Reviewed Preprint
v1 • December 29, 2025
Not revised

Reviewed Preprint
v2 • May 20, 2026
Revised by authors

For correspondence:
mcfrank@stanford.edu

Funding: See page 28

Reviewing editor: Clare Press, University College London, United Kingdom

Continuous developmental changes in word recognition support language learning across early childhood

Michael C Frank[1], Virginia A Marchman[1], Claire Augusta Bergey[2], Veronica Boyce[1], Mika Braginsky[1], George Kachergis[1], Jess Mankewitz[3], Stephan Meylan[4], Ben Prystawski[1], Nilam Ram[1,5], Robert Z Sparks[1], Adrian Steffan[6], Alvin Wei Ming Tan[1], Martin Zettersten[7]

[1]Department of Psychology, Stanford University, Stanford, United States • [2]Department of Linguistics, Stanford University, Stanford, United States • [3]Department of Psychology, University of Wisconsin-Madison, Madison, United States • [4]Department of Linguistics, University of California, Berkeley, Berkeley, United States • [5]Department of Communication, Stanford University, Stanford, United States • [6]Department of Psychology, Ludwig Maximilian University, Munich, Germany • [7]Department of Cognitive Science, University of California, San Diego, San Diego, United States

eLife Assessment

This is an **important** contribution that confirms prior evidence that word recognition - a cornerstone of development - improves across early childhood and is related to vocabulary growth. This study is distinguished by its use of a large, multi-study dataset that is uncommon in prior research on cognitive development. It provides **compelling** evidence that speed, accuracy, and consistency of word learning improve with age, and will therefore prove of interest to those studying language, and more broadly, perception and development.

https://doi.org/10.7554/eLife.109636.2.sa0

Abstract

Being a fluent language user involves recognizing words as they unfold in time. How does this skill develop over the course of early childhood? And how does facility in word recognition relate to the growth of vocabulary knowledge? We address these questions using data from Peekbank, an open database of experiments measuring children's eye movements during early word recognition. In an observational study of 26 datasets from over 2,500 children ages 6 months – 6 years, we show that word recognition becomes faster, more accurate, and less variable across development, consistent with a process of skill learning. Factor analysis reveals covariation of word recognition speed and accuracy with children's vocabulary size in cross-sectional analysis. Further, across a range of longitudinal models, speed, accuracy, and vocabulary were coupled. Children with overall faster word recognition tended to show faster vocabulary growth, though developmental growth in word recognition skill was not specifically associated with growth in vocabulary. Together, these findings support the view that word recognition is a skill that develops gradually across early childhood and that this skill is deeply intertwined with early language learning.

Introduction

Children acquiring a language are learning a body of knowledge – a set of words and the ways they are combined – but they are also learning to deploy this knowledge in the myriad complex, noisy, and fast-moving environments in which language is used. As children enter their second year, language explodes onto the scene; both vocabulary and grammatical abilities grow rapidly

and in tandem (1, 2). This growth in knowledge is also accompanied by changes in language processing efficiency: children become quicker and more accurate in recognizing words and matching them with their referents (3–5).

Yet unlike language production, which is manifest via overt behavior, evidence for word recognition – the linking of a word form to its meaning during language comprehension – is often more subtle. Very young children may not be able to point to the correct referent of a word, but they may still have some representation of word meaning (6). Eye tracking has thus emerged as a key method that allows the measurement of language comprehension with high temporal resolution: both adults and children reliably fixate the referent of a word soon after it is used (3, 7–10). This procedure measures the general construct of word recognition by operationalizing knowledge of a meaning as visual attention to a specific named referent. The relative timecourse of fixation then can provide an index of an individual comprehender's ability or be used to measure the difference between two stimulus conditions.

The version of this method that is used with children goes by many names, including the "intermodal preferential looking" paradigm and the "looking while listening" paradigm (LWL, the name we adopt here) (9, 11, 12). In LWL experiments, children are typically shown two images displayed side by side and asked to find one of them. For example, a ball and a shoe might be shown, and the child might hear "Look at the ball! Can you find it?". Accuracy is then computed as the proportion of time their eyes fixate the correct image within a fixed window after the onset of the noun ("ball" in this case). Reaction time is (typically) computed only on trials in which the child is fixating the distractor image (the shoe) at word onset; in these cases, the average time it takes for the child to shift fixation from the distractor to the target image is used as an index of processing speed. Early work using this method showed that both children's speed and accuracy increase rapidly across the second year (3, 12). Related methods have provided a window into how children process phonological (13), morphological (14), lexical (15), syntactic (16), and semantic (17, 18) information.

Familiar word recognition – as measured by LWL – is hypothesized to play a key role in language learning (19). The idea, in a nutshell, is that the faster and more accurately a child can process incoming words, the more opportunities they have for learning. Consider a child hearing the utterance "Can you put the ball in the crate?" The better the child can recognize the word "ball", the better they can use this evidence to help infer the speaker's intended meaning, allowing possible inferences about the meaning of the less familiar word, "crate" (20).

Real time language processing, including word recognition, relies heavily on predictive processing, in which comprehenders integrate expectations from prior linguistic context with noisy and ephemeral incoming signals (21, 22). The more input a child receives, the better their predictions are likely to be, and hence the more they can learn (19, 23). Indeed, measurements of children's language input at home are consistently associated with their vocabulary size (24, 25). And, in line with this predictive processing framework, one important study found that children's word recognition speed mediated the longitudinal relationship between home language input and vocabulary growth (26). Thus, word recognition is thought to play a key supporting role in ongoing word learning.

Familiar word recognition speed has also been used as an index of individual differences in early childhood (4, 19, 27–29) and beyond (30–32). Over and above measures of vocabulary size, word recognition speed at 18 months predicts children's language and cognitive abilities as measured by standardized tests administered at age 8 (27). Further, faster processing at 18 months is prospectively related to whether "late talkers" catch up to their peers or could benefit from further intervention (28). Critically, most word recognition paradigms use words that children at the target age are reported to understand and produce. They are thus not indices of vocabulary size but rather measures of how quickly and accurately the child can recognize a familiar spoken word and use it to guide their visual attention to a referent. However, it is unknown the extent to which specific responses reflect an individual child's general speed of language processing versus their familiarity of specific words.

Given the logistical hurdles involved in conducting eye-tracking experiments with young children, individual experiments typically recruit relatively small samples in a restricted range of ages. These samples provide neither the breadth of ages nor the number of participants needed to estimate how word recognition changes developmentally and how it connects with other aspects of early language development (see (30, 32) for examples of these analyses in school-aged children). To overcome these limitations, we created Peekbank, an open database of LWL data from young children, stored in a harmonized format (33). This dataset unifies and carefully curates a large amount of eye-tracking data from studies with infants and toddlers, representing cumulatively over 30 million individual measurements of children's eye movements across trials and time-points (dataset version: 2026.1). The Peekbank dataset allows us to gain an unprecedented view of the development of word recognition across a large sample of children.

We investigate two specific issues here. First, one influential theory posits that language learning is a process of skill learning, in which the child is learning the skill of fluent conversation with other language users (34, 35). In this theory, the major information processing challenge of language learning is that incoming language is ephemeral and must be processed quickly before it is lost (the "now-or-never bottleneck"). On this kind of account, we should expect to see the signatures of expertise and skill learning in word recognition, which is one of the primary skills involved in processing incoming language in real time. Accuracy should change linearly with the logarithm of age, reflecting gradual asymptotic convergence to mature levels of accuracy. In addition, we might observe what is known as the "power law of practice," the regularity found in many cases of skill learning that the logarithm of reaction time decreases with the logarithm of experience across participants (36–38, cf. 39, 40). Indeed, this pattern is predicted by an influential associative process model of early word learning (41). In our case, we expect that chronological age is a proxy for experience and so the logarithm of reaction time should decrease linearly with the logarithm of age. Finally, trial-to-trial variability in both speed and accuracy should decrease with increasing expertise, as is found in studies of motor expertise (42).

Second, previous findings have provided limited and sometimes conflicting evidence on the concurrent and predictive relations between word recognition and language learning. Initial reports showed strong prospective relationships between both speed and accuracy and later vocabulary growth (19), with replications in infants born preterm (43) and late talkers (28). Subsequent studies have primarily focused on speed of processing and found more mixed results, with reaction time measures found to be only inconsistently related to later vocabulary outcomes (4, 29, 44). A larger dataset should allow us to make a more definitive test of the presence of these relationships. Further, by examining the relationship between speed, accuracy, and vocabulary, it should be possible to assess the extent to which processing speed specifically plays a role in vocabulary growth.

Across both of these issues, the contribution of our work here lies in the detailed quantitative description of development. Nearly every theory of language learning assumes *some* role for continuous developmental change in word recognition, but these assumptions have not previously been anchored to specific measurements. Hence neither the functional form of the assumed changes nor their concurrent and predictive relationships to vocabulary have been quantified. We leverage the Peekbank dataset to accomplish these goals.

Results

We retrieved data from Peekbank, focusing on data from monolingual English-speaking children ages 6 months – 6 years and on simple word recognition trials in which children were shown two pictures of concrete objects and heard a label for an object (typically embedded in a simple carrier phrase such as "Look at the …"). While other experimental manipulations and languages are included in the database, we narrowed our sample to English-speaking children because they are well-represented across our age range and excluded manipulations which aimed to capture phenomena other than simple concrete noun reference (e.g., adjective comprehension or novel

word learning). These criteria yielded 26 datasets, including 2555 children and 4124 administrations of the LWL procedure (some datasets were longitudinal or involved multiple closely-spaced testing sessions).

Table 1 shows the characteristics of individual datasets (see also S1 Dataset Description in the Supplementary Information). The size of the combined dataset, the unified data processing pipeline, and the fact that individual studies used very similar implementations of the LWL experimental paradigm all allowed us to make a more detailed study of the development of word recognition than has previously been possible. While our analyses are exploratory in nature, they are guided by the two hypotheses outlined above: the presence of 1) signatures of skill learning in word recognition, and 2) linkages between word recognition and vocabulary.

Table 1. Characteristics of included datasets from Peekbank.

"Admins" denotes separate experimental sessions. "CDIs" refers to whether the dataset contains parent report vocabulary data from the MacArthur-Bates Communicative Development Inventory.

	Dataset Name	Pct Trials	N subjects	N admins	Mean Age	Min Age	Max Age	Avg Trials	Avg RT Trials	CDIs	longitudinal
1	Adams et al. (2018)	24.1%	69	711	23.58	13.00	38.00	18.65	7.92	x	x
2	Fernald & Marchman (2012)	20.1%	122	679	23.91	17.00	32.00	16.23	6.91	x	x
3	Weaver et al. (2024)	8.0%	141	247	15.74	13.50	23.60	18.21	6.78		x
4	Fernald et al. (2013)	7.4%	80	178	20.04	17.00	26.00	23.17	9.16	x	x
5	Fernald et al. (2006)	6.4%	63	229	19.68	15.00	25.00	15.28	6.06	x	x
6	Bergelson & Swingley (2012)	2.9%	84	84	11.76	5.98	20.83	18.96	6.74	x	
7	Yurovsky et al. (2013)	2.8%	385	385	33.77	12.20	59.51	5.89	2.63		
8	Borovsky & Peters (2019)	2.8%	79	79	18.27	17.00	20.00	19.24	0.00	x	
9	Potter & Lew Williams (2024)	2.7%	67	67	23.76	21.00	27.00	21.69	7.92		
10	Yurovsky et al. (2017)	2.6%	315	315	36.40	8.40	60.00	6.27	2.87		
11	Yurovsky & Frank (2017)	2.6%	282	282	25.64	12.59	58.65	5.91	2.79		
12	Weaver & Saffran (2026)	2.4%	64	64	18.82	18.10	20.10	21.82	8.19	x	
13	Yoon et al. (2015)	2.0%	194	194	42.30	13.20	60.00	6.72	2.98		
14	Mahr et al. (2015)	2.0%	29	29	20.83	18.10	23.80	37.00	13.62	x	
15	Garrison et al. (2020)	1.7%	35	35	14.46	12.00	18.00	27.76	9.41	x	
16	Ronfard et al. (2022)	1.3%	40	40	19.95	18.00	24.00	18.56	7.62	x	
17	Bacon & Saffran (2022)	1.3%	38	38	22.87	22.00	24.00	18.08	8.00	x	
18	Perry & Saffan (2017)	1.2%	42	42	20.45	19.00	22.00	15.45	5.43	x	
19	Pomper & Saffran (2017)	1.1%	76	76	16.70	14.00	19.00	8.71	3.38		
20	Swingley & Aslin (2002)	1.1%	50	50	15.09	14.13	16.00	11.70	3.79	x	
21	Frank et al. (2016)	0.8%	105	105	33.89	12.13	59.84	6.15	2.69		
22	Pomper & Saffran (2016)	0.8%	60	60	44.27	41.00	47.00	7.62	3.30		
23	Moore & Bergelson (2022)	0.7%	29	29	18.11	16.12	20.03	12.97	4.89	x	
24	Pomper & Saffran (2015)	0.6%	25	25	40.04	38.10	42.00	13.96	5.17		
25	Pomper & Saffran (2019)	0.4%	44	44	40.11	38.00	43.00	5.32	2.34		
26	Pomper & Saffran (2018)	0.4%	37	37	39.46	37.80	43.00	5.47	2.79		
	Total	100%	2555	4124	25.38	5.98	60.00	14.88	5.51	14	5

Speed and accuracy of word recognition increase

We began by examining developmental changes in children's word recognition. Figure 1 depicts the average timecourse of target looking at different ages across all datasets (not controlling for any variation in items and procedures across age groups). Intuitively, these timecourses show gradual increases in accuracy (more target looking; computed as the ratio of target to target plus distractor looking) and speed (faster looking to the target after hearing a label) as age increases. To characterize age gradients in speed and accuracy across children, we computed both RTs (reaction times) and accuracies (proportion looking at the target image) following standard practices in the literature (9). Reaction times were computed only on trials for which the child was fixating the distractor at the point of disambiguation (label onset), and were defined as the time from label onset to the first fixation on the target image (see S2 Reaction Times, including further details on how reaction times were computed in S2.1 and discussion of issues surrounding distinguishing "correct" vs. "incorrect" trials when computing looking-based reaction times in S2.2).

Because there is no consensus about the length of time windows for the computation of accuracy, we considered both a shorter window (from 200 – 2000 ms after noun onset) and a longer window (from 200 – 4000 ms). For each window, we averaged all fixations within the window to compute a continuous proportion of target looking between 0 (no fixation on the target during the window) and 1 (total fixation on the target during the window) on every trial. In this initial analysis, we treat observations of RT and target looking as direct measures of the constructs speed and accuracy (see S4 Test-Retest Reliability); in subsequent analyses we estimate latent variables representing these constructs.

Our first question was about the functional form of the relationships between age, speed, and accuracy (see S5 Pairwise Correlations of Main Measures for raw pairwise correlations between variables). We began by fitting linear mixed-effects models predicting speed and accuracy on each trial across the full dataset with random slopes of child age nested within study (modeling item and procedural variation across studies) and random intercepts by participant (see S8 Mixed-effects model specifications for further details on these specifications). We compared models that included both long and short accuracy windows, as well as logarithmic and linear effects of age, and logarithmic and linear transformations of RT (see S3 Checks on Data Distributional Assumptions for further analyses and discussion of these modeling choices). The best fitting model of accuracy predicted long window accuracy as a function of the logarithm of age; the best fitting model of speed predicted log RT as a function of log age as well (see S6 Functional Form Model Comparison and S7 Power Law Fits). Because long window accuracies were more correlated with other variables and showed clearer age gradients, we focus on these in our analyses.

Figure 2 shows these age gradients. Log RT decreased significantly with age, reflecting increasing speed ($\hat{\beta}$ = −0.13, 95% CI [−0. 16, −0.11], t(18.93) = −12.23, p < .001) and accuracy also increased significantly with age ($\hat{\beta}$ = 0.07, 95% CI [0.06, 0.08], t(20.17) = 13.05, p <.001). In sum, we see continuing improvements in word recognition across the full age range in our dataset that appear roughly linear in the logarithm of age. These logarithmic relationships follow theoretical expectations that both speed and accuracy should gradually asymptote to mature levels of performance, as seen in skill learning more generally (36, 38).

