, the project was impacted by COVID-19, which affected task delivery for many tasks. The CDI data gathered at 10, 12, 15, 18 and 24 months was the only measure not affected by COVID-19, as it is a parental report measure. As shown in Table 1, both comprehension and production as measured by the CDI mirrored relationships already well-established in the literature (Alcock et al., 2020). As infants got older, both word comprehension and word production improved, with comprehension outstripping production. One of the main aims of the BabyRhythm project is to identify from neural markers those infants that struggle to acquire spoken language ('late talkers'). As shown in Table 1, some infants in the sample were not yet producing any words at 24 months, while the most vocal infant was producing 609 words. This suggests that despite our relatively high-performing sample, late talkers will still be identifiable.

The infant-controlled measure of semantic knowledge, the CCT, showed a similar developmental pattern to the CDI regarding age-related improvement, with a median of 32% of items known at 18 months and 73% of items known at 24 months. We further see very strong evidence for substantial positive correlations between the CCT and the CDI at both ages tested. However, once COVID-19 arrived we could no longer administer the CCT, and only 60 infants in the sample had received the CCT at the 24-month testing visit at this point. Accordingly, the CCT at 18 months is selected as the more pragmatic experimental measure of vocabulary knowledge for our sample.

Regarding the domain of phonology, both nonword repetition and rhyming were measured. Only modest positive associations were found (see Figure 5). The rhyming measure was also impacted by COVID-19, resulting in a relatively small sample size. The nonword repetition task proved a robust measure of phonology, which may reflect the fact that it was originally designed to mirror the demands of learning new words (Gathercole, 2006). Indeed, in the developmental language disorder (DLD) literature, nonword repetition is now used as a diagnostic marker of DLD for both monolingual and multilingual children (Ahufinger et al., 2021). Our experimental NWR task also enabled us to score toddlers' phonological development in terms of different levels of phonology such as primary syllable stress, syllable and consonant phoneme accuracy. This stratification enables brain-behaviour investigations into whether some levels of phonology are more impacted by individual differences in neural entrainment than others, though we note that caution should be applied to future interpretation of the NWR stress measure, due to difference in performance related to in-person or remote administration type. On the TS hypothesis, delta band entrainment should be particularly important for the developing lexicon as it supports accurate processing of speech rhythm and prosody (Attaheri et al., 2022).

Regarding non-verbal communication or gesture, the pointing measure developed here was intended to measure individual differences in joint attention and shared communicative intent. This aim was not met in that no difference between pointing behaviour in the 'attend referent' and 'attend nothing' trials was found, however infants did differ considerably in whether they pointed at all in our paradigm. Dividing the infants by whether they pointed or not produced a robust measure of early gesture, as shown in Figure 3. Further analyses confirmed that at 12-months, our experimental measure of pointing was strongly associated with parental report of whether infants pointed via the CDI (see supplementary materials). Accordingly, pointing at 12 months is the final experimental variable retained for brain-behaviour analyses.

Investigation of each linguistic domain by sex and linguistic status produced few robust effects. We do not find strong evidence for differences dependent on whether the infant is exposed to one or multiple languages, but we should note that group size is not equal within our opportunity sample. Whilst we do see moderate evidence for no difference between monolingual and multilingual learners for some of our dependent variables, much of our evidence can be described as 'anecdotal'; more data collected from infants exposed to multiple languages would clarify differences and similarities. We do not see systematic differences between males and females across most tasks, except in the accuracy of their speech timing at 24-months, where we find strong evidence that females were more accurate than males.

In summary, the current paper documents the performance of typically developing infants on a range of language assessments, over the first two years of life. Detailed documentation of testing procedures and raw data are available open access on OSF. We identify three developmentally appropriate experimental measures that quantify infants' receptive vocabulary (CCT, 18-months), use of gesture (pointing, 12-months), and phonological skill (NWR, 24-months), that in addition to standardised vocabulary estimates from the CDI, will form the basis of brain-behaviour correlations for this sample in our further work.

Contributions

Contributed to conception and design: SR, ANC, AA, NM, SG, PAB, PB, UG
Contributed to acquisition of data: SR, ANC, AA, HOS, CG, IW, NM, SG, PAB, PB, CB, MAO
Contributed to analysis and interpretation of data: SR, ANC, AA
Drafted the article: SR, UG
Approved the submitted version for publication: SR, ANC, AA, HOS, CG, IW, NM, SG, PAB, PB, CB, MAO, UG

Competing Interests

The authors declare none.

