Infant Behavior and Development 77 (2024) 102001

Contents lists available at ScienceDirect

Infant Behavior and Development

journal homepage: www.elsevier.com/locate/inbede

Registered Report Stage II

Predicting language outcomes at 3 years using individual differences in morphological segmentation in infancy

Jinyoung Jo [a,*,1], Megha Sundara [a], Canaan Breiss [b]

[a] UCLA Department of Linguistics, 3125 Campbell Hall, Los Angeles, CA 90095-1543, United States
[b] University of Southern California, Department of Linguistics, 301E Grace Ford Salvitori Hall, Los Angeles, CA 90089, United States

ARTICLE INFO

Keywords:
Looking time
Bayesian analysis
Individual differences
Language sample analysis
Language delays

ABSTRACT

In previous research, infants' performance on speech perception tasks has been shown to predict later language outcomes, typically vocabulary size. We used Bayesian analyses to model trial-level looking time behavior of individual infants on morphological segmentation experiments. We compared the usefulness of Bayesian estimates and the raw looking time difference measures used in previous studies to predict (a) vocabulary size at 30 months and (b) outcome measures obtained from language samples elicited via a picture description task at 36 months. We found that both estimates of morphological segmentation reliably predicted expressive vocabulary at 30 months. The Bayesian estimate also credibly predicted the correct use of verb tense morphemes obtained from the language sample. We therefore conclude that the Bayesian estimate is better for indexing individual differences in segmentation tasks and more useful for predicting clinically relevant language outcomes.

1. Introduction

There is increasing evidence that performance on speech perception tasks in infancy predicts later language outcomes. Such a relationship has been demonstrated using behavioral measures obtained from habituation paradigms or the Head-turn Preference Procedure (Cristia & Seidl, 2011; Newman et al., 2006; Newman et al., 2016; Höhle et al., 2014; Singh et al., 2012; Singh, 2019) as well as neurophysiological ones (Junge, 2011; Junge et al., 2010; Junge et al., 2012; Junge & Cutler, 2014; Kidd et al., 2018; Kooijman et al., 2013; Marimon et al., 2022; Von Holzen et al., 2018; Weber et al., 2005). Specifically, in a recent meta-analysis of 18 publications with multiple methods (behavioral: conditioned head-turn, headturn preference procedure; neural: evoked response potentials), it has been shown that there is a weak, positive correlation (median coefficient: 0.31) between early speech perception skills and vocabulary measures in older infants (Cristia et al., 2014).

Establishing a relationship between speech perception skills in infancy and later language outcomes is useful for two purposes. Consistent relationships between individual variation on specific speech perception tasks with measures like vocabulary size lend support to hypotheses about the mechanisms by which infants tune into their native language. For this purpose, researchers often retrospectively compare two groups of children differing in language outcomes (e.g. larger vs. smaller vocabulary size), to determine if their performance on perception studies conducted earlier in infancy are different (Newman et al., 2006; Newman et al., 2016; Junge et al., 2012; Kidd et al., 2018; see also Singh et al., 2012). Other times, in prospective analyses, infants are divided into two groups

* Corresponding author.
E-mail addresses: jinyoungjo@ucla.edu (J. Jo), megha.sundara@humnet.ucla.edu (M. Sundara), cbreiss@usc.edu (C. Breiss).
[1] https://orcid.org/0009-0004-8174-3638

https://doi.org/10.1016/j.infbeh.2024.102001
Received 9 April 2024; Received in revised form 14 October 2024; Accepted 22 October 2024
Available online 30 October 2024

according to their performance on perception tasks, with subsequent evaluation of language outcomes (e.g. Junge et al., 2010; Kooijman et al., 2013, Von Holzen et al., 2018). Such group-level analyses are advantageous given the inherent noisiness of infant data as well as the small number of trials in infant experiments.

Variability across infants can also be indexed using individual-level analyses. In such analyses, an estimate of each infant's performance is correlated with their language outcome, typically vocabulary size (Newman et al., 2006; Singh et al., 2012; Newman et al., 2016; Singh, 2019; Wang et al., 2021; Junge et al., 2012; Kidd et al., 2018). Such correlational analyses are a necessary first step towards determining how measures obtained from speech perception tasks, collected in the first year of life, may be used for the early identification of children at risk for language delays and disorders. In this paper, we used both group- and individual-level analyses to determine how (if at all) individual variation in behavioral measures derived from speech perception tasks collected in infancy prospectively predict later language outcomes.

Table 1
Summary of studies on the relationship between infant speech perception and later language outcomes.

Study	N	Age (early → later)	Methods	Perception measure	Language outcome	Group-level / Individual-level analysis
Newman et al. (2006)	119	5–12 mos → 24 mos (Study 1)	HPP	Looking time difference	Expressive vocabulary (MCDI)	Group-level analysis
Singh et al. (2012)	40	7.5 mos, → 8–24 mos	HPP	Looking time difference	Expressive vocabulary (MCDI)	Group-level analysis and individual-level analysis
Newman et al. (2016)	96	7 mos → 24 mos	HPP	Looking time difference	Expressive vocabulary (MCDI)	Group-level analysis and individual-level analysis
Singh (2019)	17	10–11 mos → 24 and 36 mos	Habituation	Looking time difference	24 mos: Expressive vocabulary (MCDI) 36 mos: Receptive vocabulary (PPVT)	Individual-level analysis
Wang et al. (2021)	97	5–7 mos → 18, 24 mos	HPP, Habituation	Proportion of looking time	18 mos: Receptive and expressive vocabulary, (MCDI) 24 mos: Expressive vocabulary (MCDI)	Individual-level analysis
Junge et al. (2010), Junge (2011; Ch. 5), Kooijman et al. (2013)	23	7 mos →36 mos	EEG	ERP difference	Receptive language (Reynell Test voor Taalbegrip) Productive language (Schlichting Test voor Taalproductie)	Group-level analysis and individual-level analysis
Junge et al. (2012), Junge (2011; Ch. 3)	28	10 mos → 24 mos	EEG	ERP difference	Receptive and expressive vocabulary (MCDI)	Individual-level analysis
		10 mos → 12 mos			Receptive vocabulary (MCDI)	Group-level analysis and individual-level analysis
					Expressive vocabulary (MCDI)	Individual-level analysis
Kidd et al., 2018	99	9 mos → 12, 15 mos	EEG	ERP diff	Receptive vocabulary (MCDI)	Group-level analysis and individual-level analysis
Junge (2011; Ch. 6) Junge & Cutler, 2014	23	10 mos → 5 years	EEG	ERP diff	Receptive language (Reynell Test voor Taalbegrip) Productive language (Schlichting Test voor Taalproductie)	Group-level analysis
Junge (2011; Ch. 4)	25	10 mos → 16 mos	EEG	ERP diff	Receptive vocabulary (look-while-listen paradigm)	Individual-level analysis
Von Holzen et al. (2018)	22–26	8 mos → 8, 13, 16, 24 mos	EEG	ERP diff	Expressive vocabulary (French CDI)	Group-level analysis
Marimon et al. (2022)	24	9 mos → 40 mos	eye-tracking	Temporal alignment of pupillary changes with statistical or prosodic word	Expressive vocabulary and grammar (SBE−2-KT)	Individual-level analysis
Weber et al. (2005)	18	5 mos → 12 mos, 24 mos	EEG	ERP diff	Expressive vocabulary (ELFRA)	Individual-level analysis
Cristia & Seidl, 2011	24	6 mos → 24 mos	HPP	Proportion of looking time	Expressive vocabulary (MCDI)	Group-level analysis
Höhle et al. (2014)	34	4 mos → 5 years	HPP	Looking time difference	Sentence comprehension and morphological skills (SETK3–5)	Individual-level analysis

1.1. Language outcomes

In retrospective as well as prospective studies, vocabulary size, either expressive or receptive, is the most common outcome measure used to evaluate language skills of toddlers (Cristia et al., 2014; Newman et al., 2016; Singh et al., 2012; Newman et al., 2016; Singh, 2019; Wang et al., 2021; Junge et al., 2012; Kidd et al., 2018; see Junge et al., 2010; Kooijman et al., 2013 for exceptions). This is presumably due to the ease with which vocabulary can be estimated using parent questionnaires like the MacArthur Bates Communicative Development Inventory (MCDI; Fenson et al., 1993) and standardized tests like the Peabody Picture Vocabulary Test (PPVT; Dunn, 2019). We also know that vocabulary size and its composition can be used to predict language development (Hahn Arkenberg et al., 2021; Perry et al., 2023). The exclusive use of vocabulary measures as a proxy for language outcomes, however, is limiting for several reasons.