Variability of word recognition decreases

One further hallmark of increasing skill is a decrease in task-relevant variability (42). Both within and across datasets, within-individual variation in speed and accuracy decreased across the developmental range we examined (Figure 3). We fit mixed-effects models predicting the standard deviation of both speed and accuracy for each testing session for each participant, including random slopes of log age nested within dataset and random intercepts for each participant. For both speed and accuracy, within-individual variability decreased with age (speed:

Figure 1. **Timecourse of word recognition at different ages.**

The x-axis shows time (in ms) from the onset of the target label (vertical solid line). Colored lines show the average increase in proportion target looking post label onset at each age bin (in months). Age bins are larger for older children due to decreased data density. The dashed horizontal line represents chance looking. Error bands represent standard errors of the mean. Grey backgrounds highlight the short and long time windows used in subsequent analyses. The data within the figure is filtered such that at (a) participants are required to contribute at least 5 observations and (b) there must be at least 50 participants contributing to each time bin within an age group.

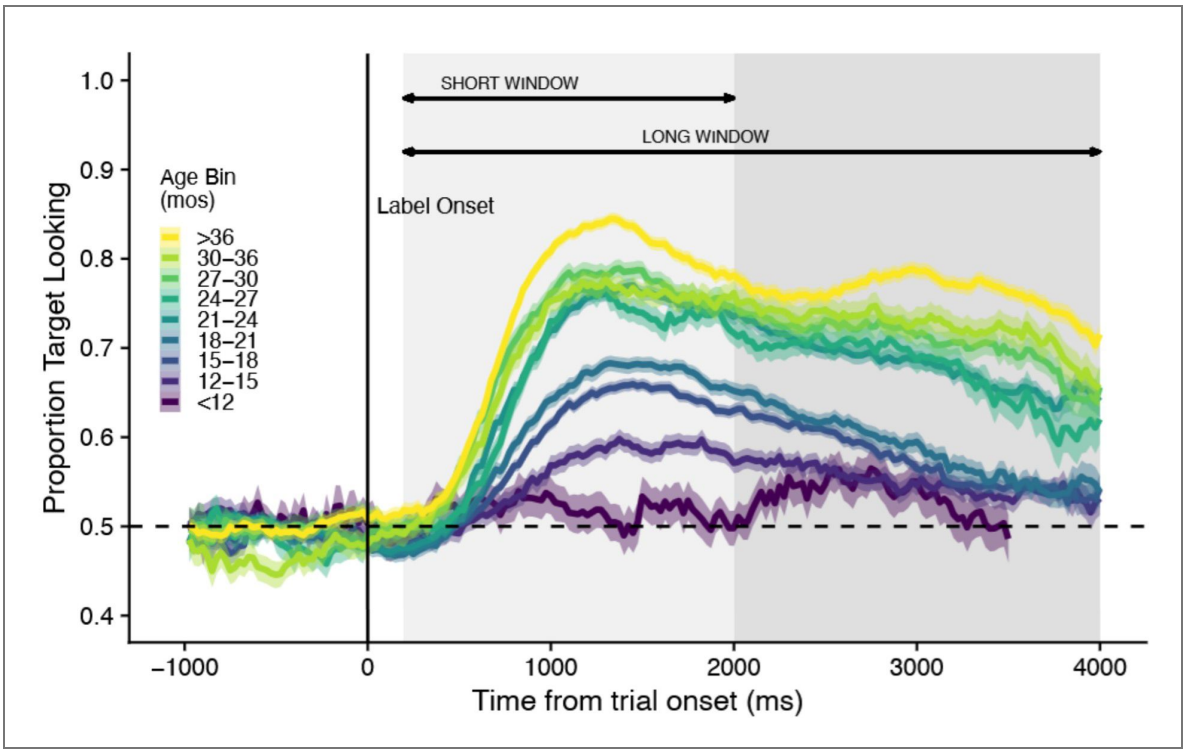
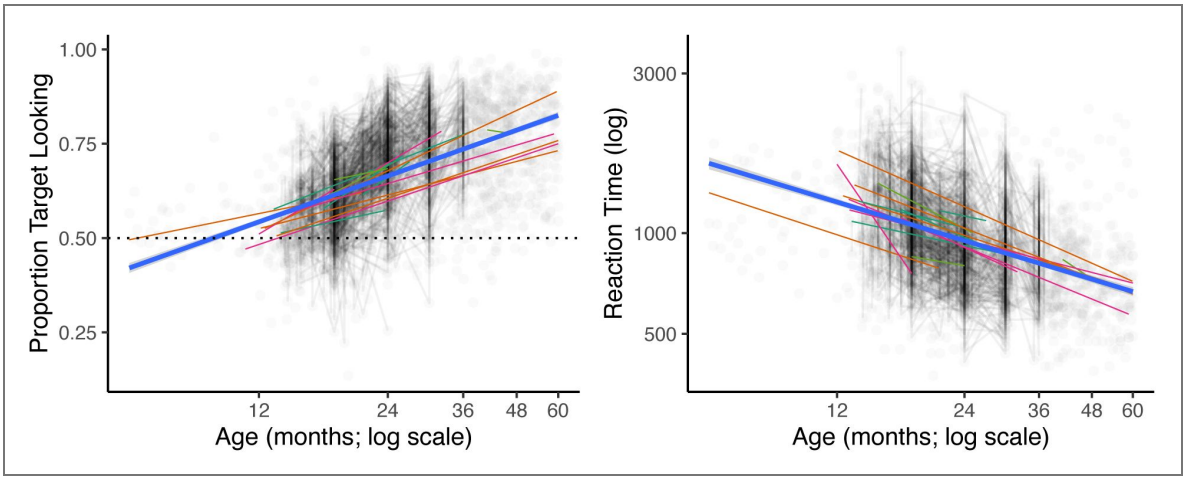


Figure 2. **Participant-level target looking and reaction time (log), plotted by age (log).**

Longitudinal datapoints are connected by lines. The solid blue line shows a linear fit and associated confidence interval. Thin colored lines show linear fits for those datasets spanning six or more months of age. The dashed line for accuracy shows chance-level looking (.5)



β = −0.05, 95% CI [−0.06, −0.03], t(16.33) = −7. 19, p < .001; accuracy: $\hat{\beta}$ = −0.04, 95% CI [−0.04, −0.03], t(12.29) = −10.45, p <.001). Thus, as well as being faster and more accurate, older children were more consistent in their real-time word recognition than younger children.

Speed and accuracy relate to vocabulary size

We were next interested in whether the various aspects of word recognition – including speed, accuracy, and the variability of each of these – were related to other aspects of early language ability. In our prior analyses, chronological age acts as a proxy for greater language experience and larger vocabulary as well as a host of other correlated developmental changes in cognition. Now we explicitly explore relations to vocabulary growth and the triadic relationship between age, word recognition, and vocabulary.

Of the studies in our database, 14 gathered parent reports about children's early vocabulary using the MacArthur-Bates Communicative Development Inventory (CDI), a popular survey instrument that provides a reliable and valid estimate of children's early vocabulary (2, 45). Different forms of the CDI can be used to measure either receptive and expressive vocabulary (for children up to 18 months) or expressive vocabulary only (for children 16 – 30 months).

We fit a series of factor analytic models to explore the dimensionality of the parent report and child LWL data. Our goal in these analyses was to understand the underlying relatedness of the various measures of word recognition and vocabulary, and in particular to assess the evidence for 1) whether the speed, accuracy, and variability measures described above all index the same underlying language processing construct and 2) the nature of the relation between this construct (or set of constructs) and early vocabulary. We initially add age as an additional variable to our models to explore whether this factor structure relates to age; later we treat age as a predictor of latent factors. We begin developing models using all data, treating each observation as independent even if it comes from a longitudinal study; this assumption is equivalent to asserting an invariant factor structure across development (for a test of this assumption, see S10 Factor Analysis on First Administrations). In subsequent models, we relax this assumption and explore longitudinal growth.

Initial exploratory factor analysis using parallel analysis to select the number of factors suggested that three factors explained substantial variance in the data (see S9 Factor Analysis). To better accommodate missing data under the assumption of data missing at random (e.g., missingness due to the age sampling schemes of the various datasets), we used confirmatory factor analysis with full information maximum likelihood to find the best set of loadings. The best fitting model was a three-factor model with factors for speed (RT and RT variability), accuracy (proportion looking to target on each trial and associated variability of this measure), and vocabulary (comprehension and production from the CDI). Fit statistics for this model were generally good (Confirmatory fit index: 0.98, RMSE: 0.06); see S11 Alternative Factor Structures).

Figure 4 shows a regression model fit to this confirmatory factor analysis, with log age predicting each latent variable. This regression model allows interpretation of the covariances between latent factors as partial correlations (controlling for age). The non-age related variance of all three latent factors was significantly related to that of the other factors. Speed and accuracy showed strong negative covariance (β = -0.89, SE = 0.03, p < .001), as expected since they are derived from the same data. Importantly, there was also weaker but significant covariation between RT and vocabulary (β = -0.35, SE = 0.04, p < .001) and accuracy and vocabulary (β = 0.45, SE = 0.03, p < .001). This model supports the idea that variation in speed and accuracy of word recognition is related to individual differences in parent-reported vocabulary beyond the effects of age. Further, the broader set of analyses support a factor structure in which speed and accuracy (and their associated variabilities) are related but distinct aspects of word recognition, rather than being measures of one single construct. These analyses treat all data as between person, however, rather than modeling change in these factors within individuals.

Figure 3. **Participant-level variability in target looking and reaction time (log RT), plotted by age (log).**

Plotting conventions are as in Figure 1.

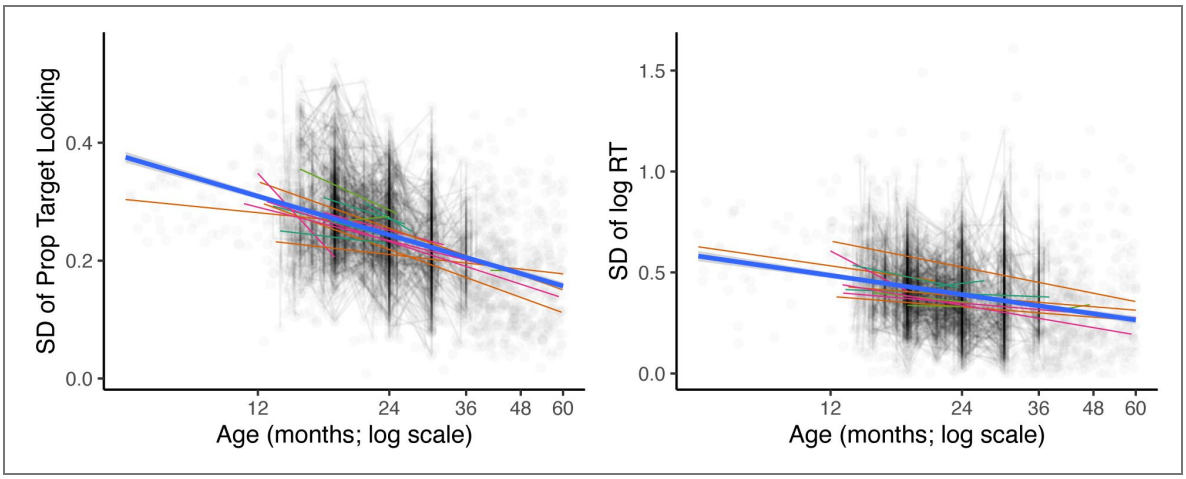
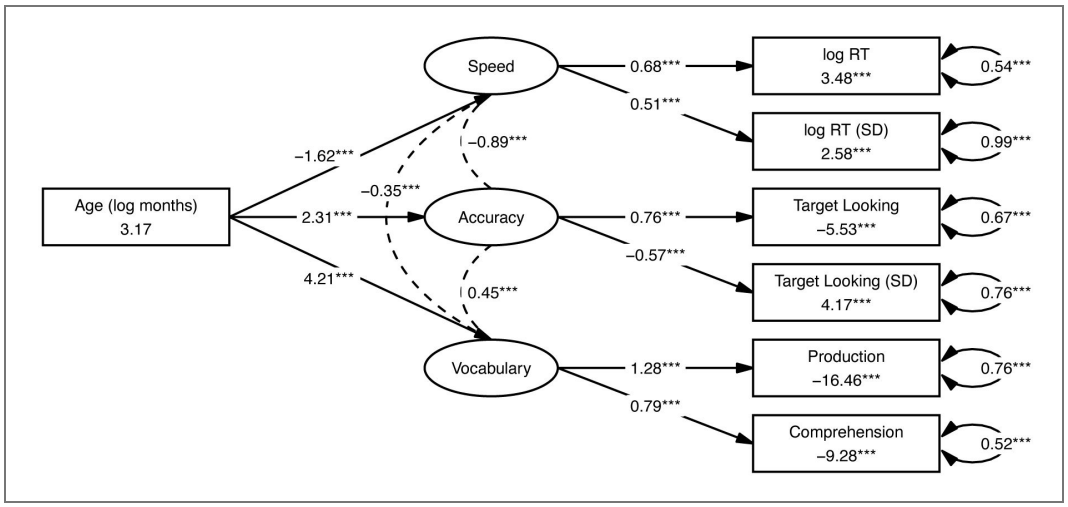


Figure 4. **Structural equation model showing the three-factor factor analysis with a regression of each latent variable on the logarithm of age.**

Observed variables are notated as squares and latent variables are notated as circles. Factor loadings and regression coefficients are shown with straight, solid lines; covariances are shown with dashed lines; residual variances are shown as solid circular connections. Stars show conventional levels of statistical significance, e.g., * indicates p < .05, ** indicates p < .01, and *** indicates p < .001. Covariances reflect age-residualized correlations between variables.



Speed of processing relates to vocabulary growth

We next investigated within-person relationships between LWL and vocabulary. In particular, we investigate two different (but not mutually exclusive) hypotheses about how word recognition skill could support word learning. First, early word recognition skill could lay a foundation for later vocabulary growth — we test this question first using a series of longitudinal growth models testing whether individual variability in processing speed predicts later increases in productive vocabulary. A second, stronger version of this hypothesis is what we call a "virtuous cycle" model of the relationship between processing speed and vocabulary growth, in which not only baseline word recognition skill, but also children's improvements in this skill are related to faster growth in vocabulary; we test this hypothesis using longitudinal structural equation models.

To investigate the first hypothesis, we began by fitting longitudinal growth models to the full dataset (though note that the same conclusions hold when restricting the data to only those children with multiple LWL sessions). We first reproduced the analysis reported in (28), in which between-person differences in longitudinal growth in productive vocabulary were predicted based on between-person differences in speed during the initial session of the study. We fit a mixed-effects model predicting growth in vocabulary as a quadratic function of age, RT at study initiation (t_0), and their interaction (as well as random effects of age nested within participant and also age nested within dataset). This model revealed a significant effect of t_0 RT ($\hat{\beta}$ = −0.14, 95% CI [−0. 19, −0.08], t(530.16) = −4.85, p < .001) and an interaction between t_0 RT and the quadratic age predictor ($\hat{\beta}$ = 2.00, 95% CI [1.04, 2.96], t(545.67) = 4.07, p <.001). This analysis suggests that children with faster initial RTs show both larger vocabularies and faster vocabulary growth over time.

We confirmed this analysis using a non-linear growth model with a logistic shape, which provides a better fit to vocabulary size within a fixed-length form than the quadratic model (see S12 Non-Linear Growth Model) (2). Figure 5 shows predictions from this model, confirming the differentiation of growth curves for children with higher and lower initial reaction time.

On the other hand, it is possible that differences in predicted growth trajectories are due to coupling between vocabulary size and language processing across the entire developmental period, rather than a predictive relationship specifically between t_0 RT and vocabulary growth (i.e., the "virtuous cycle" model). To test this relationship, we used longitudinal structural equation models. We separated the longitudinal speed, accuracy, and vocabulary data into two-month bins spanning up to 10 months from the initial measurement (i.e., t_0, ..., t_4) and fit individual growth across each of these variables. We used full-information maximum likelihood to handle the substantial missing data caused by the different longitudinal sampling schemes of studies in our dataset (see S13 SEM Longitudinal Missingness). The fitted longitudinal model is shown in Figure 6. Overall fit statistics were generally acceptable (Confirmatory fit index: 0.89, RMSE: 0.03, RMSE p-value: > .999).

Our key question of interest concerned coupling among the (latent) intercepts and slopes of these growth models. Consistent with our earlier analysis showing that faster processing is related to vocabulary growth, we saw significant between-person coupling between processing speed intercepts and vocabulary growth slopes (β = -0.18, SE = 0.06, p = .001) as well as a variety of other between-person couplings. On the other hand, there was not significant coupling between *growth* in speed and *growth* in vocabulary (β = 0.00, SE = 0.02, p = .872). This null effect could be interpreted as being consistent with these abilities growing independently, but there are other possibilities. First, the longitudinal data we had might not have allowed sufficiently precise estimates of growth slopes, or second, since vocabulary growth is non-linear, the linear model we used here might not have captured coupling among nonlinear aspects of developmental change.