Acknowledgements

We would like to thank Elika Bergelson for providing the stimuli for the Word Recognition task, Margaret Friend for help with the CCT, and the LuCiD Team for assistance with the UK-CDI. We are very grateful to all of the families who have taken part in our research, persevering even through very difficult times. We thank you for your continued support.

Funding

This project has received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 694786).

Data Accessibility Statement

All testing procedures, including acquisition and analysis SOPs, participant data, and analysis scripts, are available on OSF (link: https://osf.io/ftejv/).

Submitted: May 23, 2022 PST, Accepted: October 19, 2023 PST

References

Adams, A.-M., & Gathercole, S. E. (1995). Phonological Working Memory and Speech Production in Preschool Children. *Journal of Speech, Language, and Hearing Research*, *38*(2), 403–414. https://doi.org/10.1044/jshr.3802.403

Ahufinger, N., Berglund-Barraza, A., Cruz-Santos, A., Ferinu, L., Andreu, L., Sanz-Torrent, M., & Evans, J. L. (2021). Consistency of a Nonword Repetition Task to Discriminate Children with and without Developmental Language Disorder in Catalan–Spanish and European Portuguese Speaking Children. *Children*, *8*(2), 85. https://doi.org/10.3390/children8020085

Alcock, K., Meints, K., & Rowlan, C. (2020). *The UK Communicative Development Inventories Words and Gestures*. J&R Press Ltd.

Attaheri, A., Choisdealbha, Á. N., Di Liberto, G. M., Rocha, S., Brusini, P., Mead, N., Olawole-Scott, H., Boutris, P., Gibbon, S., Williams, I., Grey, C., Flanagan, S., & Goswami, U. (2022). Delta- and theta-band cortical tracking and phase-amplitude coupling to sung speech by infants. *NeuroImage*, *247*, 118698. https://doi.org/10.1016/j.neuroimage.2021.118698

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258. https://doi.org/10.1073/pnas.1113380109

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177.

Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: evidence from a twin study. *Journal of Child Psychology and Psychiatry*, *37*(4), 391–403. https://doi.org/10.1111/j.1469-7610.1996.tb01420.x

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy*, *2*(1), 29–49. https://doi.org/10.1207/s15327078in0201_3

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read—a causal connection. *Nature*, *301*(5899), 419–421.

Bryant, P. E., Bradley, L., Maclean, M., & Crossland, J. (1989). Nursery rhymes, phonological skills and reading. *Journal of Child Language*, *16*(2), 407–428. https://doi.org/10.1017/s0305000900010485

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), 2296. https://doi.org/10.1002/icd.2296

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), i. https://doi.org/10.2307/1166214

Dollaghan, C. A. (1994). Children's phonological neighbourhoods: half empty or half full? *Journal of Child Language*, *21*(2), 257–271. https://doi.org/10.1017/s0305000900009260

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Fenson, L., Bates, E., Dale, P. S., Marchman, V. A., Reznick, J. S., & Thal, D. J. (2007). *MacArthur-Bates Communicative Development Inventories*. Brookes Publishing Company.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i. https://doi.org/10.2307/1166093

Friend, M., & Keplinger, M. (2003). An infant-based assessment of early lexicon acquisition. *Behavior Research Methods*, *35*(2), 302–309.

Friend, M., & Keplinger, M. (2008). Reliability and validity of the Computerized Comprehension Task (CCT): data from American English and Mexican Spanish infants. *Journal of Child Language*, *35*(1), 77–98. https://doi.org/10.1017/s0305000907008264

Friend, M., Schmitt, S. A., & Simpson, A. M. (2012). Evaluating the predictive validity of the Computerized Comprehension Task: Comprehension predicts production. *Developmental Psychology*, *48*(1), 136–148. https://doi.org/10.1037/a0025511

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513–543. https://doi.org/10.1017/s0142716406060383

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, *28*(2), 200–213. https://doi.org/10.1016/0749-596x(89)90044-2

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences*, *15*(1), 3–10. https://doi.org/10.1016/j.tics.2010.10.001