First, it is not likely that performance on all speech perception tasks is related to vocabulary size. For example, Newman et al. (2006) report that infants with larger vocabularies at 24-months had performed better on word segmentation, but not language discrimination in infancy.

It is also not likely that performance on any specific speech perception task necessarily impacts only vocabulary size, and no other language outcome. For example, Kooijman et al. (2013) showed that infants who at 7 months had a better segmentation ability were better at making sentences of a similar structure to those given by the experimenter in a picture or toy description task. More generally then the choice of speech perception task as well as the language outcome measure has consequences for our hypotheses about developmental mechanisms. In this study, we investigate the relationship between two different indices of infants' speech perception skills with multiple language outcomes in addition to vocabulary size.

Second, vocabulary size measured by standardized tests is not the most preferred outcome used to assess risk for language impairment. Instead, in the clinical setting, norm-referenced standardized tests for identifying language disorders are most often supplemented by language sample analyses. Language samples are typically elicited from children in play sessions (Miller & Chapman, 1981; Watkins et al., 1995) or in a more controlled paradigm with picture descriptions (e.g., Eisenberg & Guo, 2015). Fine-grained analyses of these language samples can provide many quantitative measures that have been shown to differentiate children with and without language impairment (Hux et al., 1993; Kemp & Klee, 1997; Loeb et al., 2000) although the interrelationships among these measures themselves are not yet clear (Yang, Rosvold, et al., 2022). Quantitative measures derived from language samples include measures of lexical diversity such as type-token ratio and the number of different words used (Finestack & Satterlund, 2018; Watkins et al., 1995), mean length of utterance (e.g., Eisenberg & Lundgren, 2001; Loeb et al., 2000), the Index of Productive Syntax (Scarborough, 1990), use of verb tense marking (e.g., Bedore & Leonard, 1998; Rice et al., 1995), and percentage grammatical utterances (e.g., Eisenberg & Guo, 2013). Thus, in this paper we supplement measures of vocabulary size and composition with measures derived from language samples to better identify outcomes that could identify children at risk for impairment.

Finally, many children with small vocabularies at earlier ages catch up with their peers by age 3 (Leonard, 1998; Rescorla & Lee, 2000), especially when expressive vocabulary is considered. In comparison, the likelihood of a persistent impairment increases for those that do not catch up on production vocabulary (Rescorla & Lee, 2000; Rescorla & Schwartz, 1990), as well as for children who lag on both comprehension and production vocabulary (Law et al., 2000). Thus, language outcome measures are most informative for assessing risk of language impairment when they are evaluated at age 3, not earlier. This is a challenge because speech perception abilities are often assessed in infancy, and the 2+ years between the evaluation of speech perception abilities and language outcomes increases attrition in the sample being followed. This attrition also adds to the ubiquitous problem of small sample sizes in the developmental literature, which is particularly limiting when evaluating meaningful individual differences. Unsurprisingly then, there are relatively few published studies (Table 1) where infants were tested at 3-years or later (exceptions Höhle et al., 2014; Junge, 2011; Junge et al., 2010; Junge & Cutler, 2014; Kooijman et al., 2013; Marimon et al., 2022; Singh, 2019). Among such studies, only one (Höhle et al., 2014) had more than 30 participants.

In the present study, we evaluated two sets of language outcomes. To replicate previous findings, we evaluated vocabulary size using the MCDI at 30 months. To further our goal of developing ways to identify infants at risk for language impairment, we additionally evaluated quantitative outcome measures from a language sample analysis; the language sample was obtained from a picture description task administered at 36 months. At present (Table 1) there are no studies relating early speech perception skills to measures derived from language samples, the main outcome of interest in the present study.

1.2. Assessing individual differences in speech perception tasks

What makes relating early speech perception skills to later language outcomes challenging, both at the group and individual level, is the inherent noisiness of infant data due to the small number of trials used to test infants. Two types of solutions have been proposed to address this challenge (Houston et al., 2007; de Klerk et al., 2019). These include (a) the use of statistical techniques to get robust estimates as well as (b) novel methodologies where each infant is tested on many more trials. To generate estimates in this paper, we opted for the first approach: we used Bayesian Hierarchical Regression (de Klerk et al., 2019) to analyze data collected from the standard segmentation paradigm implemented using the classic Headturn Preference Procedure (Jusczyk & Aslin, 1995).

In all previous behavioral experiments in Table 1, including the present study, indices of individual differences were derived from looking time differences. In some experiments only the direction of difference between familiar and novel stimuli was used for group-level comparisons (Newman et al., 2006), in others vocabulary scores were correlated with the magnitude of looking time difference to demonstrate a relationship between the two (Singh et al., 2012; Singh, 2019). Finally, in still others, in order to scale across infants who differ in their absolute looking times, the proportion of looking time to one kind of trial was used to index individual performance

instead (Wang et al., 2021). In order to compare our findings with published studies, in the present study we also used looking time difference as one individual-level predictor.

However, the use of looking time differences to index individual variation – whether categorical (considering only the direction of the preference) or gradient (considering the degree of the looking time difference) – has long been contentious (Arterberry & Bornstein, 2002; Aslin, 2007; Aslin & Fiser, 2005; de Klerk et al., 2019; de Klerk et al., 2021). Conceptually, as Aslin and colleagues point out, there is no linking hypothesis that attributes a linear relationship between absolute (or proportional) differences in looking time and infants' ability to encode and detect stimulus distinctiveness. That is, these paradigms are designed to address whether infants succeed, not how well they do so. Statistically, this raises a different question: how far above 0.50 does the preference for one kind of stimulus need to be to reflect above chance performance (Arterberry & Bornstein, 2002; de Klerk et al., 2019; 2021)? Establishing a threshold to detect above chance performance is particularly challenging when examining individual-level variation in experiments with a small number of trials (as low as 2 in some experiments).

In some behavioral experiments, particularly those investigating segmentation, the number of trials can be as high as 12 (present study) or even 16 (some experiments in Newman et al., 2006), somewhat mitigating the sparse sampling of individual-level performance. With greater number of trials per individual, it is possible to fit statistical models at the level of the individual in order to quantify task performance (Houston et al., 2007). Houston et al. (2007) propose that a crucial component in this effort is the addition of a first-order autoregressive term (AR1 error structure), which allows looking time on every trial to be modeled as a function of the looking time to the previous trial. The use of an autoregressive term can thus take into account the correlation in noise between trials, and crucially, separate that noise from a potential underlying signal to get a better estimate of individual task performance. Using a non-hierarchical, frequentist approach, Houston et al. (2007) demonstrate how this can be implemented.

Most recently, de Klerk et al. (2019) use Bayesian Hierarchical Regression to improve upon Houston et al.'s (2007) approach. In Bayesian Hierarchical Regression, group and individual effects are modeled in a single analysis, instead of running regressions on each infant's looking time separately. This approach, they argue, following Gelman and colleagues (Gelman, 2006; Gelman et al., 2012; Gelman & Tuerlinckx, 2000), circumvents the possibility of chance findings due to many separate individual analyses and reduces uncertainty in estimates of individual parameters. We adapted de Klerk et al.'s (2019, 2021) approach to generate a Bayesian estimate of individual performance on a morphological segmentation task. This was our second predictor, besides the looking time difference between familiar and novel trials.