In sum, these findings provide evidence consistent with the claim that differences in processing speed are related to differences in the rate of age-related change in vocabulary (19, 26). Children with greater skill in word recognition learn words faster. However, we did not find evidence for the stronger version of this claim: in neither the non-linear growth model nor the linear SEM did

Figure 5. Growth curves from a logistic growth model showing predicted productive vocabulary growth for children with initial reaction times one SD faster than the mean (blue), at the mean (red), and one SD slower than the mean (green).

Individual longitudinal trajectories are shown in light gray. Solid lines show global model estimates and colored regions indicate 95% credible intervals.

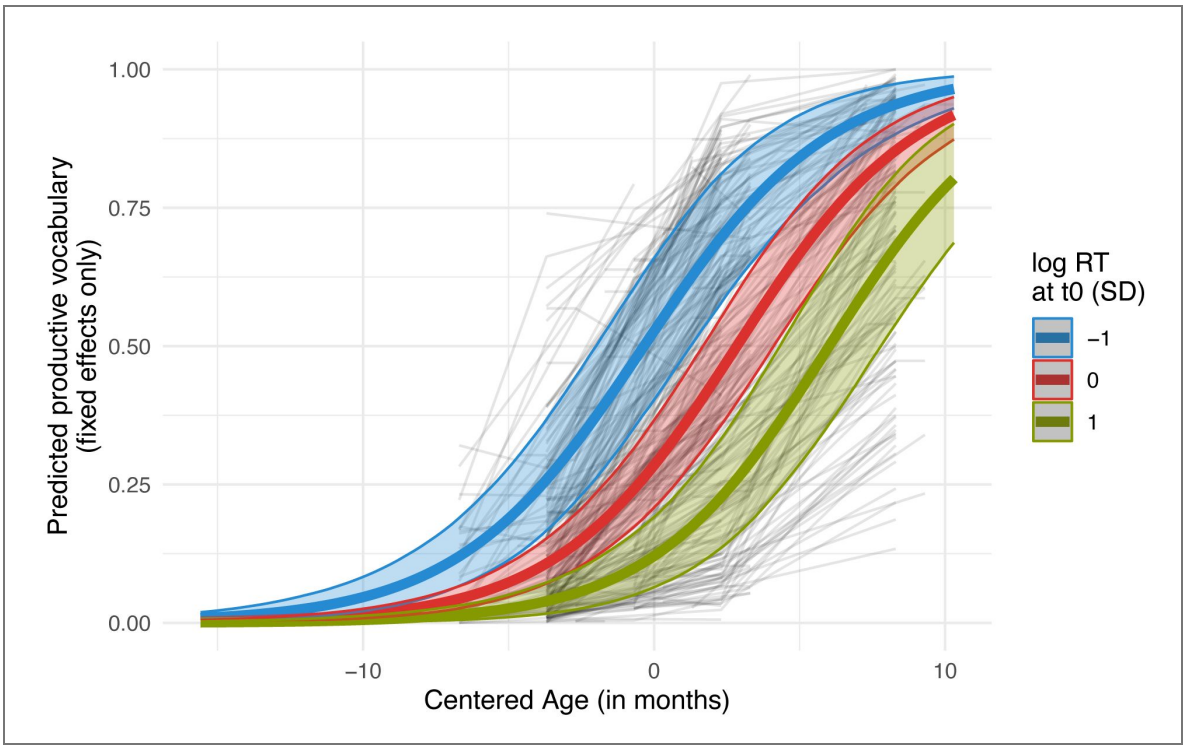
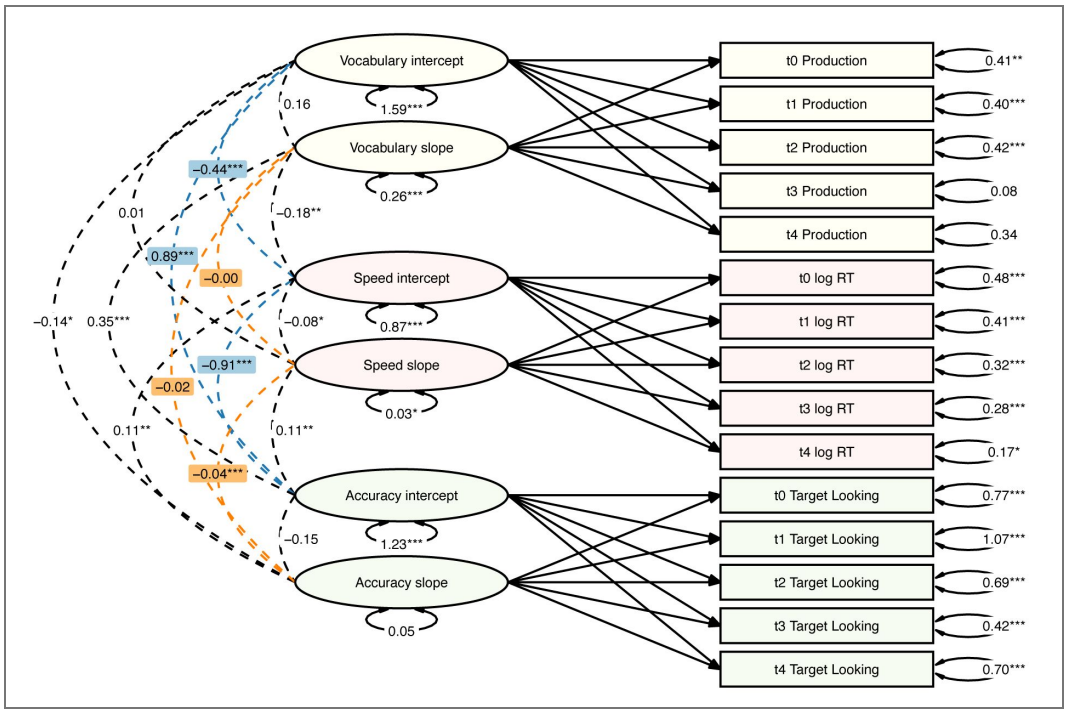


Figure 6. Structural equation model showing longitudinal couplings between growth parameters.



we find evidence that increases in speed were related to increases in vocabulary size. Thus, our findings do not support a "virtuous cycle" model in which increases in recognition specifically lead to increases in vocabulary size.

Discussion

How does word recognition change across early childhood and how does it relate to language learning? We investigated these questions using a new, large-scale dataset of developmental eye-tracking measurements compiled across many prior studies. The age gradients for speed and accuracy indicated that both improve asymptotically. Gradients for recognition speed were consistent with the log-log relationship associated with the "power law of practice," that is, with a gradual convergence to mature levels of processing efficiency. Further, the age gradient suggested that trial-to-trial variability decreases with age, consistent with both the literature on skill learning (42) and other work on developmental changes in variability (46–48). Speed and accuracy were both related to vocabulary size concurrently and processing speed was also related longitudinally to later vocabulary growth.

Together, our findings are consistent with theories that posit that language learning is a process of skill acquisition, in which children become adept at quickly converting ephemeral signals into meaning (34). This skill develops gradually over the course of early childhood and supports word learning. Further, our results point to consistency between skill development in early childhood and the continued refinement of language processing and language knowledge during middle childhood (30, 32).

By aggregating data from many pre-existing studies, we were able to overcome the limitations of prior investigations, which typically had sample sizes at least an order of magnitude smaller than ours. Our approach was to build on the time-consuming and meticulous data collection from previous infant and toddler eye-tracking studies – representing cumulatively many thousands of hours of in-lab data collection and hand-annotation of the resulting videos of child looking behavior – by harmonizing these data into a single, large-scale database. This approach illustrates how building harmonized databases can be especially powerful when composed of high-effort and high-quality datasets that are smaller in scope, maximizing the impact of previous data collection efforts and allowing us to ask broader questions about developmental change (2). In contrast to individual studies, which typically have at best the statistical power to test one or two specific contrasts, our "big data" approach provided the sample sizes necessary to explore the relationships between different variables. Because early language is so variable, these kinds of samples – with thousands, rather than dozens of children – are likely to be required to gain further insight into the psychometrics of early language learning (2, 49, 50).

Our approach is both observational and exploratory. Thus, we cannot untangle the range of different causal models that explain the variation we observed. First, early word recognition skill could lead to faster word learning, but faster children could also be faster due to their larger vocabulary and stronger lexical representations. These two causal directions could also interact reciprocally, leading to a "rich get richer" process in which children with larger vocabularies process faster, and their faster processing helps them increase their vocabulary size more rapidly. Finally, a third shared factor – perhaps general cognitive ability – could underpin both processes. Our cross-sectional data cannot distinguish these hypotheses even in principle (51), and our longitudinal data are likely too sparse to distinguish such complex causal models. Future work must also explore how the functional forms we observed here between individuals reflect processes of within-person change. Although the Peekbank dataset includes a variety of longitudinal data, most reflect a small number of measurements; denser longitudinal data collection is required to better estimate within-person growth models.

The relationships we report are derived from models that account for variation across datasets, suggesting that our qualitative conclusions are robust to cross-laboratory variation. Nevertheless, these findings are still limited in their generalizability by the convenience samples that were used in most of the studies aggregated in Peekbank. These studies typically (but not always) represent

children from well-educated parents living in university-adjacent communities. We would not expect that specific numerical parameters estimated in our aggregate convenience sample would generalize to other samples.

More broadly, our results here suggest the continued importance of the looking-while-listening paradigm as an index of children's language processing abilities. If language learning is, at least in part, a process of skill learning, then measurement of this skill in larger samples provides a critical window into understanding the remarkable process of language learning.

Materials and Methods

Data

We included information from 2555 unique participants across 26 datasets. Dataset information is given in Table 1. Although experiments in Peekbank include a variety of different experimental manipulations, we analyzed only data from standard, simple word recognition trials ("vanilla" trials); these trials were sometimes the main focus of the original studies and sometimes constituted control conditions for experiments with more complex manipulations. Requirements for being considered a standard word recognition trial included that (a) the target word was familiar (also no part-words); (b) the target word was the first point of disambiguation and appeared only once; (c) the target word was embedded in a well-formed, grammatical carrier phrase; (d) there was no informative language presented prenominally (e.g. semantically informative verbs, adjectives); (e) there were no nonsense words presented anywhere during the trial (including the carrier phrase); (f) there was no language-, speaker-, or accent-switching within trial; (g) the auditory stimulus included no intentional background noise or audio filtering; (h) both target and distractor items were familiar objects; (i) no novel visual stimuli (i.e., experimenter-created artificial items or items selected to be entirely unfamiliar) were visible; and (j) the target referent was the focal object in the target image and there were no additional focal objects competing for attention within the target image (e.g., if the target word was "orange" and the image depicted an orange on a plate, this was considered a standard trial; if however the image depicted both an apple and an orange on a plate, this was not considered standard). We focus here on English purely for practical reasons – the Peekbank dataset at present contains limited data from other languages.

We excluded trials entirely if they were missing data on more than 50% of timepoints, and excluded RTs if they were based on fewer than 50% of timepoints in the short analytic window (200 – 2000 ms). We also removed RTs shorter than 367 ms, as these were unlikely to be generated based on the specific linguistic stimulus. We then excluded participants from the analysis if they contributed fewer than four accuracy measurements or fewer than two reaction time measurements. At the participant level, these steps together led to 21.40% missingness for RTs and 8.80% missingness for long window accuracies.

Analytic methods

We used `lme4` to fit linear mixed-effects models, `brms` to fit non-linear growth models, and lavaan to fit structural equation models. Random effects structures for each model are given in text; full model specifications are available in the Supplemental Information (S8, S9, and S12) and in the reproducible code for this paper, available in the linked repository. To aid interpretability, all variables were standardized (z-scored) prior to inclusion in structural equation models.

Supplemental Information

S1. Dataset Description

Figure S1 gives the age distribution of unique participants for each separate dataset at different ages. Note that for some datasets, there are multiple administrations (i.e., experimental test sessions) for each participant. Figure S2 shows the distribution of measurement intervals for

longitudinal studies within the dataset. Table S1 has additional information on how many trials each dataset contributed and what percent of the dataset's trials were included.

Table S1. Characteristics of included datasets from Peekbank, sorted by what percent of the data the data they represent.

Percent trials refers to what percent of the trials used came from that dataset; total is the number of trials used from that dataset; and included is what percent of all trials had data that was included (based on criteria about missingness, distractor to target transition, minimum RT). LW = long-window accuracy, SW = short-window accuracy, and RT = reaction time.

	Dataset Name	Pct Trials (LW)	Pct Trials (RT)	Total Trials (LW)	Total Trials (RT)	Included (LW)	Included (SW)	Included (RT)
1	Adams et al. (2018)	24.1%	27.1%	13146	5470	86.2%	89.3%	35.9%
2	Fernald & Marchman (2012)	20.1%	22.4%	10954	4526	91.6%	94.8%	38.0%
3	Weaver et al. (2024)	8.0%	7.6%	4371	1540	88.6%	91.2%	31.3%
4	Fernald et al. (2013)	7.4%	7.9%	4055	1603	83.1%	85.7%	32.9%
5	Fernald et al. (2006)	6.4%	6.2%	3484	1254	93.1%	93.4%	34.0%
6	Bergelson & Swingley (2012)	2.9%	2.7%	1593	539	76.7%	83.5%	26.1%
7	Yurovsky et al. (2013)	2.8%	2.2%	1544	450	53.2%	55.7%	17.2%
8	Borovsky & Peters (2019)	2.8%	0.0%	1520	0	88.9%	88.9%	0.0%
9	Potter & Lew Williams (2024)	2.7%	2.6%	1453	523	89.3%	87.7%	32.1%
10	Yurovsky et al. (2017)	2.6%	2.4%	1411	493	60.4%	68.2%	21.9%
11	Yurovsky & Frank (2017)	2.6%	1.9%	1401	393	65.3%	66.6%	20.9%
12	Weaver & Saffran (2026)	2.4%	2.4%	1309	475	67.5%	69.2%	24.5%
13	Yoon et al. (2015)	2.0%	1.7%	1089	337	74.4%	77.1%	24.9%
14	Mahr et al. (2015)	2.0%	2.0%	1073	395	82.4%	82.4%	30.3%
15	Garrison et al. (2020)	1.7%	1.6%	944	320	89.8%	89.9%	30.4%
16	Ronfard et al. (2022)	1.3%	1.4%	724	282	98.5%	98.5%	38.5%
17	Bacon & Saffran (2022)	1.3%	1.5%	687	304	90.4%	89.2%	40.0%
18	Perry & Saffan (2017)	1.2%	1.1%	649	228	89.6%	89.6%	31.5%
19	Pomper & Saffran (2017)	1.1%	0.9%	592	179	75.0%	75.8%	24.5%
20	Swingley & Aslin (2002)	1.1%	0.8%	585	159	98.2%	97.8%	28.0%
21	Frank et al. (2016)	0.8%	0.7%	449	140	59.5%	66.2%	20.3%
22	Pomper & Saffran (2016)	0.8%	0.9%	457	175	95.6%	95.6%	37.9%
23	Moore & Bergelson (2022)	0.7%	0.7%	376	132	97.9%	99.2%	34.9%
24	Pomper & Saffran (2015)	0.6%	0.6%	335	124	85.7%	89.0%	31.9%
25	Pomper & Saffran (2019)	0.4%	0.4%	213	75	85.9%	89.3%	30.5%
26	Pomper & Saffran (2018)	0.4%	0.3%	197	67	90.5%	91.9%	34.4%

S2. Reaction Times

S2.1. Reaction Time Computation

Eye-tracking data are stored in Peekbank as a time series of fixations to specific areas of interest (in particular, the target and distractor on each trial). Other fixations can be to areas not in the target or distractor as well to off-screen areas. This time series has a uniform sample rate of 25ms/sample, based on resampling of the data in Peekbank to 40 Hz during preprocessing (Zettersten et al. 2023). Reaction times are computed by filtering trials to only those on which