Goswami, U., & Bryant, P. (2016). *Phonological Skills and Learning to Read*. Psychology Press.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hoff, E., Core, C., & Bridges, K. (2008). Non-word repetition assesses phonological memory and is related to vocabulary development in 20- to 24-month-olds. *Journal of Child Language*, *35*(4), 903–916. https://doi.org/10.1017/s0305000908008751

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' Detection of the Sound Patterns of Words in Fluent Speech. *Cognitive Psychology*, *29*(1), 1–23. https://doi.org/10.1006/cogp.1995.1010

Kalashnikova, M., Goswami, U., & Burnham, D. (2019). Sensitivity to amplitude envelope rise time in infancy and vocabulary development at 3 years: A significant relationship. *Developmental Science*, *22*(6), 12836. https://doi.org/10.1111/desc.12836

Kartushina, N., & Mayor, J. (2019). Word knowledge in six- to nine-month-old Norwegian infants? Not without additional frequency cues. *Royal Society Open Science*, *6*(9), 180711. https://doi.org/10.1098/rsos.180711

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831–843. https://doi.org/10.1038/nrn1533

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2007). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178. https://doi.org/10.1016/0010-0277(88)90035-2

Molnar, M., Gervain, J., & Carreiras, M. (2013). Within-rhythm Class Native Language Discrimination Abilities of Basque-Spanish Monolingual and Bilingual Infants at 3.5 Months of Age. *Infancy*, *19*(3), 326–337. https://doi.org/10.1111/infa.12041

Ortiz-Mantilla, S., & Benasich, A. A. (2013). Neonatal electrophysiological predictors of cognitive and language development. *Developmental Medicine &amp; Child Neurology*, *55*(9), 781–782. https://doi.org/10.1111/dmcn.12207

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time.' *Speech Communication*, *41*(1), 245–255. https://doi.org/10.1016/s0167-6393(02)00107-3

Ramírez-Esparza, N., García-Sierra, A., & Kuhl, P. K. (2014). Look who's talking: speech style and social context in language input to infants are linked to concurrent and future speech development. *Developmental Science*, *17*(6), 880–891. https://doi.org/10.1111/desc.12172

Rocha, S., Attaheri, A., Ní Choisdealbha, Á., Brusini, P., Flanagan, S. A., Mead, N., Boutris, P., Gibbon, S., Olawole-Scott, H., Grey, C., Williams, I., Ahmed, H., Macrae, E., & Goswami, U. (2021). *Infant sensorimotor synchronisation to speech and non-speech rhythms: A longitudinal study*. *10*(31234/osf.io/jbrga). https://doi.org/10.31234/osf.io/jbrga

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Steil, J. N., Friedrich, C. K., & Schild, U. (2021). No Evidence of Robust Noun-Referent Associations in German-Learning 6- to 14-Month-Olds. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.718742

Tomasello, M. (2000). Culture and Cognitive Development. *Current Directions in Psychological Science*, *9*(2), 37–40. https://doi.org/10.1111/1467-8721.00056

Tomasello, M. (2014). *Beyond names for things: Young children's acquisition of verbs*. Psychology Press.

Vihman, M. M., Nakai, S., DePaolis, R. A., & Hallé, P. (2004). The role of accentual pattern in early lexical representation. *Journal of Memory and Language*, *50*(3), 336–353. https://doi.org/10.1016/j.jml.2003.11.004

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63. https://doi.org/10.1016/s0163-6383(84)80022-3

Yuan, S., & Fisher, C. (2009). Really? she blicked the baby?": Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, *20*(5), 619–626.

Ziegler, J. C., & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin*, *131*(1), 3–29. https://doi.org/10.1037/0033-2909.131.1.3

Zinober, B., & Martlew, M. (1985). Developmental changes in four types of gesture in relation to acts and vocalizations from 10 to 21 months. *British Journal of Developmental Psychology*, *3*(3), 293–306. https://doi.org/10.1111/j.2044-835x.1985.tb00981.x

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/92998-language-acquisition-in-the-longitudinal-cambridge-uk-babyrhythm-cohort/attachment/194710.docx?auth_token=WElbpzzbXm7sJf3VYo6A

Supplemental Material

Download: https://collabra.scholasticahq.com/article/92998-language-acquisition-in-the-longitudinal-cambridge-uk-babyrhythm-cohort/attachment/194711.docx?auth_token=WElbpzzbXm7sJf3VYo6A