1.3. The present study

In the present study, we used 6- and 8-month-olds' performance on a morphological segmentation task as our early predictor. For this, we reanalyzed the data reported in Kim and Sundara (2021) and Sundara et al. (2021). As in the classic word segmentation paradigm, infants were familiarized with 2 of 4 suffixed target verbs in passages. This was done until infants accumulated at least 45 s of looking time to each target word. Then in the test phase, infants's recognition of the isolated stems was evaluated in 12 trials (4 stems $\times$ 3 blocks $=$ 12 trials). The dependent variable in all experiments was looking time (a proxy for listening time). All trials were infant-controlled, that is, presentation of auditory stimuli was completely contingent on infant looks; and looking times to the two familiar and two novel verbs were statistically compared. In these experiments infants at both ages demonstrated that they are able to segment verbs and relate them to stems, by listening significantly longer to familiar stems compared to novel ones. A familiarity preference is also typical in segmentation experiments using natural language stimuli with young infants (e.g. Houston & Jusczyk, 2000; Jusczyk et al., 1999; Jusczyk & Aslin, 1995). Note that in Bayesian Hierarchical Regression, used in this paper, it is not necessary to make *a priori* predictions of the direction of preference - whether for familiar or novel trials. The model can accommodate either familiarity or novelty preferences, because they are modeled simultaneously. We describe the direction of group preferences in these experiments simply to characterize the data set.

We followed de Klerk et al. (2019; 2021) in modeling trial-level performance of individual infants. Further, we extended their approach by benchmarking individual infants against group data from 238 infants aged 6-month or 8-month who were also tested on the same morphological segmentation tasks. This data, originally reported in Kim and Sundara (2021) and Sundara et al. (2021), was reused to include a larger sample beyond the subset for whom we have language outcome data. By benchmarking individuals against the group data, we can get a better estimate of task performance of individual infants relative to a cohort of peers.

In the Bayesian analysis, we modeled looking time as a function of trial-type (novel vs. familiar) and the model-inferred looking time of the previous trial (autoregressive structure) to obtain individual-level and group-level (i.e. the 9 experiments) estimates simultaneously. The model included a random intercept for each infant, and a random slope for the effect of trial-type for each infant nested within an experiment. The random intercept simply indexes individual differences in baseline looking times; it is the random slope parameter estimate that was used to represent segmentation performance at the individual-level, which we call the Bayesian Estimate.

To compare our results with previous findings of a relationship between looking time difference and vocabulary size, we assessed expressive vocabulary at 30 months - the oldest age at which norms are available (Fenson et al., 1993). In this exploratory study, we were primarily interested in the relationship between each of the two early indices of individual performance on morphological segmentation tasks - looking time difference and the Bayesian Estimate, and outcomes evaluated from the language sample. For this we elicited language samples via a picture description task when the toddlers were 3 years old (i.e., 36 months). This is the age where language outcomes are most informative for assessing risk for language impairment, as late talkers who do not catch up with their peers in production vocabulary by age 3 face a higher likelihood of being diagnosed with a long-term language impairments (Rescorla & Lee, 2000; Rescorla & Schwartz, 1990). Based on pilot testing, this was also the earliest age at which we were able to reliably collect

language samples remotely.

Ideally, language outcomes for clinical use (i) improve or increase in value as a function of child age, at least in typically developing children, (ii) reliably distinguish typically developing children from language-impaired peers and (iii) index one specific aspect of development (e.g. lexical) not confounded by others (e.g. grammatical) (Yang, Rosvold, et al., 2022). With these criteria in mind, we selected language outcome measures to provide a composite view of lexical and grammatical development in 3-year-olds. We calculated one measure of lexical development, the number of different words in the first 100 words of a language sample (NDW; Watkins et al., 1995), alongside several measures of grammatical development: mean length of utterance in morphemes (MLU; average number of morphemes per utterance), grammaticality of utterances and the use of verb tense marking, and the Index of Productive Syntax (IPSyn).

We opted not to include the more commonly used Type Token Ratio (TTR) as a proxy for lexical diversity. Although TTR is one of the two measures most commonly used by speech-language pathologists along with MLU (Finestack & Satterlund, 2018), it is extremely sensitive to the size of the language sample. Moreover, there is evidence that TTR does not increase systematically with age, even in typically developing children (Templin, 1957; Yang, Rosvold, et al., 2022), and, crucially, it does not distinguish between typically developing children and children with language-impairment (Watkins et al., 1995; Yang, Rosvold, et al., 2022). Due to these reservations, various researchers have cautioned against using TTR for clinical purposes (Charest et al., 2020; Yang, Rosvold, et al., 2022). NDW, in comparison, has been shown to increase with age and to differentiate typically developing children and those with language disorders (Klee, 1992; Watkins et al., 1995; Heilmann et al., 2010; Charest et al., 2020).

In sum, besides replicating previous findings of a relationship between early speech perception abilities indexed by looking time differences and later vocabulary size, our aim was to assess the predictive validity of the Bayesian Estimate and whether either of these early measures can inform us about language outcomes derived from a language sample analysis. Because morphological segmentation is a critical step in language acquisition, we expected to find a robust relationship between an estimate of early perception on a morphological segmentation task and clinically relevant language outcomes at age 3. Establishing such a relationship would provide an opportunity to develop methods for early identification of children at risk for language delays.

2. Methods

2.1. Subjects

We recruited 51 toddlers (female = 23) between Feb. 2021 and June 2022 to evaluate multiple language outcomes at 30 months (age mean = 29.6, range = 29–30) and at 36 months (age mean = 36.8, range = 35–38). Note that only a subset of the toddlers provided data at each age (n = 24 at 30 months; n = 36 at 36 months), as only some of the infants who participated in the segmentation experiments were 30 or 36 months old when language sample collection began. All toddlers were monolingual English speakers (percent English exposure mean = 98.5, range = 90–100). All language outcomes were evaluated remotely because of the COVID-19 pandemic. There were no parent reports of a hearing, speech or language impairment in any of the children who participated in these follow-ups. The study was conducted in accordance with the Institutional Review Board of University of California, Los Angeles (#10-001562).

2.2. Benchmarking group

To derive the Bayesian Estimate, we benchmarked the morphological segmentation performance of these 51 toddlers when they were 6- or 8-months-old against a cohort of their 238 peers (female = 119) who participated in one of the 9 morphological segmentation experiments. The behavioral data from the cohort was collected in the lab between 2013 and 2020.

2.3. Quantifying individual variation

2.3.1. Looking time difference as a proxy for individual segmentation performance

The first measure we used to index individual segmentation performance was the raw looking time difference between the familiar trials and the novel trials (available on the project OSF page). Our aim was to replicate previous studies that used looking time differences to predict later language outcomes. In segmentation studies reported here successful segmentation was demonstrated by longer looking times to familiar trials. Thus, positive looking time differences are consistent with successful segmentation. We also reran our analyses with proportion look time to familiar trials as a predictor; the pattern of results were similar to that for raw looking times, so they are not reported here.

2.3.2. Bayesian Estimates of individual segmentation performance

We modified the Stan code provided by de Klerk et al. (2019) to extend to the 9 experiments from which we draw our subjects. The Bayesian Hierarchical Model jointly infers posterior distributions over individual infants' preference for the familiar condition in each experiment, pooling information about the expected strength and variability of this preference between infants within each experiment. The model also included an autoregressive error term, where the looking time on each non-initial trial partially depended on the model-inferred looking time on the previous trial. The code for the model we implemented is available on the OSF page.

The model was fit using the *cmdstanr* package (Gabry et al., 2023) in the R programming environment (R Core Team, 2021), using a No U-Turn Sampler to draw 3500 samples in each of four chains from the posterior distribution over parameter values, the first 1000 of

which were discarded for warm-up. Priors followed those of de Klerk et al. (2019), and we carried out the same sensitivity analyses, available on the OSF page. The functions in the *posterior* R package (Bürkner et al., 2023, package version 1.4.1) and the *tidybayes* R package (Kay, 2022; package version 3.0.2) were used to manipulate and summarize the samples from the fitted models. We present the median and 95 % Credible Interval of credible values of the familiarity preference of each infant; raw values for the Bayesian Estimates for the entire cohort of 238 infants are available on the OSF page.