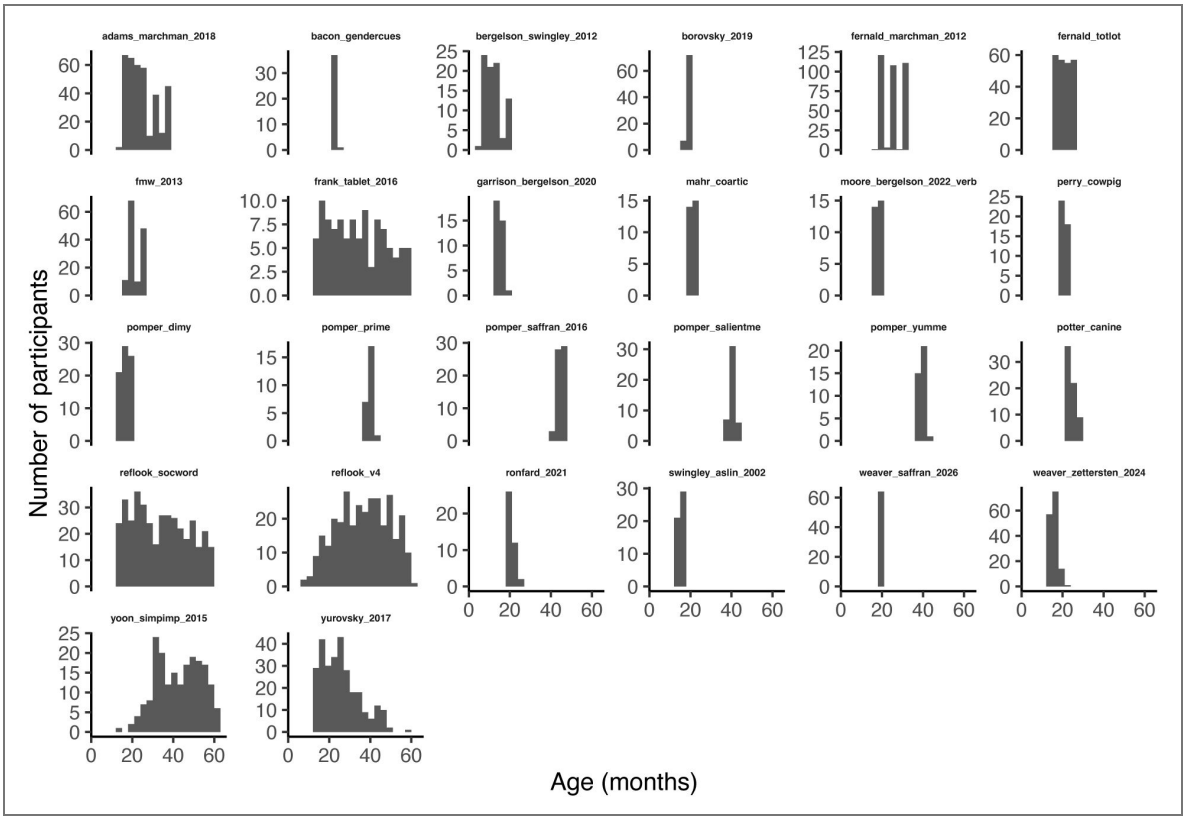


Figure S1. **Age distribution of unique participants for each dataset, using three-month bins.**

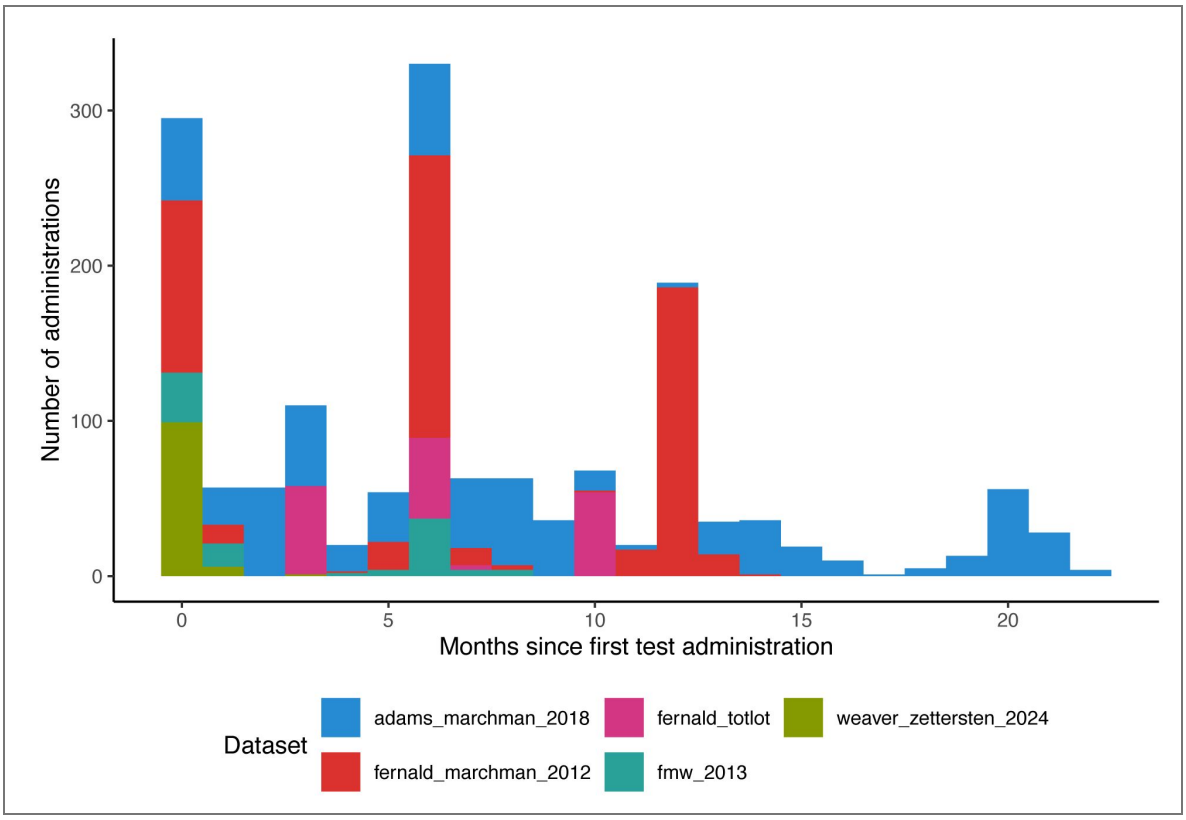


Figure S2. Distribution of retest administrations across datasets with repeated measurements, colored by dataset.

Each count indicates a retest administration (initial administrations are excluded). Administrations listed with a retest interval of 0 indicate retests within a month of the initial administration.

the child is fixating the distractor at the point of disambiguation (t =0) and then finding those trials on which the first non-missing fixation is to the target (hence excluding trials without a shift and trials on which a shift is to an off-screen location). The reaction time is then the total time from t = 0 to the first timestep during which the child fixates the target. Consistent with standard practice in the literature following Fernald et al. (2008), RTs that are shorter than 367 ms are excluded as they are too short to be considered a response to the stimulus.

S2.2. Comparison of Reaction Times for Correct and Incorrect Trials: Re-analysis of Creel (2024)

The Peekbank dataset only includes measurements of infants' looking behavior, with no measure of a final target selection. This contrasts with work in the visual-world paradigm with older children and adults, in which participants make a final explicit choice about which image matches the target label (e.g. Colby & McMurray, 2023). Having this additional response allows a clearer separation of accuracy and reaction times, because researchers can compute reaction times specifically on those trials in which participants responded correctly. This strategy helps avoid a possible mixing of reaction times for incorrect and correct responses, which might be generated by different underlying cognitive processes. A possible concern with the Peekbank datasets — and reaction times in infant looking-while-listening studies more generally — is that it is difficult to separate reaction times for correct vs. incorrect responses in the absence of an independent final choice response.

To address this concern, we investigated data from a recent large-scale word recognition study with toddlers in which eyetracking measures were collected together with a final pointing response (Creel, 2024). This dataset included 914 responses from children (2.5-6.5 years) completing a looking-while listening procedure in which they also were instructed to point to the target image. Using this dataset, we investigated the correlation between reaction times (following the same procedure as in our main analyses, i.e. focusing specifically on distractor-to-target shifts) computed over all trials and reaction times computed only for those trials in which children selected the correct referent. The results are shown in Figure S3. Reaction times (i.e., distractor to target shifts) for correct trials only were highly correlated with reaction times across all trials (r = .85, 95% CI [.82,.87], $t(479) = 34.84$, p <.001). This result suggests that having the ability to filter out incorrect trials has a minimal impact on reaction time computation, even in young children. While there is some uncertainty about how these results may generalize to infants in our younger age ranges (i.e., below 2.5 years of age), who struggle to provide reliable pointing responses, it seems reasonable to assume that our reaction time results would stay largely the same if it were possible to filter out trials on which infants make an incorrect mapping between the target label and the target image using an eyetracking-independent final choice response.

S3. Checks on Data Distributional Assumptions

Here, we check whether the distributional forms that are assumed for the distributions of RT and accuracy are a reasonable empirical fit to the data, and compare against other commonly used distributional forms. We confirm that, across the age range, the choice to use a log-normal distribution for RT and a normal distribution for accuracy is justified.

S3.1. Reaction Time

The literature focuses on the use of the Exponential-Gaussian (ex-Gaussian) distribution as well as Wald, Weibull, gamma, and log normal distributions (see for example Luce, 1986; Ratcliff, 1993; Van Zandt, 2002). All of these are 2- or 3-parameter distributions meaning that there is no necessary relationship between mean and variance.

The problem of fitting RT distributions is complex and a substantial literature exists (e.g., Ratcliff, 1979; Luce, 1986; Van Zandt, 2000; Baayen & Milin, 2010). One of the big challenges in our dataset as well as elsewhere is that distributions are conditional on factors such as participant and task, so it is challenging to draw inferences about the underlying distribution when looking at average data.

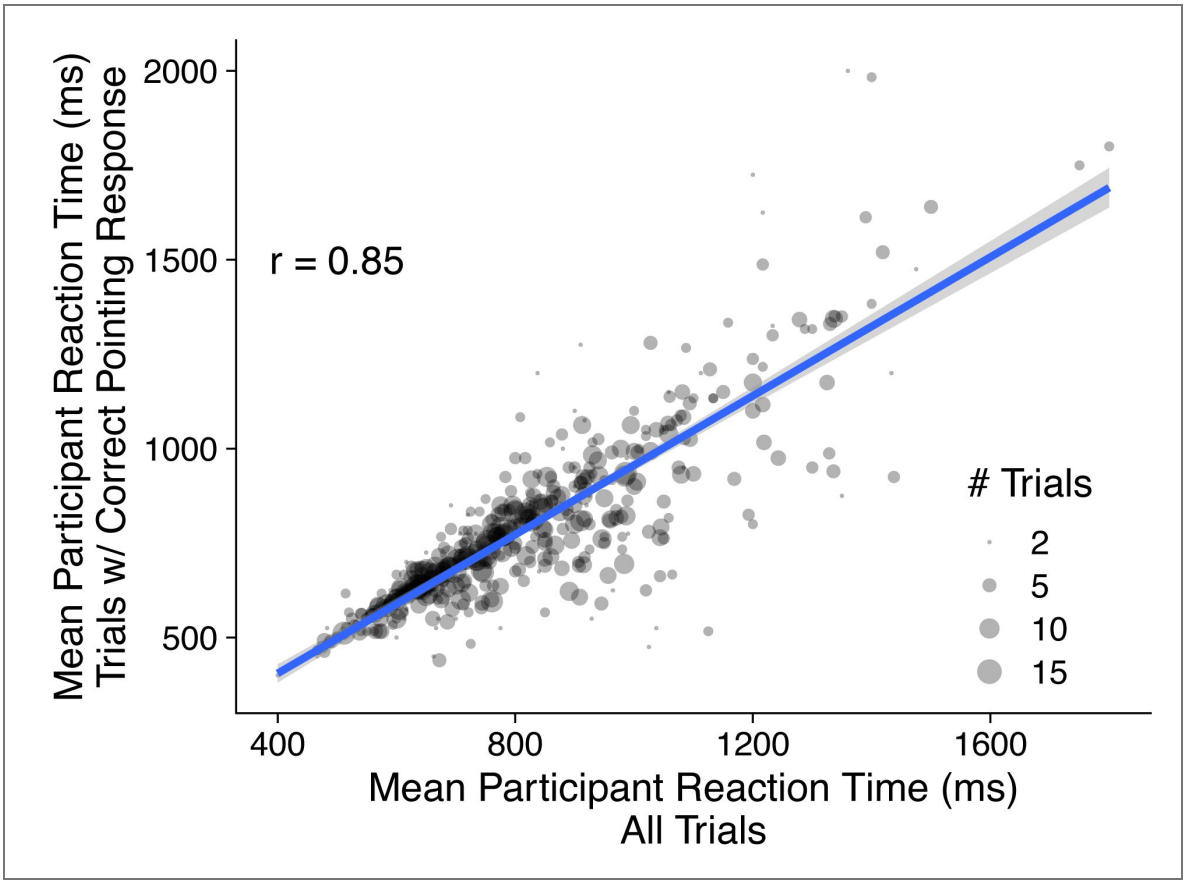


Figure S3. Correlation between reaction times on all trials and reaction times on trials where the child pointed to the correct target.

Data from Creel (2024).

That said, we find that overall the data are best fit by either an ex-Gaussian or a log normal distribution, again consistent with prior literature, giving us confidence in this conclusion. Across the full dataset, the BIC values for ex-Gaussian (46906) and log normal (46953) are quite close to one another, and better than the Wald (53949) and normal (47811) fits. When binned by age (Fig S4), younger children seem better fit for a log normal distribution and older children seem better fit by ex-Gaussian (models with lowest BICs are shown in red since significant differences can be obscured by the large scale). Figure S5 shows the RT data distribution overlaid with the corresponding log-normal distributions. Overall, we think this result generally vitiates our decision to use log-transformed RTs as our primary dependent measure.

S3.2. Accuracy

Individual trial-level accuracies are not binomial because they are an average probability of fixation over a viewing window. They are bounded at 0 and 1, but in general they tend towards the range .5 - .8 in most studies of this population. Figure S6 shows the data binned by age group with fitted gaussian distributions.

These distributions seem well fit by standard gaussians, but they are in principle bounded and so we asked whether this made a difference, using a Beta distribution (a two-parameter continuous distribution bounded at 0 and 1) for fitting. Surprisingly, across all data, the BIC values for the two distributions were very similar (-4710 for normal versus -4678 for Beta), though the normal distribution was slightly favored. Across age groups, there was heterogeneity with some groups better fit by a gaussian and others better fit by a Beta (Figure S7). Again, we feel that this result generally vitiates our approach of modeling accuracies via standard linear mixed-effects models: their distributional form is quite close to normal.

S4. Test-Retest Reliability

We examined test-retest reliability for our primary variables of interest by calculating Pearson correlations between pairs of administrations given no more than three months apart. Test-retest correlations were significant but relatively modest: $\rho_{longwindowacc}$ = 0.462, $\rho_{shotwindowacc}$ = 0.496, ρ_{rt} = 0.407. These reliabilities were biased downwards by three factors, however. First, longitudinal assessments sometimes use variable items between testing sessions, leading to item-related variance in measurement. Second, even three months can lead to substantial change in some children's language abilities, thus correlations are attenuated by true change as well as measurement error. Third, longitudinal data in the dataset come primarily from the youngest children and hence are likely to show overall higher measurement error due to variability in children's behavior and an overall lower number of trials.

S5. Pairwise Correlations of Main Measures

Table S2 shows pairwise correlations between the primary variables of interest in the dataset.