2.4. Language outcome variables

We used two follow-up tasks to assess children's language outcomes. We evaluated expressive vocabulary at 30 months and we collected language samples at 36 months to supplement norm-referenced vocabulary tests. Besides the language outcomes we report below, we also conducted the PPVT with both 30-month-olds and 36-month-olds to assess receptive vocabulary. However, we decided not to analyze the data due to the small sample sizes (n=19 for 30 months, n=16 for 36 months). The PPVT data is included in the supplementary material uploaded in the OSF.

2.4.1. Expressive vocabulary at 30 months

As in previous experiments (e.g. Kidd et al., 2018, Newman et al., 2016, Wang et al., 2021), we obtained MCDI (Fenson et al., 1993) scores from twenty-four 30-month-olds as a measure of children's expressive vocabulary. Caregivers completed the Web-CDI form online (De Mayo et al., 2021). The number of lexical items produced by the child (range = 144–676) was used as the measure of expressive vocabulary size.

2.4.2. Language sample at 36 months

We elicited language samples from thirty-six 36-month-olds using a picture description task administered by research assistants on Zoom. We followed the procedure described in Eisenberg & Guo (2013, 2015). Each child was presented 7 randomly chosen colored line drawings from a set of 15. The drawings were unrelated to one another. The order of picture presentation was also randomized. To start the session, each toddler was seated in front of the laptop facing the webcam. The experimenter used prompts provided by Eisenberg & Guo, available on the OSF page, to elicit the children's utterances. When the child was unwilling to interact with the experimenter, the caregiver was allowed to prompt the child. After the child finished describing each picture, the experimenter praised them verbally and drew stars on the screen. Each session lasted 20–25 minutes, and was video-recorded on Zoom for transcription.

The children's productions in the Zoom recordings were transcribed by undergraduate research assistants using guidelines developed by the author JJ. We used a subset of the CHAT conventions (MacWhinney, 2000) necessary to capture only the information that was crucial to our analysis (set of symbols used available on the project OSF page). JJ went through a second pass of all the transcripts to resolve any inconsistencies across transcribers.

We then used the *mor* function in CLAN to identify the morphological boundaries along with the syntactic categories of each morpheme. The *kideval* function in CLAN was used to calculate MLU, NDW and IPSyn scores. Because neither of the two predictors, looking time difference or the Bayesian measure, was useful in predicting the MLU, we do not report the analysis for MLU here, but it is available on the project OSF page for completeness.

We used the transcripts to quantify the use of verb tense morphemes (Bedore & Leonard, 1998; Leonard et al., 1992) and the grammaticality of utterances (Eisenberg & Guo, 2013; 2015) produced by individual children. For each child's transcript, two research assistants were assigned to code for both measures. The author JJ reviewed the coded transcripts and resolved any inconsistencies between the two coders. For the tense morpheme usage, we first identified obligatory contexts for the morphemes of interest (past tense *-ed*, third person present tense *-s*, copula and auxiliary *be* and auxiliary *do*), and marked a context as containing an error either when the morpheme was omitted or was produced with a tense or agreement error. The agreement rate between the two coders was 91.4 % for identifying obligatory contexts, and 90.7 % for determining whether the tense morpheme was correctly used in contexts that both coders considered as obligatory. For grammaticality of utterances, we identified utterances that contain one or more morphological, syntactic or semantic errors, following Eisenberg & Guo, (2013; 2015). The agreement rate between the two coders was 82.4 % for determining whether an utterance was eligible for scoring, and 86.7 % for determining whether the utterance was grammatical when both coders considered it eligible. The instructions provided for the research assistants, which included what counts as an utterance and what counts as an error for the grammaticality coding, are shared on the OSF page.

2.5. Analyses

We carried out group-level as well as individual-level analyses to predict expressive language vocabulary using individual measures of segmentation. Group-level analyses were done to allow comparison to published findings on vocabulary size (Cristia & Seidl, 2011; Newman et al., 2006; Newman et al., 2016; Singh et al., 2012). Specifically, 30-month-olds were split into two groups based on median scores on the individual measures of segmentation (either raw looking time difference or the Bayesian Estimate), and the subsequent group below and above the median was compared on the language outcome using Bayesian regression.

For individual-level analyses, we used the participant specific measure of segmentation (either raw looking time difference or the Bayesian Estimate) to predict the language outcome measure for that individual also using Bayesian regression implemented in the *brms* package (Bürkner, 2017, package version 2.18.0). Each model used a Normal(0,1) prior on the coefficients, while the prior for the intercepts varied depending on the model (sensitivity analyses available on the OSF page), and sampling followed the procedure described above in Section 2.3.2. We conducted separate regression analyses for the 30-month and 36-month time points. The

regression model for 30 months used expressive vocabulary as the dependent variable. For 36 months, we established four separate regression models, one for each dependent variable obtained from language samples: Number of Different Words (NDW), correct use of verb tense morphemes, grammaticality of utterances, or Index of Productive Syntax (IPSyn).

In each model, predictor variables were centered and scaled to ease cross-comparison of effect size, but we plot unscaled versions for interpretability. We report the median value for the coefficient associated with the predictors of interest along with their 95 % Credible Interval. If the credible interval includes 0, we also report the probability of an effect in the direction of the sign of the coefficient (p-direction), regardless of magnitude, by examining the proportion of samples of the posterior distribution over coefficient values that fall to one side of zero. This measure ranges from 0.5, when the posterior distribution is centered on zero, indicating equal evidence for an effect in the direction of the sign of coefficient as for one in the opposite direction, to 1 when all of the posterior samples for the coefficient lie to one side of zero, indicating strong evidence for a directional effect. In addition to assessing probability of a directional effect, we can also understand model results in further detail by examining the median and width of the posterior distribution of coefficient values to learn about the uncertainty and magnitude of the effect. In this paper we interpret p-direction values greater than 0.95 (95 %) to indicate a scientifically meaningful degree of evidence, however, with the Bayesian analysis one could evaluate the relationship between the variables in a more nuanced, graded manner using the p-direction measure, which we also provide.

3. Predicting later language outcomes using measures of morphological segmentation

Before we investigate whether either of the looking time differences or the Bayesian Estimate can predict the language outcomes, we first examined the relationship between these two predictors and the relationship among the language outcome measures. The complete analyses and accompanying figures are available on the project OSF page. There was a moderate positive correlation between the two predictors ($r = 0.57, p < .001$). As expected, we observed a moderate to high correlation between the measures of grammatical development, MLU and IPSyn ($r = 0.75$, $p < 0.001$) and percent grammatical utterances and IPSyn scores ($r = 0.54$, $p < 0.01$). Additionally, there was an inverse correlation between the lexical measure NDW and grammaticality of utterances produced at 36 months ($r = -0.41, p < 0.05$). This is perhaps not surprising because grammatical utterances tend to have a greater number of function words (e.g, *the*, *of, and*), which is likely to lower lexical diversity. Finally, the correlation between the two lexical measures - expressive vocabulary evaluated with a parental questionnaire at 30 months and NDW based on the language sample collected at 36 months was positive ($r = 0.28$), although not statistically significant, likely because of the small sample size.

3.1. Expressive vocabulary at 30 months

To replicate and extend previously published results, we evaluated the relationship between individual measures of morphological segmentation (raw looking times as well as the Bayesian Estimate) obtained between 6- and 8-months of age with productive vocabulary measured using the MCDI at 30 months. We used a negative binomial regression because the expressive vocabulary was count data represented by positive integers and exhibited overdispersion.

We first conducted a group-level analysis to determine if we could replicate findings of group differences (Cristia & Seidl, 2011; Newman et al., 2006; Newman et al., 2016; Singh et al., 2012). For this, we split the participants (n = 24) into two groups, those with looking time differences that were lower than the median (Low group), and those whose looking time difference was higher than the median (High group), as has been done previously (Cristia & Seidl, 2011). Then, the MCDI production vocabulary of toddlers in the two groups were compared (Fig. 1A). We did a similar median split for participants using the Bayesian Estimate as well (Fig. 1B).

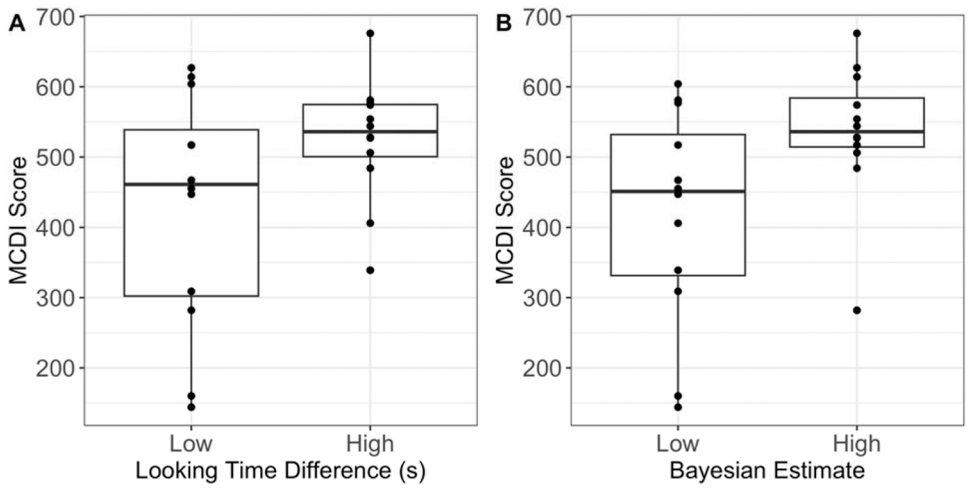

Fig. 1. Group-level analysis of MCDI score predicted by looking time difference (A) and Bayesian Estimate (B).

Results showed a near-credible group effect for the raw looking time difference (β = 0.22, [-0.07, 0.50]; p-direction: 93.7 %) as well as a credible group effect for the Bayesian Estimate (β = 0.26, [-0.02, 0.54]; p-direction: 96.7 %). The near credibility of the effect of raw looking time difference is likely due to the power (small sample size of n = 24). Therefore, we have evidence that grouping infants based on either of the two measures consistently distinguishes toddlers with low and high expressive vocabulary.

Next, for the individual-level analysis we used the continuous, by-participant measures of segmentation (both raw looking times and the Bayesian Estimate) to predict MCDI productive vocabulary scores (Fig. 2A and B respectively). Again, increase in raw looking time difference (β = 0.11, [-0.01, 0.24]; p-direction 96.9 %) as well as the Bayesian Estimate of segmentation (β = 0.13, [−0.01, 0.26]; p-direction 96.8 %) credibly predicted MCDI expressive vocabulary. The 95 % Credible Interval for both estimates did not include 0 so we can be confident that either variable is a consistent predictor of MCDI productive vocabulary at 30 months.

3.2. Number of different words in the first 100 words produced at 36 months

Next, we calculated the number of different words appearing in the first 100 words (NDW) in the language sample at 36 months (n = 34; two children excluded due to the language sample being smaller than 100 words). In two separate analyses, we used individual children's looking time differences and the Bayesian Estimates to predict their NDW (Fig. 2C and D). We used poisson regression since the dependent variable was count data, represented with positive integers, and did not exhibit overdispersion.

Looking time differences did not credibly predict NDW (β = −0.03, [-0.08, 0.03], p-direction: 83.9 %), nor did the Bayesian Estimate (β = −0.03, [-0.08, 0.02], p-direction: 86.9 %). A larger sample size is likely to have made the results credible in the negative direction. Thus, neither of the two predictors were useful in predicting lexical diversity as measured by the NDW at 36 months.

3.3. Correct use of verb tense morphemes at 36 months

We then investigated whether either of the two segmentation measures were able to predict the likelihood of a child correctly

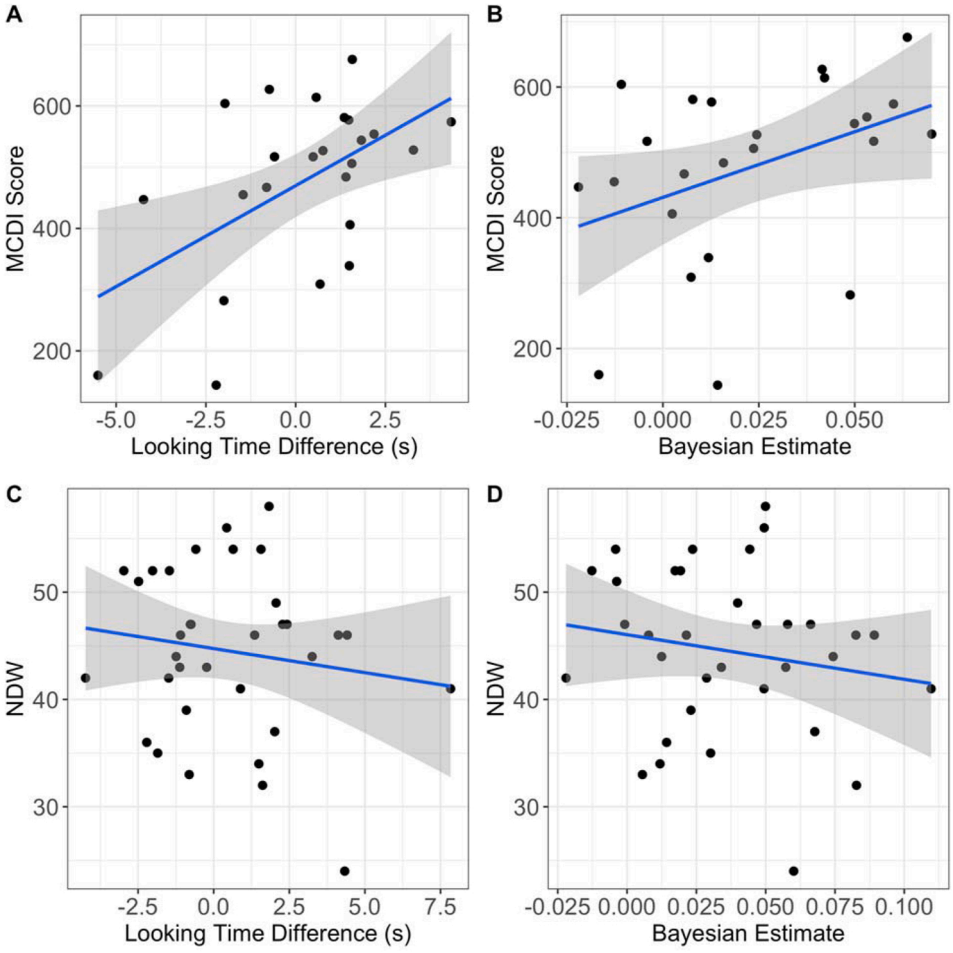

Fig. 2. Individual-level analysis of lexical measures: MCDI scores (panel A & B), and NDW (C & D) predicted by looking time difference (A & C) and the Bayesian Estimate (B & D).

producing verb tense morphemes (n = 35; one child was excluded for producing only 3 utterances). For this we used a Bayesian logistic regression model. The dependent variable was the binary outcome of whether each tense morpheme was produced correctly. The looking time difference did not credibly predict the proportion of correct instances of verb morphemes (β = 0.06 [-0.09, 0.21], p-direction: 77.3 %; Fig. 3A). However, the Bayesian Estimate credibly predicted the likelihood of correctly producing a tense morpheme (β = 0.19 [0.04, 0.34]; Fig. 3B).