Table S2. Pairwise correlations between primary variables of interest.

	age	log age	rt	log rt	long acc	short acc	prod	comp
age	1.00							
log age	0.98	1.00						
rt	-0.33	-0.35	1.00					
log rt	-0.34	-0.36	0.96	1.00				
long window accuracy	0.44	0.48	-0.48	-0.46	1.00			
short window accuracy	0.38	0.43	-0.62	-0.61	0.82	1.00		
production vocabulary	0.72	0.70	-0.31	-0.33	0.51	0.45	1.00	
comprehension vocabulary	0.42	0.42	-0.25	-0.24	0.24	0.24	0.59	1.00

Figure S4. **Goodness of fit for different distributional models for RT, split by age.**

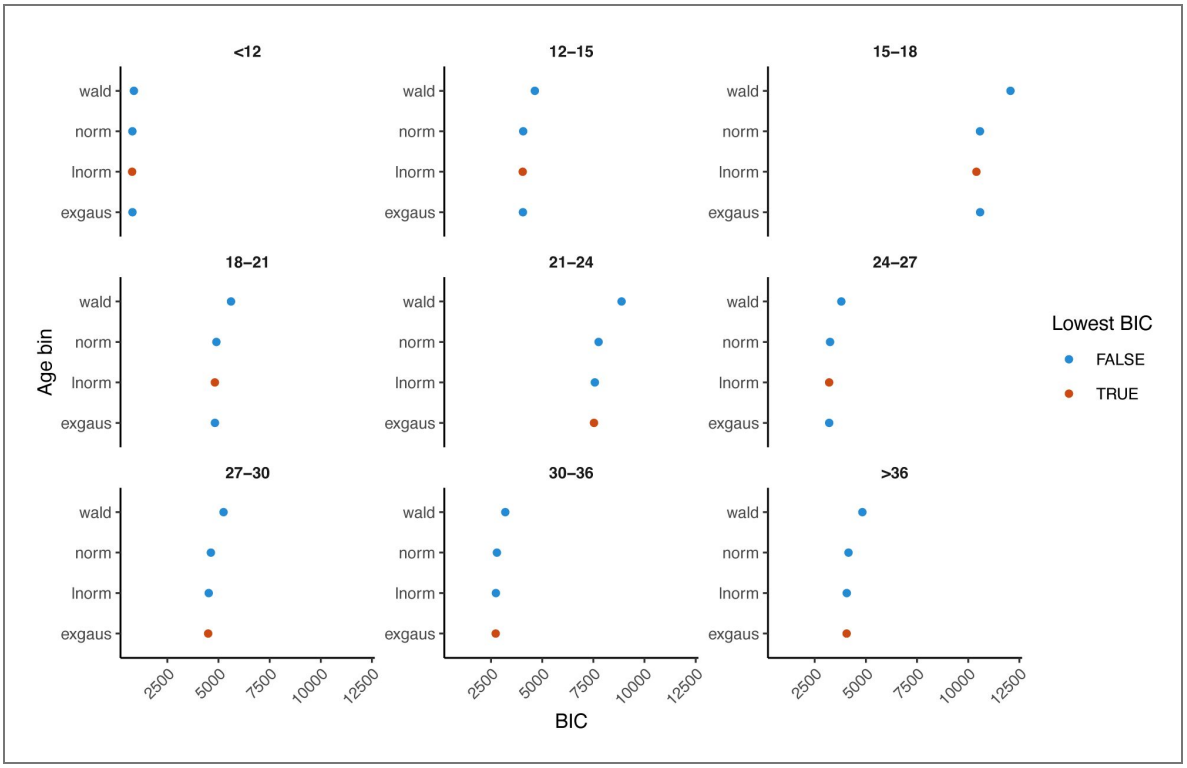


Figure S5. **Distribution of RT overlaid with a log normal distribution, split by age.**

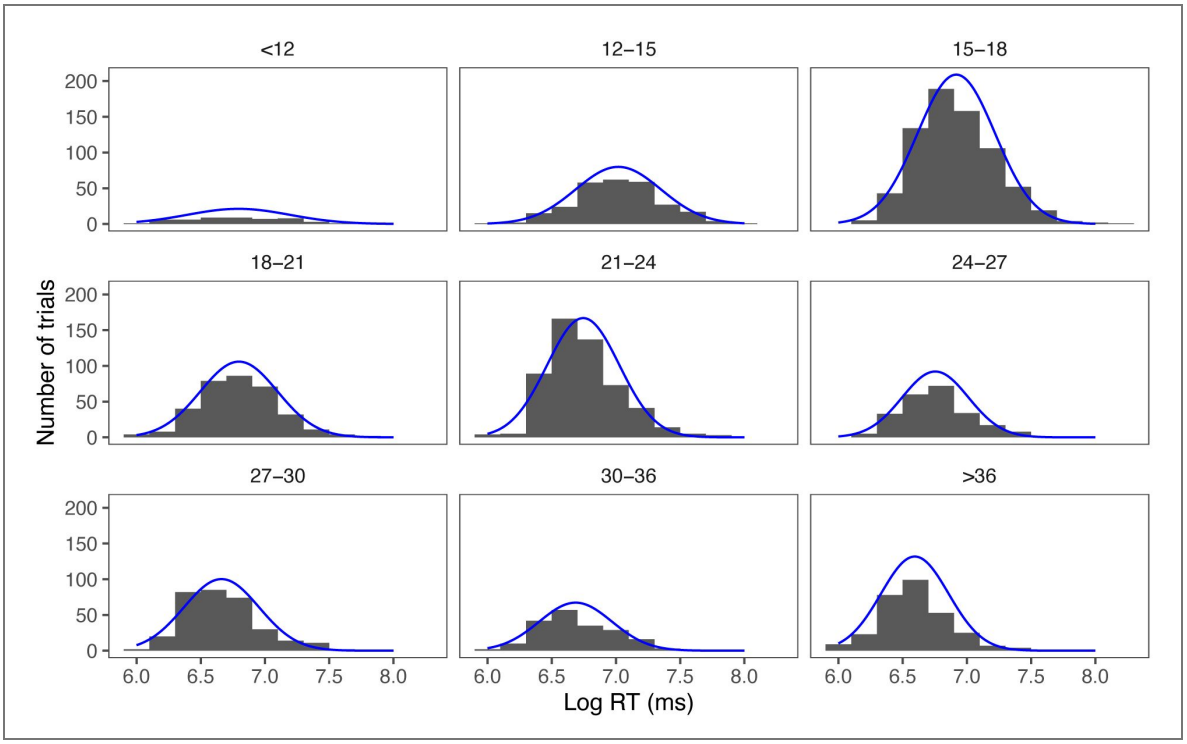


Figure S6. **Goodness of fit for different distributional models of accuracy, split by age.**

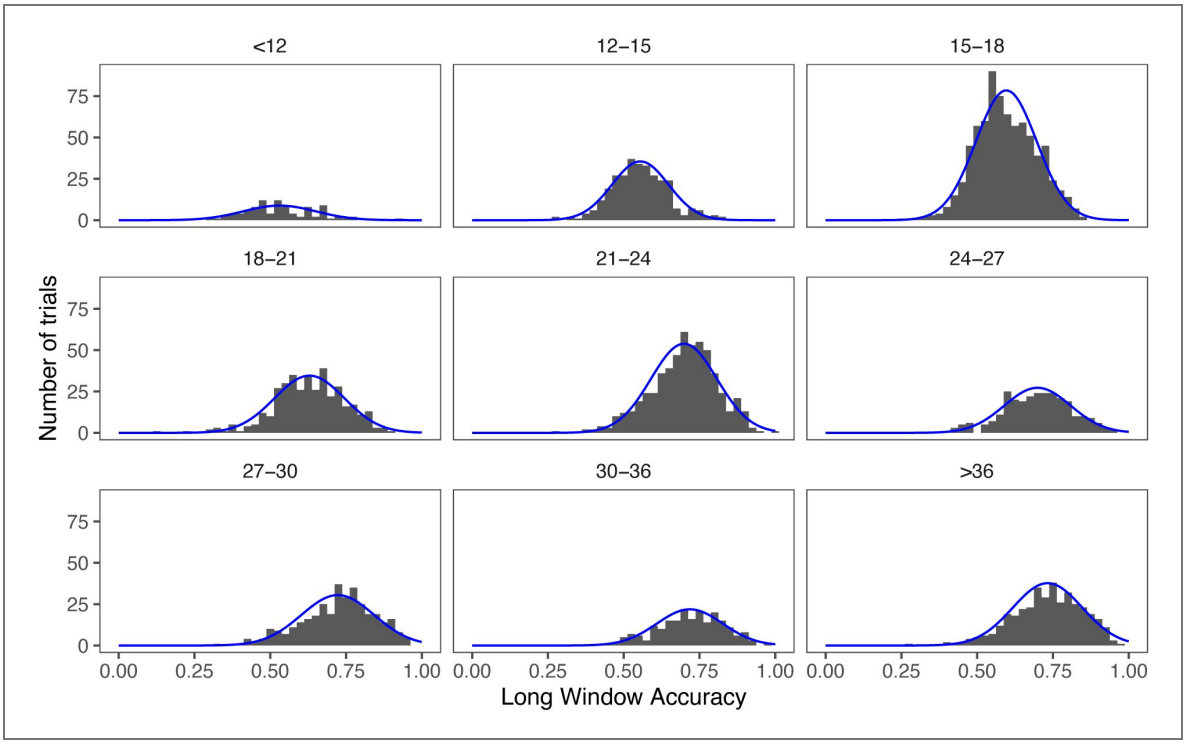
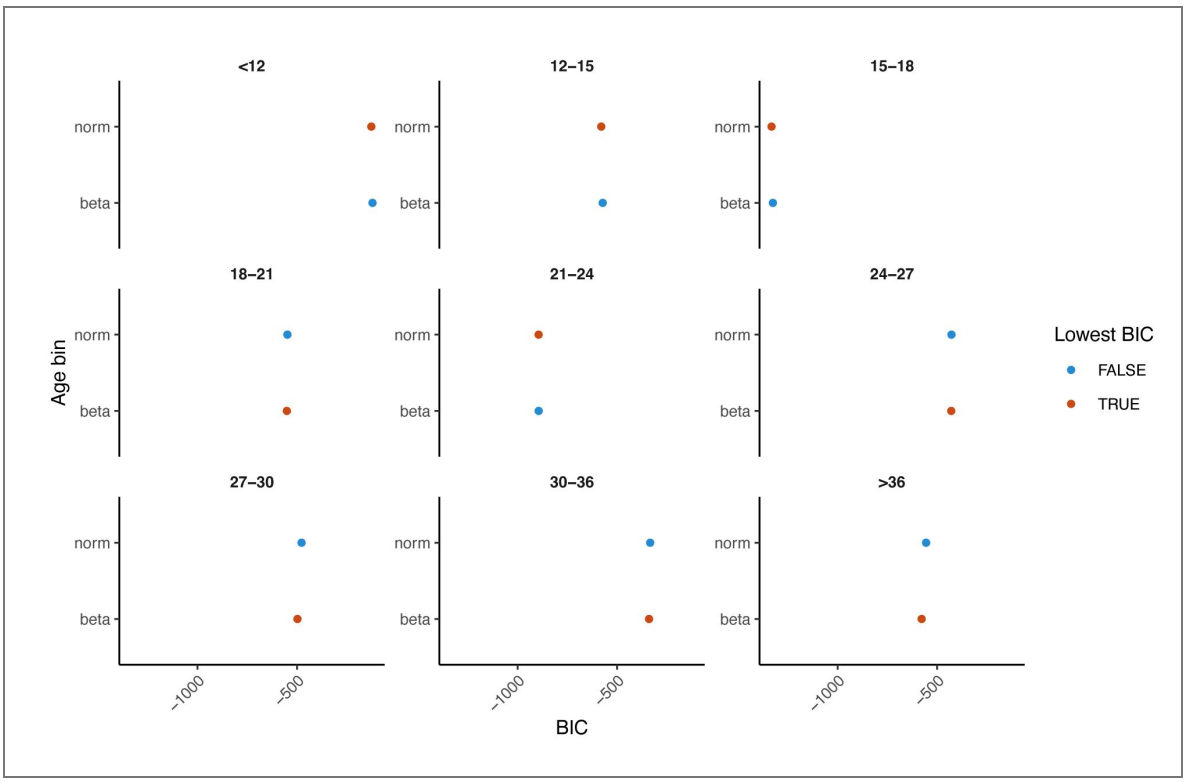


Figure S7. **Distribution of accuracies overlaid with normal distribution, split by age.**



S6. Functional Form Model Comparison

Table S3 shows model comparison measures for different models of the functional form of the relationship between accuracy and age and Table S4 shows the same for reaction time. Age gradients are estimated substantially better with long window accuracies. Note that there are a greater number of observations for short window accuracies due to less missing data. We speculate that, on average, more participants looked away from the screen towards the end of trials, leading to a greater number of exclusions of long window trials based on the 50% criterion. Note that the total percentage of trials excluded is still small for both measures: 4.8% for long window accuracy and 1.8% for short window accuracy.

Table S3. Model comparison metrics for different functional forms of the relationship between accuracy and age.

model	n. obs	sigma	logLik	AIC	BIC	REMLcrit	df.residual	r2
Long window, linear age	55337	0.272	-7012	14038	14100	14024	55330	0.124
Long window, log age	55337	0.271	-6972	13957	14020	13943	55330	0.108
Short window, linear age	57045	0.309	-14534	29082	29145	29068	57038	0.092
Short window, log age	57045	0.309	-14501	29016	29079	29002	57038	0.077

Table S4. Model comparison metrics for different functional forms of the relationship between RT and age.

model	n. obs	sigma	logLik	AIC	BIC	REMLcrit	df.residual	r2
Log RT, linear age	20689	0.458	-13844	27703	27758	27689	20682	0.232
Log RT, log age	20689	0.457	-13819	27651	27707	27637	20682	0.212
Linear RT, linear age	20689	578.837	-161573	323159	323215	323145	20682	0.197
Linear RT, log age	20689	578.338	-161546	323106	323161	323092	20682	0.185

S7. Power Law Fits

In the literature on the "law of practice", although the log-log relationship we observed is commonly present in the aggregate across individuals, the situation is substantially more complex when relationships are measured within individuals. The best fitting curves for individuals are often exponentials or delayed exponentials (Evans et al., 2018; Heathcote, Brown, & Mewhort, 2000).

With our current dataset, we unfortunately cannot specifically determine whether within-individual patterns of change conform to linear, power law, or exponential developmental patterns, because we have insufficient data about individuals' improvement across time. Thus, our current results apply to the form of the age gradient as opposed to the form of any individual's pattern of developmental change.

We believe that, unlike the skills being studied in the prior adult literature (e.g., Anderson, 1982; Heathcote, Brown, & Mewhort, 2000; Logan, 1988), language processing is being learned over the course of a child's lifetime. Thus, we do not expect to see within-paradigm changes in learning in what is a narrow period of time compared to the duration over which language processing skills are refined.

Nevertheless, here we test for other forms of the aggregate relationship between age and reaction time. In particular, we consider 1) a log~log relationship between RT and age (presented in the main text), 2) using both a log age and a linear age to predict log RT, 3) a quadratic relationship between age and RT, and 4) a cubic relationship between age and RT. As shown in Table S5, the model with a linear age term in addition to a log age term has the best fit, although the linear age term coefficient is only marginally significant (coefficients in Table S6).

Table S5. Goodness of fit comparison between different models of the relationship between age and RT.

model	n. obs	sigma	logLik	AIC	BIC	REMLcrit	df.residual	r2
Log RT, log age	18940	0.451	-12393	24801	24855	24787	18933	0.205
Log RT, log+linear age	18940	0.449	-12349	24719	24806	24697	18929	0.216
Linear RT, poly(age,2)	18940	569.500	-147594	295204	295267	295188	18932	0.180
Linear RT, poly(age,3)	18940	569.005	-147567	295151	295222	295133	18931	0.188

Table S6. Fixed effects coefficients for a model predicting log RT from both log age and linear age.

Term	Estimate	Std. Error	df	t value	*p*-value
Intercept	6.829	0.026	20.104	266.137	0.000
log age	-0.251	0.072	12.851	-3.515	0.004
linear age	0.132	0.078	13.232	1.697	0.113

These models reveal a small but significant additional linear age term over and above log age, but – because individual participant-level fits are not possible – this term can't really be used to weigh in on the debate about the precise nature of the learning pattern.

S8. Mixed-effects model specifications

Here we provide specifications for the lmer mixed-effects models used in the main text. These models are used to estimate the relationship between age and the primary variables of interest, controlling for dataset and subject-level variability.

For accuracy, 4 models were run, crossing long and short windows as the dependent variable with age or log age as the predictor.

```
long_window_accuracy ~ age_s + (age_s | dataset_name) + (1 | subject_id)
long_window_accuracy ~ log_age_s + (log_age_s | dataset_name) + (1 |
subject_id)
short_window_accuracy ~ age_s + (age_s | dataset_name) + (1 | subject_id)
short_window_accuracy ~ log_age_s + (log_age_s | dataset_name) + (1 |
subject_id)
```

For reaction time, 4 models were run, crossing rt and log rt as the dependent variable with age or log age as the predictor.

```
log_rt ~ age_s + (age_s | dataset_name) + (1 | subject_id)
log_rt ~ log_age_s + (log_age_s | dataset_name) + (1 | subject_id)
rt ~ age_s + (age_s | dataset_name) + (1 | subject_id)
rt ~ log_age_s + (log_age_s | dataset_name) + (1 | subject_id)
```

To look at the relationship between variance in the accuracy and reaction time measures and children's age, we ran two models.

```
long_window_acc_var ~ log_age_s + (log_age_s | dataset_name) + (1 |
subject_id)
log_rt_var ~ log_age_s + (log_age_s | dataset_name) + (1 | subject_id)
```

In the growth curve analysis, we fit a mixed-effects model predicting growth in vocabulary as a quadratic function of age, RT at study initiation (t0), and their interaction, using the formula below

```
prod ~ poly(age_15,2) *rt_t0 + (age | subject_id) + (1 | dataset_name)
```

S9. Factor Analysis

Figure 3 shows the result of a parallel analysis supporting the presence of three factors in the exploratory factor analysis. Table S7 shows the factor loadings for the exploratory three-factor solution using varimax rotation. The first factor is primarily driven by vocabulary measures, the second by reaction time, and the third by accuracy measures.

Table S7. Factor loadings for the exploratory three factor solution using varimax rotation.

	F1	F2	F3
RT	-0.19	0.81	-0.30
RT var	-0.10	0.81	-0.22
long window accuracy	0.33	-0.31	0.55
long window accuracy var	-0.10	0.26	-0.65
production vocabulary	0.95	-0.04	0.30
comprehension vocabulary	0.61	-0.16	-0.01
age	0.63	-0.14	0.37

The confirmatory factor analysis of this three-factor solution was fit using the following specification

```
vocab =~ prod + comp
accuracy =~ long_window_accuracy + long_window_acc_var
speed =~ log_rt + log_rt_var
```

The confirmatory factor analysis of the three-factor solution with a relation to age was fit using the following specification

```
vocab =~ prod + comp
accuracy =~ acc + acc_sd
speed =~ log_rt + log_rt_sd

vocab ~ log_age
accuracy ~ log_age
speed ~ log_age
```

The SEM with a linear growth curve used the following specification

```
accuracy_intercept =~ 1*acc_t0 + 1*acc_t1 + 1*acc_t2 + 1*acc_t3 +
1*acc_t4
accuracy_slope =~ 1*acc_t0 + 2*acc_t1 + 3*acc_t2 + 4*acc_t3 + 5*acc_t4
speed_intercept =~ 1*log_rt_t0 + 1*log_rt_t1 + 1*log_rt_t2 + 1*log_rt_t3
+ 1*log_rt_t4

speed_slope =~ 1*log_rt_t0 + 2*log_rt_t1 + 3*log_rt_t2 + 4*log_rt_t3 +
5*log_rt_t4
vocab_intercept =~ 1*prod_t0 + 1*prod_t1 + 1*prod_t2 + 1*prod_t3 +
1*prod_t4
vocab_slope =~ 1*prod_t0 + 2*prod_t1 + 3*prod_t2 + 4*prod_t3 + 5*prod_t4

accuracy_intercept ~~ NA*accuracy_intercept
accuracy_slope ~~ NA*accuracy_slope
speed_intercept ~~ NA*speed_intercept
speed_slope ~~ NA*speed_slope
```

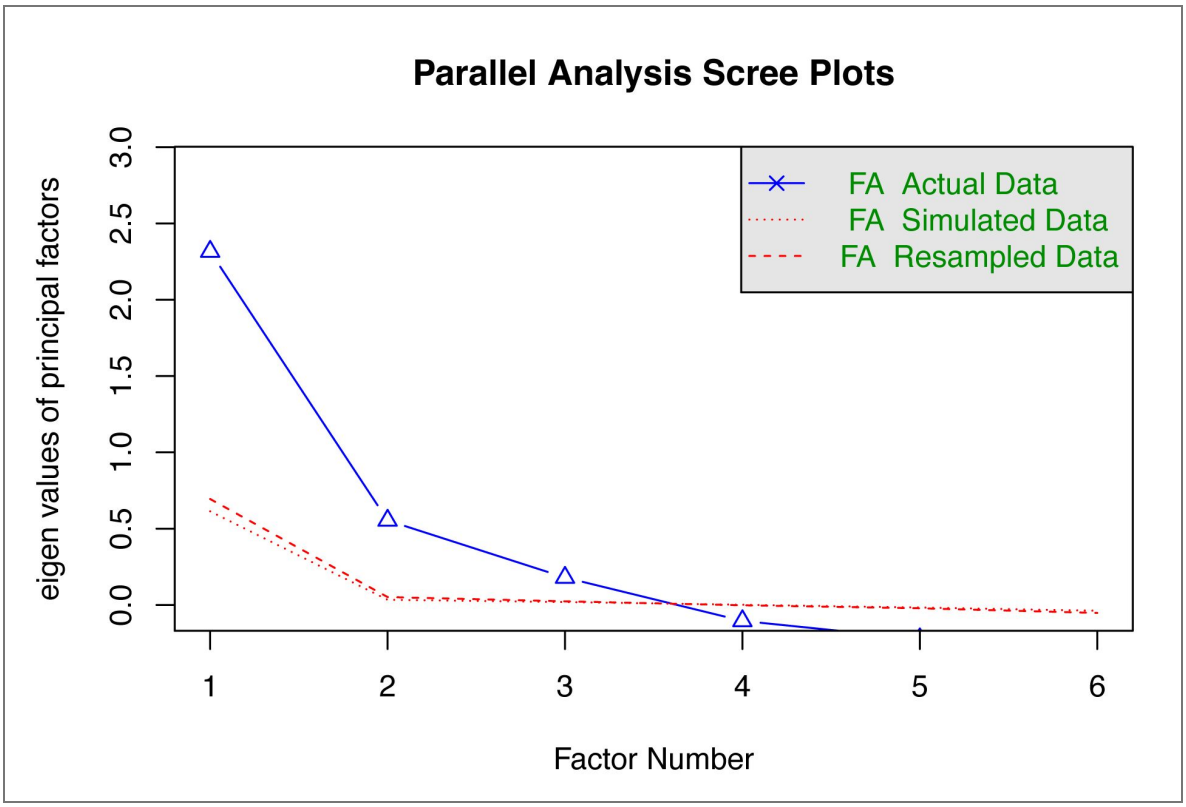


Figure S8. **Parallel analysis scree plot showing the eigenvalues for each factor, for actual, simulated, and resampled data.**

```
vocab_intercept ~~ NA*vocab_intercept
vocab_slope ~~ NA*vocab_slope
```

S10. Factor Analysis on First Administrations

As a robustness check, we tested our best factor analytic models using only cross-sectional data (filtering to the first test session in longitudinal datasets; N=1963 instead of N=3553). A comparison of all 4 models is shown in Table S8. For the three-factor CFA, the first administration model shows increased CFI (.992 instead of .972) and decreased RMSEA (.030 instead of .065). The same is true for the age-regressed three-factor CFA, which shows very good statistics on both first administrations and longitudinal data (CFI = .999 and .991, respectively and RMSEA = .009 and .037 respectively).

Table S8. Comparison of confirmatory factor analysis models on longitudinal data or first administrations only.

Model	CFI	RMSEA	$RMSEA_{Lower}$	$RMSEA_{Upper}$	TLI	SRMR	AIC	BIC
No Age, longitudinal	0.980	0.058	0.047	0.069	0.949	0.033	50821	50952
No Age, first admin	0.992	0.033	0.018	0.049	0.980	0.023	27464	27585
Age, longitudinal	0.993	0.033	0.024	0.042	0.984	0.021	48565	48717
Age, first admin	0.997	0.021	0.007	0.034	0.993	0.013	26173	26313

S11. Alternative Factor Structures

In this section, we provide comparisons between the three-factor model we report in the main text and several alternative models, including:

- a one-factor model;
- a two-factor model with vocabulary separated from speed and accuracy;
- a two-factor model with speed separated from accuracy and vocabulary; and
- a two-factor model with variability terms separated from speed, accuracy, and vocabulary.

Table S9 shows the result of these comparisons. The three-factor model shows the lowest AIC and BIC, as well as being significantly better fitting than the next-best model.

Table S9. Model comparison for alternative factor structures.

p-values show differences between adjacent models; no *p*-values are shown for comparisons between non-nested models.

	Df	AIC	BIC	Chisq	Chisq diff	RMSEA	Df diff	Pr(>Chisq)
Three-factor	6	46346.69	46475.43	91.96				
Two-factor (vocab)	8	46397.74	46514.22	147.01	55.05	0.09	2	< 0.0001
Two-factor (speed)	8	46486.96	46603.44	236.23	89.22	0.00	0	
Two-factor (variability)	8	46536.10	46652.58	285.37	49.14	0	0	
One-factor	9	46535.49	46645.84	286.76	1.39	0.01	1	0.24

S12. Non-linear Growth Models

To test for the differentiation of vocabulary growth based on initial reaction time, we used the package `brms` to fit a (Bayesian) logistic growth model to the production data. This model has two parameters for the logistic curve, a scale and an intercept. Both were allowed to interact with initial reaction time. We also included random effects of logistic intercept and scale by participant and a grouping term across datasets.

Age and initial reaction time were both mean-centered. This model showed a significant effect of initial reaction time on the intercept of the logistic growth curve, but not on its scale (see Table S10).

Table S10. Fixed effects estimates from logistic growth model.

Type	Estimate	Est.Error	l-95% CI	u-95% CI
Constant component of growth intercept	2.850	0.606	1.680	4.109
Effect of t_0 RT on growth intercept	3.187	0.604	1.970	4.369
Constant compontent of growth scale	1.121	0.058	1.009	1.242
Effect of t_0 RT on growth scale	-0.026	0.079	-0.178	0.133

The formula specification was

```
nlform <- brms::bf(
  prod ~ 1 / (1 + exp((xmid - age_c) / exp(logscale))),
  xmid ~ 1 + log_rt_0_c + (1 | dataset_name/subject_id),
  logscale ~ 1 + log_rt_0_c + (1 | dataset_name/subject_id),
  # scale ~ 1 + log_rt_0_c,
  nl = TRUE
)
```

And the priors were

```
priors <- c(
  prior(normal(0, 5), nlpar = “xmid”, coef = “Intercept”),
  prior(normal(1, 1), nlpar = “logscale”, coef = “Intercept”),
  prior(normal(0, 1), nlpar = “logscale”, coef = “log_rt_0_c”),
  prior(exponential(1), class = “sigma”),
  prior(normal(0, 2), nlpar = “xmid”, coef = “log_rt_0_c”),

  # Random effects for xmid
  prior(exponential(1), class = “sd”, nlpar = “xmid”, group =
“dataset_name”),
  prior(exponential(1), class = “sd”, nlpar = “xmid”, group =
“dataset_name:subject_id”),

  # Random effects for scale
  prior(exponential(1), class = “sd”, nlpar = “logscale”, group =
“dataset_name”),
  prior(exponential(1), class = “sd”, nlpar = “logscale”, group =
“dataset_name:subject_id”)
)
```

Age and reaction time are correlated, so to check that the effects of initial reaction time were not due to age effects, we reran the model using residualized reaction time to remove effects of age. As seen in Table S11 and Figure S9, the pattern of effects is similar for residualized reaction time as for reaction time.

Table S11. Fixed effects estimates from logistic growth model using RT residualized on age as the predictor.

Type	Estimate	Est.Error	l-95% CI	u-95% CI
Constant component of growth intercept	2.798	0.570	1.740	3.987

Effect of residualized t_0 RT on growth intercept	3.237	0.562	2.119	4.340
Constant component of growth scale	1.104	0.061	0.986	1.232
Effect of residualized t_0 RT on growth scale	0.056	0.079	-0.097	0.214

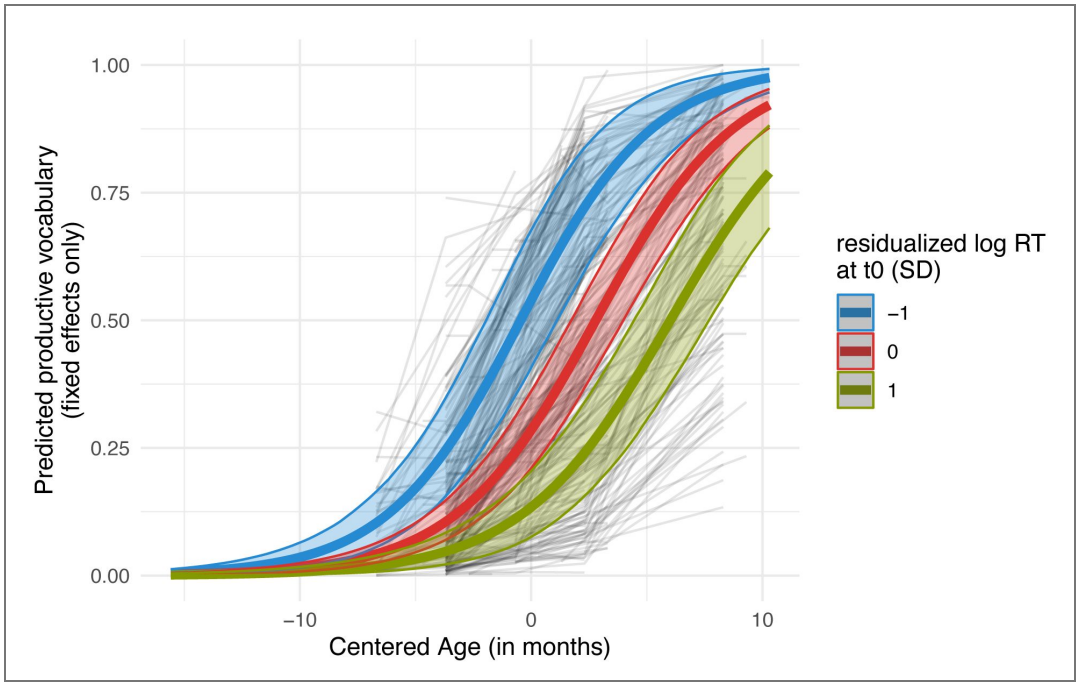


Figure S9. Growth curves from a logistic growth model showing predicted productive vocabulary growth for children based on their age-residualized initial reaction times. Predictions are shown for children with initial reaction times one SD faster than the mean for their age (blue), at the mean for their age (red), and one SD slower than the mean for their age (green). Individual longitudinal trajectories are shown in light gray. Solid lines show global model estimates and colored regions indicate 95% credible intervals.

Interpretation of growth in both this model and the linear growth model in the main text is complicated by the fact that the CDI form puts a ceiling on the total number of words that can be recorded; both the quadratic growth functions and the logistic functions come together at the form ceiling. Thus, a shift in quadratic growth in the linear model and a shift in intercept in the logistic model both point to the same overall effect, which is faster growth at the point of maximal sensitivity of the CDI. Neither model can estimate whether the overall growth trajectory is different beyond the range of the CDI. Thus, although these models might initially seem to be in conflict, we believe that they actually point to the same phenomenon, which is perhaps better described by the longitudinal SEM model reported in the main text. Children with greater skill in word recognition show an overall positive shift in the growth trajectory of vocabulary development.

S13. SEM Longitudinal Missingness

The SEM model was fit to the entire dataset, including the large mass of cross-sectional data (to anchor the estimates of t0 coefficients) and the sparse longitudinal data for each time point. We have 3%-12% of the total t0 datapoints for any given time point (see Table S12), given the sparsity of longitudinal sampling (only 6/24 of the datasets are longitudinal).

Table S12. Fraction of data present for each measure at each time point for the longitudinal SEM.

timepoint	Log RT	Accuracy	Production	Comprehension
t0	0.685	0.865	0.279	0.183

t1	0.035	0.039	0.024	0.013
t2	0.093	0.096	0.095	0.000
t3	0.028	0.029	0.027	0.000
t4	0.031	0.032	0.031	0.000
t5	0.065	0.065	0.057	0.000

Our data are MAR (missing at random) rather than MCAR (missing completely at random). This is because their missingness is due to which dataset they are part of – if they are from a cross-sectional dataset, they are by definition missing all longitudinal observations. For our analyses to be appropriate given this structure, we have to assume that the general developmental patterns we are studying are replicated across datasets. We believe that they are, and we show this statistically using our mixed-effects and non-linear mixed-effects models, which control for dataset-related variation. We also show dataset-level effects in a number of our visualizations for this same reason. The same degree of random effect specification that we can do in the mixed-effects models is not possible in the SEM model, however, purely for technical reasons. Again, this point highlights the importance of convergence across analyses.

Data availability

We retrieved all data from Peekbank release 2026.1 using the peekbankr R package. All code and data necessary to reproduce this manuscript are available at *https://github.com/peekbank/peekbank-development*.

Additional information

Funding

Funder	Grant reference number	Author
Jacobs Foundation (Foundation_JF)		Michael C Frank

Author ORCID iDs

Michael C Frank: https://orcid.org/0000-0002-7551-4378
Virginia A Marchman: https://orcid.org/0000-0001-7183-6743
Mika Braginsky: https://orcid.org/0000-0001-9039-3220
Martin Zettersten: https://orcid.org/0000-0002-0444-7059

References

1. **Bates E,** et al. (1994) Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language* **21**:85-123 https://doi.org/10.1017/s0305000900008680 | PubMed
2. **Frank MC,** Braginsky M, Yurovsky D, Marchman VA (2021) *Variability and Consistency in Early Language Learning: The Wordbank Project* Cambridge, MA: MIT Press.
3. **Fernald A,** Pinto JP, Swingley D, Weinberg A, McRoberts GW (1998) Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science* **9**:228-231 https://doi.org/10.1111/1467-9280.00044
4. **Peter MS,** et al. (2019) Does speed of processing or vocabulary size predict later language growth in toddlers?. *Cognitive Psychology* **115**:101238 https://doi.org/10.1016/j.cogpsych.2019.101238 | PubMed
5. **Bergelson E** (2020) The comprehension boost in early word learning: Older infants are better learners. *Child development perspectives* **14**:142-149 https://doi.org/10.1111/cdep.12373 | PubMed

6. **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**:3253-3258 https://doi.org/10.1073/pnas.1113380109 | PubMed
7. **Tanenhaus MK,** Spivey-Knowlton MJ, Eberhard KM, Sedivy JC (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* **268**:1632-1634 https://doi.org/10.1126/science.7777863 | PubMed
8. **Altmann GT,** Kamide Y (1999) Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* **73**:247-264 https://doi.org/10.1016/s0010-0277(99)00059-1 | PubMed
9. **Fernald A,** Zangl R, Portillo AL, Marchman VA (2008) Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In: Sekerina IA, Fernandez EM, Clahsen H (Eds). *Developmental Psycholinguistics: On-Line Methods in Children's Language Processing* Amsterdam: John Benjamins. pp. 97-135
10. **MacDonald K,** LaMarr T, Corina D, Marchman VA, Fernald A (2018) Real-time lexical comprehension in young children learning american sign language. *Developmental science* **21**:e12672 https://doi.org/10.1111/desc.12672 | PubMed
11. **Hirsh-Pasek K,** Golinkoff RM (1996) The intermodal preferential looking paradigm: A window onto emerging language comprehension. In: McDaniel D, McKee C, Cairns HS (Eds). *Methods for Assessing Children's Syntax* Cambridge, MA: The MIT Press. pp. 105-124 https://doi.org/10.7551/mitpress/4575.003.0009
12. **Reznick JS** (1990) Visual preference as a test of infant word comprehension. *Applied Psycholinguistics* **11**:145-166 https://doi.org/10.1017/s0142716400008742
13. **Mani N,** Plunkett K (2011) Phonological priming and cohort effects in toddlers. *Cognition* **121**:196-206 https://doi.org/10.1016/j.cognition.2011.06.013 | PubMed
14. **Meylan SC,** Levy RP, Bergelson E (2025) Children's expressive and receptive knowledge of the english regular plural. Developmental Psychology.
15. **Swingley D,** Aslin RN (2002) Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science* **13**:480-484 https://doi.org/10.1111/1467-9280.00485 | PubMed
16. **Trueswell JC,** Sekerina I, Hill NM, Logrip ML (1999) The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition* **73**:89-134 https://doi.org/10.1016/s0010-0277(99)00032-3 | PubMed
17. **Mani N,** Plunkett K (2010) In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological science* **21**:908-913 https://doi.org/10.1177/0956797610373371 | PubMed
18. **Bergelson E,** Aslin RN (2017) Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences* **114**:12916-12921 https://doi.org/10.1073/pnas.1712966114 | PubMed
19. **Fernald A,** Perfors A, Marchman VA (2006) Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental psychology* **42**:98 https://doi.org/10.1037/0012-1649.42.1.98 | PubMed
20. **Frank MC,** Goodman ND, Tenenbaum JB (2009) Using speakers' referential intentions to model early cross-situational word learning. *Psychological science* **20**:578-585 https://doi.org/10.1111/j.1467-9280.2009.02335.x | PubMed
21. **Pickering MJ,** Gambi C (2018) Predicting while comprehending language: A theory and review. *Psychological bulletin* **144**:1002 https://doi.org/10.1037/bul0000158 | PubMed
22. **Ryskin R,** Nieuwland MS (2023) Prediction during language comprehension: What is next?. *Trends in cognitive sciences* **27**:1032-1052 https://doi.org/10.1016/j.tics.2023.08.003 | PubMed
23. **Zettersten M** (2019) Learning by predicting: How predictive processing informs language development. In: Busse B, Moehlig-Falke R (Eds). *Patterns in Language and Linguistics: New Perspectives on a Ubiquitous Concept* Berlin: Mouton de Gruyter. pp. 255-288 https://doi.org/10.1515/9783110596656-010

24. **Hart B,** Risley TR (1995) *Meaningful differences in the everyday experience of young american children* Paul H Brookes Publishing.
25. **Anderson NJ,** Graham SA, Prime H, Jenkins JM, Madigan S (2021) Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development* **92**:484-501 https://doi.org/10.1111/cdev.13508 | PubMed
26. **Weisleder A,** Fernald A (2013) Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science* **24**:2143-2152 https://doi.org/10.1177/0956797613488145 | PubMed
27. **Marchman VA,** Fernald A (2008) Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental science* **11**:F9-16 https://doi.org/10.1111/j.1467-7687.2008.00671.x | PubMed
28. **Fernald A,** Marchman VA (2012) Individual differences in lexical processing at 18 months predict vocabulary growth in typically developing and late-talking toddlers. *Child development* **83**:203-22 https://doi.org/10.1111/j.1467-8624.2011.01692.x | PubMed
29. **Newbury J,** Klee T, Stokes SF, Moran C (2016) Interrelationships between working memory, processing speed, and language development in the age range 2–4 years. *Journal of Speech, Language, and Hearing Research* **59**:1146-1158 https://doi.org/10.1044/2016_jslhr-l-15-0322 | PubMed
30. **Colby SE,** McMurray B (2023) Efficiency of spoken word recognition slows across the adult lifespan. *Cognition* **240**:105588 https://doi.org/10.1016/j.cognition.2023.105588 | PubMed
31. **Jeppsen C,** Baxelbaum K, Tomblin B, Klein K, McMurray B (2025) The development of lexical processing: Real-time phonological competition and semantic activation in school age children. *Quarterly Journal of Experimental Psychology* **78**:437-458 https://doi.org/10.1177/17470218241244799 | PubMed
32. **McMurray B,** Apfelbaum KS, Tomblin JB (2022) The slow development of real-time processing: Spoken-word recognition as a crucible for new thinking about language acquisition and language disorders. *Current Directions in Psychological Science* **31**:305-315 https://doi.org/10.1177/09637214221078325 | PubMed
33. **Zettersten M,** et al. (2023) Peekbank: An open, large-scale repository for developmental eye-tracking data of children's word recognition. *Behavior Research Methods* **55**:2485-2500 https://doi.org/10.3758/s13428-022-01906-4 | PubMed
34. **Christiansen MH,** Chater N (2016) The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences* **39**:e62 https://doi.org/10.1017/s0140525x1500031x | PubMed
35. **Chater N,** Christiansen MH (2018) Language acquisition as skill learning. *Current opinion in behavioral sciences* **21**:205-208 https://doi.org/10.1016/j.cobeha.2018.04.001
36. **Snoddy GS** (1926) Learning and stability: A psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology* **10**:1 https://doi.org/10.1037/h0075814
37. **Kail R** (1991) Processing time declines exponentially during childhood and adolescence. *Developmental psychology* **27**:259 https://doi.org/10.1037/0012-1649.27.2.259
38. **Anderson JR** (1982) Acquisition of cognitive skill. *Psychological review* **89**:369 https://doi.org/10.1037/0033-295x.89.4.369
39. **Heathcote A,** Brown S, Mewhort DJ (2000) The power law repealed: The case for an exponential law of practice. *Psychonomic bulletin & review* **7**:185-207 https://doi.org/10.3758/bf03212979 | PubMed
40. **Evans NJ,** Brown SD, Mewhort DJ, Heathcote A (2018) Refining the law of practice. *Psychological review* **125**:592 https://doi.org/10.1037/rev0000105 | PubMed
41. **McMurray B,** Horst JS, Samuelson LK (2012) Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review* **119**:831 https://doi.org/10.1037/a0029872 | PubMed

42. **Todorov E,** Jordan MI (2002) Optimal feedback control as a theory of motor coordination. *Nature neuroscience* **5**:1226-1235 https://doi.org/10.1038/nn963 | PubMed

43. **Marchman VA,** Adams KA, Loi EC, Fernald A, Feldman HM (2016) Early language processing efficiency predicts later receptive vocabulary outcomes in children born preterm. *Child Neuropsychology* **22**:649-665 https://doi.org/10.1080/09297049.2015.1038987 | PubMed

44. **Lany J** (2018) Lexical-processing efficiency leverages novel word learning in infants and toddlers. *Developmental science* **21**:e12569 https://doi.org/10.1111/desc.12569 | PubMed

45. **Marchman VA,** Dale PS, Fenson L (2023) *MacArthur-bates communicative development inventories users guide and technical manual, third edition* Brookes.

46. **Judd N,** Klingberg T, Sjöwall D (2021) Working memory capacity, variability, and response to intervention at age 6 and its association to inattention and mathematics age 9. *Cognitive Development* **58**:101013 https://doi.org/10.1016/j.cogdev.2021.101013

47. **Galeano Weber EM,** Dirk J, Schmiedek F (2018) Variability in the precision of children's spatial working memory. *Journal of Intelligence* **6**:8 https://doi.org/10.3390/jintelligence6010008 | PubMed

48. **Kautto A,** Railo H, Mainela-Arnold E (2024) Introducing the intra-individual variability hypothesis in explaining individual differences in language development.

49. **Bergelson E,** et al. (2023) Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences of the United States of America* **120**:e2300671120 https://doi.org/10.1073/pnas.2300671120 | PubMed

50. **The ManyBabies Consortium** (2020) Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science* **3**:24-52 https://doi.org/10.1177/2515245919900

51. **VanderWeele TJ,** Batty CJ (2023) On the dimensional indeterminacy of one-wave factor analysis under causal effects. *Journal of Causal Inference* **11**:1-15 https://doi.org/10.1515/jci-2022-0074

52. **Pomper R,** Saffran JR (2015) Unpublished "prime" study: Modulating attention to different features of objects during word learning.

53. **Pomper R,** Saffran JR (2017) Unpublished "dimy" study: Do infants learn to associate diminutive forms with animates?.

54. **Pomper R,** Saffran JR (2018) More than distractors: Familiar objects influence toddlers' semantic representations in novel word learning. Poster at XXI Biennial International Congress of Infant Studies.

55. **Moore C,** Bergelson E (2022) Examining the roles of regularity and lexical class in 18–26-month-olds' representations of how words sound. *Journal of Memory and Language* **126**:104337 https://doi.org/10.1016/j.jml.2022.104337

56. **Weaver H,** Saffran JR (2026) Interrogating early word knowledge: Factors that influence the alignment between caregiver-report and experimental measures. *Developmental Science* **29**:e70088 https://doi.org/10.1111/desc.70088 | PubMed

57. **Borovsky A,** Peters RE (2019) Vocabulary size and structure affects real-time lexical recognition in 18-month-olds. *PloS one* **14**:e0219290 https://doi.org/10.1371/journal.pone.0219290 | PubMed

58. **Yurovsky D,** Wade A, Frank M (2013) Online processing of speech and social information in early word learning. In: Proceedings of the Annual Meeting of the Cognitive Science Society.

59. **Yoon EJ,** Wu YC, Frank MC (2015) Children's online processing of ad-hoc implicatures. In: Proceedings of the Annual Meeting of the Cognitive Science Society.

60. **Bacon D,** Saffran J (2022) Role of speaker gender in toddler lexical processing. *Infancy* **27**:291-300 https://doi.org/10.1111/infa.12455 | PubMed

61. **Ronfard S,** Wei R, Rowe ML (2022) Exploring the linguistic, cognitive, and social skills underlying lexical processing efficiency as measured by the looking-while-listening paradigm. *Journal of Child Language* **49**:302-325 https://doi.org/10.1017/s0305000921000106 | PubMed

62. **Perry LK,** Saffran JR (2017) Is a pink cow still a cow? Individual differences in toddlers' vocabulary knowledge and lexical representations. *Cognitive science* **41**:1090-1105 https://doi.org/10.1111/cogs.12370 | PubMed

63. **Pomper R,** Saffran JR (2016) Roses are red, socks are blue: Switching dimensions disrupts young children's language comprehension. *PloS one* **11**:e0158459 https://doi.org/10.1371/journal.pone.0158459 | PubMed

64. **Pomper R,** Saffran JR (2019) Familiar object salience affects novel word learning. *Child development* **90**:e246–e262 https://doi.org/10.1111/cdev.13053 | PubMed

65. **Frank MC,** Sugarman E, Horowitz AC, Lewis ML, Yurovsky D (2016) Using tablets to collect data from young children. *Journal of Cognition and Development* **17**:1-17 https://doi.org/10.1080/15248372.2015.1061528

66. **Weaver H,** Zettersten M, Saffran JR (2024) Becoming word meaning experts: Infants' processing of familiar words in the context of typical and atypical exemplars. *Child Development* **95**:e352–e372 https://doi.org/10.1111/cdev.14120 | PubMed

67. **Yurovsky D,** et al. (2017) Developmental changes in the speed of social attention in early word learning. unpublished manuscript.

68. **Fernald A,** Marchman VA, Weisleder A (2013) SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science* **16**:234-248 https://doi.org/10.1111/desc.12019 | PubMed

69. **Mahr T,** McMillan BTM, Saffran JR, Ellis Weismer S, Edwards J (2015) Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition* **142**:345-350 https://doi.org/10.1016/j.cognition.2015.05.009 | PubMed

70. **Yurovsky D,** Frank MC (2017) Beyond naïve cue combination: Salience and social cues in early word learning. *Devopmental Science* **20**:e12349 https://doi.org/10.1111/desc.12349 | PubMed

71. **Adams KA,** et al. (2018) Caregiver talk and medical risk as predictors of language outcomes in full term and preterm toddlers. *Child Development* **89**:1674-1690 https://doi.org/10.1111/cdev.12818 | PubMed

72. **Garrison H,** Baudet G, Breitfeld E, Aberman A, Bergelson E (2020) Familiarity plays a small role in noun comprehension at 12–18 months. *Infancy* **25**:458-477 https://doi.org/10.1111/infa.12333 | PubMed

73. **Potter C,** Lew-Williams C (2024) Frequent vs. Infrequent words shape toddlers' real-time sentence comprehension. *Journal of Child Language* **51**:1478-1488 https://doi.org/10.1017/s0305000923000387 | PubMed

74. **Baayen R. H.,** Milin P. (2010) Analyzing reaction times. *International Journal of Psychological Research* **3**:12-28 https://doi.org/10.21500/20112084.807

75. **Creel S.** (2024) Connecting the tots: Strong looking-pointing correlations in preschoolers' word learning and implications for continuity in language development. *Child Development* **96**:87-103 https://doi.org/10.1111/cdev.14157 | PubMed

76. **Logan G. D.** (1988) Toward an instance theory of automatization. *Psychological Review* **95**:492-527 https://doi.org/10.1037/0033-295x.95.4.492

77. **Luce P. A.** (1986) A computational analysis of uniqueness points in auditory word recognition. *Perception & Psychophysics* **39**:155-158 https://doi.org/10.3758/bf03212485 | PubMed

78. **Ratcliff R.** (1979) Group reaction time distributions and an analysis of distribution statistics. *Psychological bulletin* **86**:446-461 https://doi.org/10.1037/0033-2909.86.3.446 | PubMed

79. **Ratcliff R.** (1993) Methods for dealing with reaction time outliers. *Psychological Bulletin* **114**:510-532 https://doi.org/10.1037/0033-2909.114.3.510 | PubMed

80. **Van Zandt T.** (2000) How to fit a response time distribution. *Psychonomic Bulletin & Review* **7**:424-465 https://doi.org/10.3758/bf03214357 | PubMed

81. **Van Zandt T.** (2002) Analysis of response time distributions. *Stevens' Handbook of Experimental Psychology* **4**:461-516 https://doi.org/10.1002/0471214426.pas0412

Peer reviews

Reviewer #1 (Public review):

Summary:

The study examined the extent to which children's word recognition skill improves across early development, becoming faster, more accurate and less variable, and the extent to which word recognition skill is related to children's concurrent and later vocabulary knowledge.

The main strength of the study comes from the dataset which recycles previously collected data from 24 studies to examine the development of word recognition skill using data from 1963 children. This maximizes the impact of previously collected data while also allowing the study to reliably ask big picture questions on the development of word recognition skill and its relation to chronological age and vocabulary knowledge. Data analysis is rigorous, thought through and very clearly described. Data and code necessary to reproduce the manuscript are shared on the project's Github. The limitations of the study are acknowledged and the manuscript does well to tone down the causal implications of their results.

https://doi.org/10.7554/eLife.109636.2.sa1

Reviewer #2 (Public review):

Summary:

This paper presents a series of analyses of a large dataset combining many prior studies of early word recognition (Peekbank). The analyses demonstrate that the speed, accuracy and consistency of word learning improves with age. Moreover, the speed of word learning early in development was related to vocabulary growth over time.

Strengths:

A key strength of the paper is the use of a large multi-study dataset. This is particularly valuable in the field of early cognitive development, which has (due to practical limitations) often been based on small-scale studies that necessarily provide a shaky foundation for conclusions. The analyses are also well-motivated.

Weaknesses:

In an earlier version of the manuscript, the meaning of "word recognition ability" was ambiguous and could have referred to either (A) an intrinsic ability that matures, or (B) knowledge of the common, concrete words typically used in these studies that increases with experience. The revised version of the manuscript identifies these two interpretations and acknowledges that they cannot be teased apart in the current work.

https://doi.org/10.7554/eLife.109636.2.sa2

Author response:

The following is the authors' response to the original reviews

General note

We have issued a new release of the general Peekbank database, 2026.1, which includes more data integrity checks and several more datasets. As a result of this release, the underlying dataset we use in our paper has shifted slightly. The shifts represent a relatively small

proportion of the total data and thus these changes have caused only relatively minor changes to our numerical results. We also highlight that we now include a small amount of data regarding children younger than 12 months, increasing the developmental range of our analysis (see Figure 1).

> ***Reviewer 1 (Public review):***
>
> *The limitations of the study are acknowledged to some extent, but need to be improved and ensured that they run throughout the manuscript. Thus, in the discussion, the authors note that the approach is observational and exploratory, and highlight for me a key alternative explanation of the findings, namely that faster children could be faster due to their larger vocabulary, rather than faster children learning more words. Indeed, the latter explanation for the relationship is called into question, given that growth in speed was not related to growth in vocabulary. Here, the authors note that the null result may be related to the fact that they do not sufficiently precise estimates of growth slopes, rather than taking the alternative explanation seriously that there may not be as causal a link between being a faster word learner and a better word learner (learn more words).*

Thank you very much for your challenging and thoughtful comments. In hindsight we did not realize that the way we were writing about our results was ambiguous between several interpretations (one of which we endorse and one of which we do not).

We respond below to the specific suggestions about causal directionality in the longitudinal analysis, but we certainly believe that we cannot draw strong conclusions about causality from our dataset and have attempted throughout the paper to remove causal language that might have crept into our interpretation.

In response to your comments, we have made a number of key revisions aimed at qualifying and clarifying our points:

- The abstract now prominently notes that our design is observational: “In an observational study…”
- The abstract notes a positive and a negative result in the relationship between word recognition and vocabulary: “Further, across a range of longitudinal models, speed, accuracy, and vocabulary were coupled. Children with overall faster word recognition tended to show faster vocabulary growth, though developmental growth in word recognition skill was not specifically associated with growth in vocabulary.”
- The abstract removes potential casual language in the final sentence: “… these findings support the view that word recognition is a skill that develops gradually across early childhood and that this skill is deeply intertwined with early language learning.”
- A new paragraph in the Results introduces the potential hypotheses investigated via the longitudinal models.
- The final paragraph of the Results section sharpens the contrast between two possible growth hypotheses: “However, we did not find evidence for the stronger version of this claim: in neither the non-linear growth model nor the linear SEM did we find evidence that increases in speed were related to increases in vocabulary size. Thus, our findings do not support a ‘virtuous cycle’ model in which increases in recognition specifically lead to increases in vocabulary size.”

We hope these changes lead to a manuscript that better aligns with the limitations of the study.

> *This is especially since, but correct me if I’m wrong here, the current vocabulary size is not taken into consideration in the model examining vocabulary growth. Given the*

> *increasing number of studies showing that current vocabulary knowledge predicts vocabulary growth (Laing, Kalinowski et al, Siew & Vitevitch), one simple alternative explanation is that current vocabulary knowledge predicts both current word recognition skill and later vocabulary knowledge. Is there anything in the data speaking against this hypothesis?*

We think the reviewer's overall point is generally correct, as we described above, but we want to clarify a specific statistical point. The non-linear longitudinal model of vocabulary growth does in fact take into account a child's average vocabulary size. (This point feels tricky in a non-linear model but it's actually quite similar to a linear model for the purposes of this discussion). Basically, vocabulary (at all timepoints) is modeled as a function of age, with both main effects and interactions with age. Critically, each participant is also modeled as having a random intercept capturing their deviation from the average growth pattern across ages (as expressed by the fixed effects). In this model, the "main effect" (here captured by the intercept for the logistic curve in the model) that we observe for speed indicates that vocabulary growth for individuals is predicted to be faster (their curve is shifted left) if their RTs are fast. The presence of the random effects in this model thus "controls" for the fact that some participants have overall higher vocabularies (and are shifted up relative to the average growth curve).

But, we note that this model does not show an "interaction effect" (here captured by the null effect of RT on the slope parameter in the logistic model). That's one of the null effects that we now call out much more prominently in the abstract and end of the results (per our response above).

> *Equally, while the SEM examines vocabulary growth controlling for age, I wonder about the other way around. What would happen to the effect of age on word recognition skill (in the LME model, S8) if one were to add concurrent vocabulary size? So does chronological age explain word recognition skill or vocabulary knowledge? Right now, the manuscript describes this effect purely related to chronological age, but is it age per se or other cognitive abilities, including a key change across development, namely, vocabulary size? Thus, the presentation of the skill learning hypothesis suggests that age is a proxy for experience, while you actually have here a very nice proxy for experience in terms of children's vocabulary size.*

Again, thank you for engaging with this tricky set of issues. Overall, our goal is to adjust the manuscript to reflect points of agreement; in particular, we agree that age is a proxy for language experience, vocabulary, and other cognitive changes, and we have stated this explicitly now in the intro to the factor analyses: "In our prior analyses, chronological age acts as a proxy for greater language experience and larger vocabulary as well as a host of other correlated developmental changes in cognition. Now we explicitly explore relations to vocabulary growth and the triadic relationship between age, word recognition, and vocabulary."

On the statistical side, we do think that the NLME (non-linear mixed effects; the logistic growth mode) effectively controls for average vocabulary size, as described above. The longitudinal SEM also relates vocabulary growth to growth in word recognition skill. In both models, we find no evidence for coupled growth; instead the evidence points to children with higher baseline word recognition skill showing faster growth in vocabulary (speed intercept significantly related to vocabulary slope, -.14, $p < .01$) but not the reverse (vocabulary intercept not strongly related to speed slope; -.01, *ns*).

More generally, we hope our edits to the paper, detailed above, both clarify this tricky set of issues and also remove inappropriate casual language throughout.

> *Critically, while the discussion is more nuanced, the way the abstract is concluded and the way the Introduction is phrased suggest that the study is able to answer a causal question, which, as the authors themselves note, is not possible. The abstract, for instance, states that word recognition becomes faster, more accurate and less variable...consistent with a process of skill learning. And also that this skill plays a role in supporting early language learning, which is very causal language. I don't think you can really claim that you are testing the two hypotheses you suggest here. The work is definitely embedded in the context of these hypotheses, but are you really able to test them? My worry is that while the discussion is more nuanced, the extent to which this study will then be cited down the line as showing that children learn more words down the line because they are faster at recognizing words, and anything that you can do to tamper with such interpretations would be good for the literature. For me, this should not just be relegated to the discussion but should be touched upon in the abstract and Introduction.*

Thanks for pushing us to be more precise with how we frame and describe our findings. We agree with the reviewer that our findings do not warrant strong conclusions about the causal role of word recognition skill in vocabulary growth. Per our response above, we have now tried to carefully revise our language throughout the paper (in particular, in the abstract and introduction, as noted by the reviewer).

> *Finally, it would help to talk more about the mechanisms at work in any relationship between word recognition and language learning. It seems to me that this would rely on some predictive processing framework, given the description on page 4, and it would be good to make this clear (faster and more accurately you can recognize a ball, better use this evidence to infer the speaker's intended meaning).*

Thanks, this is a great point. We've revised this text and added references to predictive processing, unpacking a problematic paragraph into two:

"Familiar word recognition -- as measured by LWL -- is hypothesized to play a key role in language learning (19). The idea, in a nutshell, is that the faster and more accurately a child can process incoming words, the more opportunities they have for learning. Consider a child hearing the utterance "Can you put the ball in the crate?" The better the child can recognize the word "ball", the better they can use this evidence to help infer the speaker's intended meaning, allowing possible inferences about the meaning of the less familiar word, "crate" (20).

"Real time language processing, including word recognition, relies heavily on predictive processing, in which comprehenders integrate expectations from prior linguistic context with noisy and ephemeral incoming signals (21, 22). The more input a child receives, the better their predictions are likely to be, and hence the more they can learn (19, 23). Indeed, measurements of children's language input at home are consistently associated with their vocabulary size (24, 25). And, in line with this predictive processing framework, one important study found that children's word recognition speed mediated the longitudinal relationship between home language input and vocabulary growth (26). Thus, word recognition is thought to be a key support for ongoing word learning."

> *Equally, when referring to word recognition, it would be good to clarify what this refers to - how well a child knows what a word refers to (and in the context of LWL, what it does not refer to) or how quickly it directs attention to what is referred to.*

Thanks, we've added a capsule definition in the second paragraph, and added the sentence "This procedure [LWL] measures the general construct of word recognition by operationalizing knowledge of a meaning as visual attention to a specific named referent." We hope this clarifies the relationship between LWL and word recognition.

> *With regards to the data, I wonder if there is a clustering of kids past 24 months that is happening here, looking at Figures 1 and 2, where it seems like there is less change past the 24-month point. Is there any way to look at whether the effect of age or vocabulary on word recognition is not linear but asymptotic?*

Thanks for pointing this out; we do see what you are talking about but think it's being handled appropriately in the analysis. In Figure 1 it clearly looks like changes to RT are asymptotic – this is why we analyze the logarithm of RT throughout the paper. In Supplement S6 we show that reaction time is indeed best fit by a log-log function. Your question about Figure 2 asks whether there is further structure beyond the log-log fit; in Supplement S7 we show some analyses that suggest a polynomial fit is not better than the log-log fit; there is some small additional linear effect of age over and above the log-log fit, but it's minor and pretty hard to interpret in our view.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *Page 3. Word production may manifest in overt behaviour but need not reflect complete knowledge. A child can say the word dog and use it to refer to a cat.*

This is a good point. Since we are not able to speak to the precision of meaning representations (an important issue in its own right), we have omitted the phrase "with incomplete knowledge."

> *Page 4. The first two sentences of the paragraph beginning with word recognition ability... don't go together. The second sentence does not support the claim that word recognition plays a role in language learning.*

Thanks, we've tried to smooth out this transition as part of unpacking the role of predictive processes.

> *Page 4. "predicts children's standardized test scores years later" - make clear what test scores are here.*

We added some additional details. The specific tests were the CELF (expressive language) and the KABC (IQ), but we thought too much detail might be distracting.

> *Page 5. I love Table 1, but would like for the data to be weighted somehow. So, given that some studies had a lot more trials and more children, what percentage of the data did this study contribute? That allows a clearer view of how biased the sample is in certain studies. The x in CDIS and longitudinal could be aligned to the right. I kept wondering why there was an x near some trials.*

Thanks, we've adjusted the table to add the percentage of the total dataset (in trials) due to each study and fixed the alignment issue.

> *Page 6. 12 million individual samples: what samples are these? Individual data points per trial per time point. Making this clear would be great.*

Clarified, thanks.

> *Page 9. Your accuracy measures only seem to consider the target. From what I remember of my preferential looking days, this measure usually also includes the distractor. Why do you not do this? This is especially since you have such a wide age range, so if a 12-month-old only looks for about 50 per cent of the trial and spends that time looking at the target, that is very different from a child who looks at the screen all of the trial and spends less time looking at the target here.*

Sorry for any lack of clarity: we do in fact compute accuracy as the ratio of looking to target over looking to target plus looking to distractor. We have added this information to the parenthetical referenced above: "... accuracy (more target looking; computed as the ratio of target to target plus distractor looking)".

> *Page 12. I only found out that age was in this model by looking at S9.*

Thanks for mentioning this omission, we've clarified in the text: "We initially add age as an additional variable to our models to explore whether this factor structure relates to age; later we treat age as a predictor of latent factors."

> *Page 12. Isn't it trivial that speed and accuracy show negative covariance, especially given how you measure accuracy? Thus, if I take longer to fixate the target, I have less time to look at the target during the trial. If, however, I included the distractor in my accuracy measure, then I could still take longer to look at the target, but still look more at the target than the distractor.*

Thanks for mentioning that this covariance is not the key result of interest; that observation didn't come out in the text. Now we note that this covariation is "... as expected since they [speed and accuracy] are derived from the same data." Note per above that accuracy is computed as target / target + distractor looking; even so, your observation is correct: slower looking at the target means lower accuracy at least to some degree.

> *Page 19. If you excluded data from trials with less than 50% of timepoints, how did this vary across age? Arguably, your study has to worry less about this, given your sample size, but it would be nice to know, which you could include in the percentage of data that each study contributed to the final sample.*

Thanks, we've added this information to a new table in S1.

> ***Reviewer #2 (Public review):***
>
> *First, I wasn't entirely clear about what the authors meant by "word recognition ability". For much of the manuscript (including the use of the term "word recognition ability" itself), this comes across as an intrinsic ability or skill that improves with development. Alternatively, the speed and accuracy metrics taken from studies in Peekbank might capture children's increasing knowledge of the common, concrete words typically used in these studies. To me, this is a somewhat different construct from a general skill at recognizing words. It would be helpful if the authors could clarify which construct they intend to capture, or if it is not possible to distinguish between these constructs from the Peekbank data.*

In response to this comment and related comments above, we've added text to the first two paragraphs trying to clarify the general construct that we're talking about – recognizing the meaning of a word in real-time language comprehension. We've also clarified several times throughout the introduction that we're talking about familiar word recognition, that is, the ability to recognize specific known words. Further, we directly acknowledge the issue above in the introduction:

"Critically, most word recognition paradigms use words that children at the target age are reported to understand and produce. They are thus not indices of vocabulary size but rather measures of how quickly and accurately the child can recognize a familiar spoken word and use it to guide their visual attention to a referent. However, it is unknown the extent to which specific responses reflect an individual child's general speed of language processing versus their familiarity of specific words."

> *Second, and relatedly, if the source of the age-related improvements is increasing experience with the common concrete words used in the Peekbank studies, then one might expect word recognition and improvements with age to be related to word frequency, given that more frequent words are experienced more often. Word frequency predicts word knowledge when assessed using CDI data. Can effects of frequency be detected in Peekbank word recognition metrics? If not, why? Similarly, is the speed and accuracy of word recognition in Peekbank data related to CDI-derived word age of acquisition, and again, if not, why?*

This is a fascinating set of ideas, and one that we've pursued extensively using the Peekbank data. Unfortunately, we think it is out of scope for the current paper, which focuses on child-level metrics (including vocabulary and processing measures). Right now the current paper doesn't include any analysis of individual words.

Just to expand a bit on the problem here: unfortunately, modeling word recognition as a simple linear function of (log) word frequency is only possible in the case that distractors are held constant (e.g., "ball" always has "book" as its distractor), because distractor frequency plays an important role in the recognition process. However, in our dataset, words are paired with many different distractors across studies. This property means a fairly complex model of the LWL decision process would be necessary for a model to successfully predict effects for individual words. While such a model is an exciting research goal, it's not something we can include in the current manuscript.

> *Finally, there is a bit of a risk of the main findings of this paper coming across as a foregone conclusion. I.e., how could it be otherwise that word recognition improves with development?*
>
> ***Reviewer #2 (Recommendations for the authors):***
>
> *Regarding the feedback about the risk of the findings coming across as a foregone conclusion - perhaps a primary place in the paper where it would be useful to clarify this point is on page 6, in the paragraph beginning, "We investigate two specific hypotheses here. First, one influential theory...". Here, it might be worth clarifying whether there are alternative ideas about the emergence of word recognition in childhood that predict different patterns, so that the findings of the current paper can be framed as shedding new light on word recognition in development, rather than a confirmation of the common-sense idea that word recognition must improve over development.*

Thanks, we appreciate this feedback and it's something we've struggled with in this project. Our conclusion is that this paper does not constitute a binary hypothesis test of e.g., whether word recognition is linked to vocabulary development. Instead, we lean into the idea that there are empirical issues (rather than hypotheses) that have not been quantified sufficiently. Thus, we end the revised introduction with the following paragraph:

"Across both of these issues, the contribution of our work here lies in the detailed quantitative description of development. Nearly every theory of language learning assumes some role for continuous developmental change in word recognition, but these assumptions have not previously been anchored to specific measurements. Hence neither the functional form of the assumed changes nor their concurrent and predictive relationships to vocabulary have been quantified. We leverage the Peekbank dataset to accomplish these goals."

https://doi.org/10.7554/eLife.109636.2.sa3