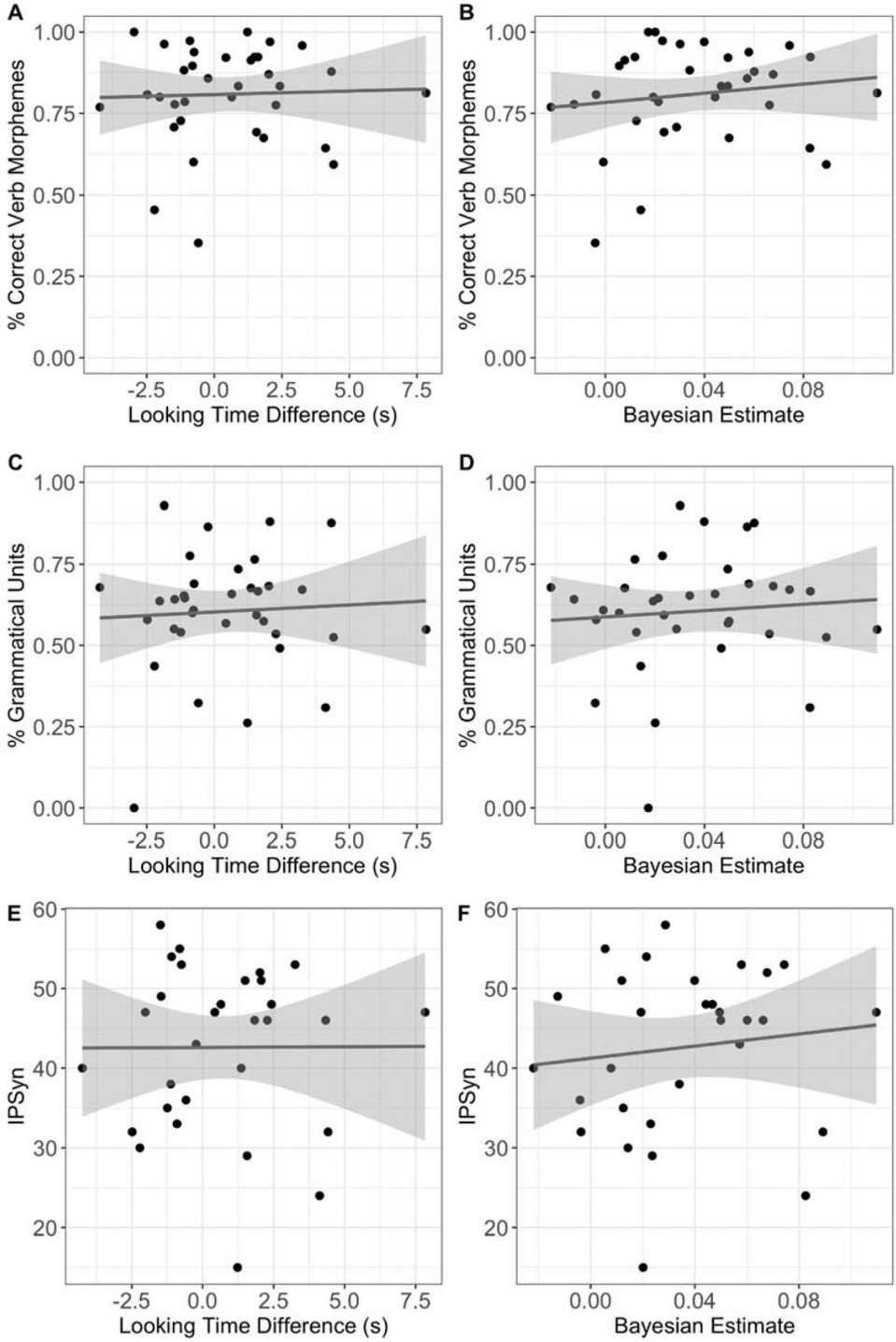

Fig. 3. Individual-level analysis of grammatical measures: number of correct verb morphemes (panel A & B), number of grammatical utterances (C & D) and IPSyn (E & F) predicted by looking time difference (A, C & E) and the Bayesian Estimate (B, D & F).

3.4. Grammaticality of utterances produced at 36 months

Next, we examined whether the two segmentation measures could predict the likelihood of a child producing a grammatical utterance (n = 35; one child was excluded for producing only 3 utterances). In the logistic regression models, the dependent variable was the binary outcome of whether each utterance was grammatical or not. The independent variable was either scaled looking time difference or the scaled Bayesian Estimate. Similar to the results reported for the correct use of verb tense morphemes, looking time differences did not credibly predict the likelihood of producing grammatical utterances ($\beta = -0.06$, [−0.15, 0.03], p-direction: 92.1 %; Fig. 3C), although this is likely to have been due to the low power - note that the effect was nearly credible with a p-direction of 92.1 %. The Bayesian Estimate did not credibly predict the likelihood of producing grammatical utterances either ($\beta = 0.03$ [-0.06, 0.12], p-direction: 75.4 %; Fig. 3D).

3.5. Index of productive syntax

Finally, we investigated whether the segmentation measures could predict individual children's IPSyn scores, a proxy for grammatical complexity obtained from the language sample at 36 months (n = 30). The number of children included in the IPSyn analysis was smaller than that for other measures obtained from the language sample analysis (see Section 5.2, 5.3 and 5.4), because fewer children met the criterion for IPSyn calculation, i.e. had 50 IPSyn-eligible utterances (see Yang, MacWhinney, et al., 2022, for details). IPSyn scores are count data with positive integers (i.e. how many of the set of syntactic constructions a language sample contains; see Scarborough, 1990, for details) with overdispersion, thus we used negative binomial regression to analyze it. Neither the looking time difference ($\beta = -0.00$ [-0.11, 0.10], p-direction: 51.3 %), nor the Bayesian Estimate ($\beta = 0.03$ [-0.07, 0.13], p-direction: 71.7 %), credibly predicted the IPSyn score (Figs. 3E and 3F, respectively).

4. Discussion

We investigated the relationship between two early indices of individual performance on morphological segmentation tasks - looking time difference and the Bayesian Estimate, with language outcomes assessed at 30 or 36 months. At 30 months we assessed expressive vocabulary as has been done previously (Cristia & Seidl, 2011; Junge, 2011; Junge et al., 2012; Marimon et al., 2022; Newman et al., 2006; Newman et al., 2016; Singh et al., 2012; Singh, 2019; Von Holzen et al., 2018; Wang et al., 2021; Weber et al., 2005), using parental language questionnaires. At 36 months, we elicited a language sample using a picture description task in a remote session. We did this because measures obtained from language samples, in addition to expressive vocabulary assessed by parental report, are used extensively in clinical diagnoses of children with language impairment. We discuss the predictive relationship between early indices of individual performance on morphological segmentation tasks and later language outcomes.

The purpose of our study was to determine the extent to which (if at all) two different indices of individual variability in morphological segmentation assessed in infancy predict language outcomes in toddlers. Our first index was the widely used raw looking time difference: although it is easy to calculate, there is no linking hypothesis that attributes a linear relationship between differences in looking time and infants' ability to encode and detect stimulus distinctiveness (Arterberry & Bornstein, 2002; Aslin, 2007; Aslin & Fiser, 2005; de Klerk et al., 2019; de Klerk et al., 2021). We replicated previous studies that show that segmentation in infancy indexed by the looking time difference is correlated with the (expressive) vocabulary size at our proximal time point of 30 months (see Cristia et al., 2014 for a meta-analysis). Additionally, we found a near credible effect such that the looking time difference may predict grammaticality of utterances at 36 months.

The second index, the Bayesian Estimate, characterizes task performance (de Klerk et al., 2021) relative to a cohort of peers. Thus, there is a clear linking hypothesis between this index and individual differences in morphological segmentation. However, it is somewhat harder to compute. To facilitate future research with Bayesian Estimates, we have shared all the code as well as the estimates and raw data from the entire cohort of 238 infants on the OSF page. The Bayesian Estimate as well credibly predicted the size of expressive vocabulary evaluated using the MCDI at our proximal time point of 30 months. Additionally, the Bayesian Estimate also consistently predicted the likelihood of correctly producing verb tense morphemes obtained from the language sample obtained at 36 months. Thus, the Bayesian Estimate had better predictive utility - qualitatively, as well as further in time, than the looking time difference.

While our study has focused on modeling the linear relationship between the measures of individual differences collected in infancy and later language outcomes, it is quite possible that the relationship between these variables is quadratic. This is particularly clearer at the upper limits of both the look time differences and the Bayesian Estimate. You can see from the figures that language outcomes of the infants with the largest look time differences or positive Bayesian Estimates were not the highest. This would be consistent with a positive relationship between the early indices and language outcomes up to a critical threshold, after which the relationship either asymptotes or decreases slightly. However, it is unusual in looking time experiments to have looking time differences (and consequently Bayesian Estimates) at this upper limit. As a result, we do not have enough subjects with early indices at extreme values to test this hypothesis reliably, and leave it for the future.

Our goal in this study was to connect individual differences in speech perception abilities measured in infancy with clinically relevant language outcomes typically measured in the third year of life. For this purpose, we chose to elicit language samples from 36-month-olds based on a picture description task – a more structured activity compared to free play, often employed to elicit spontaneous speech. Although children produce more utterances during free play, utterances collected from narrative speech or storytelling are typically longer and more complex than those elicited during free play or conversation (Mirsaleh et al., 2011; Sealey & Gilmore, 2008;

Southwood & Russell, 2004; Stalnaker & Creaghead, 1982; Wagner et al., 2000). Consistent with these findings, the elicited picture description samples from our study were also more complex than sentences produced in free play sessions (Jo & Sundara, accepted). Crucially, complex sentences are essential to evaluate production of tense and agreement marking by toddlers; the absence or reduced use of morphology is a core deficit associated with language delays and specific language impairment (Bedore & Leonard, 1998; Rice et al., 1995).

Because during pilot testing we were unable to reliably collect language samples from 30-month-olds using a picture description via Zoom, we only collected language samples with 36-month-olds. It is possible that younger infants produce more speech in a free play session. It is also possible that the younger infants produce more speech even with a picture description task, if tested in person. Because we collected language samples during the COVID-19 pandemic, we were not able to visit the family's home or invite the family to the lab to collect language samples in person. Given documented quantitative and qualitative differences in language outcomes obtained using different elicitation methods (Mirsaleh et al., 2011; Sealey & Gilmore, 2008; Southwood & Russell, 2004; Stalnaker & Creaghead, 1982; Wagner et al., 2000), we need research orthogonally manipulating elicitation methods and modality of assessment to disentangle these explanations. Such investigations will provide critical information to clinicians, given increasing interest in tele-health practices.

Our choice to evaluate morphological segmentation in infancy was also motivated by our aspiration to seek mechanistic explanations for the small, positive correlations reported in the literature between early speech perception and later language outcomes (Cristia et al., 2014). Our finding that performance on morphological segmentation tasks measured in infancy credibly predicts the correct use of verb tense morphemes in toddlers is potentially promising in this regard, as a tool to identify children at risk for language impairment. We know from previous research that correct use of verb tense morphemes at age 3 can distinguish typically developing children from those with Specific Language Impairment (Bedore & Leonard, 1998; Rice et al., 1995). It is also promising in terms of a potential link to uncover a mechanistic link. We showed that infants who were more successful at recognizing verb stems after being familiarized with suffixed verbs were more likely to correctly use verb tense morphemes at 36 months.

More generally, segmentation tasks measure infants' ability to isolate words from fluent speech, which is a prerequisite in learning word forms; therefore, task success on segmentation should logically relate to successful lexical acquisition. Infants who are not able to separate words from speech stream must rely only on words produced in isolation, which may delay word learning. This is consistent with our findings that both measures of early morphological segmentation predict expressive vocabulary size at 30 months and confirm previous findings suggesting that early perception skills and later language development are related (Cristia & Seidl, 2011; Höhle et al., 2014; Junge, 2011; Junge et al., 2010; Junge et al., 2012; Junge & Cutler, 2014; Kidd et al., 2018; Kooijman et al., 2013; Leonard et al., 1992; Marimon et al., 2022; Molfese et al., 1999; Newman et al., 2006; Newman et al., 2016; Singh et al., 2012; Singh, 2019; Sussman, 2001; Von Holzen et al., 2018; Weber et al., 2005). Interestingly, despite being correlated with expressive vocabulary to some extent, NDW – the number of different words produced by toddlers was not credibly predicted by either index of morphological segmentation. The exact ways in which expressive vocabulary size and NDW – two measures of lexical acquisition are related, is a question for future research.

Overall, our findings show that the Bayesian Estimate has better predictive validity than looking time differences for later language outcomes. The Bayesian Estimate indexing performance on a morphological segmentation task, but not the looking time difference, credibly predicted use of tense morphemes; for other language outcome measures, the two predictors fared similarly. We therefore conclude that the Bayesian Estimate is better for indexing individual differences in segmentation tasks, both conceptually and practically, and likely to be useful to predict clinically relevant language outcomes. With these findings we want to highlight the necessity of including a diversity of language outcome variables and the need for clear linking hypotheses between predictors and outcomes in future longitudinal studies. Both these pieces are a necessary first step towards determining if there are causal links between early speech perception and later language acquisition.

Ethics approval statement

This study was approved by the Institutional Review Board of University of California, Los Angeles (IRB #10-001562). A parent or legal guardian of each participant received a written description of the study online and gave informed consent before the session; each participant also assented to participating.

Funding statement

This study is funded by – NSF BCS-2028034 Sundara & Hayes, *Morphological acquisition in early infancy: Experimental and computational studies*

CRediT authorship contribution statement

JINYOUNG JO: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation. **Megha Sundara:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Canaan Breiss:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis.

Declaration of Competing Interest

None.

Data availability

The data, materials, and analysis scripts that support the findings of this study are openly available at https://osf.io/xg49b/?view_only=2032721ce25c46bab819a7dc89c0f702.

References

Arterberry, M. E., & Bornstein, M. H. (2002). Variability and its sources in infant categorization. *Infant Behavior and Development, 25*(4), 515–528.

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*(1), 48–53.

Aslin, R. N., & Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *Trends in Cognitive Sciences, 9*(3), 92–98.

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant functional analysis. *Journal of Speech, Language, Hearing Research, 41*, 1185–1192.

Bürkner, P. (2017). brms: An R Package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Bürkner P., Gabry J., Kay M., Vehtari A. (2023). *posterior*: Tools for working with posterior distributions. https://mc-stan.org/posterior/.Pathology. 28, 1275–1282.

Charest, M., Skoczylas, M. J., & Schneider, P. (2020). Properties of lexical diversity in the narratives of children with typical language development and developmental language disorder. *American Journal of Speech-Language Pathology, 29*, 1866–1882.

Cristia, A., & Seidl, A. (2011). Sensitivity to prosody at 6 months predicts vocabulary at 24 months. In N. Danis, K. Mesh, & H. Sung (Eds.), *BUCLD 35: Proceedings of the 35th Annual Boston University Conference on Language Development* (pp. 145–156). Somerville, MA: Cascadilla Press.

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.

de Klerk, M., de Bree, E., Vee, D., & Wijnen, F. (2021). Speech discrimination in infants at family risk of dyslexia: Group and individual-based analyses. *Journal of Experimental Child Psychology, 206*, Article 105066.

de Klerk, M., Veen, D., Wijnen, F., & de Bree, E. (2019). A step forward: Bayesian hierarchical modelling as a tool in assessment of individual discrimination performance. *Infant Behavior and Development, 57*, Article 101345.

De Mayo, B., Kellier, D., Braginsky, M., Bergmann, C., Hendriks, C., Rowland, C.F., Frank, M.C., & Marchman, V.A. (2021). Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories. *Language Development Research*.

Dunn, D. M. (2019). *Peabody Picture Vocabulary Test* (5th ed.). Bloomington, MN: NCS Pearson ([Measurement instrument]).

Eisenberg, S. L., & Guo, L.-Y. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools, 44*(1), 20–31.

Eisenberg, S. L., & Guo, L.-Y. (2015). Sample size for measuring grammaticality in preschool children from picture-elicited language samples. *Language, Speech, and Hearing Services in Schools, 46*(2), 81–93.

Eisenberg, Fersko, & Lundgren. (2001). The Use of MLU for Identifying Language Impairment in Preschool Children: A Review. *American Journal of Speech-Language Pathology, 10*, 323–342.

Fenson, L., Dale, P. S., Resnick, J. S., Thal, D. J., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *MacArthur communicative development inventories (CDI)*. San Diego, CA: Singular Publishing.

Finestack, L. H., & Satterlund, K. E. (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology, 27*(4), 1329–1351.

Gabry, J., Češnovar, R., Johnson, A. (2023). *cmdstanr: R Interface to 'CmdStan'*. https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

Gelman, A. (2006). Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics, 48*(3), 432–435.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness, 5*(2), 189–211.

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics, 15*(3), 373–390.

Hahn Arkenberg, R., Christ, S., & Seidl, A. (2021). Touch screen assessment of high-risk infants' word knowledge. *Canadian Journal of Speech-Language Pathology Audiology, 45*(3).

Heilmann, J. J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools, 41*(1), 84–95.

Höhle, B., Pauen, S., Hesse, V., & Weissenborn, J. (2014). Discrimination of rhythmic pattern at 4 months and language performance at 5 years: a longitudinal analysis of data from German-learning children. *Language Learning, 64*(s2), 141–164. https://doi.org/10.1111/lang.12075

Houston, D. M., Horn, D. L., Qi, R., Ting, J. Y., & Gao, S. (2007). Assessing speech discrimination in individual infants. *Infancy, 12*(2), 119–145.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1570–1582.

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24*, 84–91.

Jo, J., & Sundara, M. (accepted). Remote collection of language samples from 3-year-olds. Journal of Child Language.

Junge, C. (2011). The relevance of early word recognition: Insights from the infant brain (PhD dissertation). Radboud University Nijmegen, MPI Series in Psycholinguistics 67, Nijmegen.

Junge, C., & Cutler, A. (2014). Early word recognition and later language skills. *Brain Sciences, 4*(4), 532–559.

Junge, C., Hagoort, P., Kooijman, V., & Cutler, A. (2010). Brain potentials for word segmentation at seven months predict later language development. In K. Franich, K. Infant Speech Perception and Language Acquisition 1343 M. Iserman, & L. L. Keil (Eds.), *Proceedings of the 34th Annual Boston University Conference on Language Development* (pp. 209–220). Somerville, MA: Cascadilla Press.

Junge, C., Kooijman, V., Hagoort, P., & Cutler, A. (2012). Rapid recognition at 10 months as a predictor of language development. *Developmental Science, 15*, 463–473.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29*(1), 1–23.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159–207.

Kay, M. (2022). *tidybayes: Tidy Data and Geoms for Bayesian Models*. doi:10.5281/zenodo.1308151, R package version 3.0.2, http://mjskay.github.io/tidybayes/.

Kemp, K., & Klee, T. (1997). Clinical speech and language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13*, 161–176.

Kidd, E., Junge, C., Spokes, T., Morrison, L., & Cutler, A. (2018). Individual Differences in Infant Speech Segmentation: Achieving the Lexical Shift. *Infancy, 23*(6), 770–794.

Kim, Y. J., & Sundara, M. (2021). 6-month-olds are sensitive to English morphology. *Developmental Science*, Article e13089.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*(2), 28–41.

Kooijman, V., Junge, C., Johnson, E. K., Hagoort, P., & Cutler, A. (2013). Predictive brain signals of linguistic development. *Frontiers in Psychology, 4*.

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (2000). Prevalence and natural history of primary speech and language delay: Findings from a systematic review of the literature. *International Journal of Language Communication Disorders, 35*(2), 165–188.

Leonard, L. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

Leonard, L. B., McGregor, K. K., & Allen, G. D. (1992). Grammatical morphology and speech perception in children with specific language impairment. *Journal of Speech and Hearing Research, 35*, 1076–1085.
Loeb, D.F., Kinsler, K., & Bookbinder, L. (2000, November). Current language sampling practices in preschools. Poster presented at the Annual Convention of the American Speech-Language-Hearing Association, Washington, D.C.
MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed). Mahwah, NJ: Lawrence Erlbaum Associates.
Marimon, M., Höhle, B., & Langus, A. (2022). Pupillary entrainment reveals individual differences in cue weighting in 9-month-old German-learning infants. *Cognition, 224*, Article 105054.
Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24*, 154–161.
Mirsaleh, Y. R., Abdi, K., Rezai, H., & Kashani, P. A. (2011). A comparison between three methods of language sampling: Freeplay, narrative speech and conversation. *Iranian Rehabilitation Journal, 9*(14), 4–9.
Molfese, D. L., Molfese, V. J., & Espy, K. A. (1999). The predictive use of event-related potentials in language development and the treatment of language disorders. *Developmental Neuropsychology, 16*, 373–377.
Newman, R., Bernstein Ratner, N., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology, 42*(4), 643–655.
Newman, R. S., Rowe, M. L., & Bernstein Ratner, N. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language, 43*(5), 1158–1173.
Perry, L. K., Kucker, S. C., Horst, J. S., & Samuelson, L. K. (2023). Late bloomer or language disorder? Differences in toddler vocabulary composition associated with long-term language outcomes. *Developmental Science, 26*(4), Article e13342.
Rescorla, L., & Lee, E. (2000). Language impairment in young children. In In. T. Layton, E. Crais, & L. Watson (Eds.), *Handbook of early language impairment in children: Nature* (pp. 1–55). Albany, NY: Delmar.
Rescorla, L., & Schwartz, E. (1990). Outcome of toddlers with expressive language delay. *Applied Psycholinguistics, 11*, 393–407.
Rice, M. L., Wexler, K., & Cleave, P. L. (1995). Specific language impairment as a period of extended optional infinitiv. *Journal of Speech and Hearing Research, 38*, 850–863.
Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*(1), 1–22.
Sealey, L. R., & Gilmore, S. E. (2008). Effects of sampling context on the finite verb production of children with and without delayed language development. *Journal of Communication Disorders, 41*(3), 223–258.
Singh, L. (2019). Does infant speech perception predict later vocabulary development in bilingual infants? *Journal of Phonetics, 76*, Article 100914.
Singh, L., Reznick, J. S., & Liang, X. (2012). Infant word segmentation and childhood vocabulary development: A longitudinal analysis. *Developmental Science, 15*(4), 482–495.
Southwood, F., & Russell, A. F. (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research, 47*, 366–376.
Stalnaker, L. D., & Creaghead, N. A. (1982). An examination of language samples obtained under three experimental conditions. *Language, Speech, and Hearing Services in Schools, 13*(2), 121–128.
Sundara, M., White, J., Kim, Y. J., & Chong, A. J. (2021). Stem similarity modulates infants' acquisition of phonological alternations. *Cognition, 209*, Article 104573.
Sussman, J. E. (2001). Vowel perception by adults and children with normal language and specific language impairment: Based on steady states or transitions? *Journal of the Acoustical Society of America, 109*, 1173–1180.
Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships*. Minneapolis, MN: University of Minnesota Press. vol. 10.
Von Holzen, K., Nishibayashi, L.-L., & Nazzi, T. (2018). Consonant and vowel processing in word form segmentation: An infant ERP study. *Brain Sciences, 8*(2), 24.
Wagner, C. R., Nettelbladt, U., Sahlen, B., & Nilholm, C. (2000). Conversation versus narration in pre-school children with language impairment. *International Journal of Language and Communication Disorders, 35*, 83–93.
Wang, Y., Seidl, A., & Cristia, A. (2021). Infant speech perception and cognitive skills as predictors of later vocabulary. *Infant Behavior and Development, 62*, Article 101524.
Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech, Language, and Hearing Research, 38*(6), 1349–1355.
Weber, C., Hahne, A., Friedrich, M., & Friederici, A. D. (2005). Reduced stress pattern discrimination in 5-month-olds as a marker of risk for later language impairment: Neurophysiologial evidence. *Cognitive Brain Research, 25*(1), 180–187. https://doi.org/10.1016/j.cogbrainres.2005.05.007
Yang, J. S., MacWhinney, B., & Bernstein Ratner, N. (2022). The index of productive syntax: Psychometric properties and suggested modifications. *American Journal of Speech-Language Pathology, 31*, 239–256.
Yang, J. S., Rosvold, C., & Bernstein Ratner, N. (2022). Measurement of lexical diversity in children's spoken language: Computational and conceptual considerations. *Frontiers in Psychology, 13*, Article 905